CLOUD COMPUTING PAPERS FA18

Gregor von Laszewski

Geoffrey C. Fox

laszewski@gmail.com

Cloud Computing Papers

Gregor von Laszewski

Cloud Computing Papers

Refernces

1 Azure Data Services

Paul Filliman
pfillima@iu.edu
Indiana University
hid: fa18-516-06
github: ☁

Keywords: Azure

1.1 Introduction

This chapter focuses on an overview of the many data services highlights within the Microsoft Azure cloud. We detail the different relational and non-relational NoSQL databases as well as the many data analytics services.

1.2 Database Products

1.2.1 Azure SQL Database

The Azure SQL database is the cloud-based, SQL Server database as a service, relational database engine using the latest version of the SQL Server. There are many advantages to using an Azure SQL database as opposed to an on-premises SQL Server database platform. There are also many pricing level choices based on a function of hardware resources used.

1.2.1.1 Advantages

The biggest advantage to using an Azure SQL database rather than an on-premesis SQL database is scalability. Users can choose from many options of pricing model depending upon the utilization of their needs. Companies can start out with a low cost Azure SQL database having a fully managed database platform without the expense of an on-premises server and administrative costs and quickly scale up to a higher-cost pricing model with expanded system resources [1]. Other major benefits are database high availability and low administrative duties with operational database administrator or Windows server administration duties.

1.2.1.2 Pricing Models

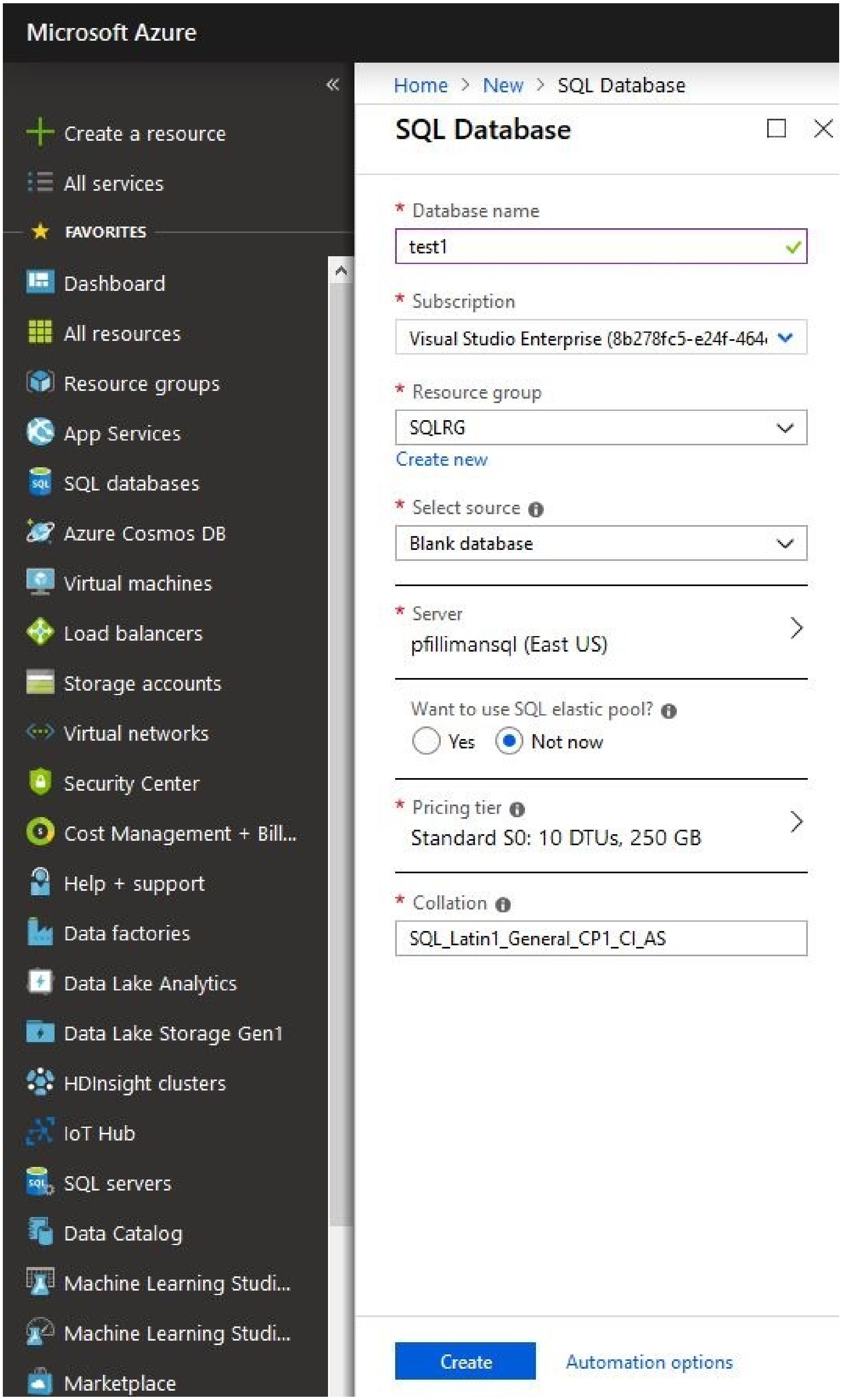
There are two purchasing models. The Database Transaction Unit (DTU) based model and the vCore based model. The DTU model uses a measure using a combination of compute, memory, and storage [2]. In the DTU purchasing model, users can choose from three different configurations, Basic, Standard, and Premium, corresponding to the extent of the needed resources. Within the vCore model, users can individually choose the compute, memory, and storage values. The Gen4 generation allows for up to 24 virtual CPU cores and 168 GB memory and the Gen5 generation allows for up to 80 virtual cores and 408 GB memory. The maximum data size within the Gen4 is 1 TB and with Gen5 it is 4 TB.

1.2.1.3 Creating an Azure SQL Database

The creation of an Azure SQL database is very easy:

1. Log in to the Azure portal
2. From the Azure portal, select ***Create a Resource***, then choose ***SQL Database*** within ***Databases***
3. Enter the name of the database to create
4. Enter the container for the resource group, create a new resource group, if desired
5. Choose if this created database will use an elastic poolcode-example
6. Select the pricing model
7. Click the ***Create*** button

Figure 1 This figure show adding an Azure SQL Database.



SQL Database

* Database name
test1 ✓

* Subscription
Visual Studio Enterprise (8b278fc5-e24f-464

* Resource group
SQLRG

* Select source
Blank database

* Server
pfillimansql (East US)

Want to use SQL elastic pool?
○ Yes ● Not now

* Pricing tier
Standard S0: 10 DTUs, 250 GB

* Collation
SQL_Latin1_General_CP1_CI_AS

Figure 1: CreateAzureSQLDatabase

Once the database has been created, we can use Microsoft Visual Studio as the development tool to the new Azure SQL database, much like an on-premesis database using SQL Server Management Studio, as shown in +???.

Figure 2 shows how to connect to an Azure SQL Database using Visual Studio.

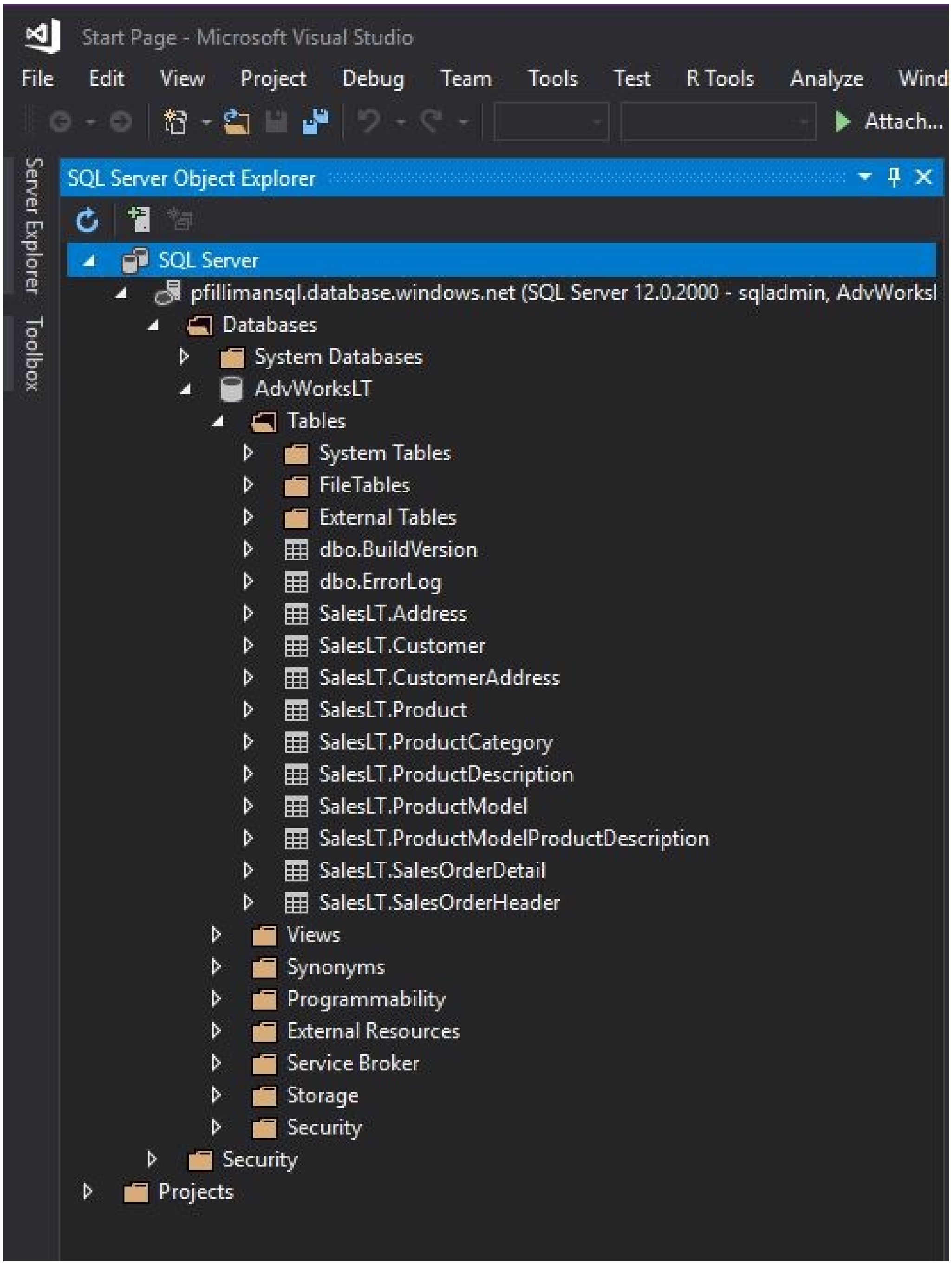


Figure 2: ConnecttoAzureSQLDatabase

1.2.2 Azure MySQL, PostgreSQL, and MariaDB Databases

Within the Azure ecosystem, it is possible to use three different open-source databases, MySQL, PostgreSQL, and MariaDB. Each of these are the cloud-based community versions of the databases. Much like Azure SQL, these have the benefits of using a cloud-based database service, for example scalability and uptime [fa18-516-06-AzureOpenSourceDB]. These Azure relational database allow users to keep using their desired open-source database platforms in the Azure cloud environment.

1.2.3 Azure Cosmos DB

The Azure Cosmos DB offers various multimodel, highly available databases for world-wide use. Cosmos DB supports many NoSQL data models including document, graph, key-value, and column-family models and is built on the ***atom-record-sequence*** data model which supports many APIs including MongoDB, Cassandra, Gremlin, and SQL [3].

Cosmos DB uses ***turnkey global distribution*** by distributing data near to where the current users are located to enable low network latency [3]. This is done through the ***multi-homing APIs*** where an application is aware of the location of the application user and can move data to the closest Azure region.

Cosmos DB service has high availability and throughput service level agreements, including a 99.999% availability and IO reads of less than 10 ms and IO writes of under 15 ms [4]. Users needing a highly available NoSQL database at a global scale, such as global web application databases, could gain from using Cosmos DB.

1.2.4 Azure SQL Data Warehouse

The Azure SQL Data Warehouse is a cloud-based, data warehouse that uses massive parallel processing for use with querying large amounts of data. The Azure SQL Data Warehouse uses Azure virtual machines for the compute nodes and Azure page blobs for storage. This separation allows for scalability for compute and storage independently [5].

One of strengths of Azure SQL Data Warehouse is its ability to ingest modern data sources, for example datalakes and Hadoop as shown in the figure below. With the ability of using Polybase, a user can query non-relation as well as relation data sources that are stored in Azure SQL Data Warehouse [6]. Various Azure services can be used having the Azure SQL Data Warehouse as a source, including Azure Analysis Services, other Azure SQL Data Warehouses, and Azure SQL Databases.

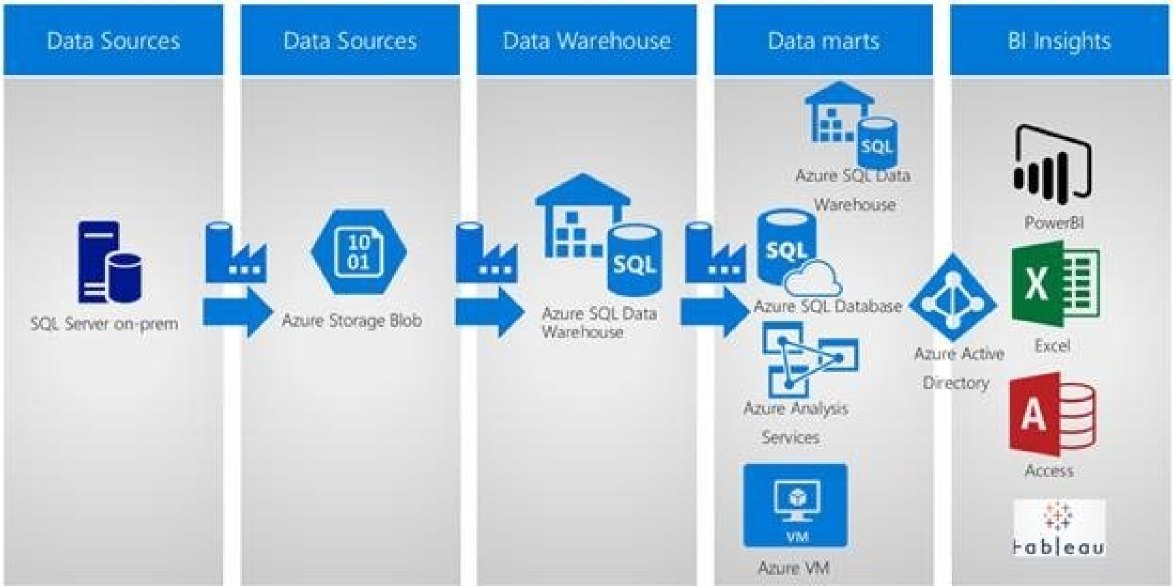


Figure 3: AzureDataWarehouse[7]

Another strength is the ability to only use this service during a particular time of day or week. If the data warehouse user only need access during a regular work week, this could save cost rather than running this service all of the time. Much like Azure SQL Database described above, this has high-avilability and backup and recoverability features as well [5].

1.3 Analytics

1.3.1 Azure HDInsight

HDInsight is the Azure services for clustering Apache Hadoop, Apache Spark, Kafka, Apache HBase, Hive, and Storm and are built around the Hortonworks

Data Platform. The concepts of the Hadoop ecosystem go beyond the scope of this chapter, but this section is an overview of the different HDInsight services available and how they can be used within Azure.

Azure HDInsight services are typically used when working with massive amounts of data in the internet of things and streaming real-time analytics scenarios. There are many ways under the HDInsight umbrella to setup clusters according to business needs. The following show for example configuring clusters using Apache Spark for parallel processing or Apache Storm for use with real-time streaming analytics. Apache HBase can be clustered in Azure for businesses needing a NoSQL database to store unstructured or semi-structured data. HBase brings very large tables having billions of rows and millions of columns. Apache Kafka can also be clustered under Azure HDInsight. Apache Kafka is a popular platform for streaming pipelines [8].

The following figure shows HDInsight within a modern data warehouse. There are multiple data sources from log files, and structured and unstructured data as batch processes for the HDInsight data sources. These data are into Azure Storage or Azure Data Lake Stores. Spark and HiveQL can then be used to query the Azure storage and these can be used to build business intelligence data models, for example Azure Analysis Services models. Finally, these data can be visualized using PowerBI.

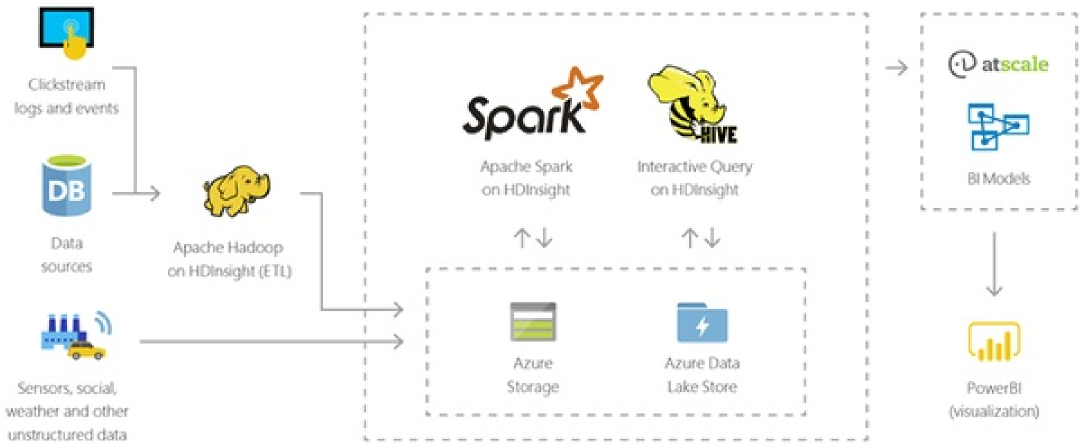

Clickstream logs and events
DB
Data sources
Sensors, social, weather and other unstructured data
Apache Hadoop on HDInsight (ETL)
Spark
Apachce Spark on HDInsight
HIVE
Interactive Query on HDInsight
Azure Storage
Azure Data Lake Store
atscale
BI Models
PowerBI (visualization)

Figure 4: ModernDataWarehouseusingHDInsight[7]

Figure 5 shows Azure HDInsight in an Internet of Things scenario. Various IoT streams can be fed into IoT hubs then read into HDInsight using the Storm, Kafka, or Spark services, then real-time visualizations or applications can be fed data from HDInsight.

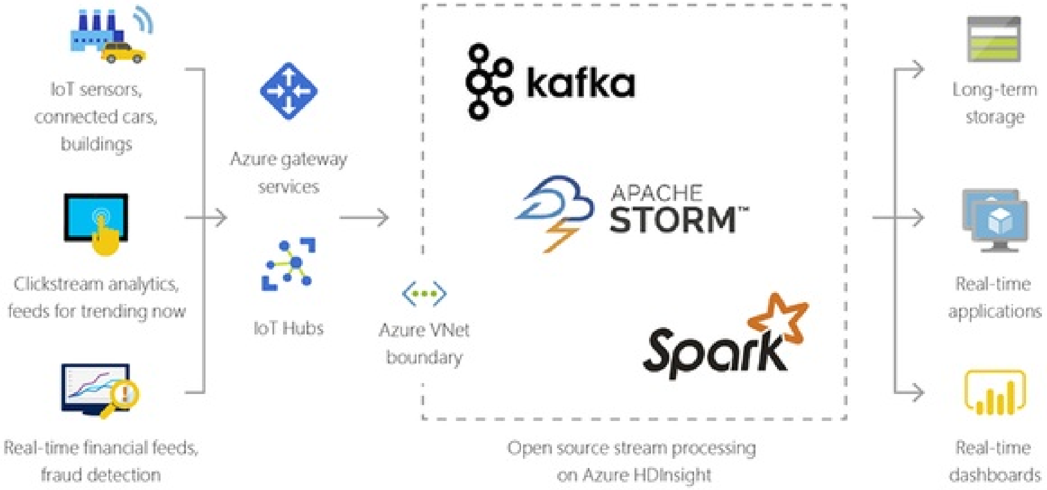

IoT sensors, connected cars, buildings
Clickstream analytics, feeds for trending now
Real-time financial feeds, fraud detection
Azure gateway services
IoT Hubs
Azure VNet boundary
kafka
APACHE STORM™
Spark
Open source stream processing on Azure HDInsight
Long-term storage
Real-time applications
Real-time dashboards

Figure 5: HDInsightinanIoTscenario[8]

One the of the strengths of HDInsight is that these services are available in Azure without the work of implementing these clusters in on-premesis servers and also having seamless integration with other Azure services. These services have high performance, five nines (99.999%) SLA and can be used on a per-use basis therefore cutting costs of permanent uptime.

1.3.2 Azure Stream Analytics

The Azure Stream Analytics service processes output from various IoT sources and can be used to analyze real-time data. Real-time data analytics is needed when data is in movement, for example, in cases such as detecting fraudlent bank transactions before the account is deducted. In past analytic systems, where an ETL load happened once per day, this system could not detect this transaction in real-time. Azure Stream Analytics is the service that manages these continuous real-time output.

Azure Stream Analytics is a part of the Azure IoT suite and ingest data from the Azure Iot Hub as well as Azure Event Hubs, Blob storage, and other relational or non-relational data sources. Once ingested into Azure Stream Analytics, real-time analytics can be gained using machine learning algorithms, for example detecting a fraudulent bank transaction. The data output from Azure Stream Analytics can also be loaded into other and uses is a part of the Iot.

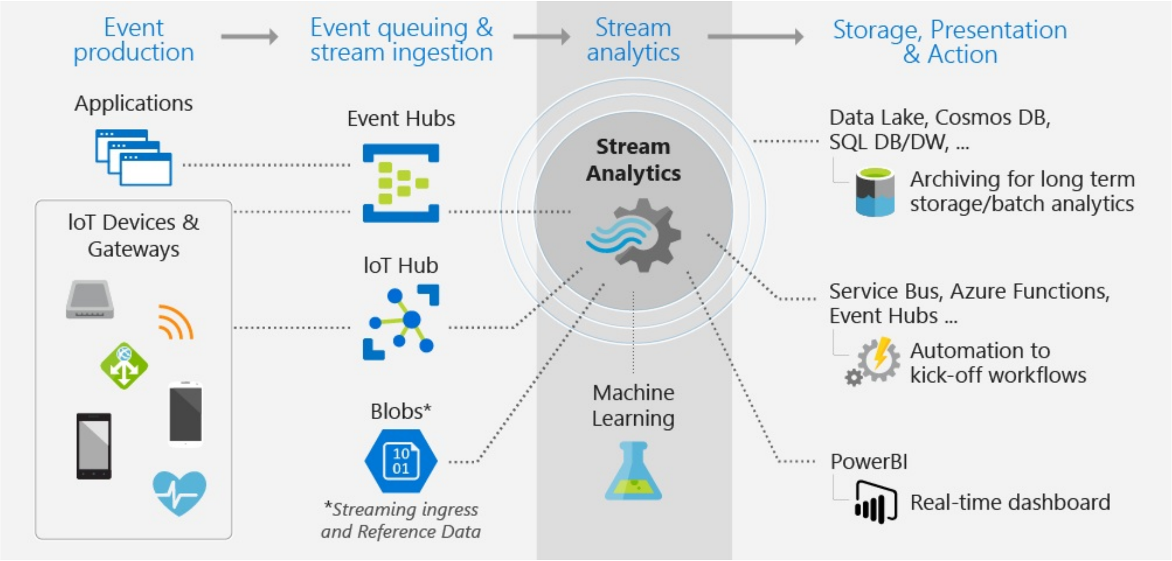

Event production
Event queuing & stream ingestion
Stream analytics
Storage, Presentation & Action
Applications
IoT Devices & Gateways
Event Hubs
IoT Hub
Blobs*
*Streaming ingress and Reference Data
Stream Analytics
Machine Learning
Data Lake, Cosmos DB, SQL DB/DW, ...
Archiving for long term storage/batch analytics
Service Bus, Azure Functions, Event Hubs ...
Automation to kick-off workflows
PowerBI
Real-time dashboard

Figure 6: AzureStreamAnalytics[9]

There are three basic parts to using Azure Stream Analytics. The first part is creating a stream job which designates the data source and uses a query language similar to SQL to make any transaformations on the incoming data. The third step is specifying where to output the data.

1.3.3 Azure Data Lake Store and Data Lake Analytics

Data lakes are scalable repositories of data stored in its original format. The Azure Data Lake Store allows users to store data within a Hadoop Distributed File System (HDFS) -compliant file system for use with big data analytics. Azure Data Lake is a cost-effective way to store scalable unstructured data in secure, active-directory environment [10].

The latest release of Azure Data Lake in June, 2018, named Gen2, is multimodal in that there is both BLOB object storage and now file system storage. This version has a Hadoop file system with hierarchical directories which allows for higher performance than a flat object namespace. This new feature in Gen2 can eliminate unneeded REST service calls, for example in moving files. Instead of separate REST service calls for copying a file to a new location and another for deleting the file from its original location, with Gen2 this process can be done in a single operation using file system storage [11].

Together with Azure Data Lake is Azure Data Lake Analytics. This service provides methods for running analytics job at a pay per use cost. The creation of data lake analytics jobs can be done using Visual Studio and U-SQL to load and transform data. Azure data lake analytics can also be used with data sources from Azure SQL Database, Azure Storage, and Azure SQL Data Warehouse, as well as the Azure Data Lake Store [12].

1.3.4 Azure Data Factory

Azure Data Factory is the integration engine within Microsoft Azure. This data service is responsible for automated movement of both structured and unstructured data within Azure and on-premisis data repositories. This work is accomplished by source and target connections together with pipelines between those connections and activities. Azure Data Factory can run in typical data warehouse environments as an extract transform and load workflow using the Azure-SSIS runtime as well as with big data workflows using unstructured data Azure HDInsight or Azure Data Lake [13].

A pipeline is a task within a data factory that comprises activities. For example, a pipeline can be used as a copy task or a data transformation task. Pipelines can be scheduled as a one-time event, hourly, daily, etc.

An activity within Data Factory is either a copy utility or a data transformation utility. A data copy utility has numerous sources and targets which can move data between cloud and on-premisis relational and NoSQL databases. A data transformation utility can manipulate the data from the previously mentioned data stores using Data Lake U-SQL queries, an HDInsight Hive or Pig activities [14].

2 Cloud and Data Privacy

Varun Joshi
vajoshi@iu.edu
Indiana University
hid: fa18-516-08
github: ☁

Learning Objectives

- Learn about data privacy in Cloud infrastructure
- European Union’s General Data Protection Regulation and how it effects Cloud computing for data Privacy
- Learn about major cloud vendors data privacy readiness
- Shift in choice of cloud infrastructure with data privacy as priority

2.1 Introduction

In this chapter we discuss the problem of data privacy in multitenant cloud infrastructure. The personal data of cloud users and data from businesses which deal in personal user data and use cloud technology, is stored and processed in cloud infrastructure. How this data is transferred between entities, how it is exposed to be used and retained, the policies for data purging and how users can control their personal data has become a mandatory policy decision for cloud vendors. With the advent of GDPR for European Union’s personal data protection, the cloud computing usage for storing and processing personal user data is changing. Data privacy in general has become the driving decision for many businesses in choosing and opearting on cloud. In subsequent sections, we will learn about cloud data privacy in the wake of GDPR and how it effects businesses across the globe.

2.1.1 GDPR Compliance

European Union’s General Data Protection Regulation (GDPR) came in to effect on May 25, 2018.

The core of the GDPR comliance is to protect EU citizens from privacy and data breaches [15]. It aims to give back the control of personal data to citizens and residents.

We may wonder that GDPR is applicable only for protecting EU citizens and the organizations based outside of EU need not be GDPR compliant. However, GDPR applies to any organization with business in EU and collect,store and process data of EU citizens. With the digital age and the organizations moving towards cloud computing, the GDPR brings new challenges both for cloud computing vendors who have data centers in EU as well as for organizations like Uber, Visa, Apple and many more who are ubiquitous in their business models and deal with EU citizens personal data.

[16] lists personal data as defined in Article 4 of GDPR:

> ***"personal data means any information relating to an identified or identifiable natural person (data subject); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;"***

In simpler language it could be any information like identification number (US equivalent for SSN),phone number, address, birth date, IP address etc. which can uniquely identify a person.

GDPR compliance poses strict fines for any personal data breach. Fines are either up to EUR 20 million or 4% of annual revenue, which ever is higher.

To summarize, GDPR compliance will protect EU citizens personal data and will enforce organizations worldwide to be GDPR compliant. Under GDPR, it is mandatory for organizations to disclose how and when personal data is collected and where it is stored, what it is used for and how the data will be erased when the data subject is no longer needed or chosses to opt out. Organizations need to explain in simplest terms to the data subject about usage of their personal information and give them option to opt out if they choose to do so. In case of data hack, the data subjects should be notified immediately when the hack happens.

In technical terms to be GDPR compliant, specifically for cloud computing use case, cloud users and cloud providers will collectively be responsible for encryption of data (both at rest and in transit), should have ability to monitor the usage,authorized access control of data, choose data storage - geolocation/type of storage/data access, dedicated data protection officers and industry certifications for data security. Privacy by design is now a legal obligation.

Now that we are familiar with the GDPR compliance, in next sections we will look into it’s impact specifically for cloud computing platform and data privacy in cloud data centers. Before that, let’s define Data Processor and Data controller with respect to cloud and as related to GDPR.

2.1.2 Data Processor vs Data Controller

Cloud solutions like AWS, Azure, GCP are all considered data processors because they offer resources and infrastructure to porcess the data.

Organizations,authority or agency which collect and direct the personal data and define mandate on how the collected personal data is processed are known as data controllers.

For example an organization like Airbnb collects personal data of its users and customers and controls the usage of the personal data like how and where it is stored, how it is used - such as providing rental suggestions or generating analytics on usage statistics. Airbnb uses AWS as it’s operation infrastructure. In this example AWS is data controller where as AWS acts as data processor.

AWS also acts as data controller for the data it collects like user account registration information of its customers, administration, service access etc.

Collection, storage, recording, organizing, structuring, alteration, consultation, retrieval, sharing, restriction and erasing and destruction all come under personal data processing.

GDPR defines collective responsibility for both data processors and data controllers for safeguarding personal data.

The defining of roles extends further if there is a third-party involved between cloud solution vendors and cloud users. For example if a company is using services of a third-party and the third-party is using cloud solution vendors directly then the roles should be understood clearly for being GDPR compliant and avoiding audit and fines.

Now we have understood the difference between data processors and data controllers, let’s look in to its impact on cloud computing by relating it to GDPR.

2.1.3 Impact On Cloud Computing

GDPR imposes collective responsibility on data controllers and data processors for personal data protection. Organizations or cloud users who deal with the personal data of their customers or consumers of their applications should be careful in choosing a cloud solution which is GDPR compliant and provides infrastructure and services options which are GDPR compliant. Data controllers should have options to define data privacy and security operations within the cloud infrastructure. Taking example of AWS as data processor, the resources like EC2, EBS, Amazon VPC all offer operations mechanism for a data controller to configure for robust data privacy and security. At the same time, AWS as a data processor needs to disclose in its contract with the data controller the options it provides for data storage and region and site for each chosen services.

Since data protection is a collective responsibility and design by principle, the data controller will have to keep the following check list [17] when choosing a cloud solution provider:

- Options to configure resources and desired settings as related to the data Privacy
- Ability to get snapshot of the current configurations in cloud
- On demand retrieval of configurations
- Historical logging
- Ability to get automatically notified of any changes in configuration
- View how resources communicate in cloud infrastructure

- Ability of cloud provider to encrypt the data either in transit or rest
- Options to set data access controls - granular access to data, multi factor authentication, geo-restrictions on data access

In summary, data controllers should define and are responsible for defining all data privacy and security rules when using a cloud infrastructure and resources. Data processors are responsible for providing resources and services to be GDPR compliant. If the processing happens with a third-party, the information needs to be disclosed and again the responsibility is collective.

Complexity the compliance may cause changes to how the cloud computing infrastructure is used by the organizations. The debate would be between private or public cloud for services which deal with the personal data. Let’s look more in detail in next section.

2.1.4 Public or Private Cloud

The definition of Private and Public cloud as defined by NIST:

- Private cloud : The cloud infrastructure is operated solely for an organization. It may be managed by the organization or a third party and may exist on premise or off premise [18]. The enterprise is solely responsible for managing and scaling the infrastructure as needed. Examples are AT&T, Cisco, T-Mobile have their own private cloud hosted in their own data centers.

- Public cloud : The cloud infrastructure is made available to the general public or a large industry group and is owned by an organization selling cloud services [18]. The enterprises and their data are virtually separated in the data center and enterprises have their own virtual private cloud network (example Amazon VPC). The cloud provider is responsible for managing and scaling the infrastructure at the data center. The enterprises can choose to individually upgrade and update the resources for patching, security etc. Examples are AWS, GCP, Azure.

Based on the above concepts, there could be multiple iterations of the cloud solution.

- Hybrid cloud is one such solution where there is a mix usage of on premise data center and public cloud data center. The use case could be based on mission critical applications, data privacy, need for on demand scalability, high availability etc. NIST defines hybrid cloud as : The cloud infrastructure is a composition of two or more clouds (private,community, or public) that remain unique entities but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load-balancing between clouds) [18].
- Managed cloud services provides enterprise an option to outsource the management of the cloud infrastructure and services to a third-party company like Rackspace or Expedient. Again the managed cloud service option depends on the need of usage. Managed services can be leveraged on private cloud, public cloud or companies like Rackspace, Expedient provide their own data centers which can be dedicated completely to an enterprise for their infrastructure needs and managed by the providers. Rackspace, for example, uses Openstack software for their managed cloud services solution.

Now with the knowledge of above concepts, it’s clear to define the cloud strategy. Keeping in mind the requirements of GDPR compliance and the responsibility of data controllers and data processors, a cloud solution can be chosen which is secured, robust, optimized and cost-effective.

Highly secured and sensitive data, for example HIPAA, can be managed in a private cloud or hybrid cloud. Other sensitive personal data which requires services of third-party for analytics generation like movie recommendation apps, shopping recommendation, election surveys, likes, social mining etc. can leverage public cloud scaling in a virtual private network utilizing GDPR compliant cloud data processor and rules and security defined by data controllers.

One important consideration is while using opensource solution like Openstack. Openstack can be used in a managed cloud service setting or independently for private cloud solution. The key is to use open source resources and their configurations which provide robust data security for compute, storage, network etc. which are integrated in Openstack software[19].

How GDPR and other data privacy compliances will shift the revenue model of major cloud vendors will be an interesting trend to observer. The trend will also relfect choice of enterprises for cloud solution provider in their journey to achieve less overhead of maintaining data centers, achieving scalability and at the same time protecting the interests of data subjects.

2.1.5 Common Vendors GDPR Readiness

Major cloud solution vendors like AWS, GCP and Azure are GDPR compliant and offer resources, services and configurations which are GDPR ready. Other vendors offering specifically SaaS and PaaS are also GDPR compliant. Privacy statements of vendors has also been updated to reflect their GDPR readiness. Refer the following for major vendors GDPR readiness:

- AWS [20]
- GCP [21]
- Azure [22]

Example of updated privacy policy in the wake of GDPR:

- Redhat [23]

Important takeaways from the RedHat privacy statement [23] is that Redhat claims the following rights:

- The right to access your personal data;
- The right to rectify the personal data we hold about you;
- The right to erase your personal data;
- The right to restrict our use of your personal data;
- The right to object to our use of your personal data;
- The right to receive your personal data in a usable electronic format and transmit it to a third party (also known as the right of data portability); and
- The right to lodge a complaint with your local data protection authority;

Above privacy statement example shows how RedHat provides users with control and rights to access their data stored in cloud which RedHat collects through its website and any other website owned by RedHat. This is the direct effect of GDPR and mandating data privacy policy in general.

2.2 Conclusion

With businesses moving their IT infrastructure to cloud and data privacy becoming a driving decision in choosing appropriate cloud strategy, the introduction of GDPR compliance is a benchmark to change the usage and selection of cloud computing solution. As more and more personal data is stored and processed in cloud, we may see new regulations or enhancemnet to existing ones improving data privacy and introducing more flexibility for cloud infrastructure.

3 Cloud Security Alliance (CSA)

github: ☁

Richa Rastogi fa18-516-17

3.1 Introduction to CSA

The Cloud Security Alliance (CSA) is a nonprofit organization that provides a variety of security resources to institutions including guidelines, education and best practices for adoption. There are many different audiences and variety of cloud and compute configurations. As this is the case, the CSA has a diverse membership, a variety of research areas and a breadth of recommendations and guidelines. In this chapter CSA topics include:

- About the CSA
- Guiding Principles
- History
- Research areas
- Membership
- Best ways to leverage CSA

3.2 About the CSA

The Cloud Security Alliance (CSA) is dedicated to defining and raising awareness of best practices to help ensure a secure cloud computing environment. The CSA leverages the expertise of industry practitioners, associations and governments, corporations and individual members. They offer research, education, certification, events, and products specific to cloud security that benefit the entire community. They essentially provide a forum through which these different parties can work together to create and maintain these recommendations for a trusted cloud ecosystem [24]. The industry group also provides security education and offers guidance to companies in different stages of cloud adoption. They also offer certification programs for cloud security providers and manage a global consulting program that allows participants to work with a network of qualified cloud security professionals [24].

3.3 Guiding Principles

The Cloud Security Alliance members are united by the following objectives according to thier LinkedIn page:

> - ***"Promote a common level of understanding between the consumers and providers of***
> - ***Cloud computing regarding the necessary security requirements and attestation of assurance.***
> - ***Promote independent research into best practices for cloud computing security.***
> - ***Launch awareness campaigns and educational programs on the appropriate uses of cloud computing and cloud security solutions.***
> - ***Create consensus lists of issues and guidance for cloud security assurance" [25].***

3.4 History

The Cloud Security Alliance was forged due to the need to have to have security best practices in a cloud computing environment [26]. As the concept of cloud computing increased in popularity in 2008, so did the surrounding issues and opportunities. The information security community took note and in November at an ISSA CISO Forum in Las Vegas the concept of the Cloud Security Alliance was born. After acknowledging emerging trends, proactive participants outlined the initial mission and strategy of the CSA. What followed was a series of meetings with industry leaders later that year which codified the foundation of CSA as we know it today.

3.5 Research Areas

The CSA currently has working groups that cover 38 domains of Cloud Security. These working groups publish a variety of white papers, reports, tools, trainings, and services that benefit the cloud security community.

For instance, the Blockchain working group meets every other week to discuss current events that surround Blockchain and any potential security implication. In addition, they published multiple documents outlining successful Blockchain projects, how to use the technology to secure the Internet of Things and a reference glossary of terms for the industry's benefit.

3.6 Membership

The Cloud Security Alliance employs roughly sixteen full-time and contract staff worldwide. It has over 400 active volunteers participating in research at any time. The CSA is a member-driven organization and individuals who are interested in cloud computing and have the experience to assist in making it more secure receive a complimentary individual membership based on a minimum level of participation [27].

The Cloud Security Alliance has a network of chapters worldwide. Chapters are separate legal entities from the Cloud Security Alliance, but operate within guidelines set down by the Cloud Security Alliance In the United States. Chapters are encouraged to hold local meetings and participate in areas of research. Chapter activities are coordinated by the Cloud Security Alliance worldwide [27].

3.7 Best ways to leverage CSA

3.7.1 Security Guidance Publication

The Security Guidance for Critical Areas of Focus in Cloud Computing documentation provides guidance to support business goals while mitigating the risks associated with the adoption of cloud computing technology. It is derived from focused research, participation from the Cloud Security Alliance members, working groups, and industry experts. It incorporates new cloud technology, reflects on real-world cloud security practices, integrates the latest Cloud Security Alliance research projects, and offers guidance for related technologies [28].

These security guidelines use NIST industry standards and are easy to read with models and illustrations that make key concepts more digestible for most readers.

3.7.2 Training and Certifications

3.7.2.1 CSA STAR

This is an industry-recognized security assurance in the cloud. STAR encompasses key principles of transparency, rigorous auditing, and harmonization of standards. According to the CSA website,

STAR consists of three levels of assurance (Self Assessment, 3rd party certification and continuous auditing), based upon:

> - ***"The CSA Cloud Controls Matrix (CCM)***
> - ***The Consensus Assessments Initiative Questionnaire (CAIQ)***
> - ***The CSA Code of Conduct for GDPR Compliance" [29].***

3.7.2.2 Certificate of Cloud Security Knowledge (CCSK)

Since Cloud Security Alliance first released the Certificate of Cloud Security Knowledge (CCSK) in 2010, thousands of IT and security professionals have used it to upgrade their skills. Certification Magazine listed CCSK at number one on the Average Salary Survey 2016 [30].

Accoring to thier CCSP page,

> ***“Certified Cloud Security Professional (CCSP) The CCSP is a global credential that represents the highest standard for cloud security expertise. It was co-created by Cloud Security Alliance and (ISC)² — leading stewards for information security and cloud computing security” [31].***

3.7.2.3 CSA Global Consulting Program

> ***“The Cloud Security Alliance Global Consulting Program***

> ***(CSA GCP) allows cloud users to work with a network of trusted security professionals and organizations that offer qualified professional services based on CSA best practices. These providers bring with them a broad understanding of the challenges organizations face when moving to the cloud" [32].***

3.7.3 Working Groups

As mentioned above working groups are a great way to learn and share with industry experts on topics you or your institution are focused on [33].

4 Guided Analytics Using KNIME

Anna Heine
avheine@iu.edu
Indiana University
hid: fa18-523-52
github:

Keywords: KNIME, workflow, workbench

4.1 Introduction

KNIME [34] stands for KoNstanz Information MinEr and is an open source data analytics software that creates services and applications for data science projects. KNIME allows its users to create visual workflows with a user-friendly drag and drop graphical interface that depletes the need for any programming. However, KNIME does allow implementation of other scripting languages such as Python [34] or R [34] that creates connections to abilities within Apache Spark or other machine learning tools. KNIME allows imports of datasets from a variety of formats, some of which include CSV [35] , PDF [35] , JSON [36] and more. The workflows and visualizations that KNIME produces allows export in many of these formats as well. It also supports several unstructured data types from images, documents, and certain networks. KNIME operates by a node system that includes embedded modules that help its users build their workflow. With this node system, users can make changes at every step of their analysis to ensure the most current version. KNIME also provides detailed visualizations from a set of defined graphs and charts which can lead to predictive analyses and machine learning implementations. Users can shape their data by a variety of mathematical models such as statistical tests, standard deviations, and means. Users can even select specific features for use in possible machine learning datasets and apply filters to mark out some of the data if needed.

KNIME is a platform that can perform intense data analytics on a graphical user interface and incorporate a user-friendly workflow. It incorporates large or small data sets and even projects as broad as deep learning. KNIME is diverse in that its users do not necessarily need to know any coding languages to use it. KNIME is a process-oriented, single base workflow with basic input/output manipulations. KNIME is an open source platform that uses thousands of its documented nodes within the node repository for use in the KNIME workbench. A node is a single processing point of data manipulations within your workflow. A workflow is described as a sequence of steps a user follows in their platform that is used to complete their final product. The collection of nodes that creates a KNIME workbench is able to be executed locally or within the KNIME web portal on its own server. The workflow that KNIME follows first begins with data collection, data cleaning, data integration, and finally, feature extraction. This workflow allows for large files such as a CSV to be accessed through the web portal and it can therefore be manipulated through several wizards [37].

4.2 KNIME Used in Big Data

KNIME is useful in Big Data applications because it aids in the process of guided analytics. Guided analytics is a process that functions by providing automation to data science projects. This usage brings to light some of the features of big data that are sometimes hidden when visualization tools are unused. Depending on the wizard the user selects, the data can then be viewed in a formatted table using forward feature selection methods. This means that the user can then select their data based on the most accurate correlating information. This will, of course, be varied based on the type of information that is entered into the table. The next step in the data-selection process is a second screen with a wizard that asks the user to choose a target variable. The next screen will then show to offer a selection of algorithms which the user can use to entrain the dataset. After choosing an algorithm, the user may browse from a list of displayed visualizations that include their dataset such as bar graphs, an ROC curve, and more. The user can then download their data model in a PMML format, which is universally configurable in any enterprise application. The steps to download and set up your KNIME platform is quite simple. First, go to the KNIME website and obtain the download. Then, install KNIME and set its working directory for KNIME to store its files in. To set up a KNIME workflow, go to the File menu on the platform and choose New. Give your workflow a name and then click Finish. This process establishes a basic, empty workflow in which users can drag nodes from the repositories on the left-hand side into the workflow space. To increase user collaboration and support, KNIME also includes a Workflow Hub [38]. This hub allows users to share their workflows and make comments or suggest improvements to their designs. The components of a KNIME server after installation are hosted on the same machine. The components include: a workflow repository, the executor, the server. Figure 7 shows the simple architecture of KNIME's server for a single user. Figure 8 shows the basic KNIME platform setup with graphs, repositories, and the workbench.

