Institut de Recherche
en Informatique de Toulouse
CNRS - INP - UT3 - UT1 - UT2J

Pyramid

Data Engineering

Evolution of Data Management Systems:

Fundamental Concepts, Methods and Applications

Emerit. Prof. Abdelkader Hameurlain

hameurlain@irit.fr

Informatics Research Institute of Toulouse IRIT

Pyramid Team*

Paul Sabatier University PSU
Toulouse , France

*** Query Processing & Optimization in Parallel & Large-scale Distributed Environments**

1. Introduction (1/2) : Main Problems of Data Management [Sto 98, Ozsu 11, …]

"Data needs to be: <Captured, Cleaned, Stored, Queried, Processed and Turned in Knowledge>"

- **Data Modelling & Semantic**
- **Query Processing & Optimization (OLAP)**
- **Concurrency Control/Transactions (OLTP)**
- **Replication & Caching**
- **Cost Models**
- **Security & Privacy**
- **Monitoring Services**
- **Resource Discovery**
- **Autonomic Data Management (self-tuning, self-repairing, …), …**
- **…**

➡ Data Management Systems DMS

1. Introduction (2/2) : Evolution of Data Management Systems [Gra 96]

➧ *"The present without past has not future"* **Fernand Braudel**

▶ <Concept ➔ Systems: Objective> [Ham 13]

- ....
- File Management Systems FMS: *Storage Device Independence*
- Uni-processor DB Systems DBMS [Codd 70]: *Prog-Data Independence*
- Parallel DBMS [Dew 92, Val 93]: *High Perf., Scalable & Data Availability*
- Distributed DBMS [Ozs 11]: *Transparency of Location, Frag., Replication*
- Data Integration Systems [Wie 92]: *Uniform Access to Data Sources*

  Characteristics =<Distribution, *Heterogeneity, Autonomy*>
- Data Grid Systems [Fos 04]: *Sharing of Available Resources*
- Mobile Database Systems : *Decentralized Control & Scalability*
- Cloud Data Mana. Systems [Aba 09, Sto 10]: *Economic Models*

  Characteristics *=<Elasticity, Fault-Tolerant >*

➧ **Evolution or Crossroad ?**

Evolution of Data Management Systems

I. From File Mana. Systems FMS to Database MS DBMS

- Motivations, Objectives, Files Organizations & Drawbacks
- Databases & Rel. DBMS: Motivations & Objectives

II. Parallel Relational DBMS

- Motivations Objectives, Characteristics and Challenges
- Parallel Query Processing
- Optimization of Data Communications: Plague of Parallelism

III. From Distributed DBMS to Data Integration Systems DIS

- Motivations , Objectives & Designing of Distributed DB
- Distributed Query Processing & Soft. Architecture
- Mediator-Wrappers Architecture & Query Processing Methodologies

IV. Cloud Data Management Systems CDMS

- Motivations, Objectives & Main Characteristics of CDMS
- Classification of CDMSs : 3 Generations (G1, G2 & G3)
- Advantages & Weakness of MR Systems & Parallel DBMSs
- Comparison between Parallel DBMSs & MR Systems

V. Conclusion & References

I.1. File Management Systems (1/2)

- **File Concept**

  ➡ ***Program and Storage Device <u>Independence</u>***

  [Storage] <File> [Program/Application]

  ▶ **Software Eng. Requirements**

- **File Organizations: 4 types**
  - **< Sequential /Indexed > Organizations**
  - **< Hashing/Relative> Organizations**

I.2. File Management Systems (2/2)

Access Methods AM

- Sequential AM
- Key AM :=<Indexed/Hashing> AM

Drawbacks of FMS

- Data description must be done in each program
- Relationships/Links between files are materialized (➔ New files)

➔ **Database Concept**

I.3. Database and DBMS (1/2)

Concept of Database DB: Motivations

- Separation between Data Structures (DB Schema) and Program
- Prog-Data Independence = <Physical & Logical> Independence

Fundamental Objectives of a DB

- Separation of Data Description and Data Manipulation
- Data Independence: Logical & Physical
- Procedural & Declarative Interfaces/Languages
- Query Processing and Optimization
- Data Integrity/Sharing/Privacy/Security
- Easy Data Administration
- …

I.4. Database and DBMS (2/2)

- **Database Management System DBMS** [Del 80, Date 86, Mir 02, Ull 89]
  - Software allowing users to interact with a DB
  - Implementation of main objectives of a DB

- **Main Functions/Tools of DBMS**
  - Data Description ➔ DDL (Data Models : Concept. , Logical, Phys.)
  - Data Manipulation ➔ DML (Querying and Updating)
  - Data Integrity/Sharing (Transaction & Concurrency)/Security
  - Data Administration, ...
  - ....

  ➡ DB Design, Languages, and Methods (Query Processing, Transaction & Concurrency Control, Integrity, Security, Administration).

- **DB Models:** <Hierarchical, Network, Relational & Object>

I.5. Relational DB and Relational DBMS [Codd 70] (1/3)

■ Main Characteristics of Rel. DB

- **Structured Data**: Relation Concept to describe <Entities & Links>

  ➔ Data Model Definition
- Stored Data on Disk ➡ Input/Output Management
- Relational Algebra: Commutative, Internal Law
- From Procedural to Declarative Languages: SQL [Cham76], QUEL [Sto 76], QBE [Zlo77], ...

▶ The System will find the (near) Optimal Access Path

➡ **Optimizer** [Sel 79, Wong 76, Gan 92, …]

I.6. Relational DBMS: Query Optimization [Sel 79] (2/3)

- **Problem Position [Gan 92]:**

  q ∈ Query , p ∈ {Execution Plans}, Cost_p (q):

  - Find p calculating q such as Cost_p (q) is minimum
  - Objective : Find the best trade-off between

    Min (Response Time) & Min (Optimization Cost)

- **Optimizer Structure= < St, Sp, C> [Gan 92]**
  - St: Search Strategies (→ Intelligence)
    - <Physical Optim., Parallelization, Resource Allocation, …>
  - Sp: Search Space (→ Control)
    - Data Structures/Queries: Linear Spaces, Bushy Space
    - Type/Nature of Queries
  - C: Cost Models (→ Knowledge)
    - <Metrics, System Environment Description>

I.7. Limitations of Uni-proc. Query Optimization Methods wrt Decision Support Systems /OLAP (RDBMS) (3/3)

- **Complex Queries***: Number of Joins >6*
- **Size of Research Space [Tan 91]:** *Very Large (e.g. 2^{N-1})*
- **Optimization Cost [Lan 91]:** *can be very expansive (e.g. Deterministic Strategies)*
- **Optimal Execution Plan:** *not guaranteed (e.g. Randomized Strategies)*

➡ *Requirements in:* **High Performance HP & Resource Availability**

➡ **Introducing a New Dimension:** ***Parallelism***

▶ **Parallel Relational Database Systems [Dew 92]**