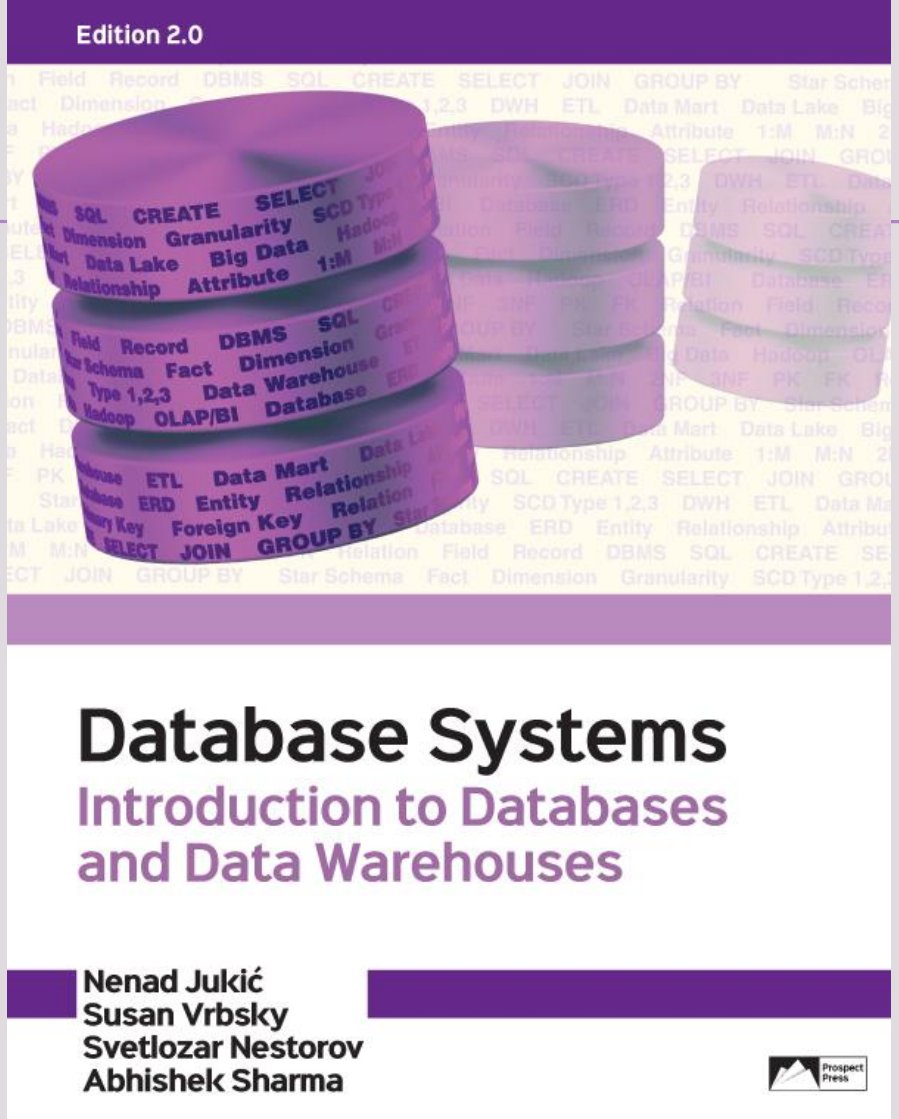
For exclusive use with:



CHAPTER 10 - Big Data and Data Lakes

Copyright (c) 2020 Nenad Jukic and Prospect Press

INTRODUCTION

- Broad categories of **large corporate data sets**:
  - **Operational Database** Data Sets
  - **Data Warehouse** Data Sets
  - **Big Data** Sets

Copyright (c) 2020 Nenad Jukic and Prospect Press

INTRODUCTION

- Examples of sources generating large amounts of data:
  - Smartphones that broadcast their location every few seconds
  - Chips built into cars that perform thousands of diagnostic tests per second
  - Cameras that are continuously recording public and private spaces
  - Radio-frequency identification (RFID) tags that send a wireless signal that is read at various points as the carts of merchandise to which the RFID tags are attached travel in the supply chain from the manufacturer to the customer

Copyright (c) 2020 Nenad Jukic and Prospect Press

SPEAKER NOTES

These are all examples that fit into category of Big Data. Large amounts of data that do not easily fit into operational databases or data warehouses.

BIG DATA

- Big data
  - Data sets in corporations and organizations that contain **massive volumes** of **diverse and rapidly growing data** that, unlike operational databases and data warehouses, are **not formally modeled for querying and retrieval** and are **not accompanied with detailed metadata**

Copyright (c) 2020 Nenad Jukic and Prospect Press

BIG DATA

- Big data
  - Typically **unstructured** or **semi-structured**
  - **Unstructured data**
    - no explicit structure
    - e.g. text document
  - **Semi-structured data**
    - some structure (but not completely structured)
    - e.g. collection of e-mails has a simple structure that distinguishes the sender, recipient, subject, and body of the message, but the message itself is unstructured

Copyright (c) 2020 Nenad Jukic and Prospect Press

Troy University Resources

BIG DATA

- Big data
  - Unstructured data, semi-structured data, low-structure data, and lightly structured data are all various terms that describe data with little or no formal metadata
  - The lack of formal metadata in big data sets is caused by the fact that, unlike operational databases and data warehouses, the repositories of big data are not created through the process of formal database modeling
  - Consequently, big data does not easily fit into tables with rows and columns

Copyright (c) 2020 Nenad Jukic and Prospect Press

V's od BIG DATA

- In the literature, big data has been characterized with a varying number of **V's**
  - Three, four, five, six, or even seven V's: **Volume**, **Variety**, **Velocity**, **Veracity**, **Variability**, **Value**, and **Visualization**
  - While **expressive**, describing big data with **simple V's can also be somewhat confusing and misleading**
  - These V's stated on their own, without context or an example, can be applied to operational databases and data warehouses as well, and therefore are **not completely and exclusively descriptive of big data**

Copyright (c) 2020 Nenad Jukic and Prospect Press

Troy University Resources

V's od BIG DATA

- In the literature, big data has been characterized with a varying number of **V's**
  - **Volume**
    - Large size of data
  - **Variety**
    - Abundance of different types of data sources
  - **Velocity**
    - High speed of incoming data
  - **Veracity**
    - Data quality issues in data

Copyright (c) 2020 Nenad Jukic and Prospect Press

V's od BIG DATA

- In the literature, big data has been characterized with a varying number of **V's**
  - **Variability**
    - Possibility of interpreting the data in different ways
  - **Value**
    - Usefulness and actionability of data
  - **Visualization**
    - Necessity for illustrative and rich visualization of data

Copyright (c) 2020 Nenad Jukic and Prospect Press

Troy University Resources

V's od BIG DATA

- V's of Big Data
  - There are examples of V's that are applicable to operational databases, data warehouses and big data
  - In order to avoid vagueness in describing big data with simple V's, a more precise definition of big data using such V's is needed

Copyright (c) 2020 Nenad Jukic and Prospect Press

SPEAKER NOTES

Examples given on Pages 318-319.