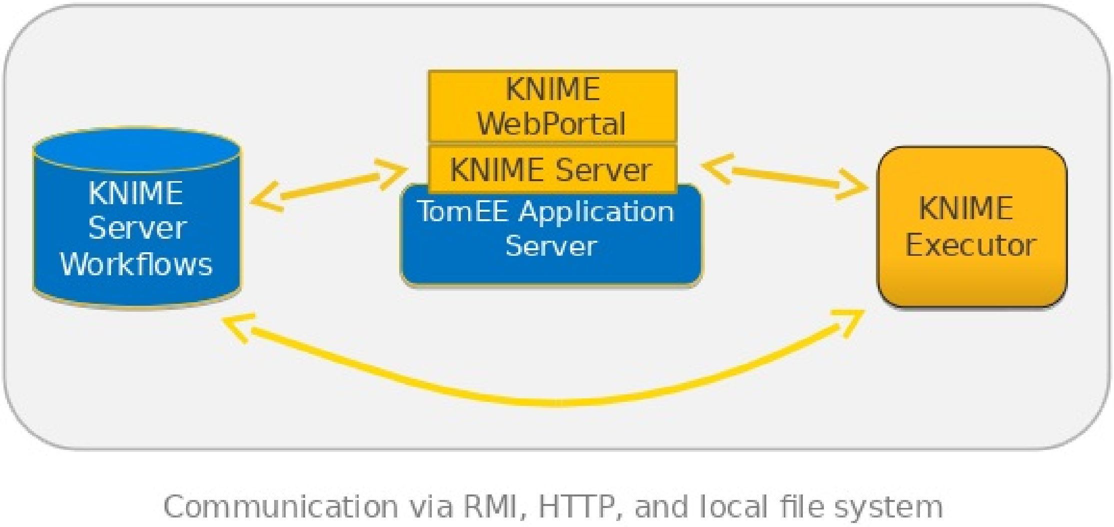


Figure 7: KNIME Diagram [39]

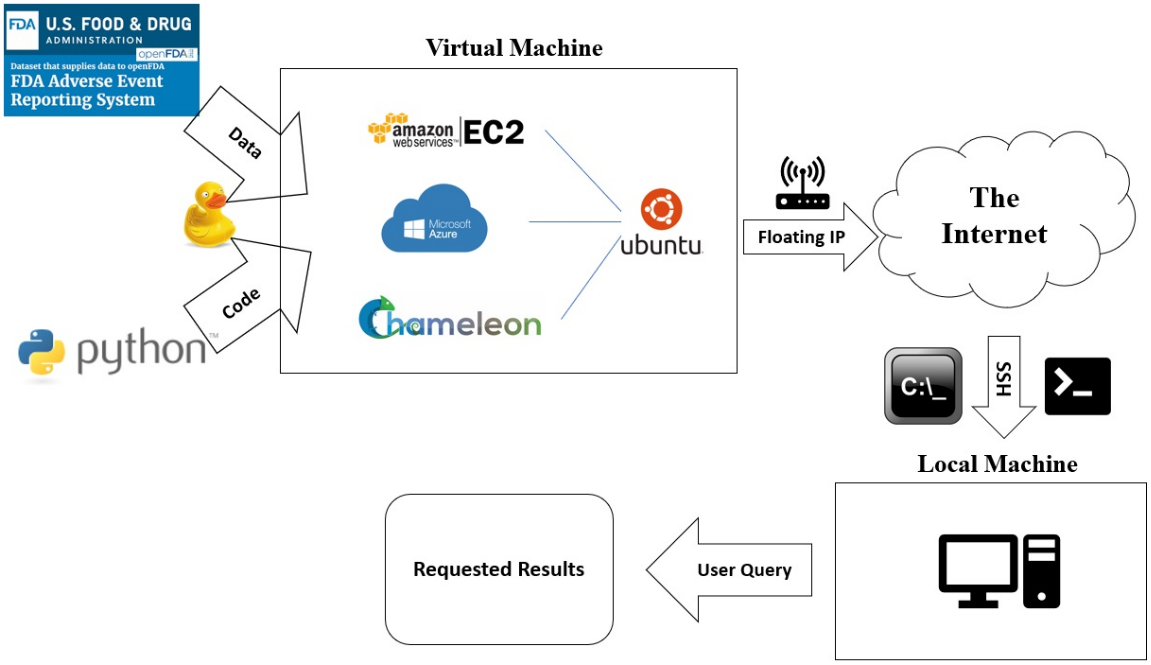


Figure 8: KNIME Architecture [40]

An example of big data analysis hosted by KNIME is a store trying to compare its products sold over multiple store locations. The first step in the visualization process would be to import your data. Within the workflow, users can view differences or find a possible correlation in their dataset by searching for linear correlation in the repository. After choosing this feature, the user must connect the data set to the linear correlation node via a line on the workflow grid. The execute option can then be chosen to view the correlation matrix. The user can hover over a specific cell to select the feature you want to use for further prediction. The next few steps are used in visualization and analysis of user data. Under the Views tab, the user can search for different graphs or plots. A scatter plot, in this case, would be a great way to visualize data from multiple items within a store. You then must drag and drop it and connect it within the workflow like before. You can then configure just how many rows that you would like to look at. Figure 9 shows an example of user analysis by the drag and drop method.

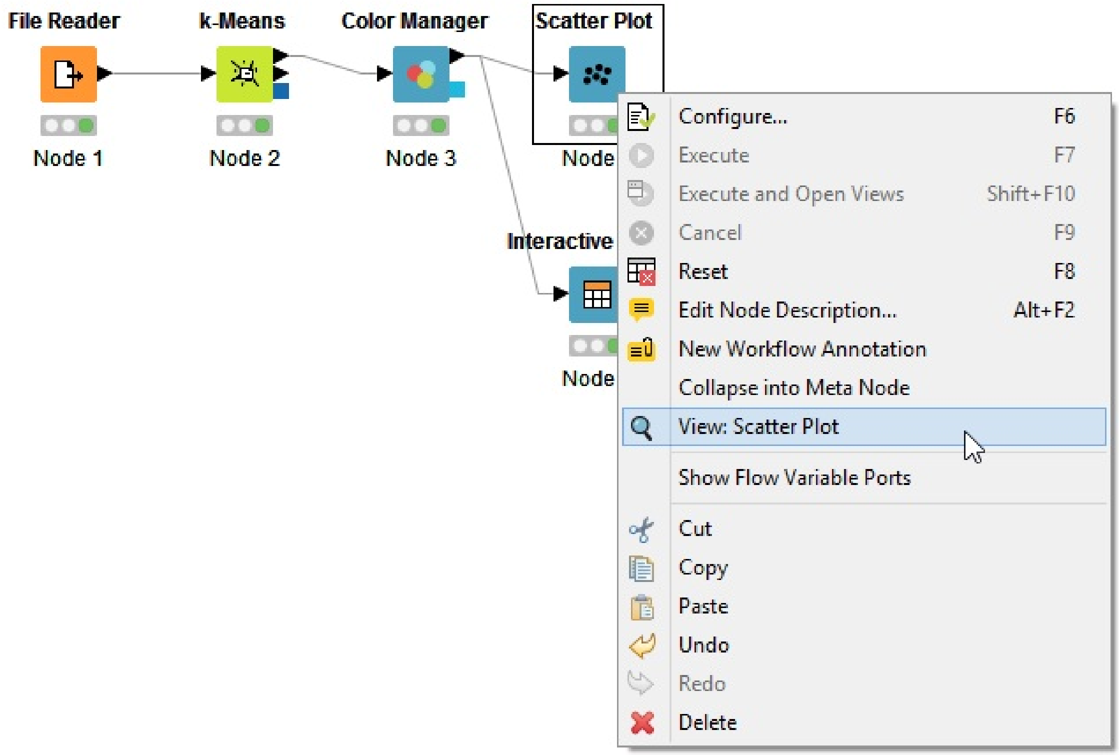


Figure 9: KNIME K-Means [41]

KNIME also includes nodes that can show missing values from certain datasets. KNIME includes a special node in which users may find imputations in their dataset. This is displayed as Missing Values in the output portal. From here, you can choose from a variety of options that allow handling of these imputations. For example, with strings you can move forward and backwards in between rows, create a custom row, or remove a row. You can also manipulate numerical data values by performing several mathematical functions. The basic limitations for KNIME include visualizations that are not extremely neat or detailed as other software. The software's updates sometimes cause user issue and result in necessary re-installation. As this is not as popular of a program as Python or other editing platforms, the community does not have as rich of a support system, therefore, users sometimes struggle with researching issues.

KNIME has the ability to be integrated with other techonlogies for larger open-source projects. These cloud services allow for user's projects to be analyzed even further. For example, KNIME can be used with Amazon AWS [42] and Azure [43]. KNIME's platform can be hosted on Microsoft Azure Cloud Services. Azure allows KNIME to perform its analytical, machine learning, and deep learning tasks on its integrated server. This application can be downloaded from Azure's Marketplace. KNIME can also be incorporated with Amazon AWS. When KNIME is connected to AWS resources, users can leverage the memory available while connected to the relational database service to construct SQL queries visually. KNIME's Analytic Platform is a free service for all who use it. However, if you are using KNIME on a cloud service such as Azure or AWS, there are often subscription fees associated. Students and other organizations can receive discounts or allocated amounts for a specified time of use. KNIME is a data analysis software platform that allows for easy read and manipulation of large datasets that can ultimately be used to make inferences and predictions. Its user-friendly interface allows for a broad integration of users and sometimes more efficient workflows. KNIME has several applications for its users such as data modeling, machine learning, predictive analysis, and more. After visualization, users can extract specific features from their data and implement it into a model of their choice, which can then be exported as a CSV file.

5 Microservices and Kafka

Chaitanya Kakarala
ckakara@iu.edu
Indiana University
hid: fa18-523-53
github: ☁

Keywords: Microservices, Kafka, Python

5.1 Introduction

Apche Kafka [44] is a distributed streaming platform which works on a subscribe-publish model. Data flows through a streaming channel also known as a Kafka cluster by either subscribing or publishing the topics. The Primary objective of Kafka is to persist the message data so multiple consumers can access the same and provide horizontal scaling. With the increase in demand for Agile methodology in software development where the user stories should be completed in a given sprint, there is great need of shrinking down the applications to smaller units. These smaller units are otherwise known as Microservices. These Microservices are loosely coupled and the instructions inside them are light weight. Microservices are autonomous by nature and they can be plugged in any host and bring them up provided the software and hardware requirements are met. As a result, small independent teams can work on these Microservices in parallel and deploy them independently. Depending on the complexity of an application there could be hundreds of Microservices defined and each of them might interact with each other. In other words, an event occurred on one Microservice could start another Microservice. With the interactions between these Microservices increasing, it’s hard trace the connection between them and hence creating a technical debt to mitigate the issue. Apache Kafka is a solution to mitigate this issue.

5.2 Architecture

The unit of data within Kafka is message. These messages are nothing but an array of bytes and Kafka is least worried about the content of these messages. Optionally a message can have a key which is again an array of text whose hash value determines the partition the message will be written to. Doing so will guarantee that the messages with same hash value will be stored into the same partition. Messages can also be sent in batches which in other words, a bunch of messages sent all at once. That leaves us with questions like, what are these messages? Where are they stored? who uses these messages? Messages in Kafka are classified into Topics. Topics are nothing but a group of partitions (Can also be described as disk space) where a collection of similar messages are stored. Messages will be appended to these partitions and will be read from beginning to end fashion. The Partitions can be hosted by different servers which makes the topic scale horizontally. All the partitions for a topic is often termed as Stream. Figure 10 describes four partitions of a single topic.

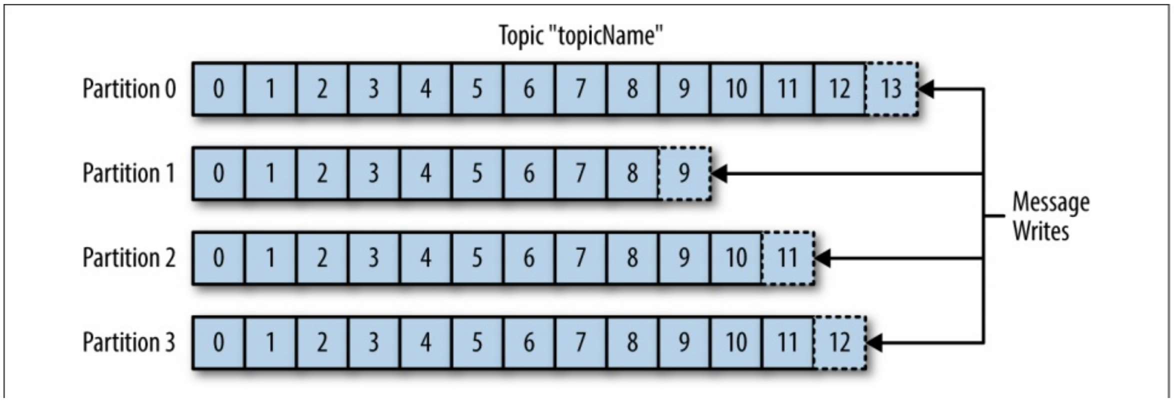


Figure 10: Representation of topic with multiple partitions [45].

There are basically two users of Kafka system. They are Producers and Consumers. Producers create messages to a specific topic. Producers are also termed as publishers. Producers by default does not care which partition they are writing the message to. However, in some cases the hash value of the key decided the partition and ensures all the messages for the same key reside in the same partition. Consumers read messages from the partitions in the order they were published by the producers. Consumers are also termed as subscribers. While reading the messages from partitions, consumers store the offset to to keep track of the read messages. By storing the offset, the system can be re-strated from the point of failure without starting all over again. Consumers are bundled together as a consumer group that restricts a given partition to be read by a unique consumer. Consumer groups helps scaling the consumers horizontally. Figure 11 illustrates on how consumer group works.

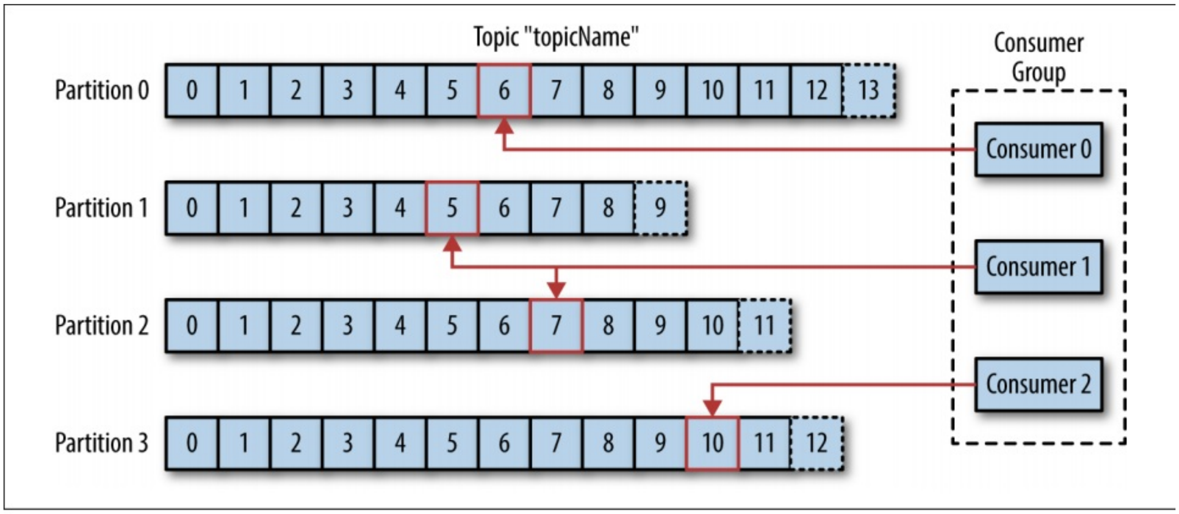


Figure 11: A consumer group reading from a topic [45].

A single kafka server is called as Broker. Each broker receives messages from producers and write them to the partitions on the disk. They will then save the offset for each message in a partition They also respond to the consumer programs for data requests from partitions and commit the same. Kafka is designed to have multiple brokers and collection of all of them is termed as a Kafka cluster. Each cluster can have multiple brokers where the leader broker replicates the data to others. Replication of data helps in durability of data even when one of the broker failed working. Figure 12 explains how multiple brokers are replicated in a kafka cluster.

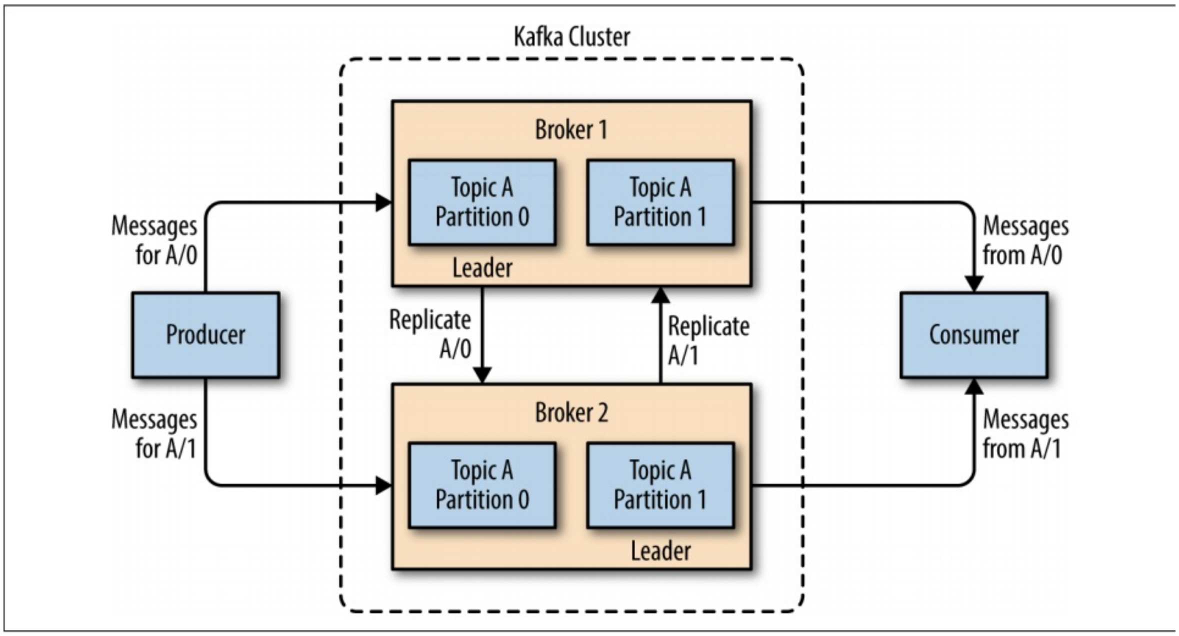


Figure 12: Representation of partitions in a cluster [45].

The major aspect of kafka is the data retention in the partitions. By default the messages in the partitions will be retained for a period of time or Size. For example, the messages in a topic can be retained for one week or until the partition reaches 1 GB. The default behavior can be overridden for topics by changing their settings. Kafka also supports multiple clusters communicating across multiple data centers.

5.3 Installation and Starting Kafka

- Kafka Installation:
    - Kafka tar file can be obtained from [46]. Please download and save it on the server. Please be aware that kafka requires Java to be installed on the server.
    - Untar the downloaded file using below commands

```
tar -xzf kafka_2.11-1.1.0.tgz
```

- If the java version in your server is having a LTS (Long Time Support) then below fix is needed in kafka-run-class.sh located in bin folder under the kafka home directory. This is a known fix and kafka is working to address this issue for future releases [47].

Change below line

```
JAVA_MAJOR_VERSION=$($JAVA -version 2>&1 | sed -E -n 's/.* version “([^.-]*).*”/\1/p’)
```

to

```
JAVA_MAJOR_VERSION=$($JAVA -version 2>&1 | sed -E -n 's/.* version "([^.-]*).*/\1/p')
```

- Start the zookeeper server using the below command. You need to be in kafka home directory to be able to successfully execute the below command.

```
bin/zookeeper-server-start.sh config/zookeeper.properties
```

- Start the Kafka server using the below command. You need to be in kafka home directory to be able to successfully execute the below command.

```
bin/kafka-server-start.sh config/server.properties
```

5.4 Use Cases

Kafka was primarily built in LinkedIn to optimize their activity logging system that has multiple Microservices. When the user does an action from the frontend, messages streams through the kafka cluster which are then subscribed by the backend consumer process. Kafka is also used in organizations for collecting the logs and metrics from their microservices. Since Kafka is analogous to the commit log in databases, it can be used to monitor the databases as an event occurs. Kafka is also used in the applications which are developed to process data in streams.

5.5 Acknowledgement

The author would like to thank Professor Gregor von Laszewski and associate instructors for their help and guidance.

6 Apache NiFi

Daniel Hinders , Nhi Tran
dhinders@iu.edu, nytran@iu.edu
Indiana University
hid: fa18-523-56 , fa18-523-83
github: ☁

Keywords: NiFi, NSA, Data Stream, ETL

6.1 Apache NiFi Introduction

NiFi is a customizable tool for building flexible data flows while preserving data provenance and security [48]. NiFi provides the ability to build or alter an ETL flow with a few clicks. NiFi builds Gets, Converts, and Pulls in a GUI and allows the user to build and customize the flow [49]. This flexibility and usability is key to NiFi's value in a big data world where stovepipes and inflexibility are frequently challenges.

As pointed out in [50] NiFi is a tool for:

> - ***Moving data between systems, including modern systems such as social media sources, AWS cloud server, Hadoop, MongoDB, and so on***
> - ***Delivering data to analytics platforms***
> - ***Format Conversion, extracting/parsing data***
> - ***Data or files routing decisions***
> - ***Real-time data streaming***

> ***NiFi is not recommended for:***

> - ***Distributed Computation***
> - ***Complex Event Processing***
> - ***Join/ Aggregated Functions***

6.2 Big Data Challenges and NiFi

Big data can be a fantastic source of information for decision making and business process definition and actualization. However, the complexity of individual datasets, the variability of dataset structure and composition, and the sheer volume of data are challenges to truly leveraging big data in the real world. This is a multifaceted problem with many inherently overlapping challenges. ETL or Extract, Transform, and Load encompass a number of potential tasks such as harvesting and moving data into a database from some other location and/or and cleaning, normalizing or even structuring data. In a case where a single dataset emerges from an ETL process and the data is somewhat structured and located somewhere predictably accessible, then we can start to leverage analytics or visualization tools to understand the data and use it to make decisions and learn things. Furthermore, productization and dissemination of that data is fairly straightforward.

But this is rarely where the real use case for big data solutions ends. The bigger challenge is dealing with disparate datasets and connecting points of information in a multi-sourced dataset environment. Consolidation of disparate data is therefore extremely important. Furthermore leveraging the correctly sourced data out of consolidated data store environment and then loaded this data into the correct product is challenging.

Apache NiFi is an application that seeks to address this big data problem. NiFi is a tool that has emerged from a unique background as a tool created by the National Security Agency then curated and improved by the open source community.

6.3 NiFi History

NiFi was first developed at the National Security Agency but was released as an open source project to the public.

> ***"NiFi was submitted to The Apache Software Foundation (ASF) in November 2014 as part of the NSA Technology Transfer Program" [51].***

Since then, Apache Foundation has used its volunteer organization to grow and mature the project [49].

6.4 NiFi Features

NiFi incorporates a straightforward User Interface (UI) to engineer traceable data provenance with configurable components. NiFi offers up the ability to custom build processors and incorporate them into a highly customizable flow. Through

> ***“data routing, transformation, and system mediation logic” [48],***

NiFi seeks to automate data flow in a big data environment and gives architects the ability to keep data flowing between evolving systems quickly. Amongst a host of features, NiFi offers, one sticks out as particularly important because of the challenges associated with what the feature addresses: data errors, data inconsistency, and data irregularity handling. NiFi provides users with the ability to incorporate in the flow, processes to catch these non-happy path realities in big data. As new situations are discovered, a user can quickly build ***if-then*** forks in the process to catch, store, or resolve the data issues.

NiFi’s main features are:

- ***Guaranteed delivery***: use purpose-built persistent write-ahead log and content repository to ensure guaranteed delivery in an effective way [48] [52]
- ***Web-based user interface***: easy to use web-based GUI with drag and drop features that allows users to build, schedule, control, and monitor data flow[48] [52]
- ***Provenance***: provide the ability to track data flows through the systems with audit trail and traceability functionalities [48] [52]
- ***Queue Prioritization***: provide the ability to configure and prioritize job flow and determine the order of events [48] [52]
- ***Secure***: provide and support multiple security protocols and encryptions, as well as authorization management [48] [52]
- ***Extensibility***: provide flexibility by allowing pre-built and built-your-

own extension the be integrated [48] [52]

- ***Scalability***: supports scale-out by clustering architecture as well as scale-up and scale-down [48] [52]

6.5 NiFi Architecture

Figure 13 shows the main components in NiFi architecture [52].

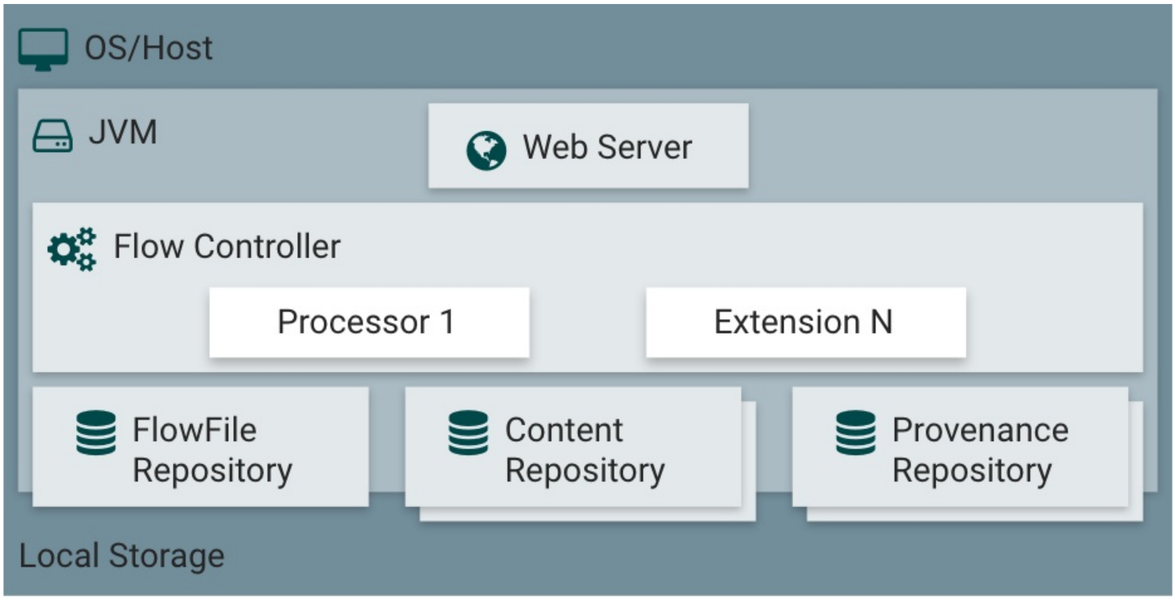


Figure 13: NiFi Architecture [52]

From the top down, NiFi is web browser accessible by a NiFi hosted Web Server. NiFi processor operations are managed through the Flow Controller and the three repositories; FlowFile, Content, and Provenance work to process data on and off disk and in a NiFi flow. NiFi is hosted in the Java Virtual Machine environment or JVM [52].

6.5.1 Web Server

NiFi's easy-to-use graphic user interface(GUI) is hosted on the Web Server within the JVM [52].

6.5.2 Flow Controller

NiFi central operations hub is the Flow Controller. Treads are managed and allocated to the processors and the FlowFiles are passed through and managed through the Flow Controller [53].

6.5.3 FlowFile Repository

Files in an active NiFi flow are tracked in a write-ahead log so that as data moved through the flow NiFi can keep track of what is known about files as they pass through [52].

6.5.4 Content Repository

The real data for a flow file is in the NiFi content repository. NiFi uses simple blocks of data in a file system to store this FlowFile data [52]. Multiple file systems can be used in order to increase speed with multiple volumes being utilized.

6.5.5 Provenance Repository

In NiFi, the provenance repository stores historic event data. The provenance data about flows is indexed to enable search of the records [53].

6.5.6 Processors

NiFi provides more than 260 processors and more than 48 controller services for users to integrate into a flow from the graphic user interface(GUI) of Nifi [54]. Processors are base on underlying controller services in the java virtual machine. Controller services can be centered around a security implementation, database CRUD (create, read, updates, and deletes), and many other foundational areas. Users can create custom processors from existing controller services or create a customer controller service as well [54].

6.5.6.1 Processor Examples

- Get
    - Examples: GetFTP, GetMongo, GetTCP, etc. [52]
    - Similar input type processors: Consume, Extract, Fetch, Listen,

etc.

Nifi provides dozens of ***Get*** processor options and many other similar input type processors. A ***Get*** processor is commonly used to pick up a file or data and launch a FlowFile. The ***Get*** file processer setup typically gives configuration options to point to a host, set timing increments for polling and timeouts, set proxy settings, and more [52].

- Convert
    - Examples: ConvertJSONToSQL, Convert Record, ConvertExceltoCSVProcessor, etc. [52]
    - Similar transformation type processors: Evaluate, Merge, Split, etc.

Once data is in the flow, NiFi provides dozens of processors to manipulate or transform data. The ***Convert*** processors can be configured to the expected schema or type from the ***Get*** processor and transform, edit, thin, enrich, or many other functions on the data in the flow [52].

- Put
    - Examples: PutFile, PutFTP, PutSQL, PutElasticSearch, PutAzureBlobStorage, etc. [52]
    - Similar output type processors: Publish, etc.

A critical part of a flow in NiFi is pushing the right data out of the flow into the right spot. There are dozens of ***Put*** processors that can be configured to set the directory to write files too. Additional configuration options are specific to the destination type to include SSL configuration, cache options, batching options, and many other configuration options based on the destination type [52].

6.5.7 NiFi Clusters

NiFi can also be integrated with ZooKeeper to operate within a cluster. Figure 14 shows how ZooKeeper manages NiFi’s nodes by determining the primary node, Zookeeper Coordinator, and failover node [52]. Each of the nodes performs the same tasks but processes different dataset(s) [52].

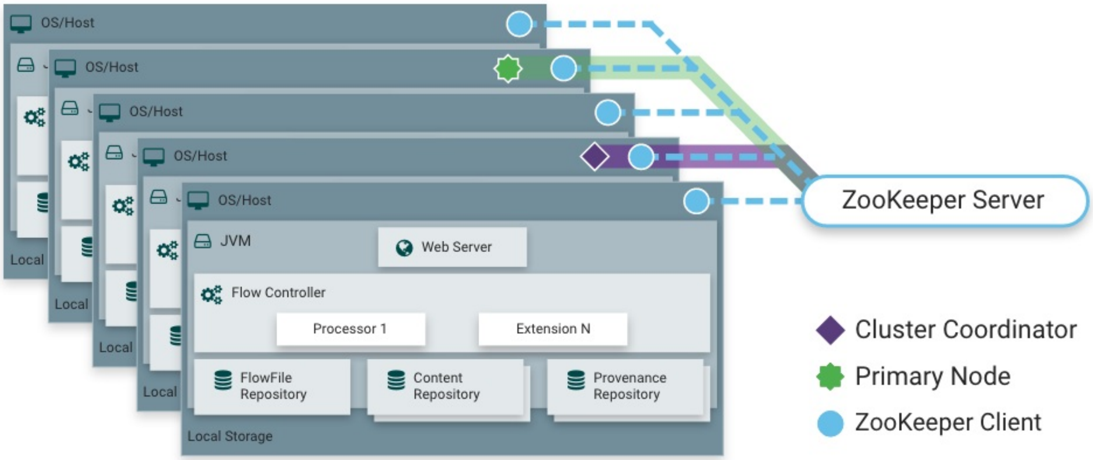


Figure 14: NiFi Cluster Architecture [52]

6.6 NiFi Download, Installing and Getting Started

NiFi can be downloaded and installed from its Downloads Page [55] with Linux/Mac ***tarball*** option, or ***zip*** file option for Windows, or Homebrew option for Mac [56].

For Window

Double-click to run `run-nifi.bat` file from NiFi `bin` subfolder within the installed folder [56].

For Linux/Mac OS X Users

Use Terminal to run `bin/nifi.sh`. An application will run and will be shutdown when the command is terminated [56].

NiFi as a Service in Linux and Max OS X

To install NiFi as a Service, run the command `bin/nifi.sh install <service_name>`. Without specifying specific `<service_name>`, nifi service name will default to ***nifi*** [56].

To start NiFi service after installation, run sudo `service <service_name> start` . To stop, run `sudo service <service_name> stop` [56].

Once NiFi has been started, the GUI can be accessed using a web browser via http://localhost:8080/nifi . The port and hostname can be configured and changed depending on which server or setting in `conf/nifi.properties` is used [56].

6.7 Use Case

6.7.1 File Transfer and Routing at MasterCard

MasterCard is a card payment and technology company that connects digital transactions globally. Figure 15 shows one of the use cases for NiFi at MasterCard, which is a file transfer mechanism [57].

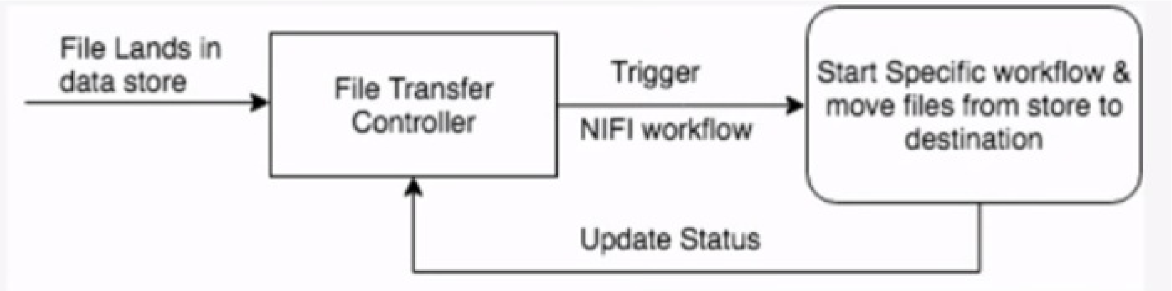


Figure 15: Mastercard NiFi Flow [57]

Batch processing is still a major part of MasterCard's ecosystem which requires multiple formatted flat files being created, transferred, and picked up by applications [57]. MasterCard uses NiFi's file transfer features to convert files source into data stream(s) and perform specific workflow to direct data into various target systems [57]. Target system could be a messaging systems, Hadoop landing zone, databases. NiFi can also feed data and trigger a map-reduce or spark jobs after transfer [57].

MasterCard provided a demo which demonstrates the use case of them using NiFi to call web services from a file transfer controller, the data flow then has a mechanism to determine which process groups NiFi should distribute data into based on file name/ format logic [57]. The process groups contain workflows that can either feed data to a different system, to Hadoop, or to Postgres database [57]. Once each process flow is completed, the process status will be captured and reported into a Status Handler process [57].

6.7.2 Streaming Analytics Solutions at OpenText Magellan

OpenText Magellan is an artificial intelligence product that supports machine learning and advanced analytics. At OpenText Magellan organization, NiFi was utilized as part of their streaming analytics infrastructure to allow continuous process and real-time analysis [58]. OpenText Magellan's infrastructure involves source applications, NiFi, Apache Spark, Python, R, Scala, and other Magellan BI and Reporting tools [58]. Figure 16 shows a typical process of streaming analytics process at OpenText Magellan, which involves six steps: (1) Data Acquisition, (2) Data Routing, (3) Streaming Processing, (4) Machine Learning, (5) Prediction Results, and (6) Actionable Insights [58].

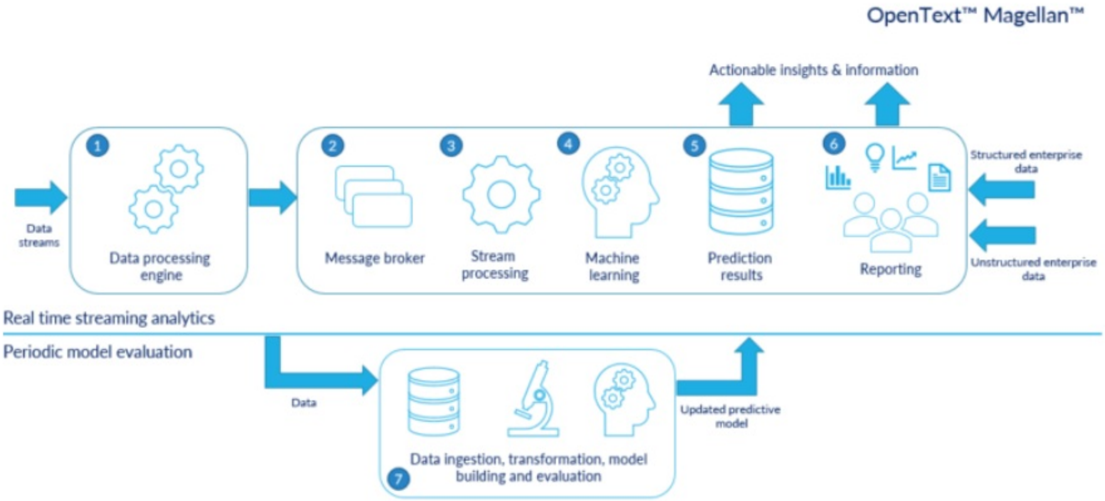

OpenText™ Magellan™
Actionable insights & information
1
2
3
4
5
6
Data streams
Data processing engine
Message broker
Stream processing
Machine learning
Prediction results
Reporting
Structured enterprise data
Unstructured enterprise data
Real time streaming analytics
Periodic model evaluation
Data
Updated predictive model
Data ingestion, transformation, model building and evaluation
7

Figure 16: Opentext Magellan NiFi Flow [58]

NiFi is used during the first Data Acquisition steps to collect data from multiple sources such as smart devices, social media, online transactions, and log monitoring [58]. The real-time data can then be combined with other historical data or other data sources before being feed into a downstream system [58]. Data is then being streamed by Kafka in Data Routing step and then being read and applied business rules by Spark Streaming API before it is being stored in a data lake [58]. Spark Streaming API will apply machine learning prediction model in Machine Learning step and then being saved in Prediction Results [58]. One the result is created, organizations can take quick decisions to provide business benefits and insights [58].

As a result, the organization was able to create low-cost solutions that has the flexibility and extensibility of open source software.

6.7.3 Social Competitive Intelligence Application at Compose

Compose is an IBM company launched in 2010 that offer databases as a service on the cloud that is production ready and is easy to manage. NiFi in being used in Compose as part of their Competitive Intelligence infrastructure that involves other software such as Twitter, IBM Watson, Redis, and MongoDB [59]. NiFi was used to extract filtered Twitter Stream data and attributes and send tweet data to IBM Watson for Sentiment analysis, as well as updating Redis for dashboards and reporting purpose and at the same time store all data in MongoDB [59] .

6.7.4 Real Time Streaming Architecture at Ford

Ford is an automobiles manufacturing company in the United States. Being a large company, data are stored and generated constantly in many applications within the enterprise such as assembly plants data, vehicle sensor data, dealership data, vehicle diagnostic data, and so on [60]. Ford came up with a solution called Real Time Streaming Architecture (RTSA) to allow data being flow between systems in real-time with proper data governance [60].

Ford's data are being sourced from Open XC which contains vehicle and phone application data into a private cloud via Cloud Foundry WebSocket or Event Hub [60]. Data from Websocket are streamed via Kafka into a cloud-based NiFi cluster together with the Event Hub data [60]. From the cloud-based NiFi Cluster, the combined data then flows to a private in-house NiFi cluster in Ford's data center and then publish to Kafka for downstream system distributions or being stored in Hadoop [60].

6.8 Work Breakdown

- Nhi Tran fa18-523-83

Use Case, NiFi Architecture image, NiFi Cluster Architecture's image, NiFi Download Installing and Getting Started

- Daniel Hinders fa18-523-56

NiFi Introduction, Big Data Challenges and NiFi, NiFi History, NiFi Architecture

- Both

NiFi Features

7 PyTorch

Divya Rajendran
divrajen@iu.edu
Indiana University
hid: fa18-523-57
github: ☁

Keywords: Deep Learning, gradient descents, Python, Neural Networks, Tensors, Computational Graph, Autograd, Auto Differentiation, Backpropagation

PyTorch [61] is a python based deep learning framework used for scientific calculations. It is popular among Neural Nets and Deep Learning developers due to its faster implementation of the algorithms and the maximum flexibility of its use. It is also used as a replacement for NumPy [62], an existing package, available in Python on scientific computing. The reason being, PyTorch mimicked most of NumPy’s functionality with an addition of increased speed by making use of the Graphical Processing Unit (GPU) [63].

Being written in a commonly used language by Machine Learning and Artificial Intelligence developers, Python, PyTorch has been gaining popularity since its inception in 2016 [64]. It is also less complex and easy to use when compared to existing Deep Learning frameworks like TensorFlow [65], Keras [66], Caffe [67], Chainer [68],MXNet [69], CNTK [70], Deeplearning4j [71].

PyTorch has been developed by the Artificial Intelligence group at Facebook [64] and is a successor framework of Torch [72] and has been built on it. Torch is a computing framework for scientific calculations wrapped in Lua, a programming language written in a general-purpose programming language C [73]. This framework, Torch, runs even in constrained platforms through LuaJIT [74] a platform specific compiler. It is used extensively to implement machine learning and deep learning algorithms [64]. It has a plethora of packages commonly used for Machine Learning, Signal Processing, Computer Vision among others, these have been inherited into PyTorch as well [72].

PyTorch seamlessly integrates all the packages which Torch offers and builds all of its functionality using Python, making Python its integral part. This makes the implementation of algorithms even faster than Torch. The main package of Torch and PyTorch is ***torch*** using which we can train neural networks, define the loss function and calculate the gradients for loss function [64].

7.1 Background

Before we start using PyTorch, we need to have a background or working knowledge on the below concepts.

7.1.1 Deep Learning

Deep Learning [75] is a branch of Machine Learning which takes its inspiration from the function and structure of a human brain [76]. It uses a subset of machine learning algorithms which processes input data in multiple layers through feature extraction and transformation and predicts the output labels [75]. It is being vastly used in different fields of computer vision, audio, and video signal processing, natural language and speech recognition and such [75].

7.1.2 Neural Networks

Neural Nets [77] is a collection of various connected nodes called neurons mimicking the neuron structure in the human brain. Each neuron receives an input, processes this input by performing a set of operations and then sends it to a next layer in the neural nets. Each layer has a weight and bias associated with it, on which a computation is done. This processing is based on some pre-defined function, a gradient descent algorithm, which transforms the input in each layer and this entire process is repeated a huge number of times until the error calculated is diminished [77].

7.1.3 Tensors

Tensor [78] is an inbuilt data structure in PyTorch and can be defined as a matrix of matrices or can be defined as a multi-dimensional array with dimensions greater than 3. So a tensor with 3 dimensions is called a 3-D tensor, a tensor with 4 dimensions is called a 4-D tensor and so on [78]. This tensor is used on a GPU which accelerates the computing process and calculation time on matrix operations when compared to existing NumPy’s ndarrays [63].

7.1.4 Computational Graph

A computational graph [79] is an internal representation of the operations performed during the neural nets training. It is also called a data graph and is also an inbuilt data structure in PyTorch. It consists of a set of nodes and edges, with nodes representing each operation and edges representing the values being sent from each operation from one layer to another layer in neural networks [79].

7.1.5 Auto Differentiation

Auto Differentiation [80] is a series of techniques used to numerically calculate the derivative of the transformation and loss functions defined in our neural networks [80].

7.1.6 Backpropagation

Backpropagation [81] is a technique in neural networks which calculates the gradient values for the peaks and troughs of the loss function and send the error values obtained to the previous layer going in the reverse or backward direction [81].

7.1.7 Autograd

Autograd [82] is a function in the ***torch*** library of PyTorch which calculates the gradients of the transformation and loss functions used in neural networks. The function Autograd uses a technique called tape-based Auto-Differentiation [83], a kind of the Auto-Differentiation technique. This technique saves the operations performed in each layer of the neural network in a reverse order, mimicking how a tape recorder works. This function also saves the gradients calculated in each layer of neural nets and replays them in a reverse order. Autograd’s implementation in PyTorch is faster than the same implementation in existing frameworks like TensorFlow [83]. We also use this function to train weights for neural networks through backpropagation [82].

7.2 Getting Started

If you have never used PyTorch before you need to install PyTorch. Check [84] for more instructions on the system requirements and different ways to install PyTorch [85].

In your system terminal or command prompt, enter the below line to install the PyTorch library.

```
$ pip install torch torchvision
```

You would see a message in your terminal or command prompt that the package has been installed.

To use PyTorch we need to first import a library called ***torch***. A sample initialization of tensors using PyTorch is done as below.

```
from __future__ import print_function
import torch

#### Creating an empty matrix
torch.empty(4, 3)

#### Creating a randomly initialized matrix
torch.rand(4, 3)

#### Constructing a tensor
torch.tensor(torch.rand(4, 3))
```

Step by step instructions for using PyTorch can be found at [86].

7.3 Implementation

PyTorch can be initialized using the package ***torch***. It contains the below functions.

1. ***torch.nn*** is a library of functions used to train or build the neural networks [87].
2. ***torch.nn.Linear*** is a function which applies a transformation, linear in nature, on the incoming values [87].
3. ***torch.nn.Sequential*** is a function used to initialize a model with a linear-stack of layers [87].

4. ***torch.nn.MSELoss()*** is used to initialize the loss function and also calculates the error between the input and the output layer in the neural nets [87].
5. ***torch.optim*** is a function which defines an optimizer algorithm which updates the weights at each layer [87].

Let us see an example of how to use these above functions to train a neural network as below.

7.3.1 Define a Neural Network

In this example, we would learn how to create and initialize a neural net model with two layers and how to apply an optimizer function on this neural network. This example is copied from [88].

```
###########################################################################
#    Title: Classifying Text with Neural Networks and Pytorch
#    Author: Mesquita, Déborah
#    Date: Oct 25, 2017
#    Code version: 1
#    Availability: https://github.com/dmesquita/understanding_pytorch_nn
#
###########################################################################
import torch
import torch.nn as nn

class OurNet(nn.Module):
 def __init__(self, input_size, hidden_size, num_classes):
     super(Net, self).__init__()
     self.layer_1 = nn.Linear(n_inputs,hidden_size, bias=True)
     self.relu = nn.ReLU()
     self.layer_2 = nn.Linear(hidden_size, hidden_size, bias=True)
     self.output_layer = nn.Linear(hidden_size, num_classes, bias=True)

 def forward(self, x):
     out = self.layer_1(x)
     out = self.relu(out)
     out = self.layer_2(out)
     out = self.relu(out)
     out = self.output_layer(out)
     return out
```

The previous code initiates the neural network with two hidden layers and one output layer. The function ***nn.Linear*** transforms the input layer data linearly, by multiplying the data with weights and adding a bias value. The transformation of our neural network is taken through the function ***forward*** which we defined above. In this function, we call the initialized values and functions and transform the output and return this value. [88]

To update the weights for our neural network, we use the optimizer algorithm ***Adaptive Moment Estimation (Adam)*** through the package ***torch.optim***. This optimizer holds the state of the object and the components based on the gradient computation. To calculate the loss we use a function called ***torch.nn.CrossEntropyLoss***. Below is a sample code on constructing the optimizer [88].

7.3.2 Constructing an optimizer

The step ***OurNet()*** is our initialization step for the neural nets we created. We get the initialized neural nets parameters and use it to initialize our optimizer function.

```
net = OurNet(input_size, hidden_size, num_classes)
optimizer = torch.optim.Adam(net.parameters(), lr=learning_rate)
criterion = nn.CrossEntropyLoss()
```

Our next step would be to load a dataset from sklearn datasets, or any other set of datasets and use it to test our functions written. The entire code for initializing neural networks, loss, and optimizer functions, including training and testing our models while applying them to a dataset of our choice is available can be found at [89].

7.4 Advantages of PyTorch

When we compare PyTorch over the existing frameworks like TensorFlow, Caffe, Keras, Chainer and such, the below are the most promising advantages.

1. Ramp Up Time is the time taken to execute all the threads or layers and their iterations. This time for code execution using PyTorch is much faster than its competitor TensorFlow [90], in that it uses dynamic creation of graphs rather than the static ones in TensorFlow [90]. Here, the compilation time for the code is much smaller for PyTorch, it uses GPU to increase the speed of execution and the graph is built during run-time, making it significantly faster than TensorFlow [90].
2. Debugging in PyTorch is easy as the underlying language is Python,

which is a common language used by developers and it is quite easier when compared to TensorFlow. We can use print statements to keep track of what values our variables take and to identify where our code fails [90].

3. Data Loading is much faster in PyTorch as the APIs for loading the data are designed in a manner to best utilize its parallelizing data loading capability. One can use either NumPy, Pandas or any other library of choice and can load data quite faster as PyTorch utilizes GPU [90].
4. PyTorch is highly extensible in that it has many custom extensions and it is easier to implement them for both GPU and CPU versions of its code [90]. It can be extended using NumPy, Scipy [91] and many other libraries [92]. Examples of extending PyTorch through Scipy and NumPy can be found at [92].

7.5 Drawbacks of PyTorch

When we compare PyTorch over its competitor like TensorFlow [93] we identify the below areas where TensorFlow outperforms PyTorch [90].

1. Coverage of functionality is less in PyTorch when compared to TensorFlow. The functions like NumPy’s flip along a dimension, checking NaN values, fast fourier transformation are not readily available in PyTorch whereas these functions and many higher functions are available in TensorFlow [90].
2. Serialization [94] can be defined as the process of translating the input data into a format which can be easily transferred across different platforms [94]. This capability in TensorFlow is better than PyTorch that it even is capable of saving the graphs can be saved as well. These graphs can easily be loaded into different languages like C++ and Java. This enables the deployments to not depend on Python alone [90].
3. Deployment is an activity which makes a code available to be used on a system where the code is placed. Since the code developed in TensorFlow can be easily saved in a format which can be used in different languages, its code can be easily deployed even in mobile applications [90].

8 Caffe - A Deep Learning Framework

Pramod Duvvuri
vduvvuri@iu.edu
Indiana University
hid: fa18-523-58
github: ☁

Keywords: Artifical Neural Network, C++, Computer Vision, Deep Learning, Machine Learning, Python

8.1 Introduction

The amount of data generated has increased exponentially and so did the advancement of computing power, both these have led us to the era of deep learning. This paper aims to summarize a deep learning framework known as ***Caffe*** ??? which was developed by a post-doctorate student ***Yangqing Jia*** at the University of California, Berkeley in 2013. It is written in C++ and is known for its fast execution and its Python interface allows it to be used by the vast majority of Python users. The framework has then been open-sourced, allowing many users to use, develop and contribute to improve the framework. The deep learning [95] revolution has led to the need for state-of-the-art implementations of Artificial Neural Network (ANN) architectures. These architectures are too hard to code from scratch for most people even with a conceptual understanding. The first deep learning framework to gain popularity was ***Theano*** ???. Theano was developed at the University of Montreal in 2007. It was primarily used by academic researchers at the university. Theano was built using Python which essentially made it slower for larger models, for production-grade models speed is imperative. This meant there was a need for a new popular and fast deep learning framework, especially in computer vision. Caffe was built using C++ and this made it very fast and ideally suitable for deployment in production. Caffe ??? at the time of public release or open sourcing had the best implementation of a Convolutional Neural Network [96], which is primarily used in solving computer vision problems. This public release made all the computer vision researchers and other people in the computer vision community adopt Caffe.

8.1.1 Artifical Neural Network

Neural Network is a machine learning algorithm which mimics the human nervous system. It consists of various nodes or artificial neurons that are interconnected and perform machine learning tasks. The advancements in computation and the introduction of Graphical Processing Units [97] (GPUs) have made it feasible for us to run such sophisticated algorithms. Neural networks have performed exceptionally well on data in comparison to other industry standard machine learning algorithms, which is why they have been adopted by both the academia and the industry. They require far more training data or examples than other algorithms and also require a considerable amount of computational resources to run.

8.1.2 Computer Vision

Computer vision [98] primarily consists of computers trying to extract or understand meaningful information from images or videos from the real world. It is a vast field that consists of many domains under it. The main goal of computer vision is to build a system that can mimic the human vision or visualize an image and understand the context and semantics. The input for such a system can take multiple forms such as a single image, sequence of images or a video or multi-dimensional data. Some of the most common are object recognition, object tracking, image segmentation, image processing. The deep learning revolution has essentially revitalized the field of computer vision. Many problems which were considered impractical have been solved using deep learning. Artificial Intelligence [99] and Computer Vision have a lot of common topics. Quite a few of these problems such as pattern recognition in vision were solved with the help of artificial intelligence and this made computer vision an integral part of artificial intelligence.

8.1.3 Deep Learning

Deep learning is a set of techniques or architectures that use Deep Neural Networks (DNNs) to solve problems in various fields such as computer vision, signal processing [100], natural language processing [101]. These DNNs can be used to solve any type of machine learning problem. Deep Neural Networks are Neural networks with more than two layers. There are usually two main types of in machine learning:

1. ***Supervised Learning***: In supervised learning, the data used to train our machine learning model is categorized into various categories also known as labels. The model is trying to learn how to categorize our data into these categories.

2. ***Unsupervised Learning***: In unsupervised learning, there are no categories. The algorithms are trying to find patterns or similarities in the data we give as input.

Any deep learning architecture at its core consists of a perceptron. A perceptron [102] is a machine learning algorithm which was made to mimic the function of a human neuron. Deep learning architectures use multiple such perceptrons or nodes as layers which is the reason these are referred to as deep learning architectures. In computer vision problems each layer serves a specific purpose and the combination of all these layers aids in solving specific problems.

8.2 Installation

To use Caffe it is recommended to install a containerized image of it. This can be done using the help of Docker ???. The official Caffe image can be found on DockerHub and can be installed using the GUI. It can also be installed using the following command with docker already running on your local machine. In your docker terminal please paste the following command:

```
$ docker run -ti bvlc/caffe:cpu caffe --version
```

```
caffe version 1.0.0
```

As indicated the latest version is 1.0.0, the above command is mostly used if the machine does not contain a GPU [97]. If your machine contains a dedicated GPU then another command can be used to install Caffe using Docker. In your docker terminal please paste the following command:

```
$ nvidia-docker run -ti bvlc/caffe:gpu caffe --version
```

```
caffe version 1.0.0
```

With Caffe now installed it can be used with an Interactive Python notebook. The below command must be used to launch an IPython notebook in the docker terminal and then import caffe in the interactive Python (IPython) notebook ??? before we can write any code in Caffe.

```
$ docker run -ti bvlc/caffe:cpu ipython
[1] import caffe
```

8.3 Caffe Tutorial

In this section, we try to solve the MNIST [103] classification problem using Caffe. We shall define files necessary to train a model that classifies hand written digits and recognizes them. Before we can run our model we must define the below files in the folder where Caffe is installed. The below code defines the different layers and the loss function in each of them.

```
###########################################################################
#    Title: MNIST Classification using Caffe
#    Author: GitHub
#    Availability: https://github.com/BVLC/caffe/tree/master/examples/mnist
#    Filename: lenet_train.prototxt
###########################################################################
name: "LeNet"
layer {
name: "data"
type: "Input"
top: "data"
input_param { shape: { dim: 64 dim: 1 dim: 28 dim: 28 } }
}

layer {
name: "conv1"
type: "Convolution"
bottom: "data"
top: "conv1"
param {
lr_mult: 1
}

param {
lr_mult: 2
}

convolution_param {
num_output: 20
kernel_size: 5
stride: 1
weight_filler {
type: "xavier"
}
```

```
bias_filler {
type: "constant"
}
}
}

layer {
name: "pool1"
type: "Pooling"
bottom: "conv1"
top: "pool1"
pooling_param {
pool: MAX
kernel_size: 2
stride: 2
}
}

layer {
name: "conv2"
type: "Convolution"
bottom: "pool1"
top: "conv2"
param {
lr_mult: 1
}
param {
lr_mult: 2
}
convolution_param {
num_output: 50
kernel_size: 5
stride: 1
weight_filler {
type: "xavier"
}
bias_filler {
type: "constant"
}
}
}

layer {
name: "pool2"
type: "Pooling"
bottom: "conv2"
top: "pool2"
pooling_param {
pool: MAX
kernel_size: 2
stride: 2
}
}

layer {
name: "ip1"
type: "InnerProduct"
bottom: "pool2"
top: "ip1"
param {
lr_mult: 1
}
param {
lr_mult: 2
```

```
}
inner_product_param {
num_output: 500
weight_filler {
type: "xavier"
}
bias_filler {
type: "constant"
}
}
}

layer {
name: "relu1"
type: "ReLU"
bottom: "ip1"
top: "ip1"
}

layer {
name: "ip2"
type: "InnerProduct"
bottom: "ip1"
top: "ip2"
param {
lr_mult: 1
}
param {
lr_mult: 2
}
inner_product_param {
num_output: 10
weight_filler {
type: "xavier"
}
bias_filler {
type: "constant"
}
}
}

layer {
name: "prob"
type: "Softmax"
bottom: "ip2"
top: "prob"
}
```

```
###############################################################
#    Title: MNIST Classification using Caffe
#    Author: GitHub
#    Availability:
#    * https://github.com/BVLC/caffe/tree/master/examples/mnist
#    Filename: lenet_solver.prototxt
###############################################################
# The train/test net protocol buffer definition
net: "examples/mnist/lenet_train_test.prototxt"
# test_iter specifies how many forward passes the test should carry out.
# In the case of MNIST, we have test batch size 100 and 100 test iterations,
# covering the full 10,000 testing images.
test_iter: 100
# Carry out testing every 500 training iterations.
test_interval: 500
# The base learning rate, momentum and the weight decay of the network.
```

```
base_lr: 0.01
momentum: 0.9
weight_decay: 0.0005
# The learning rate policy
lr_policy: "inv"
gamma: 0.0001
power: 0.75
# Display every 100 iterations
display: 100
# The maximum number of iterations
max_iter: 10000
# snapshot intermediate results
snapshot: 5000
snapshot_prefix: "examples/mnist/lenet"
# solver mode: CPU or GPU
solver_mode: CPU
```

```
cd $CAFFE_ROOT
./examples/mnist/train_lenet.sh
```

8.4 Architecture

The Caffe architecture mainly consists of layers or it had a layer-wise design all designed and built from scratch using C++ and the CUDA ??? architecture with various interfaces to write code in MATLAB [104] and Python. This architecture at the time of its creation was considered really good but since then newer deep learning framework's such as ***Tensorflow*** which was created by Google has a much flexible design. This flexible design is with respect to the various nodes in the Artificial Neural Network [105] (ANN) since ANNs primarily consist of layers and each layer has multiple nodes. The ability to flexibly design these nodes was very important to the researchers since this helped them achieve higher accuracy rates for their model and this helped with benchmarking and comparison with other similar models aimed to solve similar tasks.

8.5 Applications

Some of industry grade production levels applications ??? of Caffe are:

- Facebook used Caffe to generate alternate texts who people who are visually challenged. All the photos uploaded to Facebook were run through a caffe model to generate such text. Facebook also used caffe to detect objectionable content. As the amount of data on social media increases so has the need for a protocol to regulate and report objectionable content risen.

- Pinterest used Caffe for object detection in the images. All the images that were uploaded were run through a model for object detection. The Caffe deep learning model for visual search could search over billions of images in just under 250 milliseconds. As many as 4 million images are uploaded onto Pinterest on a daily basis.

- Yahoo used Caffe for user recommendations in Japan. The news feed on Yahoo had stories curated using a caffe model and also made restaurant suggestion using photos. Yahoo also had models to automatically arrange photos of users into albums.

8.6 Limitations and Comparisons

Caffe was developed by a post-doctorate student and then open sourced in 2013, the deep learning revolution had just begun when Caffe was launched. After its initial launch, there were more than 150 developers who were actively contributing to the framework. At the same time, large companies such as Amazon, Facebook, Google, Microsoft had all begun working on deep learning frameworks which suited their needs and fit perfectly in their respective technology stacks. The public launch of Tensorflow [65] by Google made a lot of people adopt it quickly since Google had consistently invested more time and money in the development and maintenance of this framework. Currently, there are more 1500 people who actively contribute to the Tensorflow framework. ***PyTorch*** [64] is another popular deep learning framework which was developed in 2017 at Facebook and had dynamic graph computational ability which was lacking in Tensorflow. PyTorch was received very well by the research community since it combines two of the most popular languages used by the artificial intelligence community Torch and Python. Caffe main strength was its implementation of a fast CNN [96] and ready to use GPU [97] support. Although CNNs could be used for Natural language processing [101] (NLP) tasks there were other deep learning architectures which were more suitable for NLP related tasks. Other deep learning frameworks such as Theano and Torch had better implementations of these architectures hence they were preferred to Caffe. This meant that caffe’s usability was very limited outside computer vision tasks. Caffe does not offer multi-GPU support. The exponential rise in our data meant we needed to use multiple powerful GPUs to train our models. Hence modern frameworks such as Tensorflow and PyTorch were built to serve as an all-purpose deep learning framework that had the state-of-the-art implementations of the latest deep learning architectures and also had out of the box multi-GPU support. Gradually people adopted these frameworks over Caffe. These frameworks had a much more robust architecture. Caffe was restricted by the format for input and output. It only supports one output format called HDF5 [106]. Caffe had less documentation and fewer hands-on tutorials which made it less developer friendly [107].

8.7 SUMMARY

Caffe was mainly intended to support vision tasks and was not suitable for other tasks such as speech recognition, language modeling and time series data. This made its applications and also usability limited. But the lack of proper documentation and examples made it harder to adopt for the community. All modern deep learning frameworks were built to overcome the limitations of Caffe. They also borrowed its excellent CNN implementation. More people have since then moved to Tensorflow for academic research. Caffe has contributed a lot the computer vision and deep learning communities. Caffe helped these communities make valuable contributions to research with its fast execution times. Caffe2 [108] was a project that was started at Facebook after the success of Caffe. Caffe2 was open sourced by Facebook in April 2017. By the end of March 2018, Caffe2 was merged with PyTorch by Facebook. These days the choice of a deep learning framework, when you have huge amounts of data, is either PyTorch or Tensorflow. Both Google and Facebook constantly keep updating these frameworks from the feedback they receive from the developer community to make it more developer friendly with the ability to visualize the computational graphs [79]. The goal now is to make deep learning more accessible to everybody and reduce the steep learning curves when it comes to these deep learning frameworks and caffe has contributed invaluably to achieve this goal.

9 Natural Language Applications and Challenges within Big Data

Jay Stockwell
jaystock@iu.edu
Indiana University
hid: fa18-523-61
github: ☁

Keywords: Natural Language Processing, Natural Language Understanding, Natural Language Toolkit, Deep Learning, SyntaxNet, Part of Speech tagging, Hadoop

9.1 Introduction

Organizations have recently begun to harness the immense power of big data and how the concept can prove to be a beneficial component. The term big data used to be a scary term that elicited feelings of consternation and anxiety, but with organizations experiencing exponential growth in data volume, the term has become mainstream and widely accepted. Big data is largely unstructured text and constantly in a state of enormous flux, which is why NLP offers many opportunities to tap into this vast data resource [109]. This paper will explore the evolving, and sometimes challenging relationship between NLP and Big Data, the problems that NLP can solve, which applications are leveraged, and how the data can be transformed into a presentable format for consumption.

Big data “describes the growing volume of structured and unstructured, multi-source information that is too large for traditional applications to handle [109].” The volume of today’s data is on an unprecedented growth trajectory due to the ample methods to collect and analyze data. The Internet of Things, mobile devices, sensors, cameras, and software logs are just some examples of non-traditional methods of data collection are contribute to the abundance of information today ???. There’s no end in sight to exponential growth that is projected over the next several years.

Natural Language Processing is a relatively new concept and is gaining momentum in the use of text analysis and presentation. Elizabeth Liddy from Syracuse University provides a great definition:

> ***"Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications" [110].***

Essentially, NLP's goal is to achieve a level of text analysis and processing that mimicks very closely the way that humans process language. While NLP has truly made significant progress over the last several years, the challenge of deciphering the exact context and inference of text is still an area of which NLP is still improving [110].

9.2 Natural Language Challenges

As previously mentioned, there are many challenges that face NLP today. Some of the challenges pertain to how NLP parses sentences, deals with cultural differences, language translation, conversational issues (ie. whether the statement is a question or answer), and sentiment. In order for NLP to function in an effective manner within Big Data, there needs to be a process put in place to assist with these issues.

Parsing, the ability to deconstruct a sentence into its parts, is a major challenge facing NLP. Parsing can lead to ambiguity because it can be difficult to detect the correct syntax, and/or the exact interpretation of each word. Prepositions within a sentence can cause confusion because it can be hard to determine which word is being modified. as explained by Armando Viera and Bernadete Ribeiro:

> ***"..the sentence"Alice drove down the street in her car" has at least two possible dependency parses. The first corresponds to the (correct) interpretation where Alice is driving in her car; the second corresponds to the (absurd but possible) interpretation where the street is located in her car. The***

> ***ambiguity arises because the preposition in can modify either drove or street. [111]."***

Part of Speech tagging is another NLP concern. Part of speech tagging is the process of assigning a part of speech definition to a word within a sentence. [112]. Common parts of speech include nouns, verbs, adjectives, and adverbs. Challenges can arise due to the lack of context. There are many words that have multiple meanings, which can lead to ambiguity when computers try to assign a descriptor. For example, the word chair can have multiple meanings; you can chair(verb) or lead a meeting, or you can sit in a chair (noun) [112].

Another challenge is Natural Language Understanding (NLU). NLU pertains to the ability of a computer to comprehend the sentence structure and intended meaning of human languages, which in turn allows humans to fully communicate and interact with machines using real sentences [113]. This concept is gained interest in recent years due to the many possible ways this can be leveraged within applications on a commercial scale.

With the rise of Artificial Intelligence (AI), NLU has risen in complexity due to the extraordinary level of computations involved.
The concept is considered one of the most challenging problems in the AI world. These types of problems require intense computational effort and also require human intervention and resources as they cannot be solved by machines alone. Challenging AI problems are known as AI-Hard or AI-complete.

> ***"In the field of artificial intelligence, the most difficult problems are informally known as AI-complete or AI-hard, implying that the difficulty of these computational problems is equivalent to that of solving the central artificial intelligence problem—making computers as intelligent as people, or strong AI [114]. To call a problem AI-complete reflects an attitude that it would not be solved by a simple specific algorithm [115]."***

In order to NLU to properly function, machines must be programmed to understand text. NLU must follow all the text translation rules that any human would follow if they were reading through any type or text document. Machines must have the ability to mimick human abilities and human intellectual skills such as reason, common sense, and intuition that relate to how humans perceive language and social intelligence concepts [115]. NLU requires an enormous amount of work in order to prove effective. NLU applications require extensive data gathering and subject matter investigation and research in order to properly train the system to perform [116].

9.3 Natural Language Processing Solutions

There are some solutions in place to address the challenges facing NLP. With respect to the parsing problem, there is a relatively new application designed by Google called SyntaxNet. SyntaxNet is based on the TensorFlow open source library readily available to users for designing deep learning models. With SyntaxNet, Google employed a normalized neural network model that provides an output of possible syntactical possibilities or hypotheses given a group of words [111]. SyntaxNet runs the model multiple times and discards hypotheses that are ranked lower and appear to be unlikely candidates. As far as parsing, SyntaxNet has developed a reputation for being the best parser, being known to sometimes exceed human accuracy, and recently made available in 40 languages. [111].

The latest version of SyntaxNet can be trained against a separate data set, and individuals have a good deal of freedom in tweaking the parameters of the model to better fit the particular nuances of their datasets. SyntaxNet includes a model built specifically for the English language entitled Parsy McParseface and can be used to analyze English texts right out the box [117].

Members of the Stanford University’s Natural Language Processing Group has developed a part of speech tagger that works remarkably well. The application runs on Java and is somewhat memory intensive, requiring upwards of 60-200 Mb of memory to function efficiently, with around 1 Gb recommended in order to train a dataset [118]. The latest download contains three distinct tagger models for the English Language as well as an Arabic, German, Chinese, and French model. These models can be retrained on any language. During an experiment against Penn Treebank WSJ data, the Stanford POS tagger returned an impressive per-position tag accuracy of 97.24% [119].

As NLU continues to become more mainstream, many companies are embedding proprietary NLU algorithms within their products to further enhance their overall NLP capabilities. Some of the companies and their associated applications are Apple (Siri), Google(GoogleNow, Google Search), Microsoft(Cortana), and IBM (Watson, DeepQA) [116]. These products are just the beginning of a new wave of technology that will surely influence how organizations incorporate NLP into their business processes and decisions.

In order for NLP to succeed, it needs to be trained against very large datasets or corpuses. A corpus is a collection of written texts used for research or investigation purposes. Two examples of widely used corpuses are:

> ***"The Google n-gram corpus, a trillion word database containing phrases (up to 5 words long) occurring on public Web pages. The USENET corpus, a 25 billion word (compressed) corpus containing public USENET postings on 47,680 English language, non-binary-file newsgroups between Oct 2005 and Jan 2010 [120]."***

In order to write effective programs against such large big data sets, a new platform has been developed entitled Natural Language Toolkit (NLTK). The platform is completely open source and supported by a strong community of users. Users can leverage the platform to build applications using the Python programming language. NLTK provides user with a simplistic interface that can access over 50 different corpora data sets linguistic and lexical data resources [121]. Users can also leverage several built-in libraries for classification, tokenization, tagging, parsing, semantic reasoning, and powerful wrappers to utilize with NLP libraries [121]. NLTK comes with a comprehensive learning guide to assist users, specifically those with little to no Python experiense, to get up to speed with the various syntax commands that are used within NLP.

Some organizations have started to leverage Hadoop, an open source processing framework that works extremely well against very large textual datasets that require enormous amounts of computational power. Hadoop can also act as a data management application providing massive storage space and can handle numerous concurrent processing tasks. Hadoop's ability to work with various kinds of structured and unstructured data make it an ideal application for NLP, as it can handle an exorbitant amount of data from sources such as internet clickstream records, web server application logs, social media sites such as facebook and twitter, customer emails, and sensor information from the internet of things [122]. The aforementioned deep learning algoritms can be set up to run in a Hadoop environment to leverage its extraordinary size and computational power.

In recent years, with the precipitous rise of data science and the use of algorithms for predictive analysis among other areas, one concept in particular, Deep Learning, has emerged as a possible solution to answering some of the aforementioned challenges facing NLP.

> ***“Deep learning (DL) has had a tremendous impact on natural language processing (NLP). After image and audio, probably this is the area where DL has unleashed the most transformative forces. For example, almost all projects related to NLP at Stanford University, one of the most respected institutions working on this area, involve DL research [111].”***

Deep learning, or Deep Neural Networks as it is sometimes referred to, is a branch of AI. The deep neural networks go through a series of tranformation steps on the inputted data and leverages the what it learns to build a comprehensive statistical model [123]. The model will continue to run through a series of iterations until it returns an acceptable level of accuracy. Deep learning algorithms can be trained against big data sets just like other standard algorithms and have proven to be effective in this manner. For example, "Trained on movie subtitles, language models are able to generate basic answers to questions about object colors or facts [111]. Deep Learning networks can be coded using the NLTK and Hadoop platforms mentioned above to tackle the numerous challenges facing NLP today.

9.4 Conclusion

In conclusion, the NLP field has proven its importance in the data world by allowing for entities to analyze and evaluate many aspects of the different components of language. The challenges that have arisen from leveraging NLP such as parsing and part of speech tagging have beem met with the development of new applications such as Google’s SyntaxNet, Stanford’s Part of Speech tagger, and NLTK. Since NLP is still in it’s early stages, there will continue to be new challenges, and just like SyntaxNet, new applications will meet these challenges and make NLP even more effective. These applications leverage algorithms such as neural networks and deep learning to facilitate the effective training of datasets. Today's organizations deal with, and store enormous amounts of textual data from many different sources. Because this information is comprised of primarily text, organizations that leverage big data infrastructures and work with this type of data are starting to understand the implications that NLP provides in evaluating their growing stores of data to detect patterns, connections and trends within their various data sources [124]. Big data storage has become easier to manage due to new open source database management systems such as Hadoop. The volume of NLP data will continue to grow exponentially and new processing and storage management technologies will need to be designed to be not only scaleable, but have the capacity to work with ever changing business demands. NLP will continue to grow and we need to be ready to meet the new challenges that the emerging technology presents.

10 Natural language text processing and Language generation

Sahithya Sridhar, Prajakta Patil
sahsrid@iu.edu, patilpr@iu.edu
Indiana University, Bloomington
hid: fa18-523-67, fa18-523-65
github: ☁

keywords: Natural language processing, Machine learning, Big Data, Yelp data

> ***"The goal of NLP is to accomplish human-like language processing [125]".***

Natural language processing (NLP) forms the link between machine language (binary code) to natural languages (which we humans speak). It borrows elements from artificial intelligence, computer science and information engineering. The success of NLP is highly dependent on how one can successfully program computers to process, analyze and interpret huge amounts of natural language data. Key opportunities/challenges in this inter-disciplinary field involves speech recognition, language interpretation and generation of natural language from a machine representation. With NLP, it is possible for computers to read text, hear speech, interpret, measure sentiment and determine which parts are important [126].

10.1 Purpose of Natural language processing

NLP was initially called NLU (Natural language understanding). A natural language understanding system will be used to paraphrase the input text, translate text into another language, analyze the content of the text and draw conclusions. As NLP still cannot draw inference from the text, it is still dependent on NLU. One example of such difficulty is understanding if a certain word in a sentence is a noun or a verb. For example, the word ***leave*** can be either a noun or a verb depending on the context of the sentence [126].

It is well known [126] that quite a bit of information which exists in our world is very unstructured. The challenge is to make a computer understand this and extract data from it. Humans have been writing down things for thousands of years but computers cannot yet truly understand the language as we do. The trick is to break down the process of understanding English into smaller pieces and understanding each piece separately [126].

According to [125] NLP has two distinct focuses:

- Language processing: Produces a meaningful representation by analyzing the language. This is similar to a machine reading the text.
- Language generation: Language is produced from representation. This requires a planning capability. This is similar to a machine writing the text.

The text is manipulated for abstractions and indexed for automatic knowledge extraction. Producing text in a desirable format is one of the major research areas in NLP. Structuring of large bodies of textual information to retrieve an information is classified under the natural language text processing [125].

> ***"The central task for natural language text processing systems is the translation of potentially ambiguous natural language queries and texts into unambiguous internal representations, on which matching and retrieval can take place [125]".***

10.2 Levels of NLP

The various important terminologies of NLP are:

10.2.1 Phonology

This refers to the understanding of sound of individual words, and groups of words when spoken together in a sentence of sound. There are three types of rules used here [125]:

- ***phonetic rules:*** for sounds within words.
- ***phonemic rules:*** for variations of pronunciation when words are spoken

together.

- ***prosodic rules:*** for fluctuation in stress and intonation across a sentence.

10.2.2 Morphology

This comprises of nature of words where the different parts of the words represent smallest units. A word can be broken down into its constituent morphemes to understand its meaning [125]. This involves understanding suffixes/prefixes attached to the word, whether the word is singular or plural, the roots of the word.

10.2.3 Lexical

This interprets the meaning of the sample word. Words that have only one meaning is replaced with the semantic representation of that word. This requires the word to have a simple or a complex lexicon [125].

10.2.4 Syntactic

This rectifies the grammatical mistakes in a sentence. The output reveals a pattern that has a structural dependency between the words which in turn affects the choice of the parser [125].

10.2.5 Semantic

This determines the meaning of the word by looking at the interactions among word-level meanings in the sentence. An example of this would be trying to understand if the word ***honey*** is used as a noun or an adjective in the given sentence [125].

10.2.6 Pragmatic

This is used to extract information from the text. The main goal is the use of the context over the content of the text for better understanding. This helps in determining how the word is used than what is captured in the plain text of a sentence [127].

10.2.7 Discourse

This focuses on the properties of the text that convey the meaning by making connections between component sentences. This helps to understand how a certain set of sentences convey a part of the story.

> ***“Discourse/text structure recognition determines the functions of sentences in the text, which, in turn, adds to the meaningful representation of the text [125]”.***

The current NLP system tends to implement these models that are used for processing the lower levels as they have been well researched and implemented. In addition, the applications do not require them to process at the higher levels[125].

10.3 Natural Language generation

According to [128], this part of NLP happens in four phases:

- Identifying the goal: This defines if we want to generate language in form of written text or spoken words.
- Evaluating the situation and planning to achieve the goal: This involves identifying how the goal can be achieved by breaking it into individual tasks.
- Evaluate the available resources: This determines whether we want to generate written text or spoken word and also evaluate if this can be achieved with current communication devices.
- Execute the plans as text.

10.4 Approaches to NLP:

According to [125] some approaches of NLP are as follows:

- Statistical Approach: Various mathematical techniques are used here. Observable data is used mainly as evidence. The Hidden Markov model is a widely used statistical approach. This is used in speech recognition, parsing, statistical grammar learning etc.

- Connectionist: Models are developed based on the linguistic phenomena. This combines statistical model with other theories.
- Symbolic Approach: Performs deep analysis of linguistic phenomena based on the human developed rules and lexicons. Logic or rule based systems and semantic networks are good examples of symbolic approach.

> ***“Symbolic approaches have been used for a few decades in a variety of research areas and applications such as information extraction, text categorization, ambiguity resolution, and lexical acquisition. [125]”***

Statistical Approach has been proven to work best on lower levels of NLP. This means analysis on level of parts of words, words and sentences. Symbolic Approach works on both lower and higher levels of processing. Higher level of processing is analysis based on inferences from set of sentences and understanding of whole text in a given document [125].

10.5 Related Work

Significant research has been done on NLP since the 1940s. This has resulted in the development of tools such as Sentiment Analyzer, Parts of Speech Taggers, Emotion Detection, Semantic Role Labelling etc. This has helped in the creation of real world applications such as Google search, Apple’s Siri, Amazon’s Alexa etc. Applications such as these have further increased interest in NLP as a useful research topic [125].

Researchers working on tools like Sentiment Analyzer, parts of speech, Emotion detection, Semantic Role Labelling etc have made NLP a good topic for research [128].

- Sentiment Analyzer: Analyses the document for positive and negative words. It uses the sentiment lexicon and the sentiment pattern database to analyze the sentiments [128].
- Parts of speech: It can efficiently tag and classify words as nouns, adjectives, verbs etc. Parts of speech are assigned to tokenized data. Taggers are already present for the European languages. Research is

being done on making parts of speech taggers for other languages like Arabic, Sanskrit [129], Hindi [130] etc.

- Emotion Detection: This is just like sentiment analysis, but is used for analysing emotions on social media platforms. It categorizes statements into six groups i.e. anger, disgust, fear, happiness, sadness and surprise based on emotions in the text [128].
- Sematic Role Labelling: SRL works by giving a semantic role to a sentence just like assigning roles to words that are arguments of a verb in the sentence. This helps to understand which words indicate action being done, which ones indicate result of action and words that show who is doing the action etc [131].

10.6 Applications of NLP

NLP can be applied into various areas like Machine Translation, Email Spam detection, Information Extraction, Summarization, Question Answering etc. Some more applications of NLP are:

- Machine Translation: As the name suggests, Machine translation is translating a phrase from one language to another with the help of engines like Google Translate. The major challenge in machine translation is in keeping the meaning and the grammatical structure of the translated language intact [128].
- Text categorization: This involves splitting the large volume of data into several categories. This usually works by either looking at the email subject, the content or the sender blacklisted by receiver. It is also used to categorize communication so as to forward them to the appropriate department when used in a business setting. [128].
- Spam Detection: Various machine learning techniques like Rule Learning, Naïve Bayes, Memory based Learning, Support vector, Decision Trees, Maximum Entropy Model etc. are used to detect the spams. This is very similar to text categorization[128].
- Chatbot: A chatbot is a computer trying to mimic human like interaction /communication. We see many applications of these currently on different banking/ecommerce websites. There is lot of on-going research to make this even more capable [128].

10.7 Problems with NLP: linguistic variation and ambiguity

There are certain problems in NLP that reduce the efficacy of textual information retrieval. Linguistic variation and ambiguity are some of the problems in NLP. Linguistic variation is an issue when the same words or expressions are used to communicate the idea. Multiple interpretations is one of the main problems with Linguistic variation. Linguistic Variation provokes the omission of certain documents that are relevant and ambiguity implies when a document has duplicate words or words that are not related [132].

Example 1: A notebook was the present that the teacher gave him, when we were present in the class.

Here the word ***present*** has different meanings both as an adjective and as a noun. The word present plays different morph-syntactics depending on the situation causing ambiguity problems.

Example 2: He ate food on the car.

Ambiguity is produced here again, as this sentence could mean that he ate the food which was present in the car, or he ate food when he was driving the car.

Example 3: I went to the bank.

Here the word ***bank*** could mean a place where we save money and make transactions or the ‘bank’ of a river.

These examples show that automated process is not easy and that how complex the language is. Statistical processing of NLP and specifically machine learning has as improved understanding of learning language by training on text corpus [133].

10.8 NLP in textual information retrieval

When the user gives a query, the following task are performed as a part of NLP’s textual information retrieval [132]:

- Index is created for the descriptions of a document based on the NLP techniques.
- When a query is given by a user, the system analyses it and transforms it such that it is similar to what is represented in the document.
- The description of each document is compared by the system with the query given by the user, and those documents that have the description close to the users query are retrieved. There are different methods used to perform job of matching a query and document. Boolean method does this by trying to do an exact match. Vector space model converts the query and documents the vectors that can be stored as matrices. It then finds the similarity by calculating the cosine angle between the query and the document vector. Another method that is used to perform this is language model. Language model tries to find the probability of the document generating a query. This model depends on training the algorithm on a large corpus of text data/documents.
- The results are shown in the order of the similarity. Google uses PageRank to rank the documents in terms of similarity to query from the user.

10.9 Statistical processing of natural language

> ***"This is a very simple focus based on the bag of words. In this approach, all the words in a document are treated as its index terms. Moreover, each term is assigned a weight in function of its importance, usually determined by its appearance frequency within the document. In this way, the word's order, structure, meaning, etc., are not taken into consideration [132]".***

The document processing model involves document pre-processing and Parameterization.

- Document pre-processing: Prepares the documents by removing those elements that are superfluous. There are three basic phases here:
    - Removing headers, tags etc. from the document which are not for indexing.
    - Tokenization splits text into sentences and sentences into

    words.
    - Standardizing the text by checking for capitalized or non-capitalized letters, numerals, dates etc. Making all words lowercase helps treat words such as ‘Hi’ and ‘HI’ same.
    - Stemming the terms by reducing the words to the roots.This operation removes suffixes, prefixes etc.
- Parameterization: Assigns weights to the relevant terms present in the document.

One of the most used methods to estimate the importance of a term is the TFIDF system (Term Frequency, Inverse Document Frequency).

> ***“It is designed to calculate the importance of a term relative to its appearance frequency in a document as a function of the total appearance frequency for all of the corpus’ documents i.e. the fact that a term appears often in one document is indicative that that term is representative of the content only when that term does not appear frequently in all documents. If it appeared frequently in all documents, it would not have any discriminatory value [132]”.***

Two commonly used techniques in statistical processing are:

- Detecting N-Grams: This involves identifying compound words, proper nouns etc., to be able to process them as single words. This is done by determining the probability of the compound words like European union etc. This allows to maintain the sequence of words which is different as compared to the just bag of words which does not maintain order of words.

- Stopword List: A list of empty words, with very little semantic values. Deleting these terms avoids duplications and noise [132].

Statistical evaluation in NLP systems is used to evaluate the efficiency, accuracy and robustness. It can be done using the below methods that do it in different ways [132]:

- Descriptive Statistics: This method calculates Word error rate, Accuracy rate, Recall and Precision

- Estimation: This method calculates the confidence interval for true accuracy rate with certain probability.
- Hypothesis Testing: When different NLP systems are applied to same set of data,we want to compare their performance and suggest if one method is better than the other. This method allows us to do it by comparing certain performance parameters between two methods and performing hypothesis testing to tell if there is real statistical significance between the results or not.

10.10 Linguistic processing of natural language

In the linguistic process, the words determine how they are related and used together in making grammatical units, sentences etc. Parsers are created and applied to demonstrate the text’s syntax structure. The method used to create the parser vary.To determine the semantic structure of the words, certain tools are used. The most often used tool is the is the lexicographic database WordNet [132].

> ***"This is an annotated semantic lexicon in different languages made up of synonym groups called synsets which provide short definitions along with the different semantic relationships between synonym groups [132].***

10.11 NLP for Big data

Big data is the most text based content which is constantly growing and is quite unstructured. Every industry generates a large volume of text information, documents, notes, emails, patents, patient information etc. As most of these are text based data, NLP presents an opportunity to take advantage of this situation to reveal patterns and trends [134].

- Interactions: Interactive applications are becoming more common these days like Microsoft’s Cortana, smart phone assistants, language translation programs etc. These applications use NLP.
- Business Intelligence: NLP for big data enables the user to retrive the documents that they are looking by not limiting their search to exact

keywords. NLP enables them to search using their own words and tries to retrieve documents with that search.

- Market Research: With the growth of internet, social network is full of rich, noisy information. The brands and organizations can determine what is said about their products and services by using NLP for their market research analysis.

There are lot of NLP libraries written to process big data. Some of which are:

- CoreNLP: It was originally written in java that can also support multiple languages like python. It is well known for its speed and its precise results [135].
- TextBlob: Addition of components like sentiment analyzer becomes very easy with Textblob [135].
- Gensim: Used best for topic modeling and comparing the document similarity [135].
- Spacy: It is a new library which has a very high performance [135].
- NLTK: Most commonly used NLP library.

> ***"Natural language tool kit's (NLTK) modular structure helps comprehend the dependencies between components and get the firsthand experience with composing appropriate models for solving certain tasks [135]."***

10.12 NLP IN YELP DATA REVIEW

Yelp is a social networking site that combines business listing with social elements. It helps in finding local businesses such as restaurants where customers can leave feedback on their experience. This feedback helps the other customers of what they might expect from the place. Reviews or feedbacks are in the form of starts. Higher the starts, better the place is. The reviews also help the business to improve their standards in case there is a lower review. We can use NLP to analyze the text reviews to interpret restaurant reviews on Yelp through a sentiment analysis model.

The first step in NLP depends on the application. Voice based systems like Google Assistant or Alexa translates words into text using Hidden Markov

Models. The language and context is then understood through a series of coded grammar rules that rely on algorithms that incorporate statistical machine learning. Another important step is Semantic analysis which helps interpret human sentences logically [133].

Here we will try to predict whether a user liked a local business or not based on their review on Yelp. A simple text classifier will be built based on Python's Pandas, NLTK and Scikit-learn libraries. According to [136], the plan would be to start with a dataset containing 5000 reviews with the following info:

- ID of the restaurant under review
- ID of the posted review
- Date the review was posted
- The star rating provided.
- The text for the review.

NLTK library would be used to process the text and get basic information and insight on the data. The next step would be to visualize the data by utilizing histogram grids for every star rating. The goal would be to identify which feature of the review is useful in finding correlations in the data frame. Once we derive some useful correlations, they could be visualized. According to [137] Challenges faced in NLP text processing are:

- Scalability and portability.
- Certain techniques are too expensive.
- Not very reliable as of now.
- Speech/text processing.

10.13 Building a NLP pipeline

It would be really helpful if a computer could understand what the humans are trying to say. NLP helps the computer to read and understand all the data. By applying NLP techniques, we will be able to save a lot of time to the projects. But parsing the English language with a computer has its own complications. Hence we will breakdown the process of understanding english, into small chunks and see how it performs in understanding and giving a correct output [136].

Let’s take a paragraph: ***Delhi is the capital of India and one of the most populous city in Asia. This has been a great settlement for several kings including the Mughals. The original name was Indraprastha.***

This paragraph contains several important and useful information. It would be great if a computer could read and understand that Delhi is a city in India, it was ruled by Mughals etc. But to get there we have to train the computer on how to read the sentence [136].

For a computer to understand the text and extract data we need to do some of the following steps [136]:

10.13.1 Sentence Segmentation

First step is to break the text in the paragraph into separate sentences like:

- ***Delhi is the capital of India and one of the most populous city in Asia.***
- ***This has been a great settlement for several kings including the Mughals.***
- ***The original name was Indraprastha.***

By breaking the text in the paragraph into small sentences, it is easy for the computer to read and understand them. We can use NLP pipeline methods to read the sentences and determine what it means.

10.13.2 Word Tokenization

Once we have broken the paragraph into sentences, we can process the individual sentences. We can now break these sentences into separate words or tokens. This is called “Tokenization”. Every word including the punctuation is split apart.

Eg: ***Delhi***, ***is***, ***the***, ***capital***, ***of***, ***India***, ***and***, ***one***, ***of***, ***the***, ***most***, ***populous***, ***city***, ***in***, ***Asia***, **.**

10.13.3 Predicting parts of speech

Each token is taken individually and the part of speech for that token is determined. Finding out if the word is a noun, verb etc. helps to determine what the sentence is about. Each word is then fed into a part of speech classification model which was trained already by feeding in millions of English sentences to determine the part of speech.

Eg: ***Delhi*** is a noun and ***capital*** is a noun. So we can determine that the sentence is probably about ***Delhi***.

10.13.4 Text Lemmatization

There may be cases where a same word may appear in different forms. Text Lemmatization helps to determine the base form of each word, so that it will be easy to figure out that the words are the same if they where in different base forms.

> ***"Lemmatization is typically done by having a look-up table of the lemma forms of words based on their part of speech and possibly having some custom rules to handle words that you've never seen before [136]".***

10.13.5 Identifying Stop Words

There are lot of filler words like ***a***, ***the*** etc. These words are called stop words. The stop words are considered a noise and are usually removed before performing any statistical analysis.Here we determine how the words are related to each other.

Eg: ***Delhi***, ***capital***, ***India***, ***one***, ***most***, ***populous***, ***city***, ***Asia***, **.**

10.13.6 Dependency Parsing

> ***"The goal is to build a tree that assigns a single parent word to each word in the sentence. The root of the tree will be the main verb in the sentence . In addition to determining the parent word, the relation that exists between the word is also found out [136]."***

10.13.7 Finding noun phrases

The words that represent single idea is grouped together instead of considering every word as a single entity. The information from the dependency parse tree is taken to group the related words together. By combining the non-phrases from the sentence we get:

Eg: ***Delhi the capital most populous city …***

10.13.8 Named Entity Recognition

The aim of Named Entity Recognition is to detect and label the nouns with real world concepts. A Named Entity can Recognize people’s name, location, products, date and time, money etc.

Eg: ***Delhi***, ***India*** and ***Indraprastha*** represent places on a map. With Named Entity Recognition we will be able to detect that on a map.

10.14 Installation

1. Install the latest version of Python (avoid the 64-bit versions)
2. Install Numpy (optional)
3. Install NLTK: pip install nltk
4. Install the NLTK packages:

```
import nltk
nltk.download()
```

10.15 Example

Tokenize:

```
from nltk.tokenize import sent_tokenize, word_tokenize
line = "A quick brown fox jumps over the lazy dog"
print(word_tokenize(line))
```

```
output: "A", "quick", "brown", "fox", "jumps", "over", "the", "lazy", "dog".
```

10.16 Conclusion

Based on the user’s query and information need, NLP system represents the true meaning of what is expected. The content of the document will be searched in order to retrieve the relevant document based on the search query [125]. Alan Turing famously proposed a test to check intelligence of machines. The test measures if a machine is able to exhibit intelligent behavior/thinking like humans when it is asked same question by a human and it’s response is compared to the response of other human. This led to field of Artificial Intelligence (AI). NLP is important part of AI as it helps with communication between machine and human.

10.17 Team Members and work breakdown

- Sahithya Sridhar (fa18-523-67): Introduction, Natural language text processing, Natural Language generation
- Prajakta Patil (fa18-523-65): Related Work, Applications of NLP, NLP in Yelp data review

11 BIG DATA AND STREAMING

Manek Bahl, Sohan Udupi Rai
mbahl@iu.edu, surai@iu.edu
Indiana University
hid: fa18-523-62 fa18-523-69
github: ☁

11.1 INTRODUCTION

Data obtained in real-time from various sources are called Data Streams; processing and extracting insights from such sources is called Big Data Streaming or Real-time streaming analytics. While conventional big-data systems can handle near real-time data by performing micro-batch processing, they have still got latency issues which is not acceptable in some critical applications. We look at various technologies available today for handling data streams and see how each of these deal with the challenges associated with the task. The paper then looks at some real-life examples exploring the implementation of Big data streaming systems in various domains.

Deriving insights from data has always been the key requirement for organizations to gain edge over competitors in the market. But there are certain applications where this needs to be done within a few milliseconds for it to be useful. Recent boom in the field of Internet of Things makes it paramount for analytics systems to be able to deal with such huge data quickly and effectively. This is where Big Data Streaming has gained immense importance in the recent past.

Various domains such as eCommerce, Banking and Finance, Social media generate such data sequentially, and there is a heavy need for this data to be processed as they arrive on a row by row basis to derive actionable insights. These insights enable the companies to understand the recent consumer behavior and act accordingly to cater to their needs promptly in a timely manner. This beneficial both for the consumers, who get better service, and for organizations, to be more efficient and proactive in the business decision making [138].