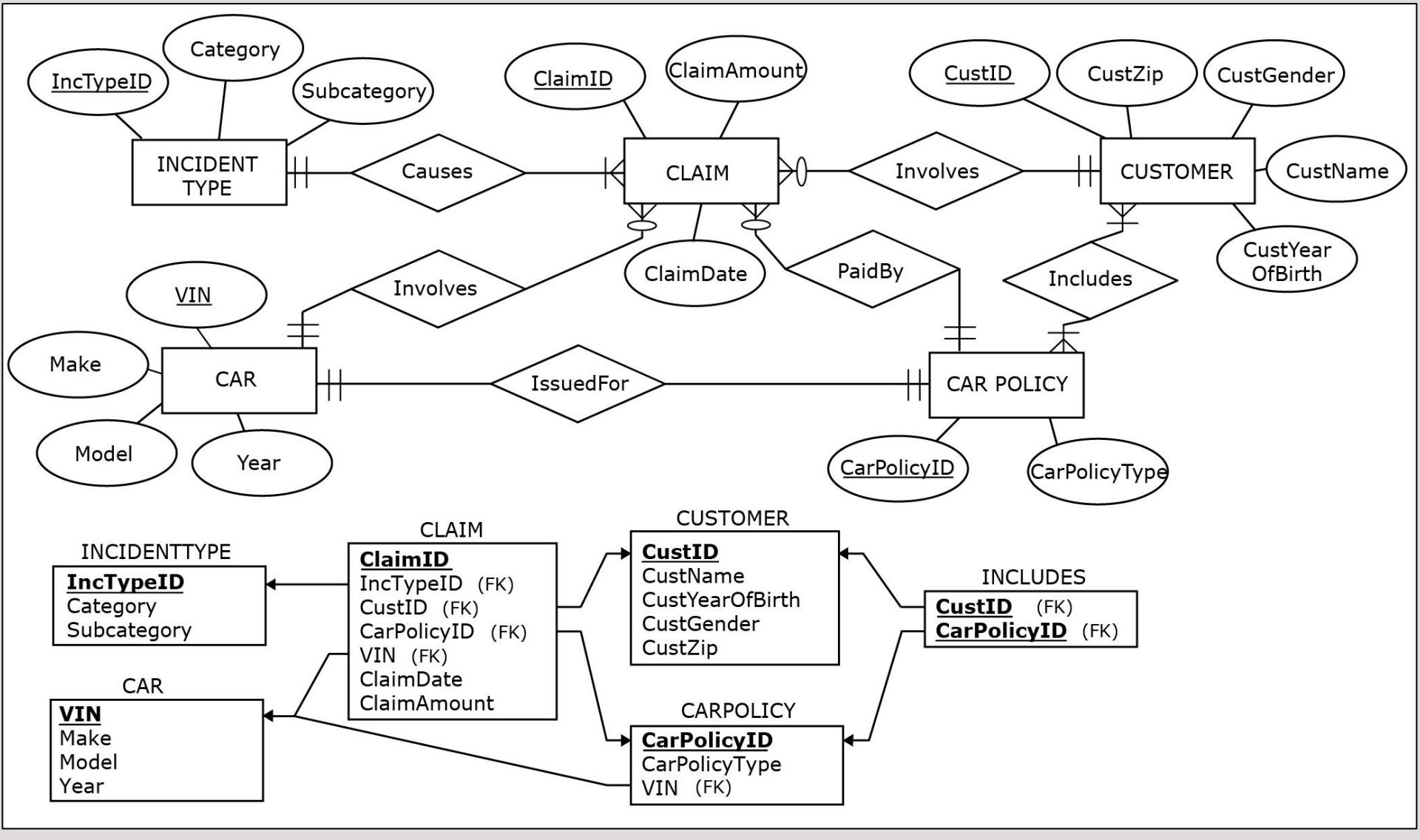
BIG DATA DEFINITION

- Big data refers to the **massive volumes** of **diverse** and **rapidly growing data** sets which **when compared with operational databases and data warehouses**:
  - *are considerably **less structured, with little or no metadata***
  - *in general, exhibit **larger Volume, Velocity,** and **Variety***
  - *have a **higher likelihood of having data quality (Veracity) issues***
  - *have a **higher possibility of different interpretations (Variability)***
  - *need a **more explorative and experimental approach to yield Value**, if any*
  - *are as likely (if not even more so) to **benefit from rich and innovative Visualization***

Copyright (c) 2020 Nenad Jukic and Prospect Press

Troy University Resources

EXAMPLE : Three Types of Data Sets in a Corporation
*Insurance Company **Operational Database***
(ER Diagram and Relational Schema)



Copyright (c) 2020 Nenad Jukic and Prospect Press

SPEAKER NOTES

This example is discussed on Pages 319-322.