11.2 Stream Processing vs Batch Processing

For a long time, the traditional way of dealing with data has been Batch Processing, wherein, the data is collected and continuously stored in a database. Insights are derived from this data by dividing it into batches and then deploying suitable algorithms. The size of the batch could be anything from as short as a day to maybe even a year, depending on the type of insights required. While this type of insight is valuable for the companies to make long term business decisions, there are certain critical applications such as fraud detection, which require analytics to be done in real-time so action can be taken immediately, before the fraudulent transaction can be completed. There arises the need for Stream Processing [139].

Advantages of Stream Processing compared to Batch Processing:

1) Data collected from continuous data sources such as traffic sensors, transaction logs and most other IOT sensors is Time Series data and Batch Processing would be tedious since there would be a need for aggregation across batches. Stream processing handles the data without the need for any such aggregation.

2) Stream processing reduces the storage requirement of a system. Since analysis is done in real-time, there is no need for the entire generated data to be stored in the system. Only the useful data can be stored.

3) Stream Processing requires less hardware capabilities compared to Batch Processing since the amount of data on which analysis is performed is small compared to the large volume of data in Batch Processing [140].

11.3 Challenges in Stream Processing

Unlike batch processing, wherein all of the incoming data is first stored and then divided into batches for processing, stream processing systems must have the capability to handle incoming data as and when it arrives. This leaves room for Data loss. Another factor that comes into play when dealing Stream data is ensuring data serialization. i.e. systems must ensure that messages must be processed in the order in which they were generated by the source. Thus, maintaining data integrity poses a great challenge for Stream processing systems.

Streaming application need to maintain the state and offset to keep track of the last message that was processed. While this is relatively simple to do, it poses a big challenge when dealing with system failure or in the case of vertical scaling since the new version of the application may pose compatibility issues [141].

11.4 Big Data Streaming Architecture and Technologies

Over the recent years, various technologies have been developed for Stream Processing. The architecture adopted by these technologies often differ from one another, making some of these more suitable for some applications than the other. Some of the prominent technologies for Stream Processing and the architecture they use, are discussed below.

11.4.1 Apache Spark

Apache Spark is an open source analytics engine that uses distributed cluster computing framework. It was developed in UC, Berkeley and the codebase was then given to Apache Software Foundation which now maintains it. Spark is known for its high performance and can be used interactively using Scala, R, Python and SQL. It provides an interface to perform various analytics functions such as Machine Learning using MLLib, Streaming using Spark Streaming and Exploratory Analysis using GraphX. Spark Streaming enables us to build streaming data applications which are scalable, fault tolerant. Spark has the capability to ingest data from various sources such as HDFS, Kafka, Flume etc. and then uses complex algorithms to process the data which can be then stored into a database or can be written out to a file. Spark Streaming receives divides the incoming data stream into batches which are then processed by the Spark Engine to produce a batch of final data. Spark Streaming brings in an ease of operation as it lets us write the streaming jobs in the same way batch jobs. The biggest advantage of using Streaming is that it has the capability to recover lost work as an inbuilt functionality. Moreover, the streaming data can be concatenated with historical data to build interactive applications. Spark uses HDFS or ZooKeeper for high availability. Currently Spark Streaming is widely being used by Yahoo, Uber, Netflix and eBay for providing real time analytics [142].

11.4.2 Apache Storm

Storm is commonly known as the “Hadoop of Real Time Processing” as it processes the streaming data as reliably as Hadoop does for batch processing. It is a distributed framework used for stream processing which was developed by a team at BackType and Twitter and was written in Clojure. The application is designed at a directed acyclic graph with spouts (input streams) and bolts (processing modules) as the vertices and the data flows in the form of tuples. The function of the spouts is to get the input streaming data and pass it to the bolts. Specialized spouts are available to retrieve data from multiple sources however Storm provides an option to create customer spouts as well. The data is then processed in the bolts and as per requirement the processed data is passed to a database or a file system [143]. Like Spark, Storm also provides high fault-tolerance, but it does not have a feature to use the same code for batch and stream processing. Current users of Storm include Twitter, Groupon, Yahoo and Spotify for its ability to create low latency distributed systems for streaming.

11.4.3 Apache Flink

Apache Flink uses a DataFlow model similar to Storm. But the main difference which makes it superior is that unlike Storm, it processes the events as and when they occur rather than processing micro-batches. This is especially useful in applications where the data stream is sporadic i.e. extremely sparse. This approach reduces amount resources needed to handle the stream. It also eliminates the effort needed to determine the optimal size of micro-batches which is done by trial and error in Storm. Flink is also more flexible compared to Storm due to the simplistic nature of its API. Its SQL API interface makes it very possible for non-programmers to deal with the data stream application. Flink can be setup in either of the two modes – Standalone or Distributed [144].

11.4.4 Apache Kafka

Apache Kafka is an open source platform for Stream processing. Its design is inspired from the Transaction logs maintained by a database management system. It is a distributed system, wherein multiple nodes known as ‘Brokers’ work together to form a cluster. Its distributed nature ensures high availability and High scalability. Which means that Kafka systems are fault tolerant and provide horizontal-scalability. Horizontal-scalability nature of Kafka is important since it ensures that additional computing power can be provided to the system by adding nodes without disrupting the ongoing operations. The main data structure used by Kafka is a commit log data structure which only allows appends. Once added to the data structure, records cannot be deleted or modified. This ensures that the data are placed in exactly the same order in which the events occurred. This data structure has an added advantage that reads and writes are done in constant time O(1) which drastically increases the speed at which Kafka can handle streaming. Read and write operations can be done simultaneously as there is no 'lock' placed on the data during a write operation. The messages stored in the Kafka nodes are divided into sub- divisions called Topics. Topics are further divided into smaller divisions called Partitions, for improved performance. The commit log-type data structure used in Kafka ensures that all messages in each partition are in the order in which they came in. It also means that Kafka provides excellent performance delivering messages at near network speed without placing the data in RAM, it uses disks to store all its records instead. Kafka stores the data for a set amount of time during which the consumer can use offsets to pull any record they want. Partitions are replicated and placed in multiple nodes of the system to ensure High-availability [145].

11.4.5 Amazon Kinesis

Kinesis is Amazon's solution to streaming data and analytics. Kinesis Data Stream is one of the key components of Kinesis platform which also includes the Kinesis Video Streams, Kinesis Data Firehose and Kinesis Data Analytics. Data streams reads the data from various input sources in the form of data records. These data records can be analyzed or stored in whichever way the application demands. The data streams start ingesting data within a second of when the data is added. The data then can be sent to any of the built-integrations or can be stored in various third-party stores such as DynamoDB or Cassandra. One advantage that Kinesis Streams offers is scaling of the applications, it allows dynamic changing of the throughput of the stream based on the volume of data expected [146].

11.4.6 Hortonworks Dataflow

Hortonworks DataFlow is an open source distributed platform capable of ingesting, storing and analyzing streaming data from multiple sources simultaneously. It provides a GUI, thus eliminating the need for the end user to have an understanding of programming. Being powered by Apache Nifi, HDF has the ability to ingest data from a variety of Streaming sources with ease. HDFs integration with Apache Ranger ensures excellent data security. HDF uses Apache Kafka as the streaming platform enabling it to process several million transactions per second while allowing users to deploy several Machine learning algorithms. The Streaming Analytics Manager provides users to perform the analytics in a simple visual fashion [147].

11.5 Industrial Use Cases of Big Data Streaming

E-Commerce: Real time transactions can be clustered and used along-with various other features such as product reviews to provide recommendations to the customers based to the latest trends. Alibaba and e-Bay are pioneers of using big data streaming to enhance their business.

Healthcare: Streaming has become an integral part of healthcare industry now with providers analyzing patient records to forecast any future health issues and recommendations to prevent them. Companies such as MyFitnessPal use Apache Spark to clean the data entered by the users to recommend healthy food diet.

Entertainment: Netflix’s application monitoring system is largely built on Amazon Kinesis which monitors all the applications within Netflix and tries to detect issues to give a high availability to its customers. Netflix also uses Apache Spark to collect all user activities on the app and analyzes it to come up with personalized recommendations.

Social Media: With various organizations paying so much attention to reviews being posted by users of various Social Websites, it has become increasingly important to be able to analyze the public sentiment. Twitter uses Apache Storm internally on its various applications for anomaly detection to provide high availability to its users.

Banking and Finance: One major use case of big data streaming in financial domain is fraud detection. If a fraudulent transaction is detected, it is imperious that corrective or preventive measures be taken in real time [148].

12 IOT and Big Data: Applications and Future Trends

Uma Kota, Jatinkumar Bhutka
umabkota@iu.edu, jdbhutka@iu.edu
Indiana University Bloomington
hid: fa18-523-71 fa18-523-59
github: ☁

Keywords:IoT, Big Data, Analytics, Smart devices fa18-523-59, fa18-523-71.

12.0.1 Introduction

Since its inception, internet has been all about collaboration between people across the globe. Games, social media, images, movies etc., available in the internet were created by people for people. It caused a revolution in the way people connected with each other and it's now woven into their lives in one or the other way. With the rise in ubiquitous computing, Internet of things has taken this technology to the next level. Physical objects were now introduced and connected to the digital world. The term IoT was coined by Kevin Ashton who envisioned a future where computers could be connected to things and by leveraging the data collected could manage the things with little human intervention [149].

In today's bigdata world, Internet of things (IoT) has established itself in different fields of life by making processes more efficient and robust. As a result, with increase in digital connections between physical objects and the internet, data generation rate is going up and the availability of vast amount of data has opened doors for different kinds of analyses with the help of the services provided by big data technologies. IoT allows a device to connect with different types of things like electronic devices, software's and sensors which exchange continuous streams of data, without any human intervention. The features offered by IoT allow companies to analyze their collected data and use it for business intelligence. Also, it can be useful to generate various models that can improve daily routine experience of the users. By 2020, Gartner has expected the IoT to connect over 20.4 billion things together ranging from mobile devices, vehicles, robots, and various industrial equipment etc [150].

The huge amount of data generated by IoT platform is not self-sufficient to generate insights and improve processes,it needs to be handled, processed, managed, integrated, analyzed in association with big data technologies, in a scalable, cost-effective way and more importantly in real time. Hence, digital world networks of physical objects and big data technologies can collaboratively be used to achieve complex tasks in the field of health care, manufacturing Industries, energy conservation, home automation, transportation system, education and research to improve quality of life.

12.0.2 Architecture

Interactions between digital mediums require special architecture and there are many architectures for IoT. The most common IoT Ecosystem architecture is as shown in Figure 17 . It consists of seven different layers [151].

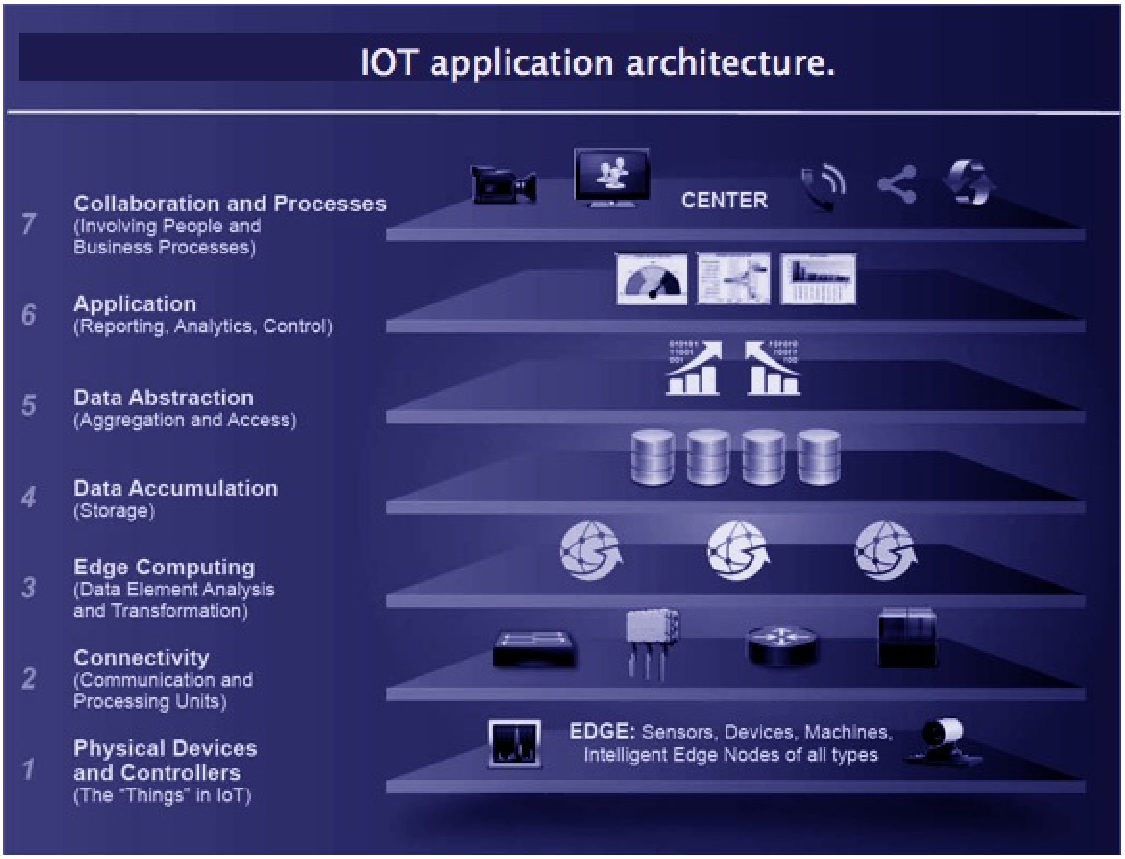

IOT application architecture.
7
Collaboration and Processes
(Involving People and Business Processes)
CENTER
6
Application
(Reporting, Analytics, Control)
5
Data Abstraction
(Aggregation and Access)
4
Data Accumulation
(Storage)
3
Edge Computing
(Data Element Analysis and Transformation)
2
Connectivity
(Communication and Processing Units)
1
Physical Devices and Controllers
(The "Things" in IoT)
EDGE: Sensors, Devices, Machines, Intelligent Edge Nodes of all types

Figure 17: IoT Architecture[151]

- ***Physical Devices and controllers:*** The lowest layer of this Internet of Things (IoT) architecture is comprised of physical devices and controllers. The layer consists of devices such as electronic gadgets, sensors and activators which input the needed data from various sources.

- ***Connectivity:*** Next layer is termed as connectivity layer which takes care of communication between sensors and processing units. These processing units try converting input data from sensors into an understandable format with help of certain protocols and perform selection on data that is to be processed further into the IoT architecture creating thresholds for incoming data.

- ***Edge computing:*** Edge Computing does the analysis of data elements. Their transformation is achieved by analyzing and filtering data before it is processed further thereby reducing huge computing and networking costs. The incoming raw data from sensors and activators can be selectively sent for further analysis. Also, one of the major concerns of Internet of Things (IoT) architecture is Data Security. Edge computing overcomes this to a considerable extent by feeding some of the sensitive data into sensor devices. Also, devices known as edge devices are coming into existence which will help in analytics and computation purposes, delivering data deliverable at a much faster speed in a robust manner. These edge devices will further help in maintaining connectivity with connected devices at source thereby allowing us to have a luxury of new smart devices.

- ***Data Accumulation:*** Data Accumulation is largely done in distributed frameworks as incoming data is in humongous volumes and variety while being input at a great velocity. Data is distributed into small sets of data using key/pair values just as in the case of data tuples, mapped and then reduced to small chunks of data before processing it further. Also, these days, organizations are heading towards PaaS (Platform as a Service) as a cloud platform for hosting data publicly and at the same time securing the data, thus customizing it to their needs.

- ***Data Abstraction:*** Different translation rules are brought into place for securing connectivity of specific devices. Also, a single model for data abstraction is created and provided to all the devices of a specific service thereby achieving integration of various devices.

- ***Application:*** The Application layer consists of reporting and analytics part of the architecture. All the efforts that were put into data accumulation, abstraction, storage, transformation, cleansing, preparing smaller chunks will be benefited only if proper analytics is performed and strategic Business Intelligence reports are generated out of this data.

- ***Collaboration and processes:*** Collaboration and processes layer is the user interface layer where people i.e., the end users and the business processes come into picture. In the end, it is the customer who engages with the business processes and as it brings both of them together, it's named as collaboration and processes layer.

12.0.3 Big Data and IoT Together

Big Data analytics helps analyze data sourced from Internet of Things (IoT) which has multiple devices connected to it. These devices can be sensors, activators, websites, social media etc. Internet of Things (IoT) has had its advancements and applications in fields such as automobile industry, health-care, transportation & logistics, education, commercialized residences etc., and incoming data from these domains is in the order of billions of gigabytes per day, at the same time it is largely diversified. Also, the velocity of the inward data flow is extremely high. A Big Data platform then takes this generated data as its input and stores it into files. Since the input data is unstructured or semi-structured, frameworks used to store this data in an intermediate place are largely distributed. Different big data tools can be used for storing this huge data such as Hadoop, Apache Hive etc. One of the most prominent ones used in the industry today is the Hadoop Architecture which has Hadoop Distributed File System (HDFS) and MapReduce as its two components to process data into small chunks for report generation and analysis purposes. The data is in a way captured, integrated, mapped into different data sets of tuples and then processed to warehouses for storage. The data warehouses help store the legacy data in them and allow generating reports on the desired data at any given point in time [152]. This is shown in Figure 18.

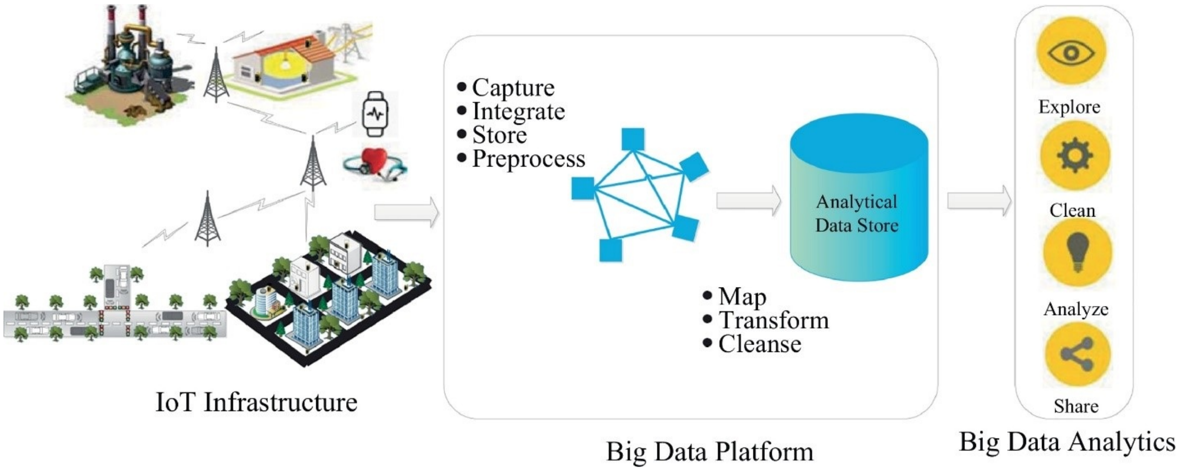


Figure 18: Big Data and IoT[152]

Big data helps businesses in coming up with inferences, insights and actionable recommendations by analyzing the unstructured or semi-structured data. Billions of devices are expected to be connected to the internet and hence for the functionalities of these devices to be held, data is needed. IoT will be a major data source for the all the analyses that will be performed by companies for growth and effective decision making, with the nexus of big data and the IOT data from the connected devices making it cheaper and faster. The process of IoT with Big Data is seen in Figure 19.

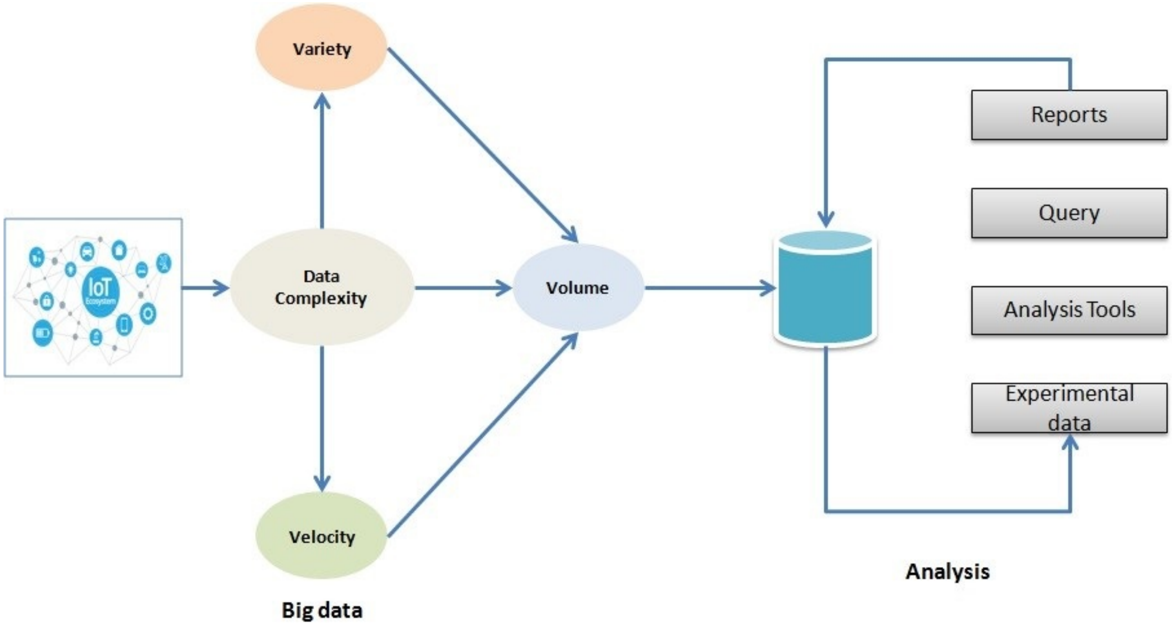


Figure 19: IoT Process[153]

- Data, which is sourced from various connected devices or from disparate data repositories is collected and stored in a big data platform.

- A big data system is often referred to as a distributed database as incoming data is sourced from different data repositories and is then stored in the form of flat files, in case of a Hadoop big data platform it's stored in the Hadoop Distributed File System (HDFS). Another big component of the Hadoop platform is Hadoop MapReduce which is used to process independent data chunks efficiently. The Hadoop MapReduce framework works on a MapReduce algorithm where data in form of files is converted into a data set using key/value pairs analogous to dictionaries and tuples. The output of a Map is then given as an input to the reduce stage where different data sets of tuples are converted into smaller chunks of data in order to be processed further.

- Reports are then generated by taking data from the concerned data warehouses using the Business Intelligence Report generation tools as per client/user specifications. The same is accomplished through complex query writing into databases. Therefore, big data platform paves way for this disparate data, collected from different ‘collected devices’ using the Internet of Things (IoT), to be leveraged in deriving different business trends and business insights.

12.0.4 Impacts of IOT on Big data

The Internet of Things (IoT) has impacted the Big data platform in a great deal and there is going to be a significant impact which lies ahead as well. Today, Humongous amounts of data are generated every hour by devices connected to the IoT and in the future there are going to be more and more devices connected to the IoT invariably making it difficult for data collection and analytics platforms to effectively and efficiently keep producing results [154], [155].

- ***Big Data Storage:*** With data being sourced from various disparate data sources and connected devices as is the case working with Internet of Things (IoT), the data storage platforms need to be very flexible as the incoming data is in the order of billions of terabytes. Also, the rate at which data is input to a data storage platform is unlike the natural data speed. One of the main features the companies seek is hosting data in cloud systems so that the data is kept publicly secure and as per the company requirements. PaaS (Platform as a Service) is one such model the companies are eyeing, to serve their purpose of effectively handling the continuous inward flow of diversified data and hosting them on cloud as PaaS provides this functionality.

- ***Big data Security:*** Major threats as far as big data security is concerned are to the Data Mining solutions as they provide the strategic solutions to the business. Hence it becomes increasingly important to secure them against the security breaches caused by fake users. There needs to be proper encryption of authentication measures of the users. Also, it can be resolved with the use of metadata (data about data) as to who accessed what type of data and so on. The other prominent security issue with Big Data is of the distributed frameworks that are used for data storage purposes. Hadoop being one such open source big data platform where more systems can be prone to the security issues as the data processed is distributed over many distributed frameworks.

- ***Big data Analytics:*** The analytics of big data consists of a lot of challenges as the inward flow of data is quite unstructured considering it is from different connected devices like the sensors and websites. Hence big data consists of three V's: Volume, huge volumes of data sourced from sensors is to be analyzed. Variety, Data is unstructured and is in the form of 3D data, 2D data, log files etc. Velocity, the speed with which the data is processed which is nothing but real-time processing and continuous data processing. Hence, we see data complexity in Big Data Analytics.

- ***Tools of big data:*** In order to draw effective insights and inferences from the data, proper big data tools need to be in place.To make sure pivotal information is extracted in form of insights from the data while secureing it, different Big data technologies are leveraged. The commonly used Big Data tools are Apache Hadoop, Apache Hive, Storm, Cloudera, Qubole etc.

12.0.5 Challenges

While IoT seems to be promising, there are many challenges associated with it. With increase in number of data sources and with these sources storing a large variety of data including private data, there is an increase in security concerns all over. Poor connectivity also seems to be an issue when the number of connected devices is in a large number. The business models that are implemented are expected to be robust and coming up with feasible, cost effective and efficient models is still a challenge. With time there may be compatibility issues between the technologies in the system as new standards may be implemented which will thereby increase the number of hardware and software devices connected. Data reliability is also a growing concern as the available data if corrupted will be leading to bad decisions. Increase in connected devices may also limit the speed with which data is collected [154], [156].

12.0.6 Applications

The Internet of Things and Big Data coming together play an important role in commercial, industrial and other applications to offer better data insights and inferences. Some of the most common application are shown in the figure Figure 20.

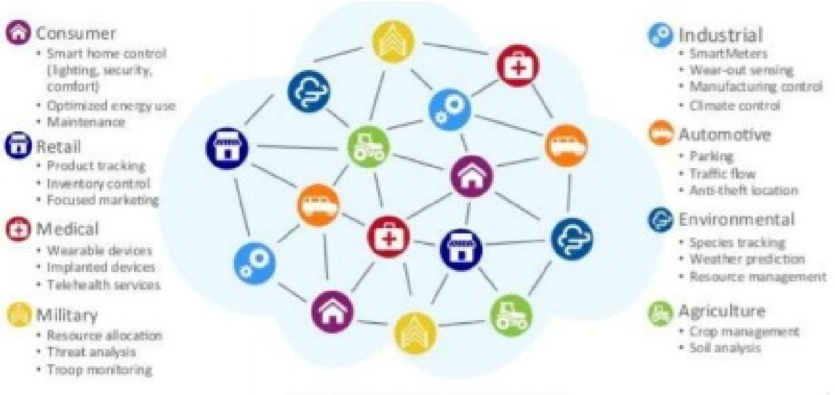


Figure 20: IoT Application [157]

- ***Application in Manufacturing Industries:*** The machinery that is enabled with the Internet of Things (IoT) assists in transmitting operation related information to the managers which helps them identify the areas of optimization and the process of automation. It is also used in facilitative management of machine tools expected to function in specified ranges of temperature. Sensors enabled by IoT can help monitor such machines sending alerts as soon as sensor senses some deviation from the original specifications. Operating machinery in the specified specification range in turn increases the efficiency of operations, diminishes the downtime of machines, thereby, helps reducing the costs and conserving energy. Another application of IoT in the manufacturing domain is the Production Flow monitoring where production lines monitoring is enabled from refining to the packaging of final products. If there is any discrepancy found in operations, then the IoT enabled monitoring of processes helps in adjusting them and managing them better from the cost of operation point of view and avoiding wastes during the production process. Further, IoT helps in tracking and tracing of inventory while delivering the products on a global basis which helps in managers getting a clear idea of the current supply chain process. We can ensure the safety of the workers in any given project by monitoring the performance indicators thereby reducing the injury rate and any relevant loss to the organization. IoT significantly helps in quality control over the life cycle of a manufacturing process [158].

- ***Application in Health care:*** The Internet of Things (IoT) has advanced its applications in the field of Healthcare by a great deal right from monitoring remotely to getting accustomed to medication. It has served as a helping hand to the doctors to make significant progress in patient's health. The hospitals have started using Internet of Things (IoT) ensuring every patient gets maximum benefits and stays healthier no matter how small the disease to be cured is. The IoT has guided the healthcare industry in improving the implementation of healthcare processes in a myriad of ways.

    - ***IoT for managing inventory***: There is a considerable amount of change in the way the hospitals have taken advantage of the IoT inventory management for the inventory control in the warehouses and the pharmacies.

    - ***IoT for optimization of healthcare workflow***: Adoption of wireless infrastructure have helped manage the throughput and perform an analysis of the existence of bottlenecks, if any, in the system and that we could get rid of the same.

    - ***IoT for integrating the medical devices***: Different ways are being looked at for integrating the devices like Fitbits with a view to obtain more data about the patient and be taken care of. Neel Ganguly, vice president and CIO at JFK Health System in Edison, New Jersey mentioned of JFK Health System starting to use devices like blood pressure cuffs, glucometers etc., with the intention of collecting data about the various signs of the patients which helped them serve the patients in a more effective manner. Just like IoT has security concerns in most of its extensive applications, the healthcare industry is no different. But continuous efforts are being made to get rid of the existing barriers as well as the potential barriers [159].

- ***Application in Energy Management of Buildings:*** The traditional building management systems look after buildings power sourced systems. The new Smart building energy management systems with help of IoT sensors offer a lot more than monitoring these systems. The data collected is also leveraged to conduct analysis on the raw data which is then transformed into insightful reports that can help make the systems way more efficient. Sensor devices that track weather and traffic data can also be connected to this system and their data can be utilized in adjusting a building's lighting, temperature etc., in real-time. The operating costs can be cut by 25% with help of these systems [160].

12.0.7 Use cases

Now, IoT and Big data is everywhere. Due to the flexibility and scalability of big data, industries in recent times have started adapting to the use of IoT. Some of the industrial use cases include,

- ***Caterpillar:*** An IoT pioneer Caterpillar, one of the leading equipment makers is a perfect example of how it has reaped the benefits of using Internet of Things (IoT). They have eased out process of identification of levels of fuels for the personnel operating the machines and the timings of replacement of the air filters using the Internet of Things (IoT) along with Augmented Reality (AR) [161].

- ***IoT based Asset Tracking System in Transportation & Logistics:***

Number of assets were able to be tracked down by the existing tracking systems and the system use to breakdown reaching its threshold after the number of assets tracked increased exponentially. Since the devices used for tracking were from diversified manufacturers, it became increasingly difficult to analyze, draw inferences from the data and provide Business Intelligence reports. The solution to this was Internet of Things (IoT) which helped built a server for processing robust messaging. Geo-fencing was brought into implementation using IoT which would allow a trigger to be made on an asset entering or leaving a area. Also, an engine with respect to analytics was developed with a view to provide better data insights and meaningful Business Intelligence reports [162].

- ***Smart metering using IoT:*** For measuring the consumption of gas, energy and water in buildings, a device capable of working with Internet called a smart meter is used. It has helped overcome the disadvantages of measuring just the overall consumption by allowing to record the level of consumption of each of the resources used thereby benefitting consumers monetary wise. In this way, Internet of Things (IoT) has helped improve the process of forecasting and binding the consumption of power [162].

- ***Predictive Maintenance using IoT:*** Millions of dollars can be saved by the companies by keeping the equipment's and assets going all the time with the help of sensors and data analytics thereby preventive maintenance [162].

- ***Airline Management using IoT:*** IoT in airline industry can be seen everywhere from baggage tracking to cabin climate control. With help of IoT, the industry is trying to increase customer satisfaction by reducing the complaints like lost baggage, flight delays and service issues as well as in cutting unnecessary operational costs.Airways like Virgin Atlantic, Etihad are manufacturing planes that are connected with help of IoT devices. Data of half a terabyte is expected to be produced over a flight journey, this will be analyzed to get the information regarding mechanical issues etc., before they may even occur. With help of such technology the flight delays, safety issues will be restricted. Delta was one of the first airlines to introduce Radio

Frequency Identification (RFID), a baggage tracking technology that leverages IoT. With help of RFID, the customers can always access their baggage location from their phone while travelling with the airlines. IoT is also being utilized in navigation of flights such that aircraft always takes an optimal route thereby optimizing the fuel use. This is being implemented by Air Asia with help of GE to cut fuel costs. Jet Blue with help of IoT has automated the check in process for customers, with a ticket and seat being issued directly to every customer not needing any customer inputs by analyzing their preferences from their data [163].

12.0.8 Future Trends

- ***Retail Industry:*** IoT promises to help retailers take better business decisions while helping them better the customer satisfaction. Omni-channel experience is going to be the future of retail and is feasible with IoT. The retail industry in future will be seamlessly integrating online stores with brick and mortar stores to make it easier for the customer to find their products while simultaneously connecting with shoppers in a more personal way. For example: Alerts can be sent to the customers when their favored location is out of a certain product with other recommendations etc [164].

- ***Health-care:*** With a proactive technology in place gathering a person's daily data, the onset of any clinical trail can be predicted there by reducing the number of emergency cases. IoT enabled technologies can make this scenario feasible in the future with help of a network of wearable devices, sensors, monitors etc., set up for monitoring health. With help of Monitoring technologies like telepresence, checkups can be done from any part of the world and the number of hours spent in hospitals can be cut down. Also, the data gathered by these technologies can help transform the check-in process by automatically sharing the past data of patient with health professionals [165].

12.0.9 Conclusion

With data coming from disparate data sources and diversified connected devices, it has given the big data analysts option to distinguish and filter data and just process the useful chunks of data further to the data warehouses and then generate Business Intelligence reports from there on. With the emergence of IoT, it has become increasingly possible to reach out to the customers at any given time of the day. Devices like Fitbits and smart devices have made it possible for the doctors to get timely updates of the patient's doing about his health thereby allowing the physicists to intervene whenever needed and provide mannerly treatment and diagnosis to their patients. Businesses have been reaping the benefits of the advancement in technology by having the luxury of digitally connecting to their clients even from the most remote locations in the world. In the years to come, the amalgam of advancement in technology i.e. the Internet of Things, big data will give more likelihood to businessmen in gathering even the most detailed information of their client and customers thereby giving themselves a better opportunity to improve their business from the perspective of effectively managing the monetary department and providing a better customer satisfaction which will invariably up their business a great deal [166].

With time and the advancements in the field of IoT and big data it can be concluded that they are two sides of the same coin bolstering one another.The Nexus of Internet of Things and Big Data can be seen as the future of business intelligence, health-care and many other industries.

12.0.10 Acknowledgments

The authors would like to thank Dr. Gregor von Laszewski for his support and suggestions to write this paper.

13 Big Data in Healthcare

Wang Tong
wangton@iu.edu
Indiana University Bloomington
hid:fa18-523-73
github: ☁

keywords: Big data, healthcare, healthcare informatics, unstructured data

13.1 Introduction

Big Data is used in the analysis of healthcare informatics, which consists of complex data sets that are hard to store, resolve, organize, process, and interpret. The benefits obtained from the application of Big Data in the healthcare system include reduced costs incurred in the delivery of healthcare services, optimal management of information, prevention of the side effects of drugs on patients, the discovery of treatments for incurable diseases, and improved service delivery. The challenges faced in the use of Big Data in the healthcare system include problems with analyzing unstructured data, focusing on correlations rather than casualty, privacy issues which act as barriers in the collection of information, data security due to hacking, and reluctance by health care centers to share patient data due to competition.

Big Data is the vast amounts of complex information which require advanced technology to assist in their capture, storage, distribution, management, and analysis. The Big Data characteristics can be explained using the 6Vs, which are velocity, volume, value, variety, variability, and veracity. The former implies the data sets are generated at very high speeds. The volume stands for large amounts of data sets that require analysis. The value suggests the vital information that is contained in the data to be analyzed. The variety stands for all types of data sets such as structured, semi-structured, and unstructured. The variability aspect entails the changes that occur during the processing of data. The veracity of the data sets collected implies that it is consistent and it can be trusted. The Big Data evolution can be attributed to the technological innovations of artificial intelligence, the internet of things, the fifth generation of the internet, and advancements in machine learning. The emergence of Big Data is essential to the growth of all fields. Therefore, the application of Big Data in the healthcare system will help to increase the speed, volume, accuracy, and efficiency in the analysis of the complex healthcare records.

13.2 Requirements of the Electronic Health Records

The electronic health records (EHR) is a system that has been developed by software engineers and contains information regarding the health care system. The data provided in the EHRs include patient information, medications, diagnoses of diseases, procedures, signs and symptoms of conditions, visit dates and times, and clinical notes of the healthcare practitioners. The data stored by the EHR system can either be structured or unstructured. The former include administrative data such as the demographics of the patients, diagnoses, and procedures. Unstructured data mainly entails the clinical notes that are taken by the doctor, and are the most efficient way for clinical documentation because they rely on human intuition. The unstructured data is difficult to analyze because of the format used since it may contain grammatical errors, abbreviations related to medicine, and spelling errors. The large amount of information about the patients and healthcare in the EHR system has led to the need for computer-based methods to help in the organization, analysis, and interpretation of the data sets. Hence, the adoption of Big Data in the EHR system has been essential due to the high speed involved in the analysis of the extensive medical health records.

13.3 The Architecture of the Electronic Health Record

The data in the electronic healthcare record system can be categorized into various forms that include genomic, clinical notes, behavior, patient sentiments, clinical reference, and administrative. Genomic data is a category which contains the DNA sequence, gene expression, and genotyping [167]. Clinical notes are unstructured data sets that include a diagnostic testing report, medical images, patient discharge summaries, and the doctor’s notes [167]. The patient sentiments contain personal information about them such as their demographics. Clinical reference data is information from text-based applications such as journals, articles, clinical research, websites, and product information. Administrative data involves information from the health care center that allows others from outside the context to understand the medical records [167]. Therefore, through the understanding of the various categories of data in the electronic healthcare record system, a proper analysis can be made on the available information.

In addition, the EHR system contains the aspect of interoperability, which implies the ability to share data between various health center facilities. The descriptive information about the patient's demographics, pharmaceuticals, and diagnoses need to be recorded in a way that allows sharing across different healthcare facilities. The institutions can implement internal encoding mechanisms that assist in the analysis of the data collected before sharing of the information with other health center facilities. Additionally, administrative data should be included in the EHR system to allow for an understanding of the information contained in the records outside the context in which they were recorded. However, the main hindrance to the health records data interoperability is the lack of a similar data standard being used by the different healthcare centers. The barrier creates the need to establish a systematic widely accepted data-encoding scheme that will enable hospitals to share descriptive information about patients contained in their EHR system. Thus, data interoperability provides for the ease of sharing EHR by hospitals, which allows research to be easily conducted.

Furthermore, the EHR system also makes it possible for data mining, which is the extraction of the information contained in the electronic health records. It entails the use of various techniques, which include regression analysis, classification, associate rule learning, and temporal data mining. Regression analysis involves the estimation of the correlation between the independent and the dependent variables. Regression linear model fitting can be used if the dependent variables adopt distributions such as binomial, normal, and Poisson. Classification involves techniques such as k-nearest neighbors and the decision trees. It is effective in clinical applications. It involves the assigning of a new observation to a class through building statistical models. Associate rule learning uses the numerous associations among clinical variables to predict the occurrence of phenomena. Temporal data mining is based on the fact that it is hard to generalize the outcome of a treatment to a given patient because of the difference in the clinical variables. Therefore, the vast amounts of complex data contained in the EHR allow for the extraction of the data using data mining.

13.4 Implementation of Big Data in Electronic Health Record

The implementation of Big Data in the EHR system has improved the ease of capture, storage, organizing, interpretation, and the analysis of the data sets. Furthermore, it has made the system more reliable, accurate, and efficient in service delivery when compared to the traditional system that was being used to keep the medical records. The benefits gained from the implementation of Big Data in the electronic health records outweigh the challenges faced. Therefore, the implementation of Big Data in EHR is a necessity that will help to make the health care system efficient in service delivery.

13.5 Benchmark of Big Data in EHR Systems

The benchmark of Big Data in the electronic health record system involves the analysis of the benefits, opportunities, and the challenges faced in the implementation process. The benefits of adopting Big Data in the EHR system include precision medicine, reduced costs incurred in treatment, and the optimization of the workflow in the healthcare center. The challenges faced include data hacking, privacy issues and the analysis of unstructured data. Thus, despite the difficulties in implementing Big Data in EHR systems, its adoption is inevitable by the hospitals.

13.5.1 Benefits of Big Data in EHR Systems

The first benefit of the adoption of Big Data in electronic health records is precision medicine. Through the evolution of Big Data, high-performance genome analysis technology was invented which allows for the collection of large amounts of genomic data. Additionally, new analytical algorithms have been designed to help analyze the genomic data collected [168]. The high-performance genome analysis technology enables the healthcare facilities to compare the genomic data of one patient to a large population of other individuals [168]. The comparison helps to establish the emergence of rare diseases. Moreover, it not only helps the healthcare facilities to develop a diagnosis of the illness but also prevent its spread to other members of the population. Nevertheless, it allows for the improved efficiency in service delivery due to the systematic collection and the analysis of genetic data. Other benefits obtained from the concept include the selection of the best treatments, avoidance of the side effects of diagnoses to the patients, and the elimination of ineffective therapies. Therefore, precision medicine has been crucial in improving efficiency in healthcare operations.

Moreover, the adoption of Big Data in the electronic healthcare records has reduced the costs incurred for medication. The technology has allowed the health care centers to notice the early signs of diseases such as cancer [168]. The earlier recognition allows the individuals to meet the low costs incurred in treating the primary phase rather than the high bills needed for the final period of the disease. Furthermore, the health facilities have reduced the expenditures incurred in the research of new drugs for the treatment of diseases because of the ease of access to information required for research purposes in the EHR. At the same time, the hospitals can administer treatment on a performance basis whereby the practitioners receive money because of the outcome of the diagnoses rather than the time spent treating a patient [168]. Hence, the adoption of Big Data in the health care system has helped to reduce the costs incurred by both the hospital and the patients during the treatment of diseases.

Besides, the adoption of Big Data in the healthcare facilities has optimized workflows in healthcare centers. It has enabled various departments within the hospital to share information using the electronic healthcare register and reduce the movements from one department to another [168]. Additionally, the practitioners can communicate their findings on patients with each other and establish the best selection of treatment that will help in the diagnoses of disease. The communication of the departments using the EHR allows for better utilization of the resources of the healthcare facility. Moreover, it also brings higher efficiency in service delivery by the health care facility to the patients [168]. Thus, the optimization in the workflow within a healthcare facility is a benefit obtained due to the implementation of Big Data in the EHR system.

13.5.2 Challenges of Big Data in EHR Systems

The first challenge faced in the implementation of Big Data in healthcare systems is the analysis of unstructured data. The test results, scanned documents, X-ray images, and progress notes are some of the examples of unstructured data found in the EHR of a healthcare facility [167]. The large volumes of unstructured data, such as medical imaging, are difficult to handle by the EHR systems. Additionally, extracting potentially useful information from the unstructured data is difficult. It is also problematic to understand the informal clinical notes because of the context in which they were written [167]. Thus, analysis of unstructured data is a key challenge faced in the implementation of Big Data in electronic health records.

Moreover, privacy issues are another challenge faced in the adoption of Big Data by the electronic healthcare records. The hospital's ethical considerations demand the protection of the patient information, and it also restricts the sharing of the data without the consent of the individuals. The aspect of privacy concerns hinders data interoperability among various healthcare facilities that would have aided in the efficiency of service delivery by the hospitals. The reason is Big Data requires data interoperability. Nevertheless, the healthcare facilities can employ privacy protection mechanisms to help protect the information of the patients from loss or unauthorized access.

Another challenge faced with the implementation of Big Data in the EHR is the aspect of data hacking. Healthcare facilities have incurred high costs due to the leakage of data [167]. Apart from being sued by the patients whose medical information is leaked, the hospitals also face the loss of their patients due to a lack of trust. Additionally, the hospital is held at ransom by the hackers who demand money in exchange for the information obtained. Nevertheless, the healthcare facilities have adopted the use of biometrics, such as fingerprints and voice recognition software, to reduce the leakage of data and improve protection of information about the patients from the electronic health record system [167]. Hence, hacking is a menace that hinders the implementation of the Big Data in the EHR system.

13.6 CONCLUSION

The adoption of Big Data in EHR is inevitable because of the benefits associated with its approval. The advantages related to Big Data include best selection of treatment for diagnoses, reduced costs incurred, optimal management of information, prevention of the side effects of drugs on patients, and discovery of treatments for incurable diseases. Despite the benefits associated with the concept, it faces challenges such as privacy issues, difficulty in the analysis of unstructured data, and data hacking. Further studies need to be carried out to establish on the ways to minimize the challenges encountered in the implementation of Big Data. Therefore, the application of Big Data in the health care systems is essential for an improvement in service delivery.

14 Big Data and Privacy

Yeyi Ma
yeyima@umail.iu.edu
Indiana University Bloomington
hid:fa18-523-74
github: ☁

Keyword: Privacy

14.1 Introduction

Big data as a smart technological aggregate of database technologies utilized at the software and hardware level and its current be used in the business, technology, governmental circles and etc. Benchmarks that have been developed in order to standardize big data, it is revealed that the technology is relatively young to employ most of these benchmarks. In respect to privacy, big data poses a major threat to peoples entitlement to confidential data by design. As such, it is necessary that big data is regulated.

Big data refers to the study and the application of complex data sets in application software and hardware. Big data is associated with certain qualities of data such as variety, velocity, volume, veracity, and value. The computer and information age has made big data even more practical and relevant in processing information. Today, big data entails the use of computerized systems which execute predictive analytics and user behavior analytics. These analytics extract value from data sets. Most processed data is available in large volumes. However, that is not a huge problem when working with big data. To computer systems, processing a single record will take the fraction of a second. Processing a million records will not take a million seconds. Instead, it might take a few extra seconds. Modern computer systems utilize economies of scale in a manner that cannot be replicated in any other domain.

The most important characteristic of big data is that it analyses large data sets in a manner that is intuitive. This explains why big data has become so popular in nearly every field in the society: government, internet search, fin-tech, business informatics, urban informatics, medicine, and travel. Big data poses a threat to privacy since the end user does not have the centralized infrastructure required to make it work flawlessly. There is always a billion-dollar company out there with access to individuals' private information for every smart device out there.