EXAMPLE : Three Types of Data Sets in a Corporation

*Insurance Company **Operational Database***

(Sample Data)

CLAIM

ClaimID	IncTypeID	CustID	CarPolicyID	VIN	ClaimDate	ClaimAmount
C-111	IT1	111	CP11	1FM1111	1-Jan-2020	$400
C-222	IT2	333	CP12	JHM2222	2-Jan-2020	$900

INCLUDES

CarPolicyID	CustID
CP11	111
CP12	222
CP12	333

CUSTOMER

CustID	CustName	CustYrOfBirth	CustGender	CustZip
111	Tina	1982	Female	60137
222	Tony	1980	Male	60611
333	Pam	1981	Female	35401

CARPOLICY

CarPolicyID	CarPolicyType	VIN
CP11	Collision	1FM1111
CP12	Full coverage	JHM2222

INCIDENTTYPE

IncType	Category	Subcategory
IT1	Multi-car collision	Fender bender
IT2	Single-car collision	Front pole hit

CAR

VIN	Make	Model	Year
1FM1111	Ford	Escape	2018
JHM2222	Honda	Civic	2019

Copyright (c) 2020 Nenad Jukic and Prospect Press

SPEAKER NOTES

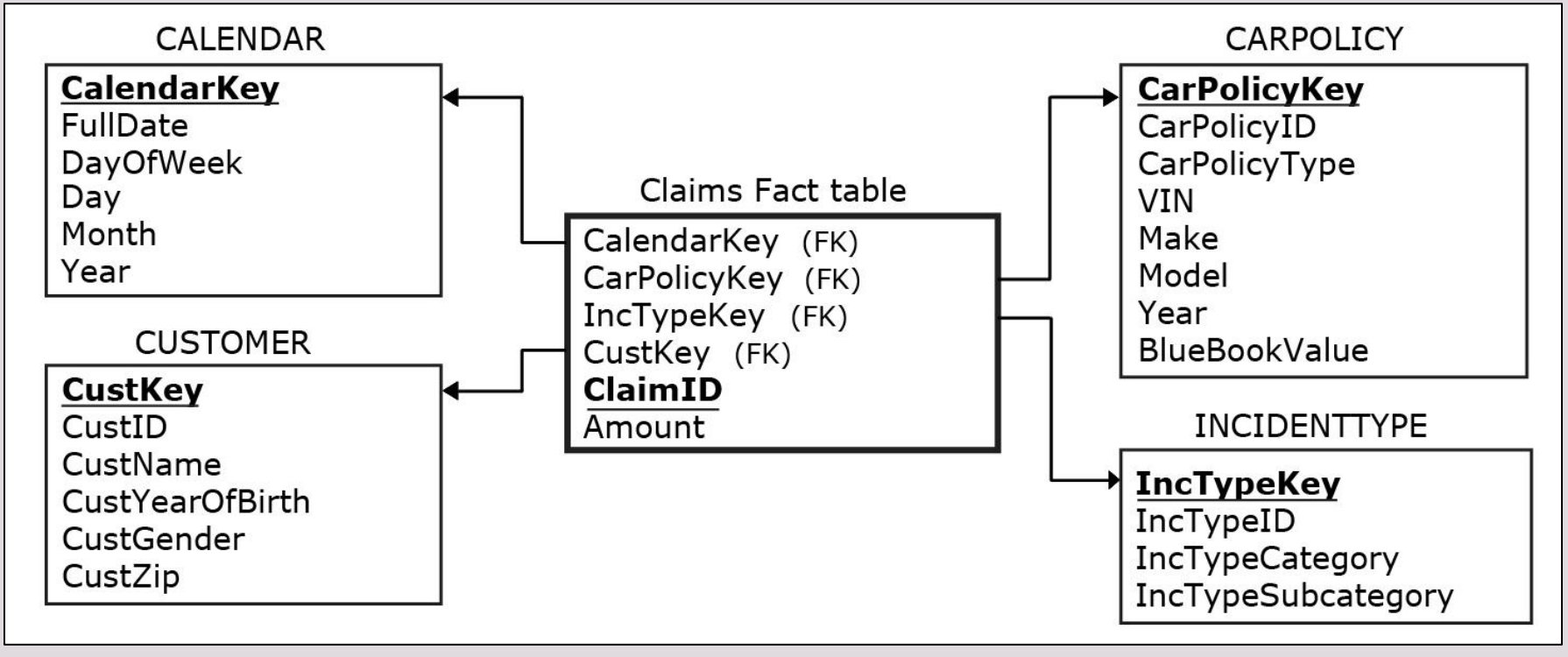
This example is discussed on Pages 319-322.

Troy University Resources

EXAMPLE : Three Types of Data Sets in a Corporation

*Insurance Company **Data Warehouse***

(Star Schema)



Copyright (c) 2020 Nenad Jukic and Prospect Press

SPEAKER NOTES

This example is discussed on Pages 319-322.

EXAMPLE : Three Types of Data Sets in a Corporation

*Insurance Company **Data Warehouse***

(Sample Data)

CALENDAR Dimension

CalendarKey	**FullDate**	**DayOfWeek**	**Day**	**Month**	**Year**
1	1/1/2020	Wednesday	1	January	2020
2	1/2/2020	Thursday	2	January	2020

CARPOLICY Dimension

CarPolicyKey	**CarPolicyID**	**CarPolicyType**	**VIN**	**Make**	**Model**	**Year**	**BlueBook Value**
1	CP11	Collision	1FM1111	Ford	Escape	2018	$16,000
2	CP12	Full coverage	JHM2222	Honda	Civic	2019	$12,000

CUSTOMER Dimension

CustKey	**CustID**	**CustName**	**CustYearOfBirth**	**CustGender**	**CustZip**
1	111	Tina	1982	Female	60137
2	222	Tony	1980	Male	60611
3	333	Pam	1981	Female	35401

INCIDENTTYPE Dimension

IncTypeKey	**IncTypeID**	**IncTypeCategory**	**IncTypeSubcategory**
1	IT1	Multi-car collision	Fender bender
2	IT2	Single-car collision	Front pole hit

CLAIMS Fact table

CalendarKey	**CarPolicyKey**	**IncTypeKey**	**CustKey**	**ClaimID**	**ClaimAmount**
1	1	1	1	C-111	$400
2	2	2	3	C-222	$900

Copyright (c) 2020 Nenad Jukic and Prospect Press

SPEAKER NOTES

This example is discussed on Pages 319-322.