14.2 Academic Theory

14.2.1 Requirements

One of the main reasons why big data is growing rapidly is the increase in the popularity of devices which fall under the internet-of-things. Therefore, gathering the data necessary for processing is relatively easy [169]. Over the past 4 decades, the technological ability to store data per capita has nearly doubled. This will only continue to grow over the next several decades.

The greatest concern for millennials and Gen-Z is who should stay in control of the big data initiatives. This raises the question of privacy. Trusting big companies such as Facebook, Apple, Google, Amazon, and Microsoft is not considered to be a credible solution. These companies own most of the big data infrastructure in the world. Moreover, they have a monopoly over all the internet-of-things devices such that they can essentially shut down all the competition by making their services and devices cheaper while harvesting user information.

The leaders of these companies are not elected by the public and are out to maximize profits. This is dangerous given that they already have a monopoly over modern communication but manage to operate as private entities capable of ideological, cultural, and religious bias. Moreover, they are not answerable to anyone other than their investors who usually only care about the bottom line. Big data promises the world numerous benefits over the next several decades. However, this comes at the cost of giving up liberties enshrined in the constitutions of most countries that have a Bill of Rights.

14.2.2 Architecture

Big data architecture is not new in the 21st century. Database management systems were popular in the 1990s and were offered by a few big companies. A company such as Wintercorp became famous for issuing intuitive big data repositories in the form of reports in this era. At the time, the biggest hard disks were only 2.5 GB. This means that the definition of the term big data keeps evolving in accordance with Kryder's Law[170]. The company has installed more of these data stores over the past 10 years. Their largest database store is more than 50 PB.

Other companies such as LexisNexis Group have been involved in developing the architecture which defines big data. In 2000, LexisNexis created a C++ system for distributing data and enforcing simple querying. This dialect uses a technology called "apply schema" in order to infer the schema of data under query.

Another important organization that has been utilizing and developing big data includes CERN[171]. This organization has been collecting big data for decades. The data is analyzed using supercomputers. In 2004, Alphabet Inc. published a paper called MapReduce which employs the same structure as CERN. The paper proposed using parallel processing to enhance speed and accuracy. In MapReduce, queries get split and distributed over nodes[172]. They are then processed in a parallel manner in what is called a Map step. The Apache open source project is one of the first to implement the MapReduce paradigm in a project called Hadoop. However, the MapReduce paradigm had certain limitations. As such, it was necessary to develop the Apache Spark. This new paradigm added the ability to add many operations as opposed to a single map.

Today, the most popular big data architecture is the MIKE 2.0. This is an open approach to Information System Management[173]. It acknowledges and addresses the need to keep revising the implications of big data. These insights were captured in a paper called "Big Data Solution Offering". Studies highlight that using multiple layers in big data is one of the ways of addressing the speed problems that have persisted in big data. In manufacturing, big data architecture is implemented as 5C. This stands for conversion, cyber, connection, configuration, and cognition. Another vital aspect of the big data architecture is data lake. This is a technology which makes it possible to shift focus from centralization to a shared model in respect to the changing information system dynamics. It also makes it possible to segregate data within a data lake in order to minimize the overhead time.

14.2.3 Implementation

The main components of big data can be summarized into three categories: analyzing techniques, databases, and visualizations. The data analysis techniques include A/B testing, natural language processing, and machine learning. On the other hand, databases include any technologies employed in the process of data storage. This covers cloud computing as well as business intelligence technologies. Visualization includes graphs, charts, videos, and other technologies which are used for displaying the output data.

In some cases, it becomes necessary to represent multidimensional big data as cubes or tensors. To do this, it is necessary to introduce an array of database sensors for high-level query and storage support. Other technologies that are necessary when implementing big data in an institution include subspace learning, tensor-based algorithms, distributed filing, HPC infrastructure, cloud infrastructure, data mining, the world wide web, and distributed databases. It is important to note that in spite of all the improvements seen in big data architecture, machine learning is relatively elementary. There are numerous challenges to machine learning. While most of these challenges are technical, a few of them are social in regards to the interaction between individuals, the law, and profitability.

Companies that have implemented big data utilize the latest and greatest when it comes to computing and storage. This means that there is a significant barrier to full utilization of big data at an individual or SME level. Companies that have embraced big data employ direct attached storage such as solid state drives and high-speed SATA inside parallel processing nodes for speed. It is quite rare for a company utilizing big data to use storage area networks or network arranged storages. These two are perceived to be slow, expensive, and difficult to use. If there is an existing technology in the market that performs faster, it needs to be implemented in big data.

Big data has numerous applications in a variety of fields. By 2010, the big data subsector was worth at least $100 billion. It was growing at the rate of 10% per year. This was about twice the rate of growth of the software industry in the same year. Many developed economies in the world are continuously using data-intensive technologies. Many people in the world have access to the internet thanks to the development in computer software and hardware[174]. The effective capacity of the world to exchange information is growing at a rate that is unprecedented. By 2014, internet traffic reached 667 exabytes by predictions. Big data is being utilized in international development, manufacturing, healthcare, media, education, the internet of things, information technology, and insurance.

14.2.4 Big Data and the Web

Big data is commonly associated to the internet of things since both concepts are based on smart data collection and manipulation. The number of people who rely on the internet for their work, entertainment, communication, travel planning, and education is increasing every year. The development of the internet is responsible for driving this shift from traditional tools to digital ones. Big data can be seen as an augmentation of the internet. This is because it creates a link between the physical world and software. Most efficient programs run with a connection to the internet for the purposes of collaboration, communication, and sharing resources. With more and more links to big data, there is a clear conduit between the world wide web and the physical world.

One the most important improvements that has taken place in technology over the past several years is the improvement in user interface that is presented to technology users. Big data seeks to fill the gap between the user and the software that takes instructions from various devices by collecting feedback without too much work from the user. This explains why the internet of things is becoming the most important component of big data. The internet of things allows physical objects to embed operating systems in order to collect data in real-time. When such information accumulates to volumes that can significantly alter or influence the manner in which an object works, big data components such as DBMS take over the records and process them yielding output that can be used adjust certain parameters of object, an apparatus, or a system. While the internet of things addresses the user interface solely, big data goes a step further and provides a processing platform that can then be used to develop useful output. Big data also allows systems to consume the feedback generated without the intervention of a human.

14.3 Benchmark and Privacy

Like any other viral and potentially useful technology, big data has various benchmarks. These benchmarks are tailored to the new technology since the existing ones have proven to be ineffective. In spite of this, the benchmarks that have been proposed have not been robust enough: BigBench, HiBench, AMP, and CloudSuite[175]. Each of these benchmarks has various merits and demerits. With all these options, the problem remains that big data is not mature enough to be tested in a standard environment. As such, it is prudent for an institution to benchmark their big data implementation using the usage scenario as opposed to these tools.

The factors that make big data efficient are the same ones that make it a major privacy risk. Big data analytics is dumb, meaning that it does not have preconceived notions about the subject on which data is being collected and processed[176]. This means that it is capable of collecting more accurate information than people perceive. For instance, big data can be used in medicine to diagnose in an objective manner. It can gather data and come up with recommendations that have been obvious but ignored due to preconceived notions, ideology, or human expectations. When it comes to big data, using algorithmic black boxes can be dangerous. Such implementations leave machines to decide what to make of inferences without the possibility of human intuition. Multinational companies that are implementing big data and using it to sell their services have handled this problem to a great extent. They have manual overrides to minimize the overreach of big data. However, these companies are not public utilities since they have a profit agenda meaning that privacy is hardly a priority.

14.4 Conclusions

Big data is replacing the raw information age that was ushered in by the growth of the internet as well as internet-based technologies in the late 20th and early 21st century. Big data makes use of numerous technologies to ensure that data is harvested in real-time and utilized in the same fashion. With the internet getting embedded in different home and personal appliances via embedded operating systems, big data will continue to become more popular. The applications of the technology are likely to outdo those of the bare internet given that user interface is one of the key areas that big data seeks to revolutionize. The applications will be seen in business, government, technology, services, travel, academia, and other sectors. However, big data poses a big threat to privacy since it eliminates the regulative human factor from the equation in an effort to offer fluid services whenever it is implemented. It is necessary that policymakers create regulations which protect technology consumers without killing innovation such as big data.

15 QlikView

Abhishek Rapelli
arapelli@iu.edu
Indiana University
hid: fa18-523-79
github: ☁

15.1 Keywords

HID fa18-523-79, QlikView, Associative In-Memory Technology, QlikView Server, QlikView Publisher

15.2 Introduction

QlikView is a popular software that was based on different approach for data discovery and analytics compared to other traditional analytical and visualization software tools. The uniqueness of QlikView is that rather than first building a query and then fetching the results, it forms associations in the data once it is loaded and then the user is prompted to explore and analyze the data with. This simplification of data exploration and the enhanced user interface has made it one of the most popular choice for traditional business analytics and reporting for businesses. QlikView, besides QlikSense is one of the software tools that was developed and maintained by Qlik software technologies company for Business Intelligence, analytics and visualization. Over the years, with more advancements and simplifications in BI tools, QlikView has gained its prominence for small to medium scale business analytics applications due to its simplicity, easiness and handiness. It adapted to the changes in the BI tools experience that businesses and Industries wanted, like for example Predictive analytics ability, advanced visualization and smart suggestions. Majority of the business analytical tasks require routine reporting and dashboard creation for presentation to managers, which is most times very repetitive and mundane. QlikView makes such work much simpler and easy for them [177].

15.3 Architecture

Before moving forward to deploy and use QlikView software, it is very much important to understand the architecture of QlikView, its products and the components. The QlikView deployment has three main infrastructure components. They are QlikView Developer (QVD), QlikView Server (QVS) and QlikView Publisher (QVP). The QVD is a desktop application for Windows operating system for designers and developers for performing operations like data retrieval, storage and processing, and also to make graphical user interface (GUI). The QVS is a component that handles the interaction or communication between the clients and QlikView applications and components. It also loads QV applications into the main memory and helps in running the user selections. The QVP is collects and loads the data from various sources, in different formats ranging from XML to CSV. It also reduces the QV application and then distributes the data to the QVS server. It is always good to separate the two components of QVS and QVP, since both function differently, handle the memory and CPU differently and have completely different roles [178].

Broadly speaking, QlikView's architecture can be viewed as the combination of two main components. They are the front end and the back end. The front end's function is to visualize the processed data whereas the back end's function is to provide security and publication mechanism for the new user documents.

15.4 The Front End

The front is the component that facilitates users to interact with the data and documents stored for data processing through the QlikView server, anywhere and at any time. The QlikView Publisher in the back end creates QlikView documents of the user, which are contained in the QlikView server of the front end. These files are stored in the formats of QVW, meta, and shared. The interaction or communication between the user or client and the server is managed through HTTPS, or QlikView Proprietary, also called QVP protocol. The QlikView server also handles the security of the client.

15.5 The Back End

The back end is the inner component of the QlikView where the QlikView documents that are created by using QlikView Developer are stored and protected. The files can be script documents for data extraction from various sources like SQL scripts, the data that is in binary form and stored within the QVD files. The most important component of the back end is the QlikView Publisher, which is responsible loading and distribution of data. The back end functions within the windows environment and required special privileges and permissions for its functioning.

15.6 Associative In-Memory Technology

Associative In-Memory technology is used by QlikView for the analyzing and processing the data. The advantage of this technology is that it stores only unique entries at a time in the memory and rest all are pointers to the main data. This makes it very fast and allows to store large amounts of data than other traditional methods. The performance and scaling is key to QlikView as the user is directly connected to the CPU of the QlikView, and hence this technology is used for speed and scaling.

The QlikView's System resources consists of computation components like CPU, RAM, etc. The QlikView Server and QlikView Publisher are the two components that use and handle these resources differently as they serve different purposes and exhibit different roles. Let us look at how these two utilize and handle these resources.

15.7 QlikView Server (QVS)

15.7.1 CPU

The QlikView Server consists of multiple CPU cores that are multi-threaded and optimized. The available cores are used linearly for processing the QV files. The QVS manages the usage of these cores for computation of these files on real-time basis by monitoring and allocating the cores efficiently. It leverages the processor for dynamic aggregation creation quickly and shows the results to the end user intuitively. Typically, the actual data that is stored in the memory or RAM is in unaggregated raw form. Thus, to perform aggregation operation on real-time bases very quickly on huge sets of data, efficient and dynamic use of computing power is needed. If the processing power is not sufficient, it may lead to waiting and the processes are done based on priority, where a waiting list is created for allocation to the cores based on priority and cores availability. This is a linear and sequential processing model.

15.7.2 Memory

The QlikView documents and files are stored in the main memory RAM, which is the primary storage for all files. The data that is stored in the RAM can either be in unaggregated form or aggregated form. QlikView memory is based on Snapshot technology, where it is continuously refreshed through a process known as reloading a QlikView document. As the QlikView document is loaded, it will establish connectivity to the data sources that are to be analyzed and the unaggregated data to be extracted and for compression, which is then stored as .QVW format in the persistent disk storage.

When a user opens an analytical application, QlikView loads the .QVW file from the persistent disk storage into the RAM. Further, QlikView only relies on the dataset that is loaded into the RAM at that time and does not care about the other data in the database. This is done because QlikView would be able to handle the process quickly, resulting in instantaneous real time response to the end user. Thus, all the data that has to be processed is placed in the RAM. RAM is the important deciding factor for QlikView about how much data it can handle at once. A Windows Operating System will consume a RAM of 500 to 1000 MB, while QlikView Server process requires a RAM of at least 100 MB on QVS.exe process.

15.7.3 Data Compression

For effective usage of RAM and quicker real time processing, we need to forgo redundant and repeating data. This can be achieved by loading on distinct data points into the RAM instead of all the data points. This not only reduces the RAM usage and scales up to accommodate huge dataset but also makes the process very quick due to non-redundancy in the dataset. This process of reducing the data points is called Data Compression and is handled by the QlikView server.

15.8 QlikView Publisher

15.8.1 CPU

QlikView Publisher is a database load engine and it creates thread for every database connection to facilitate hundred percent utilization of cores during processing. The number of cores being utilized for a process is equal to the number of databases being loaded for the process. The QVS and QPS are not placed on the same server because both use and handle CPU cores in different way as explained earlier.

15.8.2 Hard Drive

The QlikView creates a data repository of historical data that the end users have loaded from various applications. This is also known as data cache. The main advantage of this kind of data model is that it reduces the database communication and reduces the time lag. The main disadvantage of this is that the disk space needed to source files is very high. It is recommended to have at least 150 GB space of SAN drive.

15.8.3 Memory

As we noted earlier, as QlikView Publisher is a database load engine and a file distribution service. It is not just an analytics service engine. Compared to the QlikView Server, QlikView Publisher is not memory intense and hence memory is not a consideration or a factor for determining the size of the service to accommodate huge QlikView Publisher instances [179].

15.9 USES AND ADVANTAGES

QlikView is widely used by businesses for Business Intelligence, analytics, visualization and reporting task to monitor business performances on daily, weekly, month, quarterly or yearly basis. It can be used both for on-site reporting as well as client-side remote reporting on real time through internet. Hence, it can be accessed at any time or anywhere through internet. From user point of view, it is very simple and conformable UI, that is drag and drop based and hence not at all difficult or hard to master and operate quickly. The reports and visualizations can be interactive, which makes it feel good and easy for understanding to the viewer. All these advantages make it one of the most preferred BI tools in the industry [180].

15.10 Conclusion

QlikView being a web-based software tool is very simple, reliable and robust and access-able for various BI applications. It is easy to learn, work on and present compared to traditional software available in this segment. Yet, it has got lot of short coming too, where it cannot be used for large scale data analytics like Big-Data and it does not support advanced machine learning based predictive analytics. However, for small to medium-large level data sets, it is one of the best software that is available for business analytics and visualization tasks.

16 Big Data Analytics in E-commerce

Bo Li
bl15@iu.edu
Indiana University Bloomington
hid: fa18-523-85
github: ☁

Keywords: E-commerce, Consumer Behaviors, Python, TensorFlow, Big Data, Deep Learning

16.1 Abstract

In recent years, online shopping has become a more popular way of consuming. Traditional retailers are eager to find a way to maintain revenue, online retailers are also finding ways to extend the market. The rise of ‘big data’ had impacted on marketing research and practice a lot. As technology developed, we have huge data on consumer behaviors which could be very detailed and accurate, but how to mine the data is a problem. In this article, we talk about the application of big data in the consumer behaviors data, subsequently discuss how the TensorFlow can translate the data into valuable conclusions in consumers’ behaviors research.

16.2 Introduction

The Big Data has a common definition, Big Data always comes with 3V: volume, velocity, and veracity. The Internet allows us to do almost all work online and keep records of our actions. If you listen to a song in the playlist, maybe iTunes will record it as part of your individual activity log, which could be the dataset that explores your interest. If you often use Uber to commute between your home and your company, maybe they could picture your daily life including the places that you have spent time in. If you use your device to safari on the internet, your action of clicking on several links could also be recorded and researched since the action contains you using habits and preferences [181].

The online shopping data includes the consumer's all kinds of information: age, job, education, catalog preferences, price sensitivity, etc. But some of them are not presented to us directly, mining the consumer behaviors is an appropriate way to get access to the hidden information. How could we mine something meaningful to explore consumer behaviors and provide valuable insight? The answer lies in several using cases and the understanding of market research and also human psychology research.

In order to extract the knowledge behind the commercial data generated by hundreds of thousands of consumers for the use of leading managers to make the decision, it is necessary to conduct a deep analysis to the commercial data, instead of generating simple reports [182]. The deep analysis could hardly be done by SQL since the process relies on complex models. Without those models, it is impossible to get a profound understand of the commercial data [183]. People will not only need to find out what is happening now, but also need to use data to make some predictions in order to make preparations for the future events. For example, if the manager is able to predict the loss of the customer in the future, they can use discount to attract the users again [184].

In the context of big commercial data, the traditional OLAP operations are not enough anymore to meet the requirements, we also need path analysis, time series analysis, graph analysis, what-if analysis and some complex statistical models. Time series analysis, a useful method in the commercial data analysis since we have got lots of the trading historical data. The managers want to get some patterns in the data in order to fine some chances to improve the revenue. By the trend analysis, they can even predict some chances in advance. In the financial area, analysts are able to develop some software to conduct the time series analysis of the trading data, and find some profitable trading patterns. After further verification, they can use those profitable trading patterns to conduct the real trade and make profits.

16.3 Technology Background

TensorFlow is an open source software library for numerical computation using data flow graphs [185]. TensorFlow could help developers to transform from code to graph, which could benefit developer in understanding their work, and the term tensor, is generated in the process, as the tensor will go from the beginning to the end of the graph, so the technology is called TensorFlow. The process of computation could be done in CPU or GPU, as we know, in blockchain, GPU works better than CPU since the fundamental design of GPU fits better in the computation of mining coins. TensorFlow also has a data visualization module called TensorBoard, which contains the common drawing tools as well as some useful templates for the developers to visualize their data. There is no doubt that the graph will be more clear than codes especially when the structure of data is very complex. And the graph could give the readers a direct presentation of the data, which is worthwhile since it could reduce the communication cost between different developers.

There is a team in Google called Brain team, which is the initial developer of TensorFlow, there original purpose of the development of this module is to improve the efficiency of machine learning. For example, if the deep learning task is to predict a result based on a training dataset, the more layers you have, the more accuracy you will have. But more layers will cost much more time, so TensorFlow is created to solve the problem. One more important thing is that in the previous version of TensorFlow, it began to support distributed computing which means more resources could be deployed in the process so the efficiency will be boosted.

To support more developers, Python API and C APIs are also available in TensorFlow. One thing that must be pointed out is that, although TensorFlow supports many kinds of languages, the Python API is the most efficient one since Python does better in the feedback process which focuses on improving the model. And there are also more examples in Python since most of the machine learning work is done by Python.

16.4 Big Data Applications in E-commerce

Since the design of the website is getting more complex than before, the users may conduct different operations in different pages of the website, but most of them are very import to provide an essential information for us to find the customer’s preferences.

> ***“There is a way to achieve that which is called four rights. Talk to the right audience, through the right channel, with the***

> ***right message, at the right time” [186].***

> ***“Customer acquisition: Marketing will target high-value customer segments identified by behavior analytics and study behavior patterns to determine the best potential offers. Customer engagement: Behavior patterns will be used to generate personalized next-best, cross-sell and up-sell offers, while behavioral customer segmentation will be used for more general customer marketing offers. Customer retention: Behavior patterns will be used to detect possible customer churn and generate next-best retention offers” [186].***

The strategic meaning of big data is that deploying professional analysis on those meaningful datasets generated from the E-commerce trading. Some meaningful things are hidden behind the big data, mining them is the main task of the application. The improvement in the value of data is the most important part of the benefit of using big data, especially compared with the previous situation, that the data useless since the huge amount [187]. Such information has been stored in the log document, but it is too massive and fragmented to analysis it with limited technologies and techniques.

Big data could help the enterprise to have a more profound understanding by analyzing users’ behavior data, which allows the enterprise to establish strategies with more specific aims. This could make the enterprise to be competitive in the market and win more consumers’ hearts. For example, the user may want to buy a guitar for himself, and he has browsed several kinds of guitars and could hardly to make his decision. In the E-commerce domain, users have generated huge amount data about their every action: browsing products, clicking on the details of products, adding the products into the wishing list, adding the products into the cart, delete the products from the wishing list, clicking dislike product button and querying more information to the seller of the products [188]. The big data technologies could base on his operation history to find his acceptable price level and the specific version of guitar (such as with or without pick up system), then we can put such requirements into our database to find the appropriate guitars and push those potential choices to the user’s interface, which could realize an improvement in sales transformation rate.

TensorFlow could be a kind tool to analyze consumers’ behaviors since the model of recommendation is doable in TensorFlow. Due to the feature of distributed computation, the efficiency of the model is good enough for a small scale recommendation system. To have a more profound understanding of user behavior, a whole lifecycle of the user is needed to be established for the analysis of user behavior.

16.5 Features of User Behavior in the E-commerce Platform

The online platform is the main platform for e-commerce which lists the product in different ways and provides the whole chain of finishing the trade. Different from traditional commerce, online e-commerce has some special features which could be used in the big data.

There are fewer limitations for consumers in the B2C pattern since the online platform can run for 24 hours if it is well maintained. Consumers are able to conduct any operation (browsing, selecting, finishing the trade) at any time in anywhere.

The trading cost is much less than the traditional commerce pattern. For the consumer, time cost, transportation cost and delivery cost are lower than the traditional commerce pattern. Their trading action is much simplified by the online shopping systems, the trade can be done by several clicks on the mobile device.

The online product can offer a more attractive price due to the advantages of the internet. Comparing with the traditional commerce, online sellers have less item to pay for maintaining the shop. There is a lot of extra costs for the real store.

Customized service. The recommendation system is able to recommend the most wanted goods for each customer based on their user behavior and the big data technology, which is the traditional commerce cannot achieve due to the cost. The customized service can benefit a lot in the transformation between browsing and buying.

More kinds of product and no space limitations. Since the information of the product is much smaller than the product itself, and the online store can exhibit them all at the same time, so there are more choices presented for the consumers.

The information is easy to get. Every item in the online platform has been labeled by the system, so the search of the item is very convenient for consumers, the cost is significantly lower than the traditional commerce.

16.6 APPLICATIONS

Due to the hotness of machine learning and deep learning, there are a lot of applications in every domain. Search ranking and recommendation are the most common two applications.

> ***"Recommendation systems in particular benefit from specialized features describing past user behavior with items" [189].***

Just like search ranking, recommendation systems also have a problem of the balance between memorization and generalization. Memorization can be seen as the representing of the relationship between the products and users, which can be extracted as vectors. Generalization is to generate rare feature combinations in order to serve for the recommendation systems [190].

TensorFlow, with so many advantages in machine learning, is very appropriate for the recommendation system. Since the features of products could be learned by multi-labels classification, and the user's features could be learned in his historical actions in the online platform, which has the record of every consumer's trading history. When we have both features of products and users, we can establish a recommendation system by matching the two objects. Besides, the model can be judged by the dot product between the two vectors.

To establish such a recommendation system, we need to fit the TensorFlow since it is tensor that flows in the whole model. The transformation from dataset to tensor is a necessary step to conduct the model. And the users', as well as the products' character tensors, need to be transformed into the presentations of users and products by the embedding function. The next step is to generate the recommendation by the pair the presentations of users and products. Such pair of presentations contains the most match user and product calculated by the model, the vectors in the model contain all the information of the user and the product. The last step is to compare the generated score and the actual comment from users to define the result's quality, which is called loss function [191].

TensorRec scores recommendations by consuming user and item features (ids, tags, or other metadata) and building two low-dimensional vectors, a "user representation" and an "item representation". The dot product of these two vectors is the score for the relationship between that user and that item, the highest scores are predicted to be the best recommendations.

The representation function in TensorRec can be set up by developer's preferences, it could extract the features of users as well as products. It can be very convenient for developers to set the parameters independently since the scenario varies in different cases [191].

16.7 Conclusion

Information is booming in recent years, data and internet techniques are spreading in everywhere, with the significant effect on consumers' deciding pattern and purchasing pattern. The digital economic on the internet has become the focus of all the domain. For the biggest group in the digital economics in the internet, online consumers are the focus in the specific domain. How to draw the picture of the users and get the key feature of their behaviors have become a hot topic. Besides, the social network analysis is also important in the commercial data analysis. Different buyers may be divided in different groups, the members in the same group could have the same interests. Social network is the study of social entities (people in an organization, called actors), and their interactions and relationships [192]. The interactions and relationships can be represented with a network or graph, where each vertex (or node) represents an actor and each link represents a relationship. From the network we can study the properties of its structure, and the role, position and prestige of each social actor. We can also find various kinds of sub-graphs, e.g., communities formed by groups of actors [193]. Social network analysis is useful for the e-commerce because the group of buyers is essentially a virtual society, and thus a virtual social network, where each page can be regarded as a social actor and each hyperlink as a relationship. Many of the results from social networks can be adapted and extended for use in the Web context [194]. The ideas from social network analysis are indeed instrumental to the success of Web search engines.

Benefitted by the internet, we have all the records of most of the online activities of users, but the relationship between those actions and the user’s features is ambiguous. Basing on the TensorFlow, we are able to use a low-dimensional vector to represent the user’s features as well as the products’ features. The algorithm allows us to extract the key point of users as well as products, which provide a base for the recommendation of the product.

The big data technology can help us to mine the black box of the relationship between the actions and the features. Several factors which measure the user’s preferences can represent the user, and those factors are also the key parameters in the model. Once we get a clear picture of the user, we are able to customize the recommendation, which can not only improve the user’s experience and also improve the revenue of the online retailer.

17 Cloud AutoML

Keerthi Naredla
knaredla@iu.edu
Indiana University, Bloomington
hid : sp18-616-02
github: ☁

Keywords: Cloud AutoML, Neural, Google Cloud, Vision

Cloud AutoML is the state-of -the -art tool to design high-quality training and large-scale capable custom ML models. Using this, a user with no or less ML knowledge can build ML model in short span of time, low-cost hardware and infrastructure. It is a revolution by Google team to get ML to small businesses with less AI/ML expertises. It is based on 2 important technologies transfer learning and neural architecture search technology. The Neural Architecture search uses the concept of learning to learn or meta learning with an auto-regressive controller which builds a neural network on learning from feedback of child network. As it is immensely expensive to compute and execute deep neural networks, Google Cloud Services such as Google Compute Engine, Google Cloud Storage support Cloud AutoML. In additon to this, Google API's provide the ability to customize pre-trained models in Cloud AutoML. This section gives a brief introduction to Cloud AutoML, the technology behind its functioning,and the importance of Google Cloud in Cloud AutoML.

17.1 Introduction

Cloud AutoML is an innovative tool with simple graphical user interface to train and test users custom machine learning models. This is a result of collaborative effort of Google Cloud AI, Google Brain and other Google AI teams. The main purpose of developing Cloud AutoML is to enable users and businesses with limited machine learning expertise to easily build and train high quality custom ML models. It is built on Google learning to learn, transfer learning, and Neural Architecture Search technologies 195.

Cloud AutoML is a suit of Machine Learning products. Google has recently launched first product under Cloud AutoML: AutoML Vision which is a service to access a pre-trained model or create a custom ML models using Google Cloud Services, for image recognition, detecting image content, classifying images and image-based recommendation system. It offers drag-and-drop interface to upload images, train and manage models. Similarly Google is working to support integration of it's poweful API's into Cloud AutoML. Few of the API's that are in real demand are, Google Cloud Video Intelligence API which makes videos searchable, Google Text-to-Speech, Speech-to-Text which is highly used in smart-home devices, automation tasks, Natural Language API which is useful for text-analysis, extracting useful information from users, Google Translation which is useful for language detection ,conversion and Google DialogFlow which highly improves interaction and conversation with voice assistants. Thus using Cloud AutoML, a business can customize ML model according to their needs by selecting any one or a combination of these API's 196.

Just like cloud servers are used by several companies, small and big, without any knowledge of underlying complexity involved in storing, distribution and processing, the cloud automl can be used to build customized neural network that serves the purpose without actually understanding the complexity of generating a model 197. Not only that, it is time-efficient and high-accurate because the base model is already pre-trained on immense data-archives and the resources used by Google. Also, AutoML generated models run instantly on Cloud ML infrastructure leveraging the hardware and powering Google's own cloud.

With these major advantages, Cloud AutoML Vision is already in use by many good companies across the globe. For instance, shopDisney uses label-detecting feature with Cloud AutoML technology to build vision models in order to label products with Disney characters, and product categories. Also, these annotations are integrated into search engine for better product recommendations. Urban outfitters use cloud automl to automate product attribution process to recognize products like patterns and dressstyles, which is very useful in terms of accurate search results, product recommendation. Zoological Society of London are actively using AutoML Vision to categorize different animals from the images captured in order to analyze and understand animal motions, distribution and human impact on wildlife 195.

17.2 MECHNISM

This is section is based on Google’s 2017 paper on Neural Architecture Search with Reinforcement Learning, by Bareet Zoph,Quoc V.Le Google Brain Team 198. Neural Architecture Search with reinforcement learning is the basis of the Google Automl. In addition to this, the concept of Transfer learning which is to make use of pre-trained models, in order to build a custom model with small changes to base model, is the motivation for Cloud AutoML 199. Many of the previously proposed methods like Hyperparameter optimization, Neuro-evolution algorithm, sequence-to-sequence learning lack some of the core concepts like able to generate variable-length network, able to work at large-scale, learning from reward signal without any supervised/manual intervention, making use of previously learnt information or feedback framework which is also called learning to learn or meta learning.

A recurrent neural network trained with reinforcement learning generates a convolutional architecture. In this model, the concept of reinforcement learning has a RNN called controller which is used to generate a variable-length string, by constantly training the network with results of child-network on the validation-set. And the result which is in fact know as reward signal is processed through policy gradient-method to further update the controller. With increase in number of iterations of this process, the neural network grows, resulting in higher accuracy. The key additions to this neural search architecture model are: (1) Parameter-server scheme which uses distributed training of child network and allows asynchronous updates to the controller. This parallelism speed up the training process,rather than spending hours on each child network. (2) Skip connections and branching layers are used by the controller in order to increase the search space and also the complexity of the architecture as a whole rather than standard RNN.That is with skip connections the controller can decide on what input layers should link to the current layer, rather than choosing just the previous layer.

Using Neural Architecture Search the novel model built on CIFAR-10 dataset is called ConvNet for image-recoginition, has 3.65 test set error,that is 1.05x faster than best human-invented models and novel recurrent cell designed for Natural Lanaguage Processing on Penn Treebank dataset, results in 3.6 perplexity better than any previous RNN, LSTM models. Usually, to build ML models for such large datasets take not only enormous amount of time but also immense effort of ML experts would result in better architecture. But with the concept of Neural Architecture Search, a machine can generate a recurrent neural network that is far better than experts built state-of-the art models. Hence Neural Architecture Search achives building best models from scratch, with less human intervention, less time and high test set accuracy 198.

17.3 Role of Google Cloud

Google Cloud is the one of prerequisites for functioning of Cloud AI services such as Cloud ML Engine, Dialogflow,Google Cloud Job Search and Discovery and other Google API's. Using Cloud Machine Learning Engine, data scientist and ML expertise can work together to design a ML model with help of Tensorflow and then train, test the model on large scale processed data deployed on a cluster 200.

Although Cloud ML Engine supports training of the model to the extent of high accurate prediction rate, we need Google Cloud Storage for storing input data for training and testing, staging all the dependencies of the custom ML model into a trainer package, writing training artifact and storing training and prediction output files 201. Similarly, Customized Cloud Vision models which are usually deep neural networks have several millions of nodes, that need to undergo multiple training cycles to acquire high performance, high accuracy, and this results in computationally intensive even with special hardware infrastructure 202. Fei-Fei-Li, Cheif Scientist Google Cloud AI, mentioned that "Google's infrastructure is the solution to speed training times. Google has specialized ASIC, GPU and TP hardware in its cloud to accelerate training and improve the ROI with on-demand cloud resource utilization" 202.

Google's Cloud TPU:Tensor Processing Unit is crucial for lowering the time required to train comutationally intensive models. It is built with application-specific integrated circuits, and consists of 180 teraflops computing power with 64 gigabytes of high-bandwidth storage memory 203. It is flexible to shift models running on Tensorflow to Cloud TPU. And it is important to consider Cloud TPU, especially if the training dataset is huge, increasing, and model takes several cycles to achive accurate prediction, as it leverages requirement of local datacenters setup. Also with XLA just-in-time compiler and Cloud TPU hardware which has matrix unit (MXU) it is possible to train large models with very large batch size that typically takes months and years in few weeks or months. But Cloud TPU cannot be used if the neural network isn't built using Tensorflow or the main training loop of tensorflow program consists of operations, in that case GPU: Graphic Processing Units can be used instead of TPU. GPU's are also useful to accelerate machine learning workload and these can be simply added to VM instance on which model is running 204. Therefore, it is important to note that TPU's are not the only option,rather Google makes use of GPU, CPU's to run machine learning worklaods on Compute Engine when required. Thus, Gloogle Cloud Services comes with all of these resouces which play a key role in Cloud Automl and other Cloud AI products.

17.4 Conclusion

Cloud AutoML is certainly a mission to get profound concepts ML,AI, deep learning neural networks into usage for any company. Google terms this mission as "democratizing AI, point-and -click AI for all", as it is not just about leveraging the technology to build deep layers of neural network but it is also about leveraging the the infrastructure through Googgle Cloud Services especially Cloud TPU, required for getting the ML model into production in large scale and accuracy.

Moreover, with powerful Google API's such as NLP, Speech, Vision, Translation, building customized ML models becomes feasible and flexible. Thus Google's Cloud AutoML is sucessful in solving the issue of requiring highly skilled and experianced Machine learning experts to develop advanced neural networks and significant amount of time taken to build such models manaully can now be machine-generated with Neural Architecture Search with Reinforcement Learning and reused with Transfer Learning. Hence any company can have AI/ML products that could match the quality and speed of Google AI products.

The authors would like to thank Dr.Gregor von Laszewski for his support and suggestions to write this paper.

18 Amazon Relational Database Services (RDS)

Arijit Sinha
arisinha@iu.edu
Indiana University, Bloomington
hid : hid-sp18-520
github: ☁

Keywords: Amazon RDS, RDS, AWSRDS, AWS RDS

Amazon Relational Database Service also known as RDS is cloud computing platform providing a prime web service to operate with relational databases. With AWS database services, it provides an mechanism for creating/ replicating/ migrating any existing databases on AWS cloud.

18.1 Introduction

Amazon Relational database services are the web service which has been effectively used for handling and managing relational databases, which in return provides high performance, security, maximum availability and compatibility. Amazon is providing a SQL database by the service which is known as RDS, which provided key features related to easy management and tracking or monitoring. It is compatible with variety of database engines running in background including Amazon Aurora, PostgrSQL, MySQL, MariaDB, Oracle and Microsoft SQL Server.

18.2 Key features

Amazon RDS provided is the latest infrastructure platform with updated database softwares and management tools to maintain and perform databases administration with security and fault tolerant features. It check and updates the latest patches of software and ensuring the database are running on latest software and hardware. It schedule major and minor releases of the software to keep the Infrastructure platform updated. It also allows to configure with previous DB Instance version. Below are some popular features

- Scaling Storage-It can automatically increase the storage once size or volume of the databases ins reaching its maximum capacity. It can also scale storage based on high volume data getting loaded or read from the database. Based on the usage trend, it can scale the database services. It handles the read request effectively and optimally with read replicas, which is to create the replica of the database. Also, upon usage if it is not needed, it can be deleted from Management console or API and with API, it can manage the same operation as possible with management console for create, start, stop, modify, fail over, describe, authorize and add DB clusters and DB instances.

- Offers less Administrative workload-When the services are getting setup, all the databases instances are configured with its respective database engines. Amazon provides command line and management consoles for easy administration of the databases.

- Reliability-It can replicate the data to a standby instance on different Availability Zone using the Multi-AZ DB instance. It provides automates backups, user defined snapshots of the data stored on Amazon S3. In the event of failure of an hardware, it can automatically replace the instance. Upgrade from single AZ to multi AZ can occur with no latency or downtime. Once the selection is done from upgrade to Multi AZ, the snapshot of the instance is captured, after which a new instance is built from the snapshot and configuration is setup for taking or keeping the Multi AZ databases in sync [205].

- High Performance and Secure-It provides high performance using General purpose SSD storage and Provisioned IOPS SSD storage. It provides the feature of encrypting the databases using keys (AWS Key Management Services). Along with, it provides Amazon VPC for network isolation for databases on cloud to securely connect with on premise applications. With in this Amazon VPC, we may have multiple subnets with at least on the AZ zone. Data restoration or migration outside VPN is prohibited and it is not supported.

18.3 Building Blocks for Amazon RDS

DB Instances are Amazon RDS primary building blocks which is a secured database environment on AWS cloud. DB Instance can consists of multiple databases. As mentioned in above section, these DB instance can be easily managed using simple API, AWS management console and Command line interfaces to set the configurations and monitor the behavior and capabilities of relational databases. It does not need any additional database maintenance software. In these DB instances, we can have multiple databases created by many users or applications. In the background, we have DB engines interacting with DB instances. Few of the examples can be MySQL, Maria DB, PostgreSQL, Oracle and Microsoft SQL Server DB engines [206].

There are 3 types of storage available with DB instances (Magnetic, General Purpose SSD and Provisioned IOPS). Storage capacity depends on various storage type and respective database engines it been configured. Amazon RDS can select IP address range, subnets, access control list and configure routing to make it more secure and reliable. Another component provided by Amazon is IAM (Identity and Access Management), which this you can provide provision on users to create, delete, modify read any DB instances.

18.4 RDS Automated and Manual Monitoring Tools

Amazon RDS can monitored for it performance and can be reported in case of any issues or failures on DB instances, DB clusters DB Cluster Snapshots, DB parameter group or DB security group. On real time, DB instances or clusters can be monitored. It also maintains the database log files which can referred or consulted in cases of any failure or issues encountered. Amazon RDS also provided extra feature for monitoring with CloudWatch for metrics, alarms and logs, along with service health status. With Command Prompt-using below command can view performance metrics and alarm-

```
$ aws cloudwatch list-metrics --namespace AWS/RDS
```

[207]. With AWS CLI Set the alarm command

```
$ put-metric-alarm
```

[207] With API-using the CloudWatch API GetMetricStatistics with start and end time can provide detail metrics on performance and form setting up alarm with Put Metric Alarm [207] on DB Instance.

Based on the user defined baseline for performance and resource to be monitored Amazon RDS store the respective monitoring logs including your CPU, RAM, Disk Space consumption. It can monitor the network traffics and help in deciding the throughput details. With Amazon RDS, it helps in Disk space monitoring helps in taking decision, if the data needs to be purged or archived. Number of users connected to database can be monitored with kind of operation getting performed on Database with Amazon RDS console. There are configuration monitoring to identify the changes to configurations on DB instance using a service AWS Config like security, subnet, events on DB instance. Various Amazon RDS metrics dimensions can be on name of the Engine, specific DB Instances, DB Clusters with Roles and database class. It uses EBS volumes, which it automatically adjusts to upgrade and enhance performance. IOPS (SSD) storage is recommended with High workloads from Online transaction processing data. General purpose (SSD) storage is recommended for workloads with small scale on database.

It also provides the enhanced monitoring, which will be used for monitoring the health of DB instance. It will help in monitoring the operating system with process being executed details. It will capture the system level metrics for CPU, memory, file system and disk I/O. Amazon RDS can also encryption on databases with Amazon Key management services It can restore point in time data as part of recovery process.It can automatically initiate the failover process, if we can not access the primary AZ, can’t connect to primary on network, failure of storage. Database events can be integrated with another amazon service known as Amazon SNS, which can send the SMS text messages [207].

18.5 Create DB Instance-Example with MySQL

Creating a DB instance using MySQL Database engine. As prerequisite, we need to have access on AWS management console. For initial DB Instance set-up, it is managed through AWS Management Console. Search from Database section with listing name as RDS and Navigate to open Amazon RDS Console. Next, its needed to have the understanding on what configuration is needed for creating

MySQL DB instance. For example, we can have certain amount of storage and backup strategy to be build. From the AWS RDS Console, we first need to select the region, also known as Availability zone, where we can host AWS RDB activities. From AWS RDB Console, we next need to launch DB Instance from Instances menu. Next, we will provided with option to select the SQL Engine for the DB Instance. As we are taking the example for MySQL, we will select the MySQL Engine. From Console, we need to select the Database engine named MySQL and Need to select the purpose of the database instance like for production purpose, we can select the radio button as Dev/Test. In this step, we need to specify the DB details or configurations DB Instance Specifications with license, engine version, class, multi-AZ deployment, storage type, storage allocation.

Provide the Identification name of DB Instance, username and password. DB Instance can be configured with network security-Virtual private cloud, security group, subnet group. Provide the database name, port-default to 3306, DB parameter group. Along with backup strategy and monitoring capabilities cab be defined while creating.

Next launch DB Instance button on the console. These above steps have completed the creation DB Instance.The Status- creating for DB Instance changes to ready for use and then to available. Once the status is changed to available, it is ready to be connected with SQL client.

In this example, we are configuring MySQL, so we can download and connect using MySQL workbench as SQL client for MySQL database.

On MySQL Workbench-Database, Click on connect to database. We can pass the connection parameters like hostname, port -default to 3306, Username, and password.

Once connected to the database, we can perform DDL, DML statements on the database. You can connect to Read Replica as the same way with details on endpoints. DDL statement can also be performed on read Replica [208].

18.6 Paid Service

DB Instance hours, Storage per month, I/O request per month with data transfer, backup storage with provisioned IOPS per month are paid services. So the services must be deleted or stopped and avoid extra billing than usage [205].

18.7 Delete DB Instance-Example with MySQL

Once logged into Amazon RDS console, navigate to Instance Actions and hit the delete link. DB instances can be deleted after taking the final snapshot or it can be deleted with capturing final snapshot of relational database Using the CLI, there can be a option to issue commands for deleting the DB instances. Below are the CLI statements for various operating system.

```
$ aws rds delete-db-instance --db-instance-identifier mydbinstance \
--final-db-snapshot-identifier mydbinstancefinalsnapshot
```

With above command, it will first or create the final snapshot of data and then proceed with the deletion of DB instance [209].

18.8 Conclusion

Amazon RDS is provides highly optimized and high performance web services supporting multiple type of SQL databases, providing service to easily customize, configure and monitor the DB activities and administration. We have many sources of data getting generated and are gets used for multiple purpose of analysis, interpretation. This cause for highly complex and high performance databases, which as a service is provided by Amazon to maintain a relation database on cloud. With disaster recover mechanism reduces the risk of downtown and latency.

18.9 Acknowledgement

The authors would like to thank Dr.Gregor von Laszewski for his support and suggestions to write this paper.

19 Cloud Domo

Ritesh Tandon
ritandon@iu.edu
Indiana University, Bloomington
hid : hid-sp18-523
github: ☁

Keywords: Domo, API, token

19.1 Introduction

Domo is cloud based data integration platform that enables employees to engage with globally distributed data in real time. It provides flexibility to outside partners and third party vendors to integrate and collaborate with data. Domo has more than 400 data connectors. Data can be accessed directly from public or private cloud regardless if it is available on-premise or proprietary systems. Data is heart of information for any business. it is very trivial to find relevant data that is requiredby different people working in different departments of large organization. More so, the bigger challenge is to derive insights from the data after it is located. Domo transforms the way orgaizations employee access, use, analyze and share data. Domo gives power to users and helps them make decision in real time. Domo can be thought of as cloud based data operating system that has the ability to handle and process data regardless of its type and location. Domo brings different data sources spread across different locations at one central location so that it can be easily accessible for use. Domo lets user share and collaborate different data sources. Users can also visualize and report data through Domo. It also has reliable data management feature that provides high level of security, speed and scalability. Domo makes data available on any device of any size thus making it truly mobile.

19.2 Domo Inbuilt Solutions

Domo has custom inbuilt dashboard and visualization solution for different roles (such as BI, CEO, Finance, IT, Marketing, Operations, Sales and Services etc) within organization and for different industries (such as Education, Healthcare, Manufacturing, Hospitality, Retail and Transportation etc)

19.3 Data Connectors

Data Connectors is the heart of Domo. Different types of data sources, such as relational , non relational, flat files , csv Cloud App, cloud File, third party API etc can be easily connected through Domo.

Cloud App Connectors; Domo has more than 400 cloud app connectors including all famous ones such as amazon s3, AWS, Adobe analytics, Google Analytics, Facebook, Fitbit, instagram and salesforce etc.

File Connectors; Domo can also connect to data that is stored in files such as excel and/or csv files.

Database; Domo has connectors for connecting relational, non relational, SQL and NO-SQL databases such as Oracle, MS SQL, MySQL and Hadoop etc.

On Premise; Other than cloud based connectors Domo can also connect to on premise databases/files etc as long as security protocols are opened securely for connection.

Api; Domo has Dev Studio tool for creating custom apps. It is best suited for developers having web development experience (java script, css, html). Domo App CLI is the main tool that is used to create, edit and publish app designs to the Domo instance.

19.4 Connector Dev Studio (Creating custom connector)

Connector Dev Studio enables developers to create their own data connectors if they do not find any existing connector in Domo library. They may chose to use that connector exclusively for their own organization or can make it publicly available for contributing to community. Developers need to have JavaScript knowledge in order to build their own custom connector. Certain conditions are required to be met in order to build custom connector; APIs must use https, it should be REST API , it should either require No Authentication or should be able to get authenticated using OAuth 2.0 or API key or through user name and password.Build Now menu command lets the developer navigate to Connector Dev Studio Integrated development Environment (IDE). Developers have to upload their custom connector icon image, configure authentication , define reports and define data processing steps. After completing these steps developers may submit their custom connector for publishing. Domo developers will review, validate and notify developers when their connector will be ready for use [210].

19.4.0.1 Custom Connector - User Authentication

Developers have to write code block for validating API credentials and authenticating using username and password. They need to pass encoded user name and password to request header for authorization. After reading the response Developer needs to make use of authenticationSuccess() and authenticationFailed() method to let the user navigate [210].

19.4.0.2 Custom Connector - Configure Selectable Reports

Domo Custom connector also has ability to let the developers define reports that their custom connectors can contain. These Reports provides extensibility to developers of calling different API endpoints to acheive different function. This lets user of the custom connector chose reports they wish to use. These reports appears in Report dropdown menu once connector is published [210].

19.4.0.3 Custom Connector - Data processing steps

This step let the developer define data processing and transformation steps of the data that is retrieved from API endpoint call. This is performed for every report that developer has defined for their custom connector. Developers need to write script for parsing , manipulating and storing data in Domo. Data structure needs to be defined in code using datagrid.addcolumn() method and then data is added one row at a time [210].

19.4.0.4 Custom Connector - Sending data to Domo

Developers have to ensure that their data is correctly uploaded and represented in Domo. In order to do so, developers have to make use of Domo Create/Update dataset and Run Script command. Success message will confirm that data is successfully published in Domo [210].

19.4.0.5 Custom Connector - Submission of custom connector

After completing above steps, developers have to submit their custom connector for publishing using Domo Submit For Publishing command [210].

19.5 Data Flows and Transforms

Cleaning data is herculean task when dealing with dirty data that needs to be cleaned before reporting. Domo has Magic ETL tool that makes data cleaning job looks easy. It helps join, transform and tidy up data with drag and drop ease of use 211.

Domo also has SQL data flow that let the developer select data set, perform transformation operation through SQL query and generate tidy and processed output dataset. Domo also give option to run data flow whenever dataset is updated; thus making sure that final visualization and report is always based on latest clean data in almost real time.

19.6 Visualization

Domo has many inbuilt visualization template that helps user present their user story in refined visual format. These predefined template are called Cards in Domo. Horizontal bar, Vertical bar, Line, Area, Data Science, Pie and Funnel are few popular visualization categories. These individual categories contain many use full templates for e.g Data Science category has visualization template for scatter plot, box plot, predictive modeling, outliers etc to visually represent relevant data. Donut, Pie, Treemap, Funnel, Folded funnel are few of the popular visualization template under this category.

19.7 Create Domo Visualization Cards

Create data connector as needed (file, cloud, on premise, Api etc) Select the connector and create required data set by selecting table, views or by custom sql query. Select dataset and chose visualization card under respective category (Bar, Pie, Funnel, Scatter, Predictive modeling etc) Drag and drop the fields/attributes that are needed in visualization/report Apply inbuilt aggregate function on fields as needed Save card. Move to dashboard if needed. Give access and share your visualization card with concerned users.

19.8 Dev Studio

Integrated development environment that provides developers with web development experience to create custom apps that can be deployed in Domo instance easily. Development environment consists of following three main components

Domo App CLI is Used to create, edit and publish custom app on Domo environment

App Design is Custom built template that can be connected to different datasets and visualize data (This can be used when there is need of custom visualization requirement for which standard template is not available)

App Manifest is Configuration file that defines properties of custom app

19.9 Installation

First Install node.js from https://nodejs.org/en/ website. Make sure that itis installed by executing node –version command. Next step is to install CLI using below command on unix/linux platform [211].

```
npm install -g ryuu command
```

Make sure that firewall is not blocking npm registry by pining www.npmjs.com through terminal [211].

19.10 Creating simple Domo App

Command domo init on CLI terminal create basic design template.

Enter design name and starter type App. Enter myfirstdomoapp as design name and HelloWorld as starter type. This will create directory and all the necessary files that are needed for building simple app. Directory structure shown below is created [211].

```
Following project structure is created -
    app.cs
    app.js
    domo.js
    index.html
    manifest.json
```

Choose the data source to which app needs to connect to. This is optional and based on app functionality. simple custom app that does not connect to data source can be deployed on Domo instance.

From the CLI run domo dev command. This will open browser and will render myfirstdomoapp

Make styling changes in app.css and code logic changes in app.js to build UI and app functionality respectively.

19.11 Building Responsive App in Domo

Domo lets developer build app that can be displayed on device of any size, without losing quality. Domo supports principles of responsive design. Developers can build their app using this design that renders perfectly on desktop, tablet and mobile devices. Domo use container to render content of app. These container supports any of below four sizes

- Full: This is customizable and can be defined in developers app manifest file
- Large: This is defined for displaying rendering content on devices with 460x540px size
- Medium:This is defined for displaying rendering content on devices

with 225x250px size

- Small: This is defined for displaying rendering content on devices with 225x105px size

By default Domo development environment renders app content inside iframe. Developers have to open source code of iframe into new tab which can then be changed in order to test responsiveness of app content. Developers have to first create directory structure using manifest only option and then later on add responsive stylesheet, either of third party vendor or custom created css. Developer have to add javascript in order to dynamically create layout for placing app content. For e.g Row of tiles and grid can be created using java script file. Developers have to chose and decide number of tiles depending on the size of the screen content will be rendered when creating custom responsive css. Ideally for larger screen 8 tiles are used to fill in a row. For Normal desktop 6 and for tablet 4 tiles in a row are ideal [212].

19.12 API Authentication

Security of data that is transmitted over wire is of highest importance to any organization. Any public API is expected to validate and authenticate only those clients that have access. Domo API uses OAuth2.0 for authenticating and authorizing clients. Security is managed through access tokens. Only authenticated and authorized users can get tokens. For accessing Domo API through OAuth security client program must obtain ClientId and client Secret. Once authenticated; users can access API functionality through access token. Login to Domo instance and click on create new client link under user avatar icon to create client. Specify application name and description. Choose one or more from Audit, Data, Dashboard and User application scope as applicable. One has to be careful while choosing application scope; if application scope is only for accessing data one should only select Data scope else developers will get access to user, audit related information as well. Once Client Id and Client Secret is obtained, next step to obtain access token. Following request can be made to obtain access token using Id and secret [213].

```
$curl -v -u {CLIENT_ID}:{CLIENT_SECRET}
https://api.domo.com/oauth/token?
    grant_type=client_credentials&scope={SCOPE}
```

Above command provides result in JSON format. Body of JSON response contains multiple key value pairs. The most important among those are access token and expires in key. These obtained access token must be passed in header of any future request.

For e.g Below command is used for calling Domo API that gives the list of created datasets after replacing obtained access token [213].

```
$curl -v -H Authorization:'bearer {access-token}
https://api.domo.com/v1/datasets
```

Custom app using Domo API can be built as explained above.

19.13 Data API

Base url (end point) of the data API can be accessed through following command [214].

```
GET /data/v1/:alias?:queryOperators
```

Alias is the name of the dataset that is defined in manifest file. Custom query can be run using queryOperators. Aggregate functions such as count, sum, min, max, avg, filter, group by and order by etc can be used with custom query.

Format of returned data of API can be set using request accept header of XMLHttpRequest object to following formats [214].

```
array-of-objects
  csv
  excel
  json
```

Return format can also be specified in domo.get method.

19.14 Multi User API

Domo offer following end point for accessing information to all Domo instance users [215]

```
GET /domo/users/v1?includeDetails={true|false}
    &limit={int}&offset={int}
```

User details returned by API can be controlled by calling API. Such as , number of records. Developers also has control over retreiving custom user list by passing offset to API [215].

19.15 Single User API

Domo offer following end point for accessing information of single user [215].

```
GET /domo/users/v1/:userId?
    includeDetails={true|false}
```

user id of user whose details are required can be passed.It is very easy to get details of current logged in user by accessing through environment variable [215].

19.16 Sharing custom app using Domo

Custom app, visualization card, report can be easily shared with users by logging in to Domo CLI through domo login command and then publishing the custom built app on domo instance using domo publish command.

19.17 Conclusion

Domo is used as cloud based tool for real time data visualization and reporting. Through Dev Studio and public API, Domo lets the developer extends the capability of customizing visualization and build reporting template that may be used for building custom app. Domo business cloud platform offers high availability, performance and scalability for the applications that are deployed on Domo instance.

19.18 Acknowledgement

The author would like to thank Dr Gregor Von Laszewski for his suggestions and guidance on the content that is presented in paper.

20 IBM Watson Construction and its Services

Pavan Kumar Madineni
pmadinen@iu.edu
Indiana University
hid: fa18-523-82
github: ☁

20.1 Introduction

IBM Watson [216] is basically an artificial intelligence that is bringing rapid changes to the way the world works while simultaneously making businesses faster, smarter and more secure. This AI system is helping businesses utilize artificial intelligence to work at scale providing unparalleled business advantage. As the world continues to become more social, the data is ought to grow, and competitive advantage is to those who utilize the data better than their peers and can directly connect it to their business outcomes and other useful pursuits. Watson enables businesses to personalize customer experiences by streamlining processes, minimizing risks associated and kindling innovation. Watson is helping millions of engineers to seamlessly process huge volumes of data across different disciplines of a business thereby aiding to predict the decline in business, point of break down and proactively fixing them. Therefore, to play a competitive role in the field of business, there is an immense need for tools and processes, which help to collect all the required data generated across numerous platforms easily, to store, manage, manipulate, aggregate, analyze and integrate the data which help in making insightful business decisions. Watson has its services spread across varied disciplines from healthcare to automotive to telecom to education. It helps banks to deploy artificial models that act as virtual agents trained on thousands of customer inquiries helping them to provide expert service to large number of customers at one and a half times faster pace. Watson not only ensures faster pace of working but also transforms workflows.

Keywords: hid fa18-523-82, Artifical Intelligence, Machine Learning, Watson, Transfer Learning, Data Visualization

20.1.1 Watson’s Methodology of Working

Watson centralizes the data which enables the teams and business to connect to data whether the data is on the storage tapes within the organization or in the cloud or on any online file storage without any disruption and hassles [217]. With each day passing, the businesses are relying more heavily on artificial intelligence. So, the businesses demand an intelligence system that does not compromise on transparency of operations to confirm if the recommendations given are trustworthy. Watson is committed to provide the expected transparency through not making any biased recommendations while simultaneously ensuring the variance in the recommendations is also within the permissible limits. Watson as an intelligence system follows three main thumb rules in order to increase the ease of doing business; they are reducing disruptions, enriching customer interactions and making confident recommendations. All these features make Watson the most sought-after intelligence system owing to the ability of Watson to deliver smarter and more productive work [218].

20.1.2 Role of Artificial Intelligence in Building Watson

Artificial Intelligence is the process that gives a machine the power to learn adapt to new inputs after thoroughly analyzing the already known inputs along with their solutions to make better informed decisions. Machine learning is a branch of artificial intelligence that uses computer algorithms to analyze the trends in the data and make smart and intelligent decisions based what these algorithms observe in the data. This is how numerous machine learning models are built on numerous platforms across numerous disciplines [219]. A machine learning model is generally of the form of an input output device that takes in a few observed samples of inputs and provides an output after implementing pertinent computer algorithm on these inputs. A sample machine learning model can be used to predict flu outbreaks based on the volume of tweets mentioning flu-related keywords, recognizing the patterns in human mobility by analyzing the mobile phone call records, or a future forecasting model that can be used to forecast the financial success of a movie by studying the page views statistics of the Wikipedia articles about a movie. All these models or predictive examples in common illustrate the concept of quantifying and measuring human activity at a collective level to understand and build better human societies through a computational framework. Watson broadly falls under one such computational framework that works on real data to make informed decisions and build smarter and more secure human societies [220].

20.1.3 Role of Machine Learning in Building Watson

Among the existing machine learning techniques, deep learning is probably the most sophisticated machine learning technique. Watson utilizes deep learning framework to create an artificial neural network that can perpetually learn from various inputs determining whether decisions made are correct while constantly improving the quality and accuracy of results. This is what enables Watson to learn even from unstructured data that is available from the society such as photos, videos and audio files. Deep learning also enables Watson's natural language understanding capabilities thus allowing it to learn by deconstructing sentences and then analyzing and identifying the concepts and underlying relations of those sentences. Once these relations are discovered, it can decipher the context and intent of what the original sentences would want to convey [221]. Any artificial intelligence in general also needs to understand the specific language and terminology in order to decipher the jargon pertinent to that particular domain and industry but this process is absolutely cumbersome using the traditional artificial intelligence models and also requires huge volume of data and computing power to run these algorithms unintermittingly for long durations. Watson however simplifies this process drastically by applying a technique called transfer learning.

20.1.4 Role of Transfer Learning in Building Watson

Transfer learning is a machine learning technique where a model trained on one task is re-trained and on another related task. This process ensures improvement of learning in a new task through the transfer of knowledge from a related task that has already been trained and learned. Transfer learning also reduces the overall training time by alleviating the need to train the algorithm from scratch which can be achieved by feeding Watson with knowledge from an already trained model. Watson's transfer learning architecture comprises a three-layered artificial intelligence model. The bottom layer constitutes an out-of-the-box general knowledge like Wikipedia for artificial intelligence. This layer for provides the model the basic knowledge about the domain the model is trying to get trained and learn. The middle layer is prepackaged with knowledge exquisite to specific domains and industries. This layer takes care of jargon specific to domain thereby removing ambiguity associated with terms which have different meanings in different contexts. The top layer is where personalized learning takes place [222]. The model will be fed with all the training data the model is intended to learn. The model is now potentially knowledgeable to understand a company's specific risk and behavioral attributes. Transfer learning is thus an integral part of how Watson is able to accelerate business operations at a rapid pace thereby dramatically reducing the operating costs.

20.2 IBM Watson and its Services

IBM Watson is one such intelligent data analysis and visualization service on the cloud that lets anybody pose different questions and provides answers to the questions posed in natural language [223]. IBM Watson provides all the tools and services necessary to work with your data and build machine learning models at one place which makes analyzing data much simple be it a novice or an expert. IBM Watson offers numerous services namely Watson Analytics, Visual Recognition, Natural language Understanding, Speech to Text, Text to Speech, Tone Analyzer, Language Translator, Machine Learning etc [218].

20.2.1 IBM Watson Analytics

IBM Watson Analytics is one such service that enables extracting intricate data flow processes as well as convoluted relationships between different fields of data. The analytics services offered by IBM Watson not only enables you to discover novel insights about an organization but also swiftly create and share highly informative and illustrative dashboards and infographics. The Watson Analytics is also capable of analyzing data directly from social platforms like Twitter and output the sentiment of the customer or learn the best time to tweet to enhance the span of the audience. It also enables multiple users to collaborate and share visualizations and dashboards with each other.

20.2.1.1 Watson Analytics on Health Care

One of the few domains that extensively needs an artificial intelligence system about which everyone is concerned about is health care. The current approach to tackling different health problems today is flawed and grave concerns are exhibited over improvements in health care industry. Watson has partnered to build solutions that shall allow the larger health care community which involves both individual patients as well as larger health populations to be benefitted as the participants share and apply data-driven insights in real-time [224]. The medical data is growing exponentially each year with data flowing in from numerical sources such as medical and clinical research, various sensors, personal fitness trackers etc. The health industry is unable to keep up with this staggering growth of medical data. The IBM Watson health cloud brings together huge volumes of medical data into one centralized platform on the cloud. This data is then used to apply a fusion of traditional analytical techniques and Watson's advanced machine learning techniques to churn out valuable insights out of it. This is possible because of ecosystem environment in which Watson Health Cloud operates i.e. it has multiple contributors to keep the system functioning smoothly. Watson Health Cloud uses huge volumes of data, knowledge pertaining to data and perspectives and opinions of researchers and domain experts as contributors to its ecosystem. Watson has customized this ecosystem even for personalized use which make the health cloud more dynamic and efficient. Watson is also well-known for the way it handles highly confidential data such as patient related information. It makes sure to remove all personal identifiers associated with the data that is being uploaded to the cloud. This process is called De-identification which is very crucial to create safe and secure cloud environment [225]. Watson's extraordinary machine learning abilities combined with its interactive ecosystem as well as its secure de-identification has made it one of a kind health care analysis tools which is transforming the health care industry substantially.

20.2.2 IBM Watson Machine Learning

IBM Watson Machine Learning is another powerful service which is integrated to IBM Watson Studio that enables users to perform two fundamental operations of machine learning i.e. training and scoring. Training is the process of teaching an algorithm the underlying trends and behavior of data by feeding labelled data to the algorithm. A trained algorithm learns coefficients of mathematical expressions which represent the behavior of the data in the best possible way. Scoring is the process of predicting an output using the coefficients learned from the trained algorithm which is otherwise called as a predictive model. These predictions made by the algorithms after scoring enables data scientists to collaborate with data engineers to further explore the data and gain deeper insights out of it.

20.3 Conclusion

To sum up, IBM Watson is an ideal tool if one wants to perform their own analysis but perplexed where to begin. IBM Watson is a smart data discovery tool that enables you to leverage all the state-of-the-art data analytics and visualization techniques to draw valuable insights out of data almost instantly. It automates the processes such as data preparation, predictive modeling, and data visualization which are otherwise very hectic and tedious processes [226]. Watson is today considered the best artificial intelligence for most of the enterprises. It facilitates faster learning from smaller chunks of data. It is explicitly designed to make sure that data flows only from bottom to the top so that this structure aids in transfer learning process by sharing knowledge in the bottom two layers while still holding a firm control over the top layer.

21 SAS Viya

Ritu Susan Sanjay
rssanjay@iu.edu
Indiana University, Bloomington
hid : fa18-523-66
github: ☁

Keywords: SAS Viya, Data Mining, Cloud Analytic Services, Analytics

21.1 Introduction

SAS Viya is an in-memory analytics engine, with cloud enables features, providing users with accurate, quick and most importantly reliable insights. The software provides features like authorship, full-versioning, change management and a lineage viewer along with centralized admiistration enabling tracking of users, servers and job content [227].

The demand for data scientists is at an all time high. At the simplest level, it is the art of making sense from the never-ending sea of data - simple as that. To better unserstand this, we can think of developing a data mining model analogous to making a dish. You first scrape together the ingredients; this is our raw data. To improve the taste, you add a pinch of salt and other seasonings or spices; this is the creation of new predictor variables. Finally, you mix it all up together, and have a taste. If it doesn't taste right, you might want to try a couple of other methods until you've perfected the recipe; analogy here referring to iteratively building descriptive (unsupervised) or predictive (supervised) models. And just like in the show Master Chef you often end up competing with other data scientists in developing the best possible recipe, or in this case the best model [228].

In the above example, just as how the chef needs his sharp knives, a data scientist needs the right tool. The top rated software being python, spark, R, Matlab to name a few. SAS Viya is the latest enhancement of the SAS platform.

In the words of its developers:

> ***“SAS Viya addresses the complex analytical challenges of today, and effortlessly scales to meet your future needs, with cloud-enabled, > elastic in-memory processing, in a high availability, multi-user environment. It is designed to address the new, and increasingly diverse, needs of organizations with methods, access, and deployment that scale to meet burgeoning analytics use cases” [228].***

SAS 9 brought with it a user-friendly server-client web service model, all processes goverened by a resilient metadata server. It was one of the more popular platform for analytics, albeit a bit expensive, and hence was usually preferred by organizations willing to invest in analytic platforms rather than self-financed analysts. However, the advent of cloud computing blew the whole tech industry out of proportion.

> ***“SAS Viya brings a more resilient, elastic, unified, and accessible architecture, which leverages cloud-friendly microservices and a next generation analytics run-time engine” [229].***

SAS Viya enables users to explore data deeper, using the latest innovations in in-memory analytics. The methods available to users is classified in two: Data Wrangling methods and Modeling methods. Data wrangling methods include binning, transformations, SQL, clustering etc., while modelling techniques include everything from regression to text mining to neural networks. One of the prime features of using SAS for these methods is its innate ability to run all above-mentioned methods in-memory and of course take advantage of the parallel processing infrastructure [230].

SAS Viya emphasises a unified experience for data scientists and analysts alike. The new cloud analytical platform, allows programmers to execute using open-source laguages like Python, Java, and Lua. Furthermore, the platform also allows these codes to be written and executed on Jupyter notebooks [231]. Figure 21 shows a detailed view of the SAS Viya 3.4 architecture.

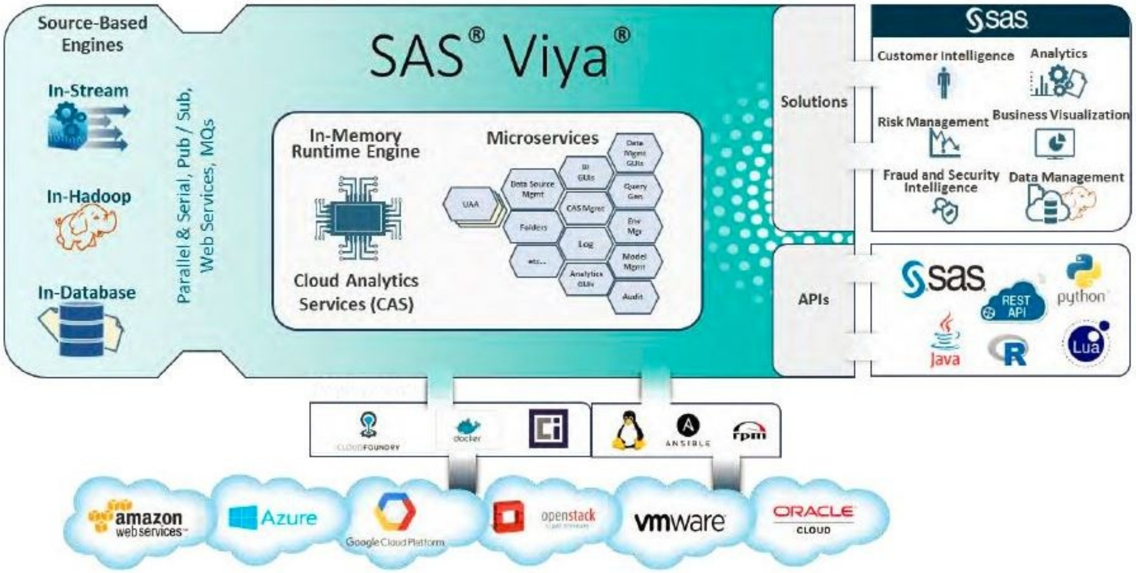


Figure 21: SAS Platform [232]

21.2 SAS Viya Components

SAS Viya 3.4 gives data scientists access to the following components:

21.2.1 SAS Cloud Analytic Services

> ***"CAS goals are to provide an analytics service with a public API accessible by many clients supported by SAS or open-source clients using plug-in modules from SAS" [233].***

SAS CAS is a platform for distributed and high-performance computing with a cloud-based RE. The many features of CAS include data sharing between sessions, security, and fault tolerance (i.e. allowing a node to fail without data loss). CAS was designed to operate fully on-cloud, either as a single host or on a cluster (private or public). CAS uses sessions to track users and offers a full Security interface to protect data at the file level, as well as the column level. The sessions provide isolation for the user, which protects the integrity of the server.The purpose of connecting to CAS is to execute server requests. A user must create a session to submit a request. The user can connect to the server either through a HTTP_based REST interface or through a ProtoBUF-based binary interface. The user must be authenticated by CAS in order to a create session [234].

The most import aspect of CAS is that all data is stored in the form of tables. These tables may be streamed from a database into the server , ESP stream or loaded from disk. Data (including metadata) in CAS is all stored and accessed through the CASLIB. The caslib is basically a container that usually has one or more instances of the CAS tables [235].

Like all cloud services, SAS Viya too concentrates on fault tolerance. Node failure is inevitable when dealing with multiple number of nodes are implemented in a system. Data is replicated across the cluster, in order to retrieve data in case one of the worker nodes fail. The new system has been dubbed the GCCOMM. The subsystem can detect failure in nodes; the controllers and workers can reconfigure the system, thus restarting the action and allowing the remaining worker nodes to access lost data from the redundant blocks [235].

21.2.2 SAS Studio

In the simplest of words, SAS Studio is an editor designed for both expert and novice programmers to write and execute SAS code in an assisted environment. Users are provided with a single interface to access all data files, programs and libraries (user-defined and in-built). The SAS studio is extremely convinient: there is no local installation involved i.e. once the software is installed you just provide the users with the url to access the software. This centralizes and simplifies regular maintenance. Other features range from (data) table analyzer, sql engine, code snippet library prompts from frequently executed codes, report generation and export in multiple formats including pdf and xml. One important feature to be noted is that the studio interface is consistant, regardless of where and how the software runs; as the IT infrasturcture is modified, the SAS Studio environment is the same [236].

21.2.3 SAS Visual Analytics

> ***“SAS Visual Analytics provides a complete platform for analytics visualization and interactive selfservice BI and reporting capabilities are combined with out-of-the-box advanced analytics to enable users to discover insights from any size and type of data > including text” [237].***

Popular analytic tools include - goal seeking and scenario analysis, automated forecasting, decision trees, text analytics and network diagrams [237].

21.2.4 SAS Visual Statistics

> ***"It provides a drag-and-drop web browser interface that empowers multiple users to explore massive data, and then interactively and iteratively create descriptive and predictive models" [238].***

Statisticians are most often faced with the challenge of vhoosing the right model that fits the data. As the amont of data to be analyzed keeps increasing, this can prove to be a very time-consuming task. SAS Visual Statistics allows users to generate reports that detail model comparision summaries, misclassification tables and ROC charts. The software includes an interactive slider that manipulates thresholds preset by the user [238].

21.2.5 SAS Visual Data Mining and Machine Learning

> ***"In addition to innovative machine learning and deep learning techniques for analyzing structured and unstructured data, it integrates > all other tasks in your analytical processes. From data preparation and exploration to model development and deployment, multiple personas work in the same, integrated environment" [239].***

The software allows multiple users to concurrently analyze data (structured or unstructured) with the Model Studio. Every project can be subdidvided using visual pipelines, displayed in a logical sequence. The SAS Visual Data Mining and Machine Learning boasts of a significantly reduced runtime, owing to the multicore architecture. Popular models included in the software are: gradient boosting, SVMs, neural networks, Gausian models, bayesian networks among others [239].

21.2.6 SAS Econometrics

> ***"SAS Econometrics supports a range of econometric model***

> ***types with a single framework. It’s fully integrated with all of the contributing analytics that coincide with econometrics, and with data preparation, exploration, presentation and reporting capabilities in SAS that are essential to successful econometric analysis” [240].***

The in-memory analytics feature of SAS Viya ensures that iterative and repetitive tasks can be run quickly without re-loading data. The software also provides a wide-range of tools for modelling business scenarios. Simulations and forecasting techniques can be easily implemented as the software makes use of the SAS VIya Engine, ensuring high-availability and the ability to code using open-source languages [240].

21.2.7 SAS Visual Forecasting

> ***“SAS Visual Forecasting provides a resilient, distributed and optimized generic time series analysis scripting environment for cloud computing” [241].***

> ***“Users can range from analysts responsible for the creation of the forecasts to the managers and directors responsible for overseeing > the forecasting and planning processes” [241].***

The software itself recommends the most suitable model and additionally models are selected based not on how well they fit past data but on how well they can be used to predict the future. The interface allows for training and modelling data mining algorithms including neural networks. For example, the Multstage Forecasting node (for regression and time series included) creates a forecast combining signals obtained from different models [241].

21.2.8 SAS Visual Text Analytics

SAS Visual Text Analytics seeks to bring together concepts of NLP [242] along with machine learning and data mining techniques to derive insights from unstructured data. The accuracy of an analytical model may be increased by utilizing a combination of machine learning techniques and rule-based approaches. Users also have the ability to build their own custom search engine, provided by microservice architecture and built in APIs. The text-analytics pipeline makes available five types of nodes : Text Parsing, Concepts, Catagories, Sentiment and Topics. The software also features automatic extraction of features identified by topics generated by the machine [243].

21.2.9 SAS Optimization

SAS Optimization was designed for industry experts who utilize operations research and optimization techniques to create decision-making models to solve problems.

> ***"SAS Optimization provides a powerful, intuitive algebraic optimization modeling language and an array of algorithms. This involves a range of models, including linear, mixed-integer linear, nonlinear, quadratic, and network optimization, as well as solve constraint satisfaction problems" [244].***

Models are executed efficiently since the software runs on the Viya Engine. Notable models on the SAS Optimization are : Local Search Optimization, Constraint Programming, Multistart Algorithm and the Decomposition Algorithm.

21.3 Deployment

> ***"SAS Viya has undergone rigorous performance testing with various hardware combinations. In addition to being tested on high- performing Intel Xeon E3-E7 series microprocessors, SAS Viya has also been tested with newer Intel chips, such as Intel Xeon Scalable > Processors. SAS Viya also supports 64-bit AMD chipsets (thirty-two-bit chipsets are not supported)" [245].***

It is necessary to note that a seperete independent host is needed if SAS 9.4 exists on the system (co-installation is not possible). Also, if the existing SAS software on the system is SAS 9.3, note that many of the features on SAS 9.3 are not supported if the Java version has been updated to Java 8 or plus. The hardware requirements for a programming only environment also differs from a full deployment.

21.3.1 System Requirements

It is first necessary to understand the difference beyween the two deployment types: full deployment and programming-only deployment. The full deployment includes all the features SAS Viya has to offer and is usually the default mode. However, it is also possible to deploy only a subset of the features; the programming-only deployment excludes the SAS Drive and a number of the graphical features [246]. The hardware requirements for a programming only and a full deployment differ. When determining the specifications of the host, three components are to be kept in mind: CAS server,programming runtime and the service layer [245].

1. CAS Server : Before installation a key point to be taken to account is the required amount of RAM. This may vary depending on activity level of users in the SAS software environment and the data (load) to be processed. However, less that 1 gegabyte of RAM is mandatory for CAS SErver startup [245].

2. Programming Runtime : The CAS license procured determines the number of CPU cores required for your environment. On that note, if the CAS license specifies N cores, then the user is entitled to the same number of cores on their setup. SAS however, specifies a minimum of two cores and at least four cores for optimal performance. The software also necessiates a minimum of 4 gegabyte RAM for the programming environment [245].

3. Service Layer : The service layer primarily are the components required for a full deployment ideally. These also include services that support other SAS softwares. Aditionally, it includes supporting services for the SAS VIYa analytics software, namely the Core Services host [245].

21.3.2 Installation

The first step in the installation involves setting up the accounts. The user account for both the CAS as well as the postgreSQL requires the SAS credentials to be specified. Instructions on how to set this up is extensively detailed in the setup manual (the link to the latest version of the same is provided at the end of this section). If the user seeks to set up the full deployment, then changes may be made to modify the postgreSQL settings to specify personal ports and directories. Further to this, the CAS Server Monitor port may also be changed along with modifications to the kerberos [247] settings. The final step after the configuration files have been modified is to simply run the SAS Viya setup batch file through the command prompt folling the regular intructions [248].

After completing the installation of SAS Viya,it is necessary to configure the connection to your identity provider before your users can access SAS Environment Manager and SAS Visual Analytics. The final step then remains to create a backup configuration [249]. Note that this is also different depending on the type of deployment (i.e. programming-only or full deployment).

The latest version of the deployment guide may be accessed at : SAS Viya Deployment Guide ☁

21.4 Sample Illustration

This section consists of an example detailing how easy it is to create a model in SAS Viya. For the example, we consider a dataset consisting of the political opinion poll from the Annual National Election Survey, consucted once every four years. Thermometer measures are used in many surveys to rank opinions. These variables range from 0 (very cold, or unfavorable feeling) to 100 (very warm, favorable feeling). This example examines the multivariate relationship of the preference for the Democratic variable against current economy condition, religious attendence and how better off the respondant is compared to the previous year.

Step 1: Start up the SAS Viya Service

SAS Viya starts up with a user friendly interface. The left panel details all about the data and the models that may be created and executed using the engine. While the right panel houses options for modifying or altering data variables and adjusting model parameters. Figure 22 shows a screenshot of the start menu.

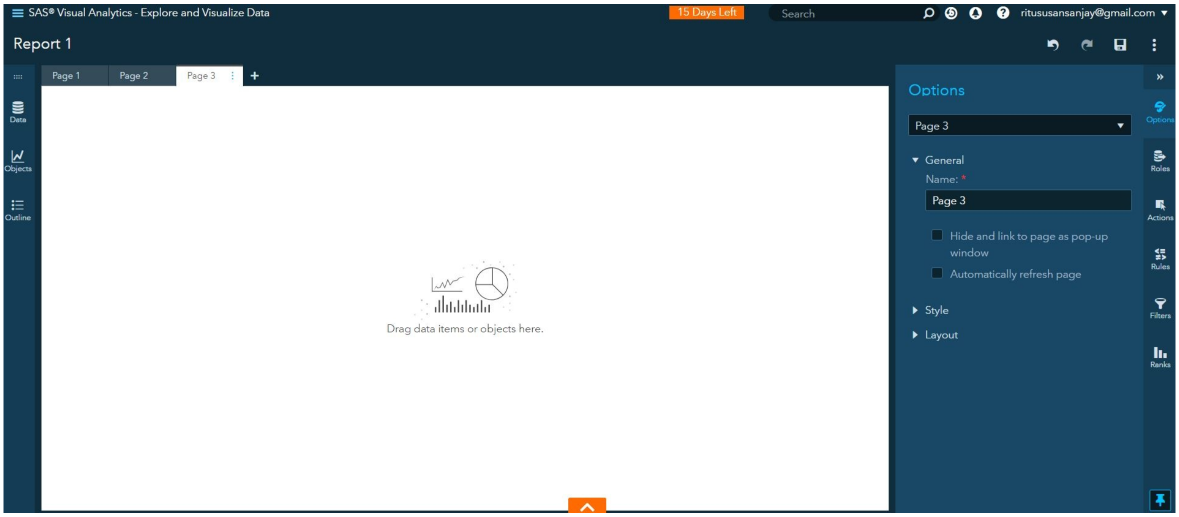
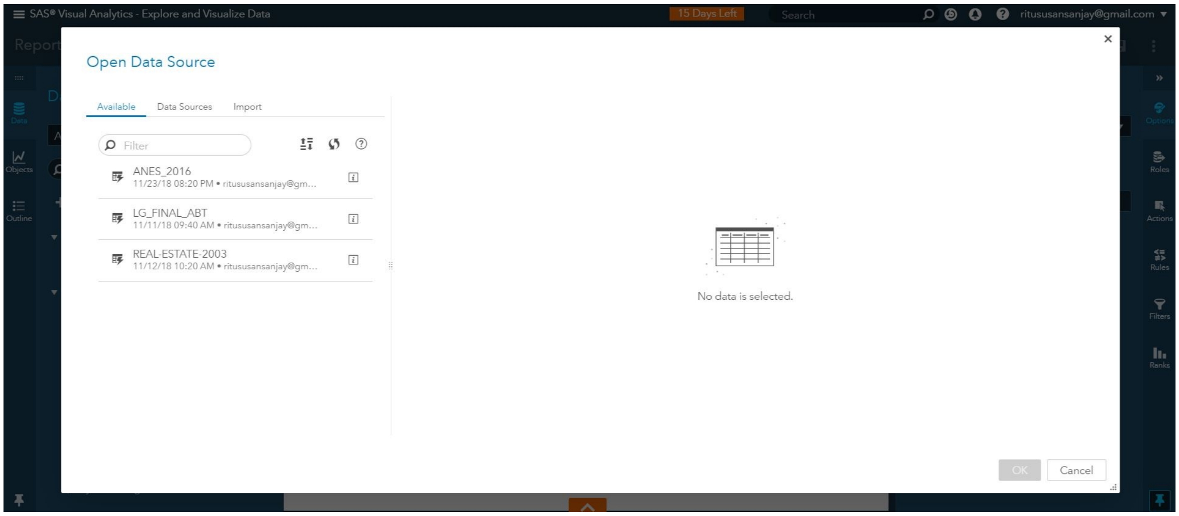

SAS® Visual Analytics - Explore and Visualize Data
15 Days Left
Search
ritususansanjay@gmail.com
Report 1
Page 1
Page 2
Page 3
Data
Objects
Outline
Drag data items or objects here.
Options
Page 3
General
Name:
Page 3
Hide and link to page as pop-up window
Automatically refresh page
Style
Layout
Options
Roles
Actions
Rules
Filters
Ranks

Figure 22: SAS Viya Start Page [250]

Step 2: Import data

Figure 23 shows how datasets may be imported onto the SAS library.



ANES_2016
11/23/18 08:20 PM • ritususansanjay@gm...

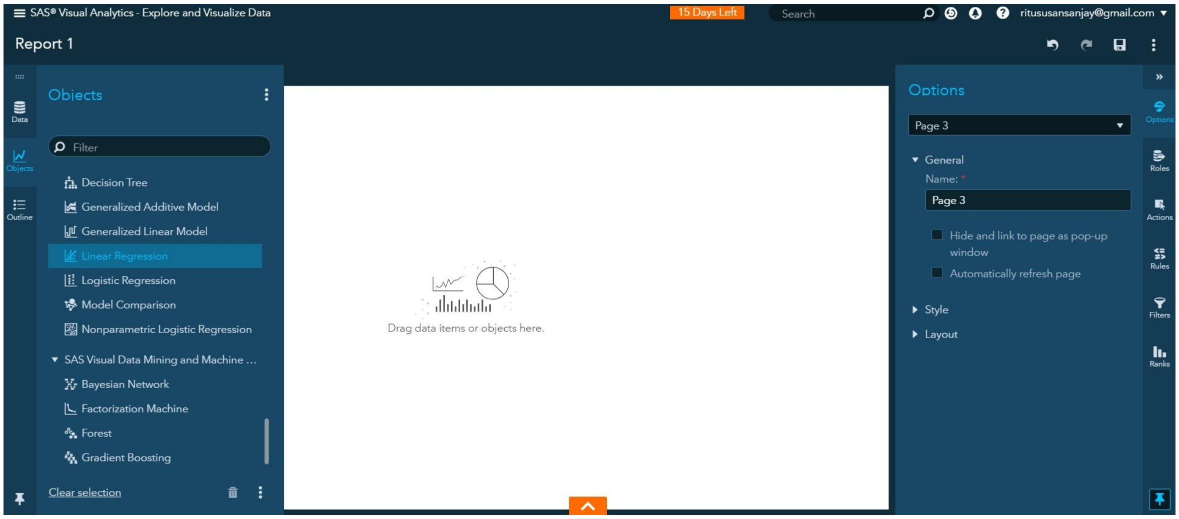
LG_FINAL_ABT
11/11/18 09:40 AM • ritususansanjay@gm...

REAL-ESTATE-2003
11/12/18 10:20 AM • ritususansanjay@gm...

Figure 23: SAS Viya Import Data [251]

Step 3: Add model object i.e. linear regression object

To create a linear regression model, all you have to so is drag and drop the desired object into the analysis screen. Figure 24 demonstrates how a model may be added.



Name: * Page 3

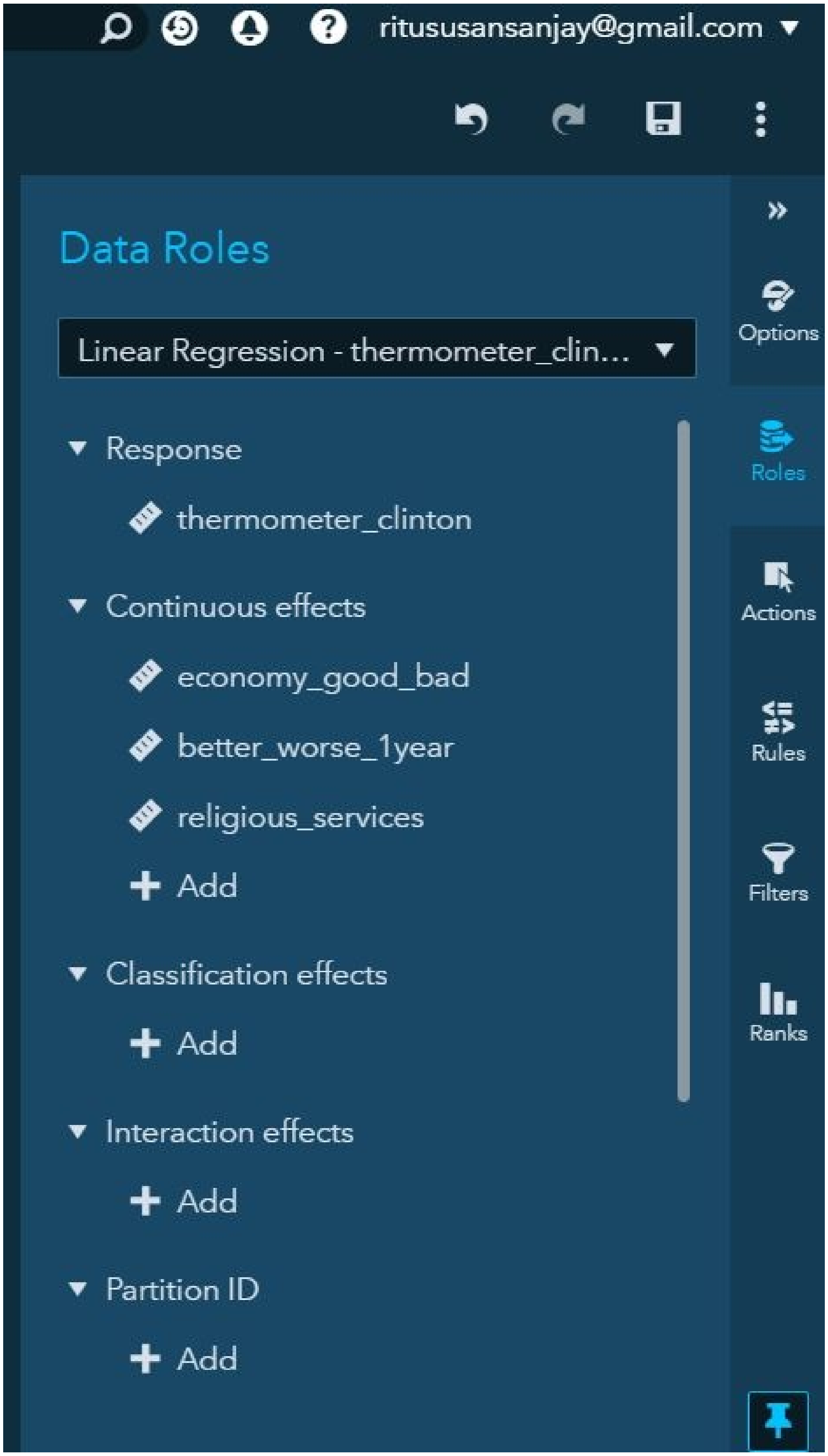
□ Hide and link to page as pop-up window

□ Automatically refresh page

Figure 24: SAS Viya Add Data Object [252]

Step 4: Specify roles on the right options panel

Figure 25 demnstrates how roles and rules may be applied to the model.



Data Roles

Linear Regression - thermometer_clin...

Response
- thermometer_clinton

Continuous effects
- economy_good_bad
- better_worse_1year
- religious_services
- + Add

Classification effects
- + Add

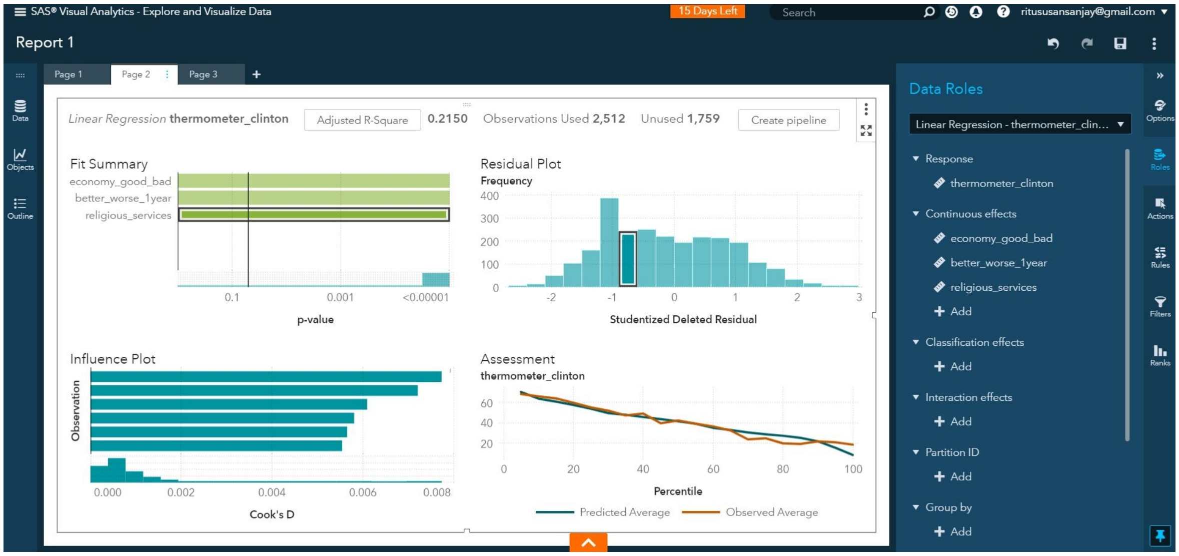
Interaction effects
- + Add

Partition ID
- + Add

Figure 25: SAS Viya Add Variable Roles [253]

Step 5: Modify parameters if necessary to improve model

The model can then be interpreted using measures like the adjusted r-square, that predicts approximately 21.5% of the variation in the variable. Looking at other measures like the F-statistic (very high) and p-value (very low) implies that the model is statistically significant. Figure 26 displays the final result.