Troy University Resources

EXAMPLE : Three Types of Data Sets in a Corporation

*Insurance Company **Big Data Set***

(Sample Data generated by a Telemetry Device)

```
. . .
1.1.2020; 10:00:00
[(41.879632,-89.064769),0],[(41.879632,-89.064769),0],[(41.879632,-89.064769),0],
[(41.898933, -89.061731),3],[(41.900747,-89.064872),6], . . .
. . .
1.1.2020; 13:25:00
[(41.898239,-87.628423),0],[(41.898239,-87.628423),0],[(41.897311,-87.627119),5],
[(41.897105,-87.626878),12],[(41.896594,-87.626792),18], . . .
. . .
1.1.2020; 18:33:05
[(41.791636,-87.582976),0],[(41.792100,-87.581989),7],[(41.794307,-87.583534),16],
[(41.793188,-87.580057),27],[(41.794083,-87.579628),39], . . .
. . .
```


Copyright (c) 2020 Nenad Jukic and Prospect Press

SPEAKER NOTES

This example is discussed on Pages 319-322.

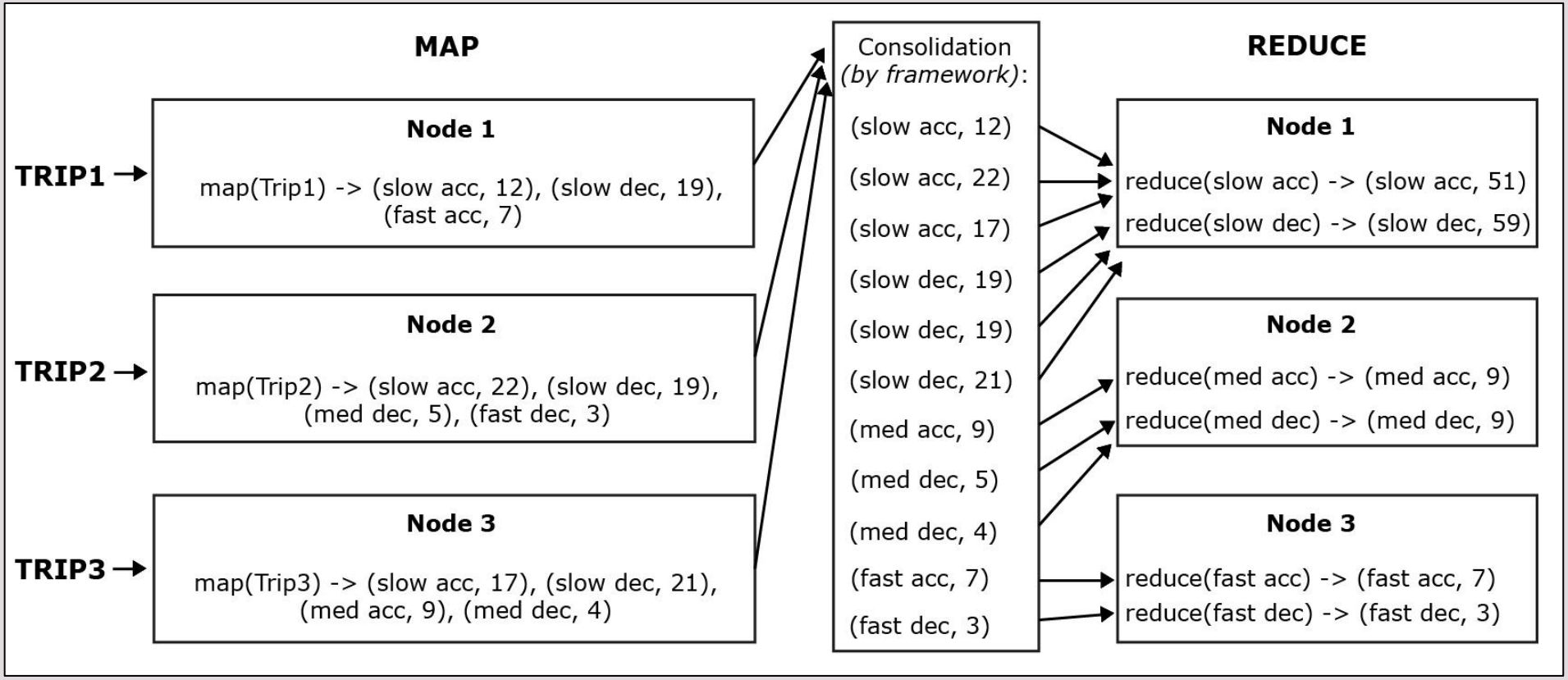
MAPREDUCE AND HADOOP

- Standard database and data warehousing technologies **cannot** adequately
  - Capture the unstructuredness of big data
  - Process big data sets and achieve satisfactory performance handling complex queries over big data
- MapReduce
  - An approach for dealing with big data
    - Map – maps every record into a key-value pair
    - Reduce - gathers all records with the same key and generates a single output record for every key
  - Hadoop
    - Open-source MapReduce implementation

Copyright (c) 2020 Nenad Jukic and Prospect Press

Troy University Resources

EXAMPLE : MapReduce Processing *Insurance Company* ***Big Data Set*** *(Telemetry Data)*



Copyright (c) 2020 Nenad Jukic and Prospect Press

SPEAKER NOTES

This example is discussed on Pages 322-324.