Linear Regression thermometer_clinton

Adjusted R-Square 0.2150 Observations Used 2,512 Unused 1,759

Figure 26: Linear Regression Results [254]

21.5 Conclusion

Big data speaks volumes when applied to find solutions to challenging ‘everyday’ problems. We capture and store far more data than is actually used; the true potentials of big data are just being realized [255]. Today data is the raw material generated and consumed by businesses, governments and scientific researchers. Given the right tools and the computing power, data can open up a whole new world of insights. The new cloud-based analytic software offered by SAS helps create well-defined models and generate results, giving way to new ideas. However, SAS Viya is just one of the many tools among thousands offered today and choosing the right tool depends on the users’ goals.

22 Utilizing Python Matplotlib Package for Data Visualization of In Cancer Clinical Trials

Evan Beall
ebbeall@iu.edu
Indiana University
hid: fa18-523-80
github: ☁

Keywords: Python, Visualization, Matplotlib, Cancer, Oncology

22.1 Introduction

Cancer clinical trial research is an ever-evolving field. The pharmaceutical companies that carry out these trials employee individuals of various backgrounds to carry these trials out successfully. The individuals required to run these trials have backgrounds that range from scientific, business, operational, etc. In order for these various teams to work together efficiently, the business units involved in this pursuit will need to be able to communicate effectively. Individuals from these varied business units will have different specializations and prior knowledge. To accommodate the varied backgrounds between groups communication can be aided by visualizations. Visualizations provide a method that allows all business units involved regardless of background to have a general understanding of research progress.

In the last several years, cancer research data has been ported into Electronic Data Capture (EDC) systems [256]. These EDC systems function only as databases to have data entry performed and hold data. These systems have no native way of creating visualizations. Due to the lack of native visualization capability, data needs to be extracted and run through software that can quickly and effectively create visualizations of large data sets. One such tool is the Matplotlib package that is available via Python programming.

Cancer research is beginning to utilize big data technology like Matplotlib to help to analyze, standardize, and communicate results. Matplotlib is a open source library within the Python environment. This environment provides simple but extremely powerful 2D and 3D visualizations of large amounts of data [257]. Matplotlib is currently not widely used in clinical trial research but could become an extremely powerful tool within the clinical trial ecosystem. This tool would be especially helpful for scientists and data managers to be able to display their trial progress to those that do not come from a scientific background. The types of visualizations available within this library range from basic scatterplots that require little outside knowledge to interpret; all the way up to EEG plotting capabilities that require specific medical knowledge to interpret. An example of the EEG plotting capabilities can be seen in Figure 27.

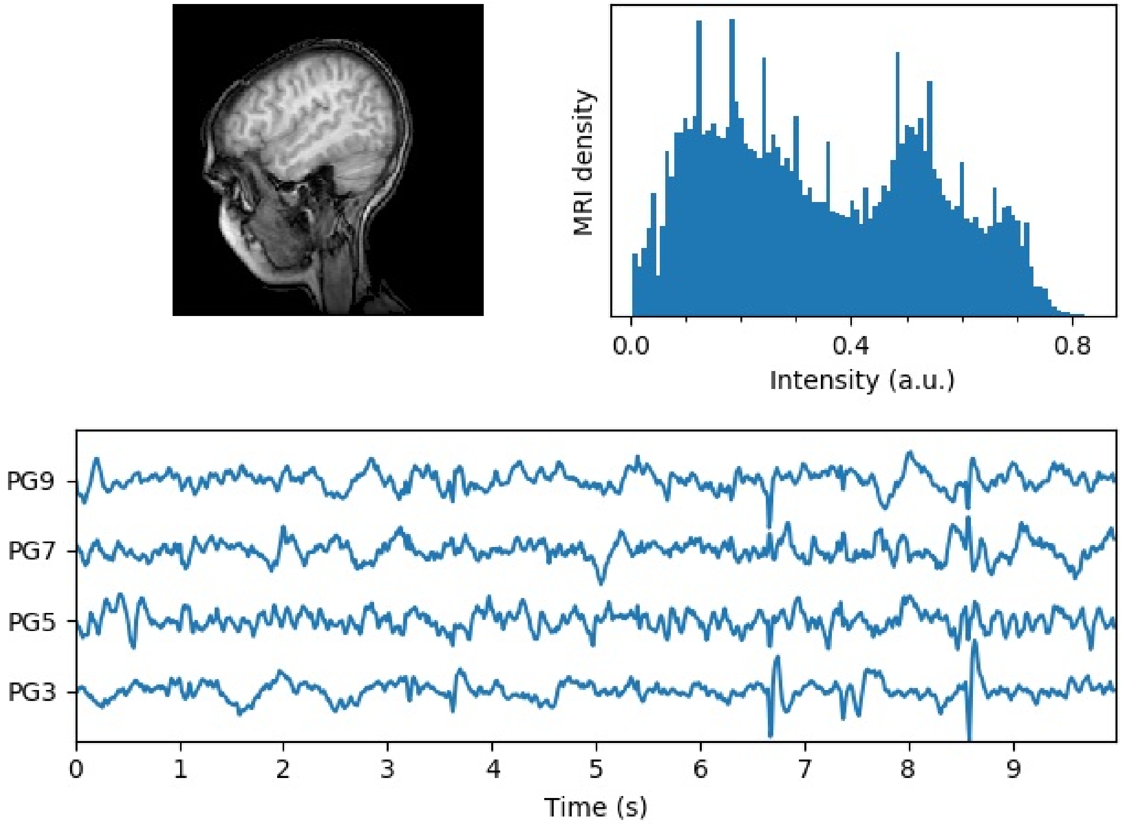


Figure 27: Matplot EEG [258]

22.2 Architecture of Electronic Data Capture Systems

Database structure of an electronic data capture system can vary widely across industry pharmaceutical companies [256]. The data present in each of these systems will vary based on several factors. These factors include: phase of trial, therapy being investigated, type of lesion being research, etc. While a trial is actively accruing and treating patients, the data within the EDC system is unstructured and it is nearly impossible to create comparisons between other trials [256]. The goal of this research is to prove that these new drugs provide benefit to the general public and allow patients to glean those benefits.

In the United States the FDA is an organization in place to guard the general public from harmful and useless new treatments being brought to market. Every new treatment resulting from a clinical trial needs to submit its data to the FDA to gain approval to market the drug. To accomplish this, the clinical trial research community has created a structure that standardizes data once it has been extracted from the electronic data capture systems used in a clinical trial. This standardization structure is called Study Data Tabulation Model or SDTM [259]. SDTM provides a standard structure for both human clinical and nonclinical studies to be submitted to the FDA for approval. In 2004, SDTM was chosen by the FDA as the standard that would be utilized for all submissions for drug approval [259].

Clinical trials can vary widely regarding what observations are collected throughout the life of the trial. SDTM provides a set of defined variables that each of these observations will need to fit into. Each of these observations are broken down by topic, timing, qualifiers, and identifiers depending on the type of observation that is being assessed [259]. Each observation is then sorted into a domain. Domains are groups of variables or observations that are related by a topic-specific commonality or scientific commonality. In general, each domain correlates to a corresponding dataset, however, some domains can be spread across multiple datasets. Examples of datasets that are used in Oncology clinical trials are: DM (Demographics), AE (Adverse Events), etc [259]. Utilizing SDTM coded databases allows for data coming out of electronic data capture systems to be compared. In turn, this allows for the FDA to compare across clinical trials to assess the scientific backing of each submission to the FDA. The FDA is then able to better analyze the efficacy of the research and if the general public would benefit from having this product available in the American healthcare marketplace. This standardization also allows for comparisons to be drawn between research done all over the world.

22.3 Matplotlib Use Case for Clinical Trials Visualizations

Oncology clinical trial research and clinical trial research in general generates an enormous amount of data. Clinical trials can include thousands of patients with health data for multiple years at a time. Managing this massive amount of data is not a small task. Thousands of individuals are involved in the procurement, entry, cleaning, and manipulation of these databases before a clinical trial can be called a success or failure. The cost to bring a drug to market is currently estimated to be about 648 million dollars. Along with this, it is estimated that the cost to run one clinical trial depending on its phase could be anywhere between 10 million to 40 million dollars [260].

When this many people from differing backgrounds are working together to carry out each of these trials, it is necessary to have a concise and universally understood way of communicating trial progress and goals. This is where matplotlib's visualizations tools can be an extremely powerful asset. The visualizations created via Python's matplotlib library would provide researchers a way to communicate with other business units in ways that do not require oncology knowledge. Utilizing visualization tools allows for workers with different backgrounds to communicate in a universal manner. Some examples of visualizations that may be helpful during trials would be identifying how many patients are responding to treatment positively, showing a graph of how long patients are remaining on the investigation therapy, or showing a graph of the characteristics of tumors that are being studied. An visualization that plots survival status of patients on treatment can be seen in Figure 28 below:

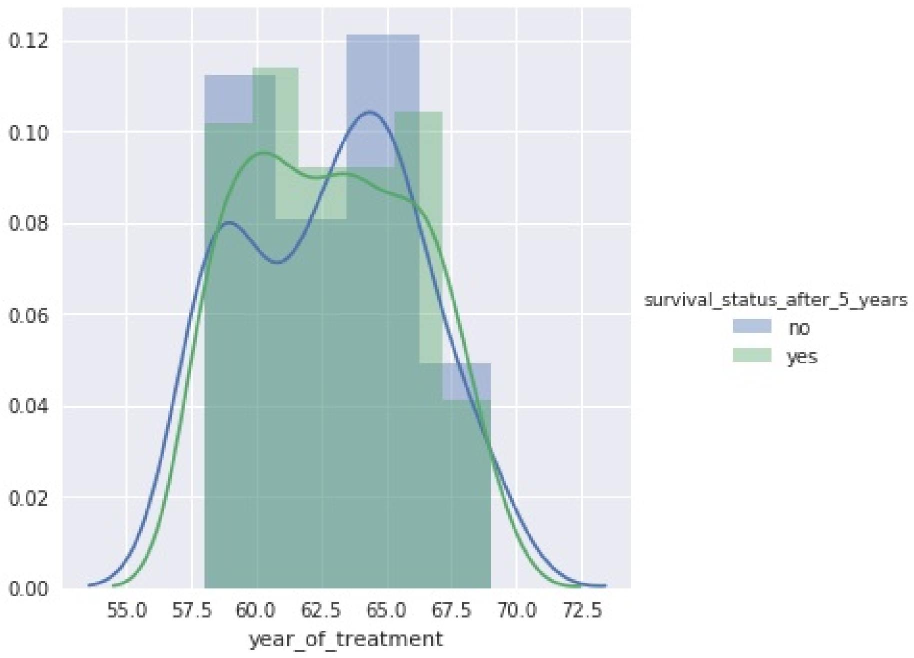


Figure 28: Survival Status [261]

Communicating outcomes such as those seen in Figure 28 early on in the study can inform other business' units decisions. The survival status layout shown in the image above can be interpreted by the medical team to determine patients benefit gained from the study. Medical staff might be able to determine trial futility earlier on in the study utilizing visualizations such as this. These individuals are then able to pass this information on to other internal teams while utilizing the visualization to communicate their analysis. Following this, the budgeting departments might be able to make decisions on where to increase or decrease funding. Increasing the availability of visualizations to everyone involved in a clinical trial would allow for better communication and potentially save millions of dollars throughout the life of a study. These visualizations would provide a method for each business unit to have robust oversight throughout the trial.

Most of the current electronic data capture systems are built in SQL based relational databases. One such SQL based system is an Oracle based database called Inform [256]. Databases like Inform do not provide great tracking or visualization tools for people of all backgrounds to be able to understand how research is progressing. Most individuals involved in clinical trials research are not well versed in data manipulation, statistical analysis, or querying databases.

This is the most critical reason why accessibility of visualization tools would be a large boon to large pharmaceutical companies allowing all business units to be able to communicate with each other during the clinical trial process. Visualization tools will allow for earlier analysis of clinical trial research to occur by all business units such as budgetary, medical, and legal. The cancer research industry uses statistical teams to perform deep analysis, however, utilizing this staff is expensive and time consuming. Each time statistical analysis is run on the entire trial, it requires fully committing statistical colleagues to run analysis on every aspect of the trial. By instead utilizing visualization tools, each business unit would be able to quickly run analysis on specific aspects of the trial. Visualization tools would provide a quick glance at specific areas of interest allowing all business units to quickly determine if the study/research is going as planned. If these units have further questions regarding trial progress, then a full analysis can be performed on the study. This could save time and money for the company as a whole. Also, having a robust oversight plan would allow for mistakes and problems to be caught even earlier.

22.4 Matplotlib Architecture

Matplotlibs architecture is split into three different layers. The layers involved in producing a plot with the matplotlib library are the backend layer, artist layer, and scripting layer [262]. These layers are developed in a stack orientation in which the layers can talk to the layers below them, but the lower layers are not aware of those layers above them [262]. The backend layer involves FigureCanvas, Renderer, and Event classes. The FigureCanvas class involves the area where the plot will be drawn. The renderer class actually does the drawing of the plot. Finally, the Event class handles keyboard or mouse events that might occur during the drawing process. The Artist Layer is the middle layer of the Matplotlib stack. Within this layer is the Artist class. This class allows the user to create and customize the plot. Allowing for the system to understand the drawable area on the canvas and what type of titles and types of plots should be drawn. An example of how the Artist layer functions is displayed in both Figure 29 and Figure 30

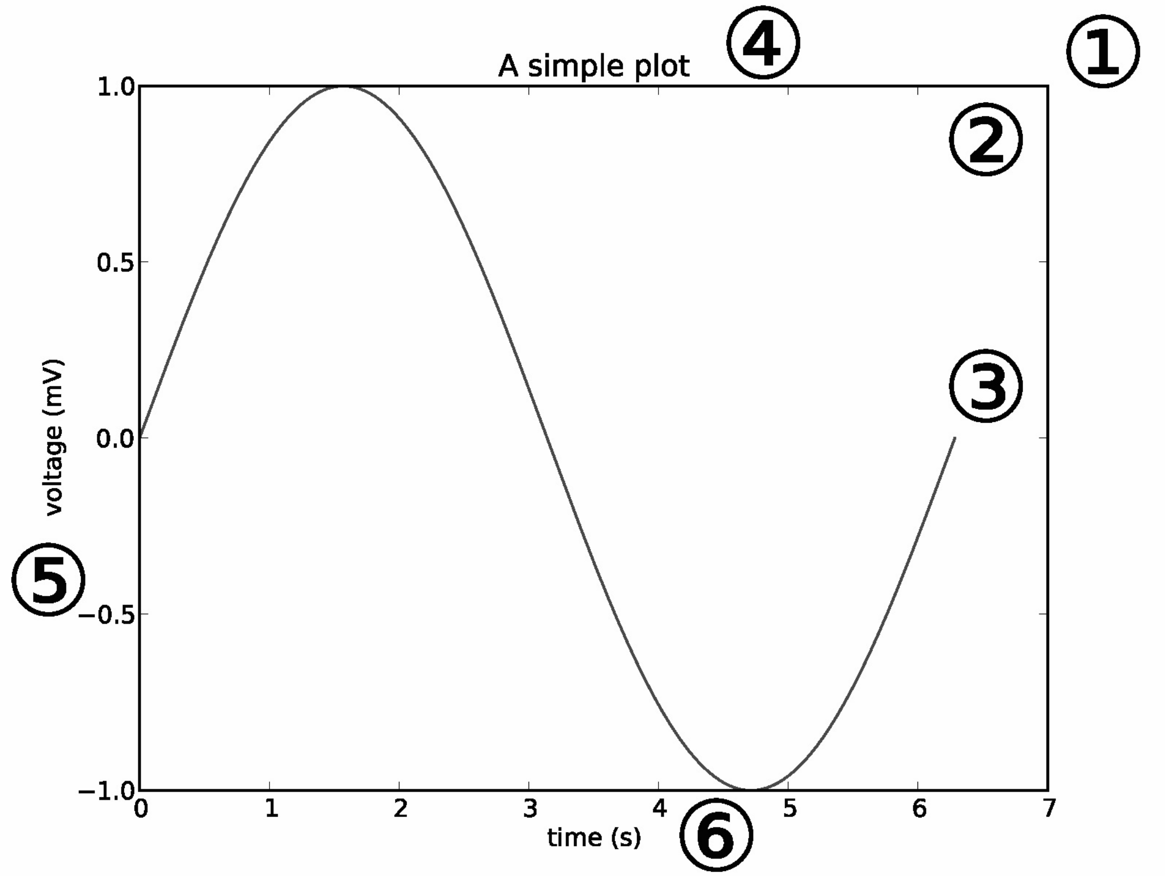


Figure 29: Artist Plot [262]

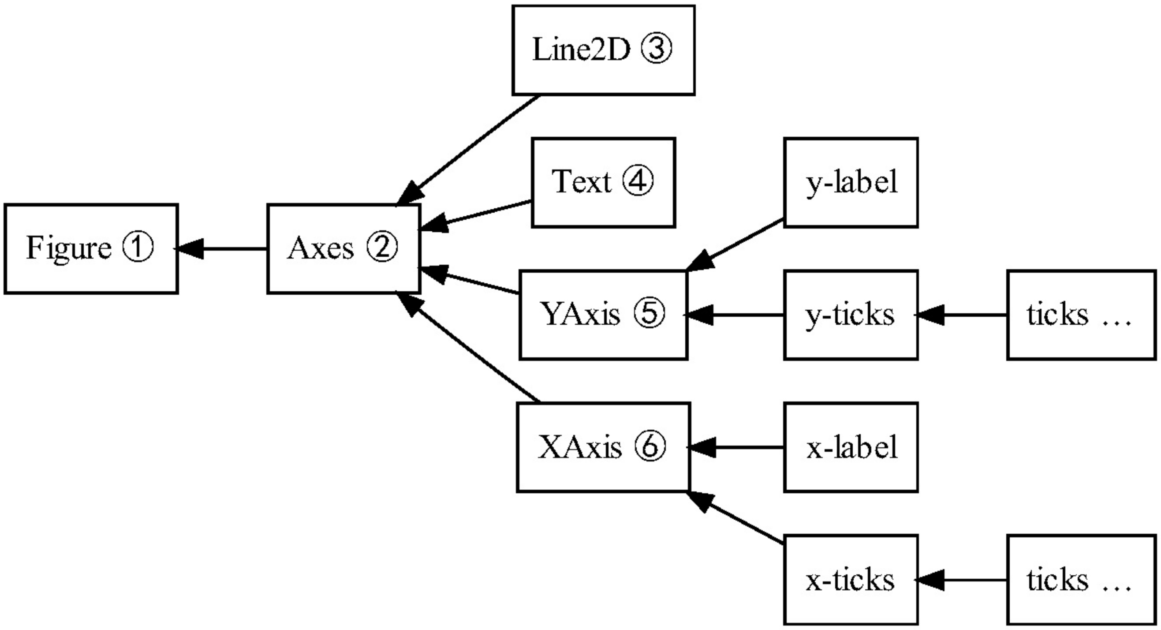


Figure 30: Artist Flow [262]

Figure 29 diplays the final output plot that is being generated by the Artist layer. Figure 30 shows the hierarchy of how this finished plot is built within the artist layer.

Finally, the Scripting Layer utilizes the pyplot API to allow the user to speak to the backend layer and make choices about what is represented in the plot that is created. This is the layer that the user will interact with and create code to influence the other layers of the library.

22.5 Matplotlib Features for Clinical Trial Research

Matplotlib is a massive library that allows for easy creation of basic plots, while also providing the functionality to create very intricate and powerful visualizations depending on your skill with the library. The most utilized package within the library is pyplot. The usefulness of the Matplotlib library is found in the versatility of visualizations that can be created. Very basic analysis can be run and presented in laymens terms to those individuals that are not versed in clinical trials research (such as the public) or more advanced visualizations can be run for internal teams.

Some plots and features that are available within the Matplotlib package are: line plots, multiple subplots, histograms, data handling, three-dimensional plotting, streamplotting, tables, scatter plots, gui widgets, log plots, EEG plots, etc [258]. There are several applications of the offered matplotlib features that would be extremely useful when running analysis in the middle of a clinical trial. Specific examples are gui widgets. This powerful tool would allow individuals with no knowledge of data handling to model outcomes if specific criteria were to happen. In oncology clinical trials and example would be: if a certain number of patient's cancer progress within a certain time point, they would choose to discontinue the trial. Another visualization that would be helpful throughout these trials would be basic line, bar, and histographs. These types of graphs require the least amount of external knowledge to interpret. With these types of charts it is easy to show how research is trending. Researchers would be able to visualize how tumor size is increasing/decreasing in the patient population over the course of a trial. Other business units would be able to clearly see if tumor size to understand and make changes based on these results.

22.6 CONCLUSION

Carrying out an oncology clinical trial successfully results in a massive output of data. This data needs to be analyzed throughout the course of the trial in order to ensure goals are being met. By analyzing the data output frequently, it will allow all business units involved to make better decisions to help both the patients and the company carrying out the trial. Current electronic data capture systems do not natively have functionality to create visualizations of the unstructured data within them. However, by utilizing the SDTM standard chosen by the FDA to give the data structure and combining it with the matplotlib Python package these problems could be remedied. The visualization created by the matplotlib package would provide pharmaceutical companies with a great way to communicate both internally and externally throughout the course of a trial. The visualizations available within matplotlib provide plotting functionality for basic to advanced plotting depending on the audience that they are intended for. Utilizing the matplotlib package throughout the course of the trial could allow pharmaceutical companies to make better decisions for patients, budgetary decisions for themselves, and provide progress updates for potential investors. These factors and much more make a great argument to start to implement the matplotlib package's visualization tools into standard practice for clinical trials research.

23 IBM Cognos Business Intelligence

Harika Putti
haputti@iu.edu
Indiana University
hid: fa18-523-81
github: ☁

23.1 Introduction

IBM Cognos [263] is a business intelligence suite that can help users to provide powerful insights to drive better business decisions. Business intelligence is an all-encompassing term that includes tools, infrastructure and applications that help in analyzing, evaluating and visualizing data. With help of BI software, one can envision relationships within the data that help organizations make knowledgeable business decisions. Using business intelligence tools, organizations can integrate their data and create reports, dashboards, metrics and scorecards to gain insights. Business intelligence suites by definition need to incorporate techniques like data processing, data modelling, querying, visualizing among other things. IBMs Cognos Business Intelligence is one among the many suites that has an extensive set of options ranging from exploration, modelling and querying to data visualization. The Cognos BI suite has multiple components including report studio, event studio, metric studio, framework manager, workspace among many others.

Keywords: hid fa18-523-81, business intelligence, BI, cognos, analytics, reports, queries

23.2 History

Cognos was a consulting and performance management company that was founded in 1969. It was acquired by IBM into its Infosphere product line in 2008 when companies like SAP, Oracle, Microsoft were fighting to become leaders in the BI market. In 2005, the company had released its Cognos 8 suite which introduced tools such as the Report studio, Query studio, Analysis studio and many others. After the acquisition by IBM, IBM Cognos 10 was released that had the capability to incorporate SPSS predictive analytics, historical and real-time analysis, better, faster and more flexible way of generating reports and dashboards. The next version of Cognos was the IBM Cognos Business Intelligence 10.2.2 that had the ability to integrate Microsoft Office with Cognos [264]. The latest edition of Cognos is the IBM Cognos Analytics 11.0 which is a state-of-the-art Analytics tool. This version of Cognos is a very powerful BI tool with ability to connect to Hadoop, an in-built AI assistant and smart data-discovery

23.3 ARCHITECTURE

Cognos has a 3-tiered architecture. Each tier separated by network firewalls.

- Tier 1: Web servers and Gateways
- Tier 2: Applications
- Tier 3: Data and Content store

Figure 31 shows the architecture for the Cognos Business intelligence suite.

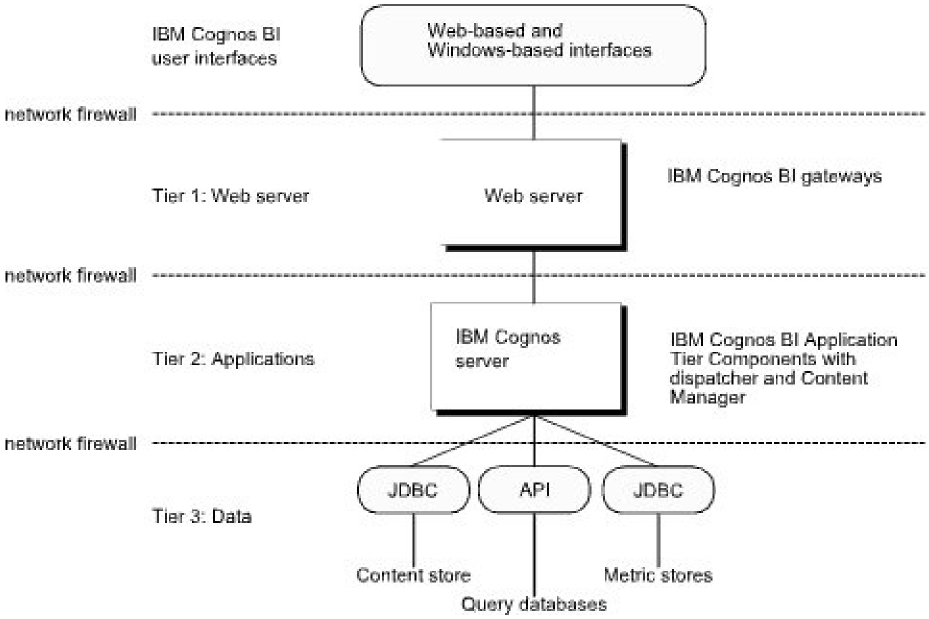


Figure 31: Architecture for cognos [265]

23.4 Components

The various components of Cognos Business intelligence [266], [267] are:

Component	Application
IBM Cognos Connection	Publishing, managing, and viewing content
IBM Cognos Insight	Managed workspaces
IBM Cognos Workspace	Interactive workspaces
IBM Cognos Workspace Advanced	Ad hoc querying and exploring data
IBM Cognos Report Studio	Creating reports
IBM Cognos Event Studio	Event management and alerting
IBM Cognos Metric Studio	Metrics and Scorecards
IBM Cognos for Microsoft Office	Cognos in Microsoft Office
IBM Cognos Query Studio	Ad hoc querying
IBM Cognos Analysis Studio	Exploring Data
IBM Framework Manager	Creating metadata models
IBM Cognos Transformer	Multi-dimensional data modeling

23.4.1 IBM Cognos Connection

IBM Cognos Connection is a web-based user interface that is used to access various cognos services such as Query Studio, Report Studio, Analysis Studio, Metric Studio etc. Using IBM Cognos connection one can access all the available reports and perform different operations such as create, run, schedule and access the reports. The admin has the ability to set up access permissions, manage data sources and provide individual and group memberships with in the interface. Users can personalize the connection as per their choice.

23.4.2 IBM Cognos Insight

Insight [268] is an individual entity with-in the Cognos family that incorporates a range of analytic abilities like ad-hoc querying and analyzing what-if scenarios. It is very user-friendly since one can perform data analysis, do off-hand querying, create dashboards with utmost ease. It’s almost like tableau where you can drag and drop the data files and insight automatically creates crosstabs and charts using something called smart metadata. Insight has the ability to create natural hierarchies between the data it receives to create OLAP cubes which makes it easier for the users to slice and dice the data as they prefer. Insight also comes with write back functionality that provides the ability to create models using prior data or new data that the user provides.

23.4.3 IBM Cognos Workspace

Workspace [269] is a web-based tool that can be used to create interactive dashboards known as workspaces using existing or new reports with in the IBM Cognos connection. This allows the users to create insightful visuals that can convey as much meaning from the information at a time as possible. It contains widgets and filters that users could use while creating the dashboards.

23.4.4 IBM Cognos Workspace Advanced

Business Intelligence is an area with a lot of options to choose from i.e. which studio, which data package etc. Most of the Cognos BI users are perplexed about the time wasted in navigating between different studios. Cognos Workspace Advanced was designed to bridge the gap between various studios. With the advent of Cognos Workspace Advanced, the redundancy of having three different studios is clearly noticed [270]. The advanced Cognos was designed in a way that it can amalgamate all the services on one platform where the users can use multiple services at the same time. It provides a single interface for querying and analysis and is very flexible in terms of presentation options.

23.4.5 IBM Cognos Report Studio

Report Studio [271] is a report composing instrument that proficient report creators and designers use to fabricate complex, various page reports against numerous databases. It’s the most used component in the IBM Cognos Business intelligence suite. With Report Studio, you can make any reports that your association requires. Report Studio offers a variety of services such as creating and formatting report using grouping, headers, footers, and other formatting options. Report Studio also enables focusing reports by filtering data and using prompts. The report studio also aids in adding value to the reports by performing different manipulations. It enhances the visual appeal of the reports though advanced formatting and exceptional data highlighting.

23.4.6 IBM Cognos Event Studio

The Cognos Event Studio [272] is predominantly used to keep track of the events in an organization to ensure that the decision-makers are notified of the upcoming events to make timely and effective decisions. An event is generally a situation that can create some impact on the business. When there are significant changes in the data, an event is detected to be taking place and agents are placed within this framework to detect the occurrence of events in organizational data. The sole purpose of an event studio is to notify the decision-makers about an event which is otherwise inconspicuous. When these agents are activated by the changes in data, the activation gets passed as notification to for further action such as sending an e-mail, adding information to the portal and running reports.

23.4.7 IBM Cognos Metric Studio

This service of IBM cognos helps to manage the performance of an organization by measuring the metrics at all the levels of the organization through creation of scorecards. The sole purpose behind creating score cards is to put performance indicators alongside the organization’s main performance measures. These scorecards then can be used to link to reports containing related information. These score cards can be customized using metric studio which helps in monitoring and analyzing metrics and projects throughout the organization [273]. Metric Studio helps in converting an organization’s strategy into relevant and measurable goals. That align every employee’s actions with a strategic plan. An environment rich of scorecards is analogous to a quick review document that shows how successful is an organization and where the flaws are to work and improve upon. It helps the decision-makers at all levels to react and plan based on the reports generated from the comparison of performance against targets while also simultaneously making a note of the current status of the business.

23.4.8 IBM Cognos Query Studio

This service of IBM Cognos is preliminarily for people with little or no training, one can quickly design, create and save reports that are not covered by the organizational reports. One can use query studio for ad-hoc reporting and can view data in hierarchies, create crosstab views and filter, sort, suppress and group the data easily without having to create any complex reports. One cannot define the properties of a data object like we can do in report studio, one cannot create multi-query or multipage reports either. Report studio offers a lot more visualizations and templates. Query studio is generally used by customers to quickly create reports that can be used as a reference to create higher level reports by the developer [274].

23.4.9 IBM Cognos Analysis Studio

Analytics studio [275] can be used to intelligently manipulate the data to understand the relationships within it. It provides support for filtering, calculating, sorting and analyzing the data. It can be used to comprehend patterns and inconsistencies, look at information, for example, points of interest to outlines, or real outcomes to planned outcomes evaluate execution by concentrating on the best or most exceedingly bad outcomes. It helps in multidimensional examination and investigation of expansive data sources. IBM Cognos Analytics is intended to enable one to report and dissect an organization’s performance rapidly and effortlessly.

23.5 Cognos Analytics

Cognos Analytics [276] is a futuristic tool which can provide analytic solutions with ease. It’s very user-friendly and has the ability to perform machine learning, pattern detection and smart visualization. It also enables sharing visualizations and reports on platforms like slack.

23.5.1 Big data and Cognos

With the increase in the amount of information that an organization can gather, the need for an integration between big-data and business intelligence tools has also increased. IBM having recognized that, created the possibility to connect to Hadoop databases.

23.6 Conclusions

In conclusion, Cognos is IBM’s business intelligence tool that can be used for a thorough scrutiny of performance of a business. This product is intended to empower business clients without programming prowess to extract corporate information, analyze and create reports using that information. Cognos empowers even a layman to create professional reports which are otherwise created only using years of expertise. Cognos provides exceptionally good implementation and deployment options that support scalability to grow along with the organization. It integrates well with other applications such as its seamless ability to integrate email and pdf functionality thereby allowing one to schedule reporting data deliveries. Cognos facilitates connection to multiple databases which can be stored within cognos’ content store.

24 OCR Technology Overview

Joao Paulo Leite
jleite@iu.edu
Indiana University
hid: fa18-523-88
github: ☁

Keywords: fa18-523-88, OCR, Optical Character Recognition, Computer Vision

24.1 Abstract

Optical Character Recognition (OCR) technology first appeared in the 1940’s and grew alongside the rise of the digital computer. It was not until the late 1950’s when OCR machines became commercially available and today this technology presents itself in both hardware devices as well as software offerings. Optical Character Recognition (OCR) was created as a way to transform text from a document into machine encode text. At a high level, an OCR system works by locating and segmenting each character, running the segmented character through a pre-processor for normalization and noise reduction, and extracting critical features to assist in the classification of each character. Once each character has been classified, the characters are regrouped and contextual information is applied to assist in word construction and to detect potential character misclassifications. While OCR technology has continued to evolve over the years into the realms of handwriting recognition, known as Intelligent Character Recognition (ICR), the main problem with these systems have been around degraded characters, which are incorrectly fragmented or joined characters, which causes issues during the segmentation process. OCR technology has far-reaching applications and is typically the first step when attempting to provide automation to document-centric processes such as image classification and data entry/indexing.

24.2 Introduction

The main principle in Optical Character Recognition (OCR) is to automatically recognize character patterns. This is accomplished by showing the system each class of pattern that can occur and providing a training set for each pattern. At the time of recognition, the system uses the previously provided examples to classify the new character to the closest match. Typical, OCR system are designed to solely transform text on a document into machine-encoded text and additional systems must be built to further extract relevant information from the document. That is to say, the process of OCR is the first step in transforming structured, semi-structured and unstructured documents into valuable and relevant information.

24.3 Optical Character Recognition

As stated in the name Optical Character Recognition, the characters that are typically trained are letters, numbers and special symbols. Each differing character is defined as its own class, and the system builds an understanding of each class utilizing examples of characters provided. The steps that are typically performed by an OCR system are threshold processing, character segmentation, character preprocessing, feature extraction, classification and post processing.

24.3.1 Threshold Processing

At its core, the OCR process expects to process a black character presented against a white background. While images coming into an OCR system could have already undergone this transformation from color image into a black and white image via a scanner, it is beneficial to perform this step before passing the image into the OCR engine to provide the highest level quality to the OCR engines. The mechanism behind this conversion analyzes each pixel to determine if it should be assigned as a black or white pixel. For color images, this thresholding can be set at a fixed level so that any faintly colored pixels can be dropped as white while truly dark colored pixels are converted to black. In the case of grayscale images, the same threshold can be set with the difference being the shades of gray presented in each pixel. Once this process is complete, the newly created black and white images are used for the remainder of the process[277].

24.3.2 Character Segmentation

Character segmentation is a critical step in the process which represents breaking the image down into logical segments. While the system can be designed to segment the image into words, typically OCR is most successful if it is segmented to the lowest common denominator, the character. Each character is defined as a contiguously connected set of pixels and a break in the connection constitutes the beginning of a new character. While this may sound like a straightforward process, problems can occur when characters are fragmented or touching. Character distortions due to image quality issues or ‘serifed’ fonts are the main culprits behind fragmented or touching characters, while noise such as marks, handwriting and dots can also contribute to challenges when attempting to segment characters. To alleviate this issue, before the characters are presented to the feature extraction phase in the process, the characters are run through the preprocessing phase in an attempt to correct some of the issues that may have manifested themselves[278].

24.3.3 Character Preprocessing

Character Preprocessing is a vital step used to clean up common defects introduced during the previous scanning or thresholding steps. The goal of preprocessing is to remove the faults that can later cause poor character recognition in the subsequent steps.

To combat these defects, the preprocessor employs a common technique called smoothing. Smoothing serves to both fill in gaps within a character (fragmentation correction) as well as thin the width of lines within a character (touching correction). When properly applied, smoothing is successful in filling in pits and removing bumps from characters, which will increase the likelihood of recognition in the following steps[279]. The preprocessor also invokes tasks for noise removal and character normalization. The noise removal task removes of specks, thin lines and other inconsistencies through the analysis of height, size and density of a grouping of pixels. If the characteristics of a particular grouping is not consistent with the characteristics found for a typical character, the grouping is deemed noise and removed as such. The normalization of characters is applied to provide a uniformly sized and oriented character, fixing issues around scaling, slanting and rotation of characters. With the character preprocessing completed, OCR is ready for feature extraction.

24.3.4 Feature Extraction

The most simplistic extraction technique is known as template matching. The technique does not use feature analysis and will only compare the input character against a known set of characters provided for each class at a pixel level. The distance between the inputted character and the set of known characters is calculated for each class. Once that comparison is completed, the class with lowest distance is assigned as the class for the input character. However, the set-back of this method is that it does not afford any flexibility around noise or font variations that have not yet been assigned[278].

Because of rigidity of the template matching technique, feature based techniques were later developed to extract significant features from a character. Some common feature extraction methods are zoning, distance profiling, and directional distribution analysis[277].

24.3.4.1 Zoning

Zoning is a technique that frames the character in a set of overlapping or non-overlapping zones. The pixel density in each zone must be calculated by taking the number of black pixels in the zone divided by the total number of pixels presented in the zone. The resulting ratio for each zone becomes the feature that describes the character.

24.3.4.2 Distance Profiling

Distance profiling is a technique that frames the character in a bounding box. The distance from the bounding box to the outer edge of the character is calculated for each of the four side (top, bottom, left and right). The resulting calculated distance becomes the feature that describes the character.

24.3.4.3 Directional Distribution

Directional distribution analysis is a technique that assigned a center point to the character. Once the center point is assigned, the weight is calculated by taking the number of black pixels found in each direction divided by the total number of pixels found in the character. The resulting ratio for each direction becomes the feature that describes the character.

Because these techniques are independent, there are possibilities to combine multiple features to increase the accuracy of recognition.

24.3.5 Classification

The classification step is the culmination of all the previous steps to obtain the desired result of assigning a character to the correct class. One such classification method that could be used is K- Nearest Neighbor.The K-Nearest Neighbor (KNN) provides a method to classify characters based on the closest features extracted in the training set. Typically regarded as a simple machine learning algorithm, KNN calculates the Euclidean distance between features value of the input character against the features value of the characters in the training set. Once the distance is calculated, the results are arranged in order and the input character is assigned the character class that corresponds to the majority of its nearest neighbors[277].

24.3.6 Post Processing

24.3.6.1 Grouping

Once all the individual characters have been successfully classified, the system can begin to group those set of characters into the next level of association. Grouping characters into logical strings of words, numbers or tokens is an easy task of considering the location of each individual character and evaluating the pixel distance (white space) to the next individual character. With machine printed text, the assumption is that distances between words are far greater than distances between characters within a word. Once grouping is complete, the system is able to leverage the newly formed words to provide error detection and logical character correction.

24.3.6.2 Error-Detection

Because individual character recognition will never be 100 percent accurate, we can utilize the context around our newly formed words from the grouping phase to increase the accuracy and detect errors around the recognition. This secondary evaluation process will be based on the systems understanding of the underlying language for which the text is written in.

24.3.6.3 Language Syntax

One form to evaluate the accuracy is to use the syntax of the language and rule out specific combinations of characters appearing in sequence. As an example, if the recognition for the three-letter word “cut” came back as “cwt”, the system would understand that the syntax of a C followed by a W and a W followed by a T is highly improbable in the English language and flag this a potential error ???.

24.3.6.4 Dictionaries

Another evaluation method that can assist with the accuracy is a dictionary lookup. Following the logic of the example above, after understanding that we have mistakenly extracted “cwt”, we can apply dictionaries to assist in correcting the error that was caused by the individual character recognition engine. Because “w” and “u” share some common characteristics, the original classification can be utilized to not only provide the highest matching character but also consider which matching characters provides the highest probability of forming a word that matches an entry in the dictionary???.

24.3.7 Conclusion

As the evolution of Optical Character Recognition systems continue to evolve, new techniques may be developed to increase the accuracy of such systems. With that said, the overall structure and process of these new systems will follow what has been outlined and discussed in this paper. This is especially true in the more challenging arena of handwritten recognition, where systems based on neural networks have begun to emerge in recent years.

25 Scikit-learn

Mark Miller
mgm3@iu.edu
Indiana University Bloomington
hid:fa18-523-63
github: ☁

Keywords: HID fa18-523-63, Scikit Learn, Machine Learning Methods, Python, Artificial Intelligence

Scikit-learn [280] is Python's inherent machine learning library. It is a robust library that intends uses object oriented programming to implement commonly used machine learning algorithms effectively, efficiently, pythonically, and swiftly. While, for specific purposes, many experts are able to implement their own algorithms that may improve upon the Scikit-learn library, it has sufficient tools and robustness that enable it to be the leading library for machine learning topics within Python.

There are built in libraries to Python that can make these tasks much simpler to understand and to implement, Scikit-learn provides one such solution [280]. Junior machine learning experts are gaining footing in the industry and are able to gain reputation, thanks to the help of many of these libraries.

As data becomes larger and larger with time, experts in the field are needed to perform this operation. Or, better stated, experts are needed to be able to program computers to perform these operations for them. A computer can process streamlined and well-defined data faster than humans in some instances, especially when reviewing the data becomes increasingly tedious. While algorithms may never be as good at recognizing what is in an image quite like a human can, they will become closer and closer to the point where advertizing will be even more targeted, devices will understand the wants of their human masters clearly, and effective decisions can be made with minimal error. Scikit-learn is not the most sophisticated a capable machine learning algorithm out there, but it is effective and easily implemented via Python.

25.1 The supervised algorithms (some of them)

The following alorithms are chosen as they are among the more common machine learning algorithms/methods that are implemented in today's data science world.

- Nearest Neighbors: Nearest neighbors is a machine learning algorithm that makes the decision of one input variable based on training data (making it a supervised algorithm) that most similarly matches itself. Once one of the training sets is identified as the closest match of inputs, it will assign the same category for the test instance. With careful parameter tuning, some scholars believe this to be a better classification method than random forests [281]
- Naive Bayes: Naive Bayes takes a look at Baysian statistics and makes one majorly naive assumption: all of the inputs are independent of each other. While this is a glaring assumption, due to ease of implementations, most issues that would arise from generally erroneous assumption are not impactful [282].
- Decision trees: Separating on different attributes of the input data, decision trees are one of the more robust machine learning algorithms in that they are able to handle a wide variety of inputs and still maintain their quality [283]. Splitting on each attributes (usually on traits that maximize the entropy of the model, to enhance effectiveness). A branch off algorithm to decision trees are random forests which use many small, randomly chosen (with replacement) trees that can use small subsets of the data to formulate better opinions in a less computationally intensive way.
- Neural network models (supervised) {284] are algorithms that are particularly good at image classification. They obtain different neurons, each of which contributes in the decision making. They are based off of the way a human neural network would work if it could be modeled accurately via code. each neural network has different levels which contribute to the decision making process.

25.2 The unsupervised algorithms (some of them)

There are many unsupervised learning algorithms supported by Scikit, here are highlighted two of them. The ones here were chosen because they are commonly used with simple implementations, which don't take an expert to implement.

- Clustering [285]: Stemming from the k-means clustering, this is designed to group the datapoints based on similarity to others in in the same dataset. The inputs are generally strictly numerical but can be n-dimmensional. Clustering converges quickly via iterative methods but is highly sensitive to initialization, making it very important to have domain knowledge and valuable visualization strategies when the data is 3-dimmensional and higher.
- Neural Network (unsupervised) [286]: much like the aforementioned neural networks, sklearn has libraries for unsupervised machine learning algorithms, which don't require training data to make decisions. In this sense, it becomes more of a clustering algorithm than a group identification.

25.3 Other Method groupings within sklearn

- Dataset transformations [280]: Oftentimes, data comes mangled and hard to use, requiring the need for effective data wrangling. scikit-learn has methods for feature extraction, preprocessing, random projection, dimmensionality reduction, and more. This makes it a valuable library for more aspects than just the machine learning algorithms themselves.
- Dataset loading utilities [280]: Scikit-learn has the utilities needed to load data as well as read it. There are Application Programming Interfaces for training datasets, real-world datasets, generated-datasets, and the tools needed to use them effectively.

25.4 Further functionalities to Scikit

There are more uses and tools in scikit-learn than what are mentioned here. It was inappropriate to just copy the user guide or man pages for these articles, even though they are good. The user guide contains valuable examples and assists with syntax. You will need basic machine learning understanding to be able to use any of these methods in this library. Once the understanding is there and basic implementations are used, microtuning and enhancing of the algorithms comes quickly and simply to an expert with a good eye. This article does not contain mentions of every method in Scikit, just a few machine learning algorithms that can be used [287].

25.5 Real world applications for scikit learn

The real world applications are nearly as endless as are the applications for machine learning and artificial intelligence. The trick is getting the data to work together, whether that be through internet of things, internet of computers, internet of people, etc. This tool can be used in many ways, ranging from sports analytics to automation of analysis of the stock exchange. Expert knowledge of this library alone can bring six-figure salaries as a machine learning engineer, which many major and minor companies alike choose to employ. [287]

References

[1] M. Scherocman, “Top 5 benefits of microsoft azure sql database.” website, 2016 [Online]. Available: https://www.interlink.com/blog/entry/top-5-benefits-of-windows-azure-sql-database

[2] Microsoft, “Azure sql database purchasing models | microsoft docs.” website, 2018 [Online]. Available: https://github.com/MicrosoftDocs/azure-docs/blob/master/articles/sql-database/sql-database-service-tiers.md

[3] Microsoft, “Welcome to azure cosmos db.” website, 2018 [Online]. Available: https://docs.microsoft.com/en-us/azure/cosmos-db/introduction

[4] Microsoft, “SLA for azure cosmos db.” website, 2018 [Online]. Available: https://azure.microsoft.com/en-us/support/legal/sla/cosmos-db/v1_2/

[5] A. Ali, “Getting started with azure sql data warehouse - part 1.” website, 2017 [Online]. Available: https://www.databasejournal.com/features/mssql/getting-started-with-azure-sql-data-warehouse-part-1.html

[6] Microsoft, “What is polybase?” website, 2018 [Online]. Available: https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-guide?view=sql-server-2017

[7] J. P. Hoang, “Common isv application patterns using azure sql data warehouse.” website, 2017 [Online]. Available: https://blogs.msdn.microsoft.com/sqlcat/2017/09/05/common-isv-application-patterns-using-azure-sql-data-warehouse/

[8] Microsoft, “What is azure hdinsight and the apache hadoop technology stack.” website, 2018 [Online]. Available: https://docs.microsoft.com/en-us/azure/hdinsight/hadoop/apache-hadoop-introduction

[9] Microsoft, “What is stream analytics?” website, 2018 [Online]. Available: https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-introduction

[10] Microsoft, “Azure data lake storage.” website, 2018 [Online]. Available: https://azure.microsoft.com/en-us/services/storage/data-lake-storage/

[11] Microsoft, “A closer look at azure data lake storage gen2.” website, 2018 [Online]. Available: https://azure.microsoft.com/en-us/blog/a-closer-look-at-azure-data-lake-storage-gen2/

[12] Microsoft, “What is azure data lake analytics?” website, 2018 [Online]. Available: https://docs.microsoft.com/en-us/azure/data-lake-analytics/data-lake-analytics-overview

[13] Microsoft, “Introduction to azure data factory.” website, 2018 [Online]. Available: https://docs.microsoft.com/en-us/azure/data-factory/introduction

[14] Microsoft, “Transform data in azure data factory.” website, 2018 [Online]. Available: https://docs.microsoft.com/en-us/azure/data-factory/transform-data

[15] GDPR.ORG, “GDPR Key Changes.” Web Page, 2018 [Online]. Available: https://eugdpr.org/the-regulation/

[16] gdpr-info.eu, “GDPR Definitions.” Web Page, 2018 [Online]. Available: https://gdpr-info.eu/art-4-gdpr/

[17] AWS, “Navigating GDPR Compliance on AWS.” Web Page, Sep-2018 [Online]. Available: https://d1.awsstatic.com/whitepapers/compliance/GDPR_Compliance_on_AWS.