V's od BIG DATA

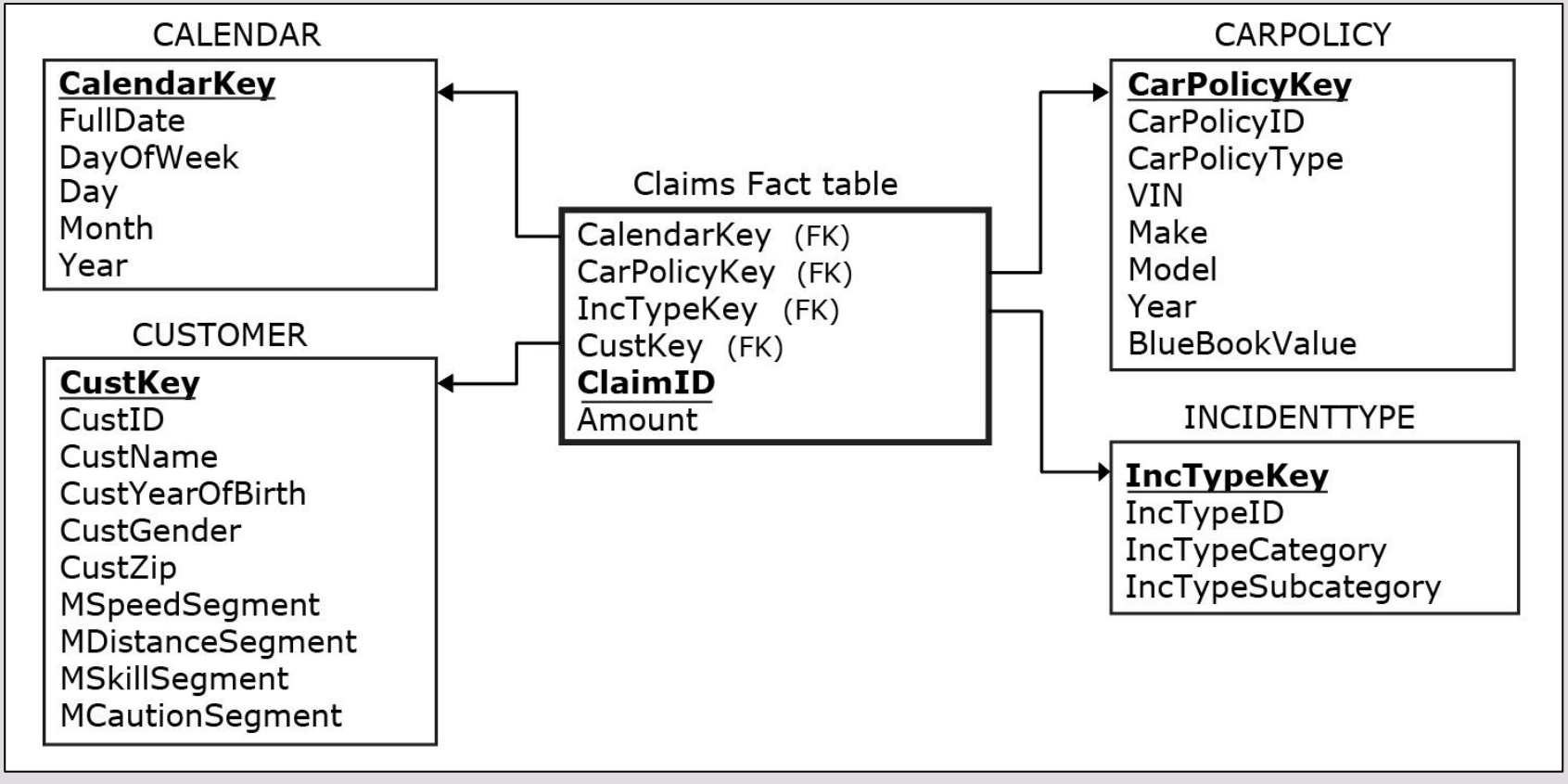
- Example – Examining V's of Big Data
  - **Volume** - consider the volume of telemetry data
  - **Variety** - consider the variety of devices capable of generating big data
  - **Velocity** - consider the velocity of telemetry data
  - **Veracity** - consider the possible data quality issues in telemetry data
  - **Variability** – consider the possibilities of interpreting telemetry data in different ways
  - **Visualization** – consider the possibilities of visualizing and superimposing telemetry data
  - **Value** – consider the value provided by telemetry data

Copyright (c) 2020 Nenad Jukic and Prospect Press

SPEAKER NOTES

This example is discussed on Page 324.

EXAMPLE : *Big Data as a Source for a Data Warehouse (Star Schema)*



Copyright (c) 2020 Nenad Jukic and Prospect Press

SPEAKER NOTES

This example is discussed on Pages 324-326.

EXAMPLE : *Big Data as a Source for a Data Warehouse (Sample Data)*

CALENDAR Dimension

CalendarKey	FullDate	DayOfWeek	Day	Month	Year
1	1/1/2020	Wednesday	1	January	2020
2	1/2/2020	Thursday	2	January	2020

CARPOLICY Dimension

CarPolicyKey	CarPolicyID	CarPolicyType	VIN	Make	Model	Year	BlueBook Value
1	CP11	Collision	1FM1111	Ford	Escape	2018	$16,000
2	CP12	Full coverage	JHM2222	Honda	Civic	2019	$12,000

CUSTOMER Dimension

CustKey	CustID	CustName	CustYearOf Birth	CustGender	CustZip	MSpeed Seg	MDistance Seg	MSkill Seg	MCaution Seg
1	111	Tina	1982	Female	60137	9	9	8	8
2	222	Tony	1980	Male	60611	8	7	6	7
3	333	Pam	1981	Female	35401	7	4	6	7

INCIDENTTYPE Dimension

IncTypeKey	IncTypeID	IncTypeCategory	IncTypeSubcategory
1	IT1	Multi-car collision	Fender bender
2	IT2	Single-car collision	Front pole hit

CLAIMS Fact table

CalendarKey	CarPolicyKey	IncTypeKey	CustKey	ClaimID	ClaimAmount
1	1	1	1	C-111	$400
2	2	2	3	C-222	$900

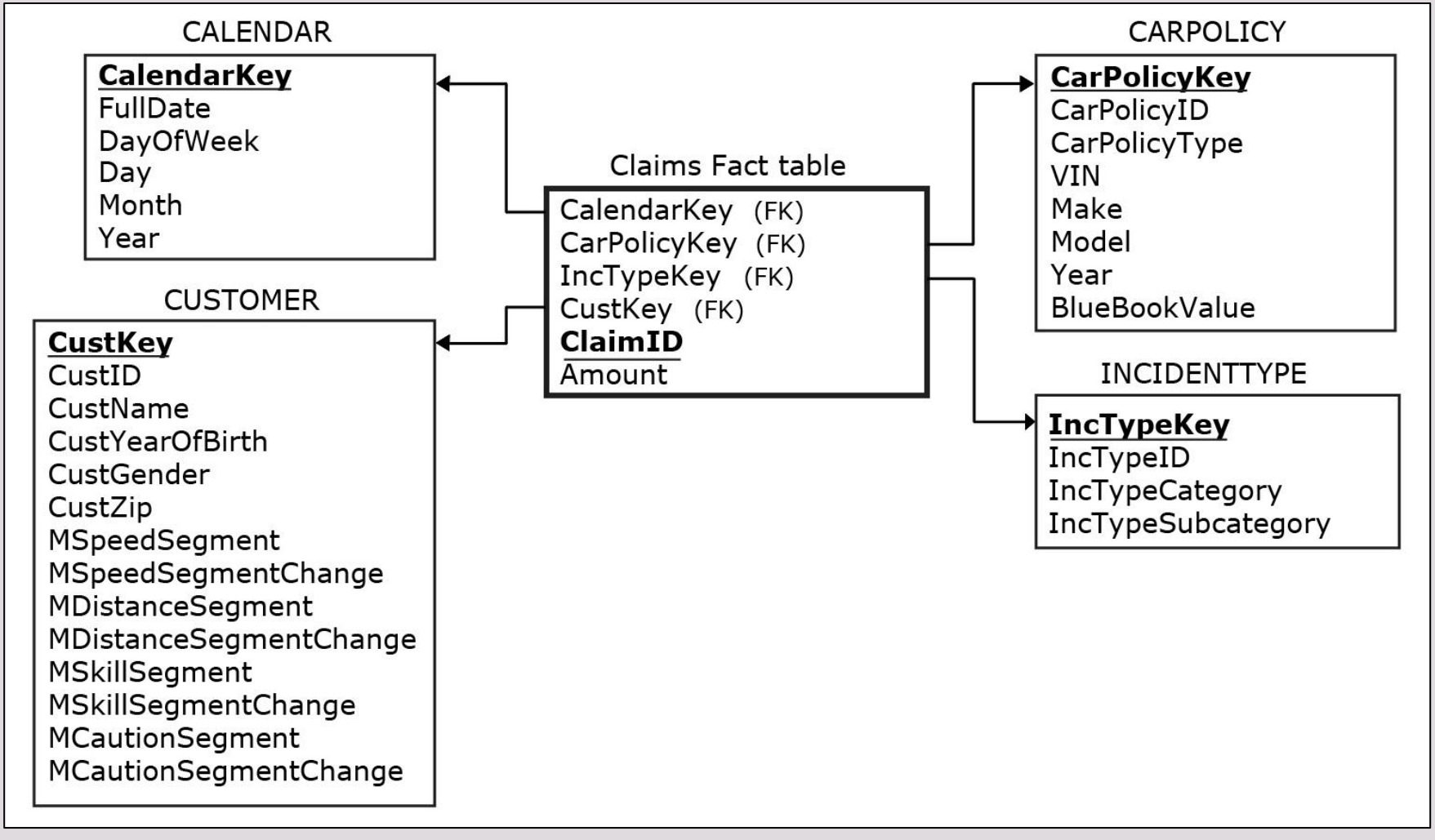
Copyright (c) 2020 Nenad Jukic and Prospect Press

SPEAKER NOTES

This example is discussed on Pages 324-326.

Troy University Resources

EXPANDED EXAMPLE : *Big Data as a Source for a Data Warehouse (Star Schema)*



Copyright (c) 2020 Nenad Jukic and Prospect Press

SPEAKER NOTES

This example is discussed on Pages 324-327.

EXPANDED EXAMPLE : *Big Data as a Source for a Data Warehouse (Sample Data)*

CALENDAR Dimension

CalendarKey	FullDate	DayOfWeek	Day	Month	Year
1	1/1/2020	Wednesday	1	January	2020
2	1/2/2020	Thursday	2	January	2020

CARPOLICY Dimension

CarPolicyKey	CarPolicyID	CarPolicyType	VIN	Make	Model	Year	BlueBook Value
1	CP11	Collision	1FM1111	Ford	Escape	2018	$16,000
2	CP12	Full coverage	JHM2222	Honda	Civic	2019	$12,000

CUSTOMER Dimension

Cust Key	CustID	Cust Name	CustYear OfBirth	Cust Gender	CustZip	MSpeed Seg	MSpeed SegCh	MDist Seg	MDist SegCh	MSkill Seg	MSkill SegCh	MCaut Seg	MCaut SegCh
1	111	Tina	1982	Female	60137	9	+1	9	0	8	-2	8	-3
2	222	Tony	1980	Male	60611	8	0	7	0	6	+1	7	0
3	333	Pam	1981	Female	35401	7	0	4	0	6	0	7	+1

INCIDENTTYPE Dimension

IncTypeKey	IncTypeID	IncTypeCategory	IncTypeSubcategory
1	IT1	Multi-car collision	Fender bender
2	IT2	Single-car collision	Front pole hit

CLAIMS Fact table

CalendarKey	CarPolicyKey	IncTypeKey	CustKey	ClaimID	ClaimAmount
1	1	1	1	C-111	$400
2	2	2	3	C-222	$900

Copyright (c) 2020 Nenad Jukic and Prospect Press

SPEAKER NOTES

This example is discussed on Pages 324-327.

Troy University Resources

CORPORATE USE OF BIG DATA

- Big data methods:
  - ***Increase*** the ability to analyze the data that the organization owns or to which it has access
    - It is a mistake to treat big data as a completely separate and distinct data issue
    - Integrated approach, where a big data initiative goes hand-in-hand with other data projects, is more beneficial
  - ***Do not replace*** database and data warehousing approaches developed for managing and utilizing formally modeled data assets
    - Instead, they allow organizations to analyze and get insight from the kinds of data that are not suited for regular database and data warehouse technologies