[18] P. Mell and T. Grance, “The NIST Definition of Cloud Computing.” Web Page, Oct-2009 [Online]. Available: https://www.nist.gov/sites/default/files/documents/itl/cloud/cloud-def-v15.pdf

[19] V. D. Somma, “Openstack compliance with GDPR.” Web Page, 2018 [Online]. Available: https://archive.fosdem.org/2018/schedule/event/vai_openstack_gdpr_compliance

[20] C. Woolf, “All AWS Services GDPR ready.” Web Page, 2018 [Online]. Available: https://aws.amazon.com/blogs/security/all-aws-services-gdpr-ready/

[21] S. Frey, “Google Cloud: Ready for the GDPR.” Web Page, 2018 [Online].

Available: https://cloud.google.com/blog/topics/inside-google-cloud/google-cloud-ready-for-gdpr

[22] Cloud App Security Team, “Assess GDPR readiness with Microsoft Cloud App Security.” Web Page, 2018 [Online]. Available: https://techcommunity.microsoft.com/t5/Enterprise-Mobility-Security/Assess-GDPR-readiness-with-Microsoft-Cloud-App-Security/ba-p/250572

[23] RedHat, “Privacy Statement.” Web Page, May-2018 [Online]. Available: https://www.redhat.com/en/about/privacy-policy

[24] Cloud Security Alliance, “About.” website, 2018 [Online]. Available: https://cloudsecurityalliance.org/about/

[25] Cloud Security Alliance, “About us.” website, 2018 [Online]. Available: https://www.linkedin.com/company/cloud-security-alliance/

[26] E. Messmer, “Cloud security alliance formed to promote best practices.” Website, Mar-2009 [Online]. Available: https://www.computerworld.com/article/2523598/security0/cloud-security-alliance-formed-to-promote-best-practices.html

[27] Cloud Security Alliance, “Chapters.” website, 2018 [Online]. Available: https://cloudsecurityalliance.org/chapters/

[28] Cloud Security Alliance, “Guidance.” website, 2018 [Online]. Available: https://cloudsecurityalliance.org/guidance/#_overview

[29] Cloud Security Alliance, “STAR.” website, 2018 [Online]. Available: https://cloudsecurityalliance.org/star/#_overview

[30] Cloud Security Alliance, “CCSA.” website, 2019 [Online]. Available: Why Obtain the CCSK?

[31] Cloud Security Alliance, “CCSP.” website, 2018 [Online]. Available: https://cloudsecurityalliance.org/education/ccsp/#_overview

[32] Cloud Security Alliance, “Global consultancy.” website, 2018 [Online]. Available: https://cloudsecurityalliance.org/global-consultancy/#_overview

[33] Cloud Security Alliance, “Groups.” website, 2018 [Online]. Available: https://cloudsecurityalliance.org/research/#_groups

[34] KNIME, “KNIME integrations.” Web page, 2018 [Online]. Available: https://www.knime.com/knime-software/knime-integrations

[35] A. Vidhya, “Building your first machine learning model using knime (no coding required!).” Web page, 2017 [Online]. Available: https://www.analyticsvidhya.com/blog/2017/08/knime-machine-learning/

[36] KNIME, “JSON processing.” Web page, 2018 [Online]. Available: https://www.knime.com/whats-new-in-knime-211#JSON

[37] U. Sewwandi, “Guided analytics using knime analytics platform.” Web page, 2018 [Online]. Available: https://towardsdatascience.com/guided-analytics-using-knime-analytics-platform-b6543ebab7e2

[38] KNIME, “KNIME workflow hub.” Web page, 2018 [Online]. Available: https://www.knime.com/whats-new-in-knime-36#knime-workflow-hub

[39] KNIME, “Distributed executors in the next major version of knime server.” Web page, 2018 [Online]. Available: https://www.knime.com/blog/distributed-executors-in-the-next-major-version-of-knime-server

[40] KNIME, “KNIME analytics platform.” Web page, 2018 [Online]. Available: https://www.knime.com/knime-software/knime-analytics-platform

[41] KNIME, “KNIME quickstart guide.” Web page, 2018 [Online]. Available: https://forge.epn-campus.eu/svn/edna/trunk/deprecated/rcp-knime/org.edna.workbench.target/knime2.1.2/org.knime.workbench.help_2.1.2.0

[42] KNIME, “KNIME on amazon web services.” Web page, 2018 [Online]. Available: https://www.knime.com/knime-software/knime-aws

[43] KNIME, “KNIME on microsoft azure.” Web page, 2018 [Online]. Available: https://www.knime.com/knime-software/knime-azure

[44] Apache, “Kafka.” Web Page [Online]. Available: https://kafka.apache.org/

[45] T. P. Neha Narkhede Gwen Shapira, ***Kafka: The definitive guide***, First. O'REILLY, 2017 [Online]. Available: https://www.confluent.io/wp-content/uploads/confluent-kafka-definitive-guide-complete.pdf

[46] Apache, "Apache kafka." Web Page, 2018 [Online]. Available: https://www.apache.org/dyn/closer.cgi?path=/kafka/2.1.0/kafka-2.1.0-src.tgz

[47] Apache, "Kafka." Web Page [Online]. Available: https://issues.apache.org/jira/browse/KAFKA-6855

[48] The Apache Software Foundation, "Apache nifi." Web page, Oct-2018 [Online]. Available: https://nifi.apache.org/

[49] A. DOKAEVA, "How to make etl simple and intuitive with nifi." Web page, Mar-2018 [Online]. Available: https://issart.com/blog/how-to-make-etl-simple-and-intuitive-with-nifi/

[50] S. Maarek, "Introduction to apache nifi (hortonworks dataflow - hdf 2.0)." Presentation [Online]. Available: https://www.udemy.com/apache-nifi/

[51] A. Bridgwater, "NSA 'nifi' big data automation project out in the open." Web Page, Jul-2015 [Online]. Available: https://www.forbes.com/sites/adrianbridgwater/2015/07/21/nsa-nifi-big-data-automation-project-out-in-the-open/#68cdd7dc55d6

[52] Apache NiFi Team, "Apache nifi overview." Web page, Oct-2018 [Online]. Available: https://nifi.apache.org/docs.html

[53] hortonworks, "Analyze transit patterns with apache nifi." Web page, Oct-2018 [Online]. Available: https://hortonworks.com/tutorial/analyze-transit-patterns-with-apache-nifi/section/1/

[54] S. Gupta, "Creating custom processors and controllers in apache nifi." Web page, May-2018 [Online]. Available: https://medium.com/hashmapinc/creating-custom-processors-and-controllers-in-apache-nifi-e14148740ea

[55] Apache NiFi Team, "Apache nifi downloads." Web page, Oct-2018 [Online]. Available: http://nifi.apache.org/download.html

[56] Apache NiFi Team, "Getting started with apache nifi." Web page, Oct-2018 [Online]. Available: https://nifi.apache.org/docs/nifi-docs/html/getting-started.html#downloading-and-installing-nifi

[57] V. Anand, "Using nifi to simplify data flow & streaming use cases @ mastercard." Presentation [Online]. Available: https://dataworkssummit.com/san-jose-2018/session/using-nifi-to-simplify-data-flow-streaming-use-cases-mastercard/

[58] S. Vaid, "Streaming analytics with opentext magellan." Web page, Aug-2018 [Online]. Available: https://blogs.opentext.com/streaming-analytics-with-opentext-magellan/

[59] Compose, "A real use case with nifi, the swiss army knife of data flow." Presentation [Online]. Available: http://mybbt.bbtconsulting.com:8069/slides/slide/a-real-use-case-with-nifi-the-swiss-army-knife-of-data-flow-121

[60] Ford, "Real time streaming architecture at ford." Presentation, Jun-2017 [Online]. Available: https://www.slideshare.net/Hadoop_Summit/real-time-streaming-architecture-at-ford

[61] A. Paszke ***et al.***, "Automatic differentiation in pytorch," 2017.

[62] NumPy, "NumPy." Website [Online]. Available: http://www.numpy.org/

[63] PyTorch, "What is pytorch?" Website [Online]. Available: https://pytorch.org/tutorials/beginner/blitz/tensor_tutorial.html

[64] Wikipedia, "PyTorch." Website [Online]. Available: https://en.wikipedia.org/wiki/PyTorch

[65] TensorFlow, "Get started." Website [Online]. Available: https://www.tensorflow.org/

[66] Keras, "Keras: The python deep learning library." Website [Online]. Available: https://keras.io/

[67] Caffe, "Caffe." Website [Online]. Available: http://caffe.berkeleyvision.org/

[68] Chainer, “Get started.” Website [Online]. Available: https://chainer.org/

[69] MXNet, “Apache mxnet (incubating).” Website [Online]. Available: https://mxnet.apache.org/

[70] Microsoft, “CNTK.” GitHub [Online]. Available: https://github.com/Microsoft/CNTK

[71] DL4J, “Quick start.” Website [Online]. Available: https://deeplearning4j.org/

[72] Torch, “Torch.” Website [Online]. Available: http://torch.ch/

[73] Wikipedia, “C programming language.” Website [Online]. Available: https://en.wikipedia.org/wiki/C_(programming_language)

[74] LuaJIT, “The luajit project.” Website [Online]. Available: http://luajit.org/

[75] Wikipedia, “Deep learning.” Website [Online]. Available: https://en.wikipedia.org/wiki/Deep_learning

[76] J. Brownlee, “What is deep learning?” Website [Online]. Available: https://machinelearningmastery.com/what-is-deep-learning/

[77] Wikipedia, “Artificial neural network.” Website [Online]. Available: https://en.wikipedia.org/wiki/Artificial_neural_network

[78] M. Mayo, “WTF is a tensor?!?” Website [Online]. Available: https://www.kdnuggets.com/2018/05/wtf-tensor.html

[79] A. M. Giancarlo Zaccone Md. Rezaul Karim, “Computational graphs.” Website [Online]. Available: https://www.packtpub.com/mapt/book/big_data_and_business_intelligence/9781
graphs

[80] Wikipedia, “Automatic differentiation.” Website [Online]. Available: https://en.wikipedia.org/wiki/Automatic_differentiation

[81] Wikipedia, “Backpropagation.” Website [Online]. Available:

https://en.wikipedia.org/wiki/Backpropagation

[82] PyTorch, “Autograd mechanics.” Website [Online]. Available: https://pytorch.org/docs/stable/notes/autograd.html

[83] V. Rao, “Get started with pytorch.” Website [Online]. Available: https://developer.ibm.com/articles/cc-get-started-pytorch/

[84] PyTorch, “QUICK start locally.” Website [Online]. Available: https://pytorch.org

[85] PyTorch, “GET started.” Website [Online]. Available: https://pytorch.org/get-started/locally/

[86] PyTorch, “WELCOME to pytorch tutorials.” Website [Online]. Available: https://pytorch.org/tutorials/index.html

[87] D. Mwiti, “Introduction to pytorch for deep learning.” Website [Online]. Available: https://heartbeat.fritz.ai/introduction-to-pytorch-for-deep-learning-5b437cea90ac

[88] D. Mesquita, “How pytorch gives the big picture with deep learning.” Website [Online]. Available: https://medium.freecodecamp.org/how-pytoch-gives-the-big-picture-with-deep-learning-e4a0f372f4b6

[89] D. Mesquita, “README.” Website [Online]. Available: https://github.com/dmesquita/understanding_pytorch_nn

[90] kdnuggets, “PyTorch or tensorflow?” Website [Online]. Available: https://www.kdnuggets.com/2017/08/pytorch-tensorflow.html

[91] Scipy, “Scipy.” Website [Online]. Available: https://www.scipy.org/

[92] A. Paszke, “CREATING extensions using numpy and scipy.” Website [Online]. Available: https://pytorch.org/tutorials/advanced/numpy_extensions_tutorial.html

[93] Siftery Discover, “PyTorch alternatives.” Website [Online]. Available: https://siftery.com/pytorch/alternatives

[94] Wikipedia, “Serialization.” Website [Online]. Available: https://en.wikipedia.org/wiki/Serialization

[95] Wikipedia, Website [Online]. Available: https://en.wikipedia.org/wiki/Deep_learning

[96] Wikipedia, “Convolutional neural network.” Website [Online]. Available: https://en.wikipedia.org/wiki/Convolutional_neural_network

[97] Wikipedia, Website [Online]. Available: https://en.wikipedia.org/wiki/Graphics_processing_unit

[98] Wikipedia, “Computer vision.” Website [Online]. Available: https://en.wikipedia.org/wiki/Computer_vision

[99] Wikipedia, “Artificial intelligence.” Website [Online]. Available: https://en.wikipedia.org/wiki/Artificial_intelligence

[100] Wikipedia, “Signal processing.” Website [Online]. Available: https://en.wikipedia.org/wiki/Signal_processing

[101] Wikipedia, Website [Online]. Available: https://en.wikipedia.org/wiki/Natural_language_processing

[102] Wikipedia, “Perceptron.” Website [Online]. Available: https://en.wikipedia.org/wiki/Perceptron

[103] Wikipedia, “MNIST.” Web page [Online]. Available: https://en.wikipedia.org/wiki/MNIST_database

[104] Wikipedia, “MATLAB.” Website [Online]. Available: https://en.wikipedia.org/wiki/MATLAB

[105] Wikipedia, “Artificial neural network.” Website [Online]. Available: https://en.wikipedia.org/wiki/Artificial_neural_network

[106] Wikipedia, “HDF5.” Website [Online]. Available: https://en.wikipedia.org/wiki/Hierarchical_Data_Format

[107] P. P. Subhasis Das Hou Yunqing, “Quora-answer-3.” Website [Online]. Available: https://www.quora.com/What-are-the-pros-and-cons-of-Caffe-the-deep-learning-framework

[108] Facebook, “Caffe2.” Web page [Online]. Available: https://github.com/pytorch/pytorch/tree/master/caffe2

[109] “NLP for big data: What everyone should know?” Web page [Online]. Available: https://www.expertsystem.com/nlp-big-data-everyone-know/

[110] E. Liddy, “Encyclopedia of library and information sciences,” 2nd ed., Marcel Decker, Inc., 2001 [Online]. Available: https://surface.syr.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1019&context=cnlp

[111] A. Vieira and B. Ribeiro, ***Natural language processing and speech. In: Introduction to deep learning business applications for developers***. Apress, Berkeley CA., 2018.

[112] Robin, “Part-of-speech tagging.” Web page, Dec-2009 [Online]. Available: http://language.worldofcomputing.net/pos-tagging/parts-of-speech-tagging.html#

[113] Gartner, “Natural-language understanding.” Web page, 2018 [Online]. Available: https://www.gartner.com/it-glossary/nlu-natural-language-understanding/

[114] S. C. Shapiro, “Encyclopedia of artificial intelligence,” 2nd ed., S. C. Shapiro, Ed. John Wiley & Sons, Inc., 1992, pp. 54–57 [Online]. Available: https://cse.buffalo.edu/~shapiro/Papers/ai.pdf

[115] Wikipedia, “AI-complete.” Web page [Online]. Available: https://en.wikipedia.org/wiki/AI-complete

[116] M. Clark, “Understanding nlu: A cheatsheet for beginners.” Web page, Apr-2017 [Online]. Available: https://info.contactsolutions.com/digital-engagement-blog/understanding-nlu-cheat-sheet-for-beginners

[117] S. Petrov, “Announcing syntaxnet: The world’s most accurate parser goes

open source.” Web page, May-2016 [Online]. Available: https://ai.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html

[118] “Stanford log linear part-of-speech tagger.” Web page [Online]. Available: https://nlp.stanford.edu/software/tagger.shtml

[119] K. Toutanova, D. Klein, C. Manning, and Y. Singer, “Feature-rich part-of-speech tagging with a cyclic dependency network,” in ***Proceedings of hlt-naacl 2003***, 2003, pp. 252–259 [Online]. Available: https://nlp.stanford.edu/~manning/papers/tagging.pdf

[120] N. Madnani and J. Lin, “Natural language processing with apache hadoop and python.” Web page, Mar-2010 [Online]. Available: https://blog.cloudera.com/blog/2010/03/natural-language-processing-with-hadoop-and-python/

[121] N. Project, “Natural language toolkit.” Web page, 2017 [Online]. Available: https://www.nltk.org/

[122] M. Rouse, “Hadoop.” Web page, Apr-2018 [Online]. Available: https://searchdatamanagement.techtarget.com/definition/Hadoop

[123] M. Rouse, “Deep learning (deep neural network).” Web page, 2018 [Online]. Available: https://searchenterpriseai.techtarget.com/definition/deep-learning-deep-neural-network

[124] E. Systems, “NLP for big data: What everyone should know?” Web page, Aug-2016 [Online]. Available: https://www.expertsystem.com/nlp-big-data-everyone-know/

[125] E. D. Liddy, “Enhanced text retrieval using natural language processing,” ***Bulletin of the American Society for Information Science and Technology***, vol. 24, no. 4, pp. 14–16, 1998.

[126] website [Online]. Available: https://www.sas.com/en_us/insights/analytics/what-is-natural-language-processing-nlp.html

[127] G. G. Chaudhary, “Natural language processing,” ***Dept. of Computer and***

Information Sciences University of Strathclyde, Glasgow G1 1XH, UK, 2003.

[128] D. Khurana, A. Koli, K. Khatter, and S. Singh, “Natural language processing: State of the art, current trends and challenges,” ***CoRR***, vol. abs/1708.05148, 2017.

[129] N. Tapaswi and S. Jain, “Treebank based deep grammar acquisition and part-of-speech tagging for sanskrit sentences,” in ***Software engineering (conseg), 2012 csi sixth international conference on***, 2012, pp. 1–4.

[130] P. Ranjan and H. V. S. S. A. Basu, “Part of speech tagging and local word grouping techniques for natural language parsing in hindi,” in ***Proceedings of the 1st international conference on natural language processing (icon 2003)***, 2003.

[131] M. Palmer, D. Gildea, and P. Kingsbury, “The proposition bank: An annotated corpus of semantic roles,” ***Computational linguistics***, vol. 31, no. 1, pp. 71–106, 2005.

[132] M. Vallez and R. Pedraza-Jimenez, “Natural language processing in textual information retrieval and related topics,” 2011.

[133] “Natural language processing.” website [Online]. Available: https://en.wikipedia.org/wiki/Natural_language_processing

[134] chirag, “NLP for big data: How natural language processing is poised to revolutionise big data analytics.” website, Jun-2017 [Online]. Available: https://huddle.eurostarsoftwaretesting.com/nlp-for-big-data-how-nlp-will-revolutionise-big-data-analytics/

[135] V. Fedak, “5 heroic tools for natural language processing.” website, Jan-2018 [Online]. Available: https://towardsdatascience.com/5-heroic-tools-for-natural-language-processing-7f3c1f8fc9f0

[136] A. Geitgey, “Natural language processing is fun!” website, Jul-2018 [Online]. Available: https://medium.com/@ageitgey/natural-language-processing-is-fun-9a0bff37854e

[137] T. Mills, “What is natural language processing and what is it used for?”

website, Jul-2018 [Online]. Available: https://www.forbes.com/forbes/welcome/?toURL=https://www.forbes.com/sites/forbestechcouncil/2018/07/02/what-is-natural-language-processing-and-what-is-it-used-for/&refURL=https://www.google.com/&referrer=https://www.google.com/

[138] Amazon, Web page [Online]. Available: https://aws.amazon.com/streaming-data/

[139] G. Vaseekaran, "Big data battle : Batch processing vs stream processing." Web page, Oct-2017 [Online]. Available: https://medium.com/@gowthamy/big-data-battle-batch-processing-vs-stream-processing-5d94600d8103

[140] S. Perera, "A gentle introduction to stream processing." Web page, Apr-2018 [Online]. Available: https://medium.com/stream-processing/what-is-stream-processing-1eadfca11b97

[141] R. Vadai, "Challenges processing data in real-time using conventional big data solutions." Web page, Mar-2017 [Online]. Available: https://codelook.com/challenges-with-processing-data-in-real-time-using-conventional-big-data-solutions-bb602b33da0c

[142] "Spark streaming." Web page [Online]. Available: https://spark.apache.org/streaming/

[143] A. C. Oliver, "Storm or spark: Choose your real-time weapon." Web page, Dec-2014 [Online]. Available: https://www.infoworld.com/article/2854894/application-development/spark-and-storm-for-real-time-computation.html

[144] K. Khare, "What makes apache flink the best choice for streaming applications?" Web page, Apr-2018 [Online]. Available: https://hackernoon.com/what-makes-apache-flink-the-best- choice-for-streaming-applications-fc377858a53

[145] S. Kozlovski, "Thorough introduction to apache kafka." Web page, Dec-2017 [Online]. Available: https://hackernoon.com/thorough-introduction-to-apache-kafka-6fbf2989bbc1

[146] "Amazon kinesis data streams." Web page [Online]. Available: https://aws.amazon.com/kinesis/data-streams/

[147] "Hortonworks dataflow (hdf)." Web page [Online]. Available: https://hortonworks.com/products/data-platforms/hdf/

[148] "Top 5 apache spark use cases." Web page, Jun-2016 [Online]. Available: https://www.dezyre.com/article/top-5-apache-spark-use-cases/271

[149] Wikipedia, "Kevin ashton." Wikipedia, Aug-2018 [Online]. Available: https://en.wikipedia.org/wiki/Kevin_Ashton

[150] Gartner, "Gartner." Web Page, 2017 [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2017-02-07-gartner-says-8-billion-connected-things-will-be-in-use-in-2017-up-31-percent-from-2016

[151] M. Ligade, "Architecture for iot applications." Web Page, 2016 [Online]. Available: https://medium.com/@maheshwar.ligade/architecture-for-iot-applications-d50ece031d38

[152] E. Ahmed ***et al.***, "The role of big data analytics in internet of things." Paper, 2017 [Online]. Available: https://www.researchgate.net/publication/317617290_The_role_of_big_data_ana

[153] A. Verma, "Internet of things and big data – better together." Web Page, 2018 [Online]. Available: https://www.whizlabs.com/blog/iot-and-big-data

[154] K. Khan, "Future iot and big data." Web Page, 2017 [Online]. Available: https://www.researchgate.net/publication/316890756_Future_IOT_and_Big_data

[155] A. Monnappa, "How big data is powering the internet of things (iot) revolution." Web Page, 2017 [Online]. Available: https://www.simplilearn.com/how-big-data-powering-internet-of-things-iot-revolution-article

[156] A. K. Zainab Alansari Nor Barul and J. Alshaer, "Challenges of internet of things and big data integration." Web Page, 2018 [Online]. Available: https://arxiv.org/ftp/arxiv/papers/1806/1806.08953.pdf

[157] A. Grilo, “Internet of things: An introduction.” Web Page, 2018 [Online]. Available: https://fenix.tecnico.ulisboa.pt/downloadFile/1689468335603255/SER-IoT.pdf

[158] newgenapps, “8 uses, applications, and benefits of industrial iot in manufacturing.” Web Page, 2017 [Online]. Available: https://www.newgenapps.com/blog/8-uses-applications-and-benefits-of-industrial-iot-in-manufacturing

[159] Techtarget, “A guide to healthcare iot possibilities and obstacles.” Web Page, 2017 [Online]. Available: https://searchhealthit.techtarget.com/essentialguide/A-guide-to-healthcare-IoT-possibilities-and-obstacles

[160] I. Innovation, “How iot technology is changing building energy management systems.” Web Page [Online]. Available: https://internet-of-things-innovation.com/insights/the-blog/how-iot-technology-is-changing-building-energy-management-systems/#.W_ZYzuhKhPY

[161] B. Marr, “IoT and big data at caterpillar: How predictive maintenance saves millions of dollars.” Web Page, 2017 [Online]. Available: https://www.forbes.com/sites/bernardmarr/2017/02/07/iot-and-big-data-at-caterpillar-how-predictive-maintenance-saves-millions-of-dollars/#50d19cc97240

[162] IoTONE, “Accelerating the industrial internet of things.” Web Page, 2018 [Online]. Available: https://www.iotone.com/usecases

[163] M. Drummond, “5 great ways airlines are using the internet of things.” Web Page, Aug-2016 [Online]. Available: https://w3.accelya.com/blog/5-great-ways-airlines-are-using-the-internet-of-things

[164] D. Newman, “The iots impact on the future of retail.” Web Page, Feb-2018 [Online]. Available: https://www.forbes.com/sites/danielnewman/2018/02/20/the-iots-impact-on-the-future-of-retail/#711cacb07b1a

[165] G. Christopher, “IoT centred healthcare.” Web Page, Jul-2016 [Online].

Available: https://www.computerworlduk.com/iot/iot-centred-healthcare-system-3643726

[166] F. Online, "The future of iot and big data." Web Page, 2018 [Online]. Available: https://financesonline.com/future-iot-big-data

[167] L. Wang and C. A. Alexander, "Big data in medical applications and health care," ***American Medical Journal***, 2015 [Online]. Available: http://thescipub.com/pdf/10.3844/amjsp.2015.1.8

[168] T. H. subgroup, "Needs, opportunities and challenges of the health sector," Big Data Value Association, 2016 [Online]. Available: http://www.bdva.eu/sites/default/files/Big%20Data%20Technologies%20in%20H

[169] F. Provost and T. Fawcett, "Data science and its relationship to big data and data-driven decision making," ***Published Online***, vol. 1, no. 1, pp. 51–59, 2013 [Online]. Available: https://doi.org/10.1089/big.2013.1508

[170] C. Walter, "Kryder's law," ***Scientific American***, pp. 32–33, Aug. 2005 [Online]. Available: https://www.scientificamerican.com/article/kryders-law/

[171] V. Marx, "Biology: The big challenges of big data," ***International Journal of Science***, 2013 [Online]. Available: http://pic.b.qs1401.com/42548/pdf/bigbioldata_nature13.pdf

[172] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," ***Commun. ACM***, vol. 51, no. 1, pp. 107–113, Jan. 2008 [Online]. Available: http://doi.acm.org/10.1145/1327452.1327492

[173] J. S. Ward and A. Barker, "Undefined by data: A survey of big data definitions," 2013 [Online]. Available: https://arxiv.org/pdf/1309.5821.pdf

[174] B. Wellman and M. Gulia, "Net surfers don't ride alone: Virtual communities as communities," ***Communities and cyberspace***, 1997 [Online]. Available: http://groups.chass.utoronto.ca/netlab/wp-content/uploads/2012/05/Net-Surfers-Dont-Ride-Alone-Virtual-Community-as-Community.pdf

[175] C. L. Lei Wang Jianfeng Zhan and B. Qiu, "BigDataBench: A big data

benchmark suite from internet services." 20th IEEE International Symposium On High Performance Computer Architecture (HPCA-2014), Orlando, Florida, USA, Feb-2014 [Online]. Available: https://arxiv.org/pdf/1401.1406.pdf

[176] S. Rattay, "Profiling algorithms and content targeting - an exploration of the filter bubble phenomenon," Master's thesis, Malmö University, 2014 [Online]. Available: https://muep.mau.se/bitstream/handle/2043/21535/Rattay_Exploration_of_the%2 sequence=2

[177] Wikipedia, "Qlikview." Website [Online]. Available: https://en.wikipedia.org/wiki/Qlik

[178] E. Mathews, "QLIKVIEW architecture and system resource usage." Website [Online]. Available: https://www.quora.com/What-is-QlikView-and-what-is-the-future-of-ones-career-in-QlikView

[179] QlikView, "QLIKVIEW architecture and system resource usage," ***QlikView complete architecture***, 2011.

[180] Edureka, "Understand the power of qlikview's click-visualization." Website [Online]. Available: https://www.edureka.co/blog/qlikview-tutorial/

[181] S. C. Matz and O. Netzer, "Using big data as a window into consumers' psychology," ***Current Opinion in Behavioral Sciences***, vol. 18, pp. 7–12, 2017.

[182] S. Gavaris, "Use of a multiplicative model to estimate catch rate and effort from commercial data," ***Canadian Journal of Fisheries and Aquatic Sciences***, vol. 37, no. 12, pp. 2272–2275, 1980.

[183] T. W. Sidle, "Weaknesses of commercial data base management systems in engineering applications," in ***Design automation, 1980. 17th conference on***, 1980, pp. 57–61.

[184] C. J. Hoofnagle, "Big brother's little helpers: How choicepoint and other commercial data brokers collect and package your data for law enforcement," ***NCJ Int'l L. & Com. Reg.***, vol. 29, p. 595, 2003.

[185] M. Abadi ***et al.***, "Tensorflow: A system for large-scale machine learning."

in ***OSDI***, 2016, vol. 16, pp. 265–283.

[186] Datameer, “Six ways to create better customer behavior analytics.” online, Feb-2018 [Online]. Available: https://www.datameer.com/blog/six-ways-create-better-customer-behavior-analytics/

[187] L. M. Powell ***et al.***, “Field validation of secondary commercial data sources on the retail food outlet environment in the us,” ***Health & place***, vol. 17, no. 5, pp. 1122–1131, 2011.

[188] J. X. Dempsey and L. M. Flint, “Commercial data and national security,” ***Geo. Wash. L. Rev.***, vol. 72, p. 1459, 2003.

[189] P. Covington, J. Adams, and E. Sargin, “Deep neural networks for youtube recommendations,” in ***Proceedings of the 10th acm conference on recommender systems***, 2016, pp. 191–198.

[190] H.-T. Cheng ***et al.***, “Wide & deep learning for recommender systems,” in ***Proceedings of the 1st workshop on deep learning for recommender systems***, 2016, pp. 7–10.

[191] xingangzhongzhinao, “A recommendation system based on tensorflow.” online, Jan-2018 [Online]. Available: https://blog.csdn.net/shebao3333/article/details/78966926

[192] C. M. Hoehner and M. Schootman, “Concordance of commercial data sources for neighborhood-effects studies,” ***Journal of Urban Health***, vol. 87, no. 4, pp. 713–725, 2010.

[193] M. Zanker, M. Jessenitschnig, D. Jannach, and S. Gordea, “Comparing recommendation strategies in a commercial context,” ***IEEE Intelligent Systems***, vol. 22, no. 3, 2007.

[194] R. J. Birk, T. Stanley, G. I. Snyder, T. A. Hennig, M. M. Fladeland, and F. Policelli, “Government programs for research and operational uses of commercial remote sensing data,” ***Remote Sensing of Environment***, vol. 88, nos. 1-2, pp. 3–16, 2003.

[195] J. L. Fei-Fei Li, “Cloud automl: Making ai accessible to every business.”

2018 [Online]. Available: https://www.blog.google/topics/google-cloud/cloud-automl-making-ai-accessible-every-business/

[196] G. C. AI, “Cloud automl-main.” [Online]. Available: https://cloud.google.com/automl/

[197] K. Leetaru, “Google’s cloud automl and its push to democratize point and click ai for all.” 2018 [Online]. Available: https://www.forbes.com/sites/kalevleetaru/2018/01/18/googles-cloud-automl-and-its-push-to-democratize-point-and-click-ai-for-all/#1f4d5bf8c018

[198] Q. V. G. B. Barret Zoph, “Neural architecture search with reinforcement learning.” 2017.

[199] J. Browniee, “A gentle introduction to transfer learning for deep learning.” 2017 [Online]. Available: https://machinelearningmastery.com/transfer-learning-for-deep-learning/

[200] G. C. A. Team, “Training models in the cloud.” [Online]. Available: https://cloud.google.com/ml-engine/docs/tensorflow/training-steps

[201] G. C. Team, “Working with cloud storage.” [Online]. Available: https://cloud.google.com/ml-engine/docs/tensorflow/working-with-cloud-storage

[202] S. M. Patterson, “4 ways google cloud will bring ai, machine learning to the enterprise.” 2017 [Online]. Available: https://www.networkworld.com/article/3179127/cloud-computing/4-ways-google-cloud-will-bring-ai-machine-learning-to-the-enterprise.html

[203] G. Cloud, “Cloud tpu.” [Online]. Available: https://cloud.google.com/tpu/

[204] G. Cloud, “Cloud tpu documentation.” [Online]. Available: https://cloud.google.com/tpu/docs/tpus

[205] AWS, “Amazon rds faqs,” 2018 [Online]. Available: https://aws.amazon.com/rds/faqs/

[206] AWS, “Amazon rds,” 2018 [Online]. Available: https://aws.amazon.com/rds/?p=tile

[207] AWS, “Automated monitoring tools,” 2018 [Online]. Available: https://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/CHAP_Monitoring.

[208] AWS, “Create and connect to a mysql database,” 2018 [Online]. Available: https://aws.amazon.com/getting-started/tutorials/create-mysql-db/

[209] AWS, “CLI,” 2018 [Online]. Available: https://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/USER_DeleteInstar

[210] Web, “Domo,” [Online]. Available: https://developer.domo.com/docs/custom-connectors/connector-dev-studio

[211] Web, “Domo,” [Online]. Available: https://developer.domo.com/docs/dev-studio/dev-studio-quickstart

[212] Web, “Domo,” [Online]. Available: https://developer.domo.com/docs/dev-studio/building-a-responsive-app

[213] Web, “Domo,” [Online]. Available: https://developer.domo.com/docs/authentication/overview-4

[214] Web, “Domo,” [Online]. Available: https://developer.domo.com/docs/dev-studio-references/data-api

[215] Web, “Domo,” [Online]. Available: https://developer.domo.com/docs/dev-studio-references/user-api

[216] IBM, “IBM watson.” Website [Online]. Available: https://www.ibm.com/watson/

[217] I. Watson, “Build with watson.” Website [Online]. Available: https://www.ibm.com/watson/developercloud/

[218] IBM, “Getting started with watson analytics.” Website [Online]. Available: https://community.watsonanalytics.com/wp-content/uploads/2017/11/wa_tutorial-3.pdf?cm_mc_uid=96448751331815393204336&cm_mc_sid_50200000=73367031539

[219] I. Cloud, “AI openscale.” Website [Online]. Available:

https://www.ibm.com/cloud/ai-openscale

[220] I. Watson, “How does ibm watson work.” Youtube [Online]. Available: https://www.youtube.com/watch?v=r7E1TJ1HtM0&t=163s

[221] I. Cloud, “Watson machine learning.” Website [Online]. Available: https://www.ibm.com/cloud/machine-learning

[222] IBM, “IBM watson documentation.” Website [Online]. Available: https://console.bluemix.net/developer/watson/documentation?cm_mc_uid=71151039716715369550371&cm_mc_sid_50200000=18902661539

[223] I. Watson, “IBM watson services.” Website [Online]. Available: https://www.ibm.com/watson/products-services/

[224] I. Watson, “IBM watson health.” Website [Online]. Available: https://www.ibm.com/watson/health/index-1.html

[225] I. Watson, “How it works: IBM watson health.” Yotube Video [Online]. Available: https://www.youtube.com/watch?v=ZPXCF5e1_HI

[226] Wikipedia, “Watson (computer).” Website [Online]. Available: https://en.wikipedia.org/wiki/Watson_(computer)

[227] SAS Institute, ***SAS® Viya® 3.4: Overview***. SAS Institute Inc, 2018.

[228] Randy Guard, “Discovering SAS Viya : Special Collection,” ***SAS Global Users Group Proceedings***, pp. 8–10, 2017.

[229] M. Schneider, “SAS Viya: What It Means for SAS Administration,” ***SAS Global Users Group Proceedings***, 2017.

[230] Jonathan Wexler and Susan Haller and Radhikha Myneni, “An Overview of SAS Visual Data Mining and Machine Learning on SAS Viya,” ***SAS Global Users Group Proceedings***, 2017.

[231] Xiangxiang Meng and Kevin Smith, “I Am Multilingual: A Comparison of the Python, Java, Lua, and REST Interfaces to SAS Viya,” ***SAS Global Users Group Proceedings***, 2017.

[232] SAS, “SAS Platform.” 2017.

[233] J. Pendergrass, “The Architecture of the SAS® Cloud Analytic Services in SAS® Viya™,” ***SAS Global Users Group Proceedings***, pp. 10–18, 2017.

[234] J. Pendergrass, “The Architecture of the SAS® Cloud Analytic Services in SAS® Viya™,” ***SAS Global Users Group Proceedings***, pp. 1–3, 2017.

[235] J. Pendergrass, “The Architecture of the SAS® Cloud Analytic Services in SAS® Viya™,” ***SAS Global Users Group Proceedings***, pp. 4–9, 2017.

[236] SAS, ***SAS® Studio***. SAS Institute Inc, 2018.

[237] SAS, ***SAS® Visual Analytics***. SAS Institute Inc, 2018.

[238] SAS, ***SAS® Visual Statistics***. SAS Institute Inc, 2018.

[239] SAS, ***SAS® Visual Data Mining and Machine Learning***. SAS Institute Inc, 2018.

[240] SAS, ***SAS® Econometrics***. SAS Institute Inc, 2018.

[241] SAS, ***SAS® Visual Forecasting***. SAS Institute Inc, 2018.

[242] Wikipedia, “Natural Language Processing.” Webpage, Nov-2018.

[243] SAS, ***SAS® Visual Text Analytics***. SAS Institute inc, 2018.

[244] SAS, ***SAS® Optimization***. SAS Institute Inc, 2018.

[245] SAS, ***SAS® Viya® 3.4 on Windows: Deployment Guide***. SAS Institute Inc, 2018.

[246] SAS, “Getting ready to install SAS Viya 3.4 on Windows? Then read on.” Webpage, May-2018.

[247] MITKerberos Consortium, “ABOUT - THE MIT KERBEROS CONSORTIUM.” Webpage, 2017.

[248] SAS, ***SAS® Viya® 3.4 on Windows: Deployment Guide***. SAS Institute

Inc, 2018.

[249] SAS, ***SAS® Viya® 3.4 on Windows: Deployment Guide***. SAS Institute Inc, 2018.

[250] “SAS Viya Start Page.” Screenshot, Nov-2018.

[251] “SAS Viya Import Data.” Screenshot, Nov-2018.

[252] “SAS Viya Add Data Object.” Screenshot, Nov-2018.

[253] “SAS Viya Add Variable Roles.” Screenshot, Nov-2018.

[254] “Linear Regression Results.” Screenshot, Nov-2018.

[255] Victor Mayer-Schonberger and Kenneth Cukier, ***Big Data : A Revolution That Will Transform How We Live, Work and Think***. John Murray, 2013.

[256] Shah, “Electronic data capture for registries and clinical trials in orthopaedic surgery: Open source versus commercial systems,” ***Clin Orthop Relat Res***, vol. 10, p. 2664, Jul. 2010.

[257] Tutorialspoint, “Python - matplotlib.” Web Page, Nov-2018.

[258] J. D. Hunter, “Matplotlib: A 2d graphics environment,” ***Computing In Science & Engineering***, vol. 9, no. 3, pp. 90–95, 2007.

[259] CDISC, “Study data tabulation model (sdtm).” Web page, Nov-2018 [Online]. Available: https://www.cdisc.org/standards/foundational/sdtm

[260] C. Mattina, “Bringing drugs to market costs less than perviously thought, study finds,” ***Associated Journal of Managed Care***, Sep. 2017 [Online]. Available: https://www.ajmc.com/newsroom/bringing-drugs-to-market-costs-less-than-previously-thought-study-finds

[261] G. Karthik, “Haberman’s Cancer Survival: Visual Exploratory Data Analysis using Python.” Webpage, Mar-2018 [Online]. Available: https://medium.com/@gokulkarthikk/habermans-cancer-survival-visual-exploratory-data-analysis-using-python-e7dcb7ac01ed

[262] M. D. John Hunter, ***The Architecture of Open Source Applications***, vol. II. lulu.com, 2008.

[263] I. K. Centre, “IBM cognos business intelligence.” Website, 2017 [Online]. Available: https://www.ibm.com/support/knowledgecenter/en/SSEP7J_10.2.2/com.ibm.swg

[264] Wikipedia, “Cognos.” Website, 2007 [Online]. Available: https://en.wikipedia.org/wiki/Cognos

[265] I. K. Center, “IBM cognos architecture.” Website, 2014 [Online]. Available: https://www.ibm.com/support/knowledgecenter/en/SSEP7J_11.0.0/com.ibm.swg

[266] I. Community, “IBM cognos 8 bi.” Website, 2016 [Online]. Available: https://www.ibm.com/developerworks/community/blogs/8d7e4a2b-2364-4719-9f4e-aa9e24db7465/entry/ibm-cognos-bi-suite?lang=en_us

[267] I. Support, “Cognos business intelligence version 10.2 product documentation.” Website, 2014 [Online]. Available: https://www-01.ibm.com/support/docview.wss?uid=swg27024067

[268] I. K. Center, “IBM cognos insight.” Website, 2016 [Online]. Available: https://www.ibm.com/support/knowledgecenter/en/SSEP7J_10.2.2/com.ibm.swg

[269] I. K. Center, “IBM cognos workspace.” Website, 2017 [Online]. Available: https://www.ibm.com/support/knowledgecenter/SSEP7J_10.2.1/com.ibm.swg.ba

[270] I. K. Center, “IBM cognos workspace advanced.” Website, 2017 [Online]. Available: https://www.ibm.com/support/knowledgecenter/en/SSEP7J_10.2.2/com.ibm.swg

[271] I. K. Centre, “Understanding report studio.” Website, 2017 [Online]. Available: https://www.ibm.com/support/knowledgecenter/en/SSRL5J_1.0.1/com.ibm.swg.l

[272] I. K. Center, “Event studio user guide.” Webiste, 2017 [Online]. Available:

https://www.ibm.com/support/knowledgecenter/SSEP7J_11.0.0/com.ibm.swg.ba

[273] BMC, “IBM cognos bi metrics server.” Website, 2015 [Online]. Available: https://docs.bmc.com/docs/display/Configipedia/IBM+Cognos+BI+Metrics+Serv

[274] I. K. Center, “IBM cognos query studio.” Website, 2017 [Online]. Available: https://www.ibm.com/support/knowledgecenter/en/SSEP7J_10.2.2/com.ibm.swg

[275] I. K. Center, “IBM cognos analysis studio.” Website, 2017 [Online]. Available: https://www.ibm.com/support/knowledgecenter/en/SSEP7J_11.0.0/com.ibm.swg

[276] I. K. Centre, “IBM cognos analytics 11.0 documentation.” Website, 2018 [Online]. Available: https://www.ibm.com/support/knowledgecenter/SSEP7J_11.0.0/com.ibm.swg.ba

[277] A. Rosebrock, “Using tesseract ocr with python.” www [Online]. Available: https://www.pyimagesearch.com/2017/07/10/using-tesseract-ocr-python/

[278] L. Eikvil, “OCR - optical character recognition.” 1993 [Online]. Available: https://pdfs.semanticscholar.org/9484/96f9d73cab9c7b4fd5c3b656d1e5b1dc50d3

[279] P. P, “A study on preprocessing techniques for the character recognition,” Nov. 2018 [Online]. Available: https://pdfs.semanticscholar.org/2831/35b2ff5dc1510246ff2f0d3989179738d8f6.

[280] S.-l. developers, “User guide,” ***None*** [Online]. Available: http://scikit-learn.org/stable/user_guide.html

[281] S. Li, E. J. Harner, and D. A. Adjeroh, “Random knn feature selection - a fast and stable alternative to random forests,” ***BMC Bioinformatics***, vol. 12, no. 1, p. 450, Nov. 2011 [Online]. Available: https://doi.org/10.1186/1471-2105-12-450

[282] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, “Naïve bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy,” ***Applied and Environmental Microbiology***, vol. 73, no. 16, pp.

5261–5267, 2007 [Online]. Available: https://aem.asm.org/content/73/16/5261

[283] S. K. Murthy, “Automatic construction of decision trees from data: A multi-disciplinary survey,” ***Data Mining and Knowledge Discovery***, vol. 2, no. 4, pp. 345–389, Dec. 1998 [Online]. Available: https://doi.org/10.1023/A:1009744630224

[284] L. K. Hansen and P. Salamon, “Neural network ensembles,” ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, vol. 12, no. 10, pp. 993–1001, Oct. 1990.

[285] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, “An efficient k-means clustering algorithm: Analysis and implementation,” ***IEEE Transactions on Pattern Analysis & Machine Intelligence***, vol. 24, pp. 881–892, Jul. 2002 [Online]. Available: doi.ieeecomputersociety.org/10.1109/TPAMI.2002.1017616

[286] J. F. Martins, V. F. Pires, and A. J. Pires, “Unsupervised neural-network-based algorithm for an on-line diagnosis of three-phase induction motor stator fault,” ***IEEE Transactions on Industrial Electronics***, vol. 54, no. 1, pp. 259–264, Feb. 2007.

[287] N. Živković, “Real-world machine learning projects with scikit-learn.” Web Page, Aug-2018 [Online]. Available: https://www.packtpub.com/big-data-and-business-intelligence/real-world-machine-learning-projects-scikit-learn-video