Copyright (c) 2020 Nenad Jukic and Prospect Press

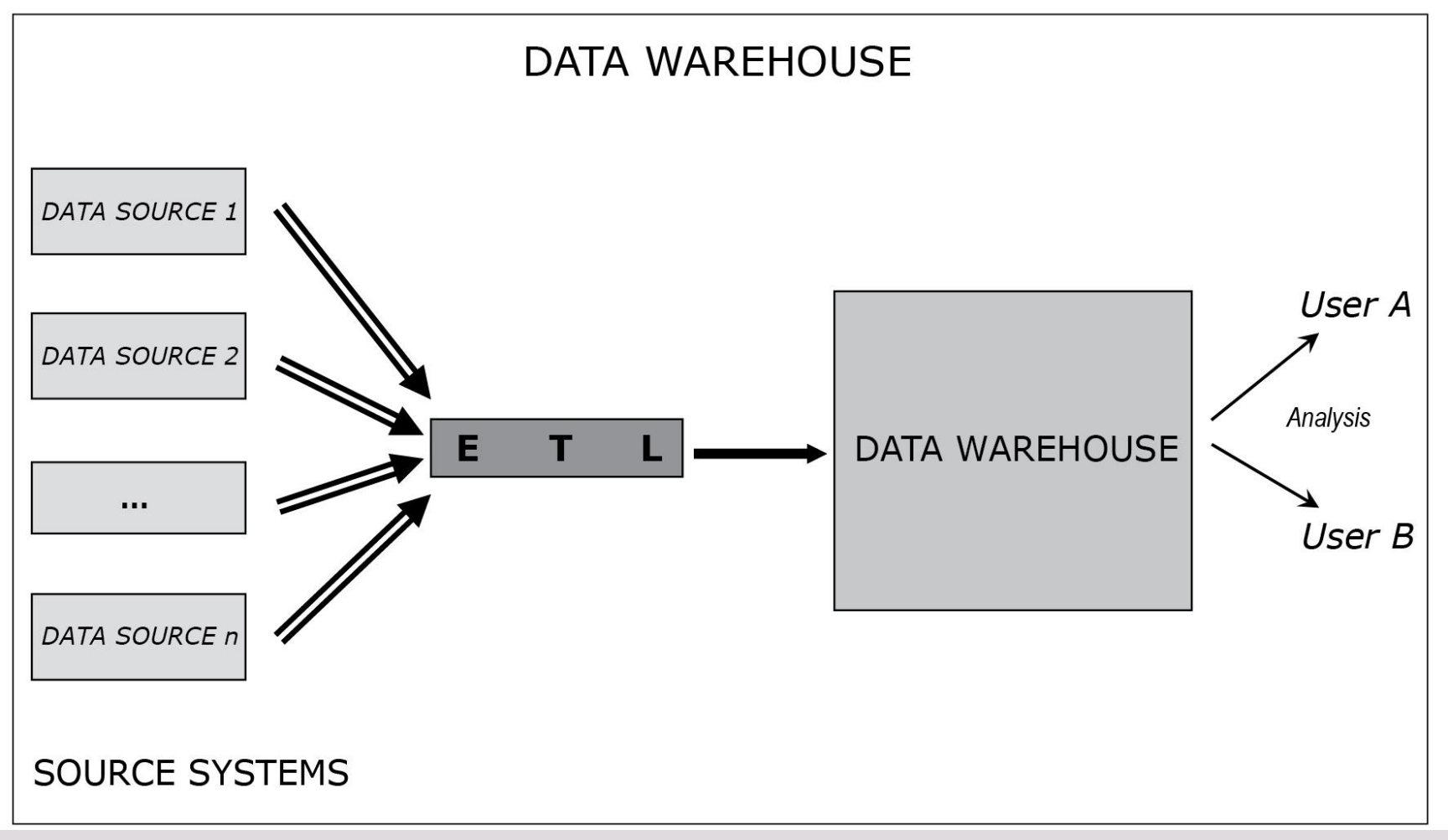
DATA LAKE

- **Data Warehouse**
  - Stores data in a predefined target structure with detailed metadata (e.g., star schemas)
    - Users retrieve and analyze data
- **Data Lake**
  - A large data pool in which the schema and data requirements are not defined until the data is queried
  - Data lake stores vast amount of raw data in its native format until it is needed
    - This approach is often referred to as **schema on read** or **schema on demand**
    - Users have to prepare and transform data, prior to analyzing it

Copyright (c) 2020 Nenad Jukic and Prospect Press

Troy University Resources

DATA LAKE vs DATA WAREHOUSE

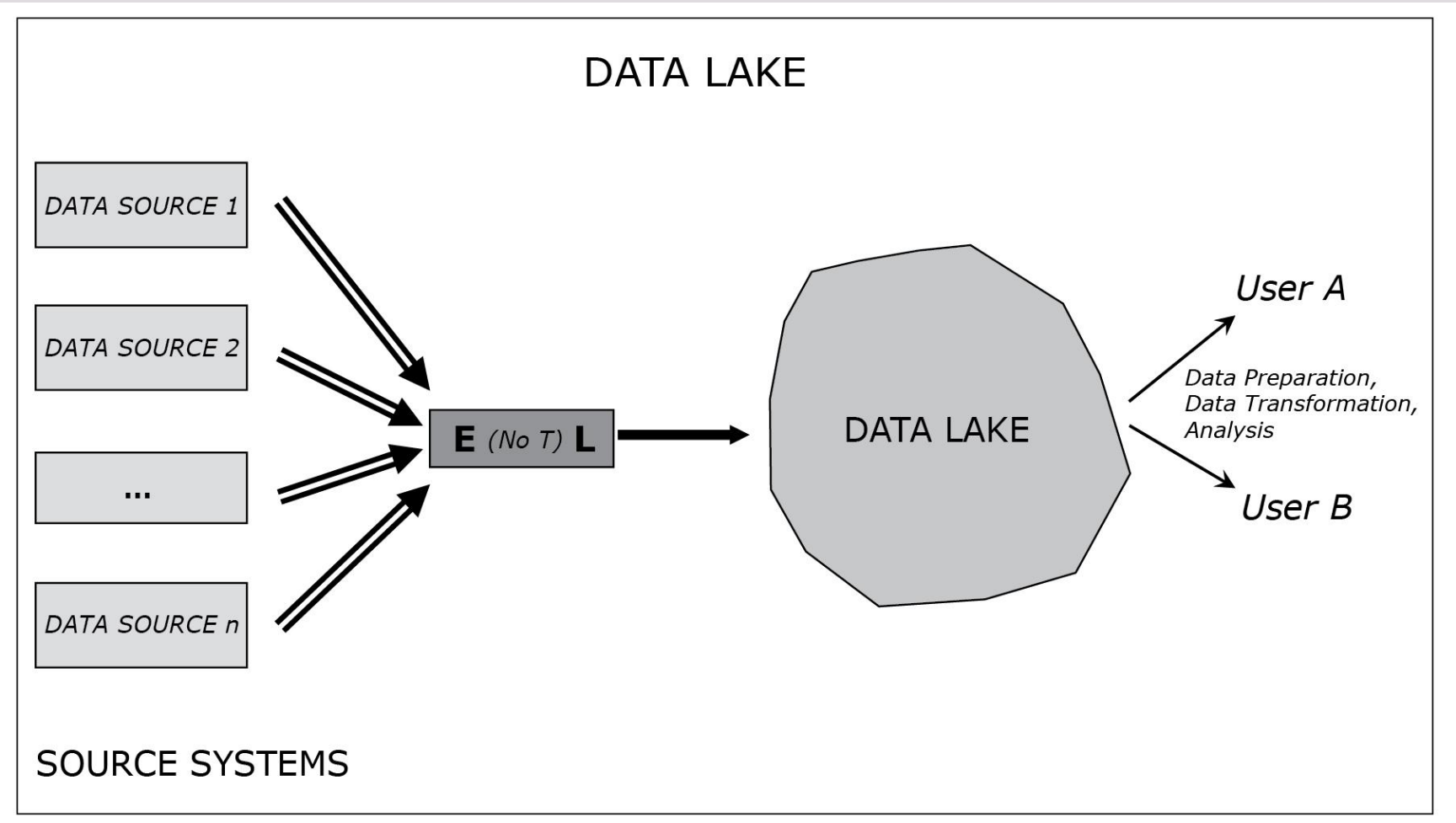


Copyright (c) 2020 Nenad Jukic and Prospect Press

SPEAKER NOTES

This example is discussed on Pages 328-329.

DATA LAKE vs DATA WAREHOUSE



Copyright (c) 2020 Nenad Jukic and Prospect Press

SPEAKER NOTES

This example is discussed on Pages 328-329.

Troy University Resources

DATA LAKE

- A suitable concept for the storage of big data

Copyright (c) 2020 Nenad Jukic and Prospect Press

DATA LAKE

- Is Data Lake a suitable alternative for a data warehouse?
  - OK, if companies *consciously* and *deliberately* want to have a *cheaper* and more flexible approach, while at the same time a *less organized* and *less powerful* approach to the analysis of *some of their structured* data
  - Not OK, if companies take a superficial view of data lakes as an *equivalent* (and yet somehow *magically cheaper*) alternative to data warehouses,
    - In such cases, companies end up with an unwieldy mass of disparate data (i.e. **data jungle** or **data swamp**)
    - This requires analysts to devote most of their efforts to locating and preparing the needed data, while sacrificing the amount of time they can devote to actual analysis

Copyright (c) 2020 Nenad Jukic and Prospect Press

Troy University Resources

DATA LAKE

- Data Warehouse vs Data Lake
  - Analogy: agriculture vs hunting/gathering

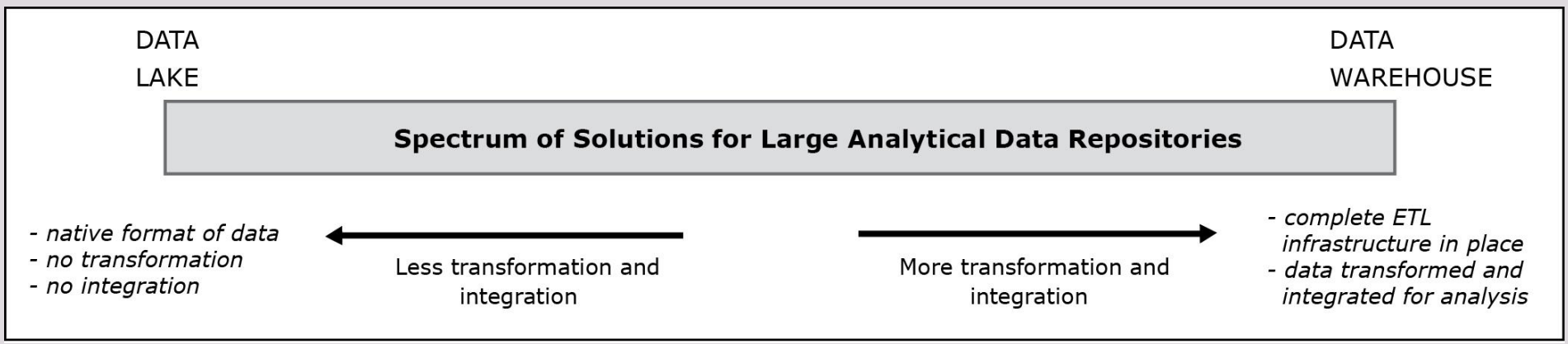
Copyright (c) 2020 Nenad Jukic and Prospect Press

SPEAKER NOTES

This example is discussed on Page 329.

DATA LAKE

- Spectrum of Solutions for Large Analytical Data Repositories



Copyright (c) 2020 Nenad Jukic and Prospect Press

SPEAKER NOTES

Fully developed data warehouse and a data lake are at the opposite ends of a broad spectrum of solutions for large analytical data repositories.

DATA LAKE

- Spectrum of Solutions for Large Analytical Data Repositories
  - A fully developed data warehouse and a data lake are at the opposite ends of a broad spectrum of solutions for large analytical data repositories
  - An example inside the spectrum
    - Quick and simple transformations of source data to merge data into a standardized form for ad-hoc use, without necessarily having elaborate star schemas and surrogate keys
  - In practice, within this spectrum, it is common to hear the terms *"data warehouse"* or *"data lake"* used loosely (by people not adhering to proper definitions of these terms) to describe solutions that do not necessarily refer to the ends of the spectrum

Copyright (c) 2020 Nenad Jukic and Prospect Press

SPEAKER NOTES

This example is discussed on Page 329.

DATA LAKE

- Spectrum of Solutions for Large Analytical Data Repositories
  - Allows for solutions that combine the use of data warehouses and data lakes
  - Example:
    - The most commonly used analytical data is kept in a data warehouse
    - As needed, this data can be combined with the data from a data lake for more complex and explorative analytical solutions

Copyright (c) 2020 Nenad Jukic and Prospect Press

SPEAKER NOTES

This example is discussed on Pages 329-330.