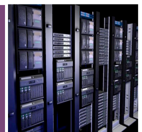
Data-Management for Data-intensive Computing



CMPSC 274: Advance Topics on Database System

Divyakant Agrawal
Department of Computer Science
University of California at Santa Barbara

Prelude

2

- Data management is at the cross-roads:
  - New models for data-intensive computing
  - Significant turmoil in terms of technological advances
  - Rapid changes have presented the data management research community with unprecedented challenges.
- Inevitable to re-examine the context in which data management evolved in **the past**.
- In the same vein, we need to explore the role of data management in **the future**.
- Course Objective: Comprehensive understanding of data management and data analysis paradigms.

Data-intensive Computing

3

- Storage and retrieval management of persistent data.
- Large-scale data analysis for data-centric decision making.

Course Organization

4

- First Half of the Course (4 weeks):
  - Persistent Stores for Enterprise Applications
- Second half of the Course (2 Weeks):
  - Persistent Stores for Internet and Web-scale Applications

SPRING'2011: CMPSC 274 4/4/11

Course Organization

5

- Second half of the course (2 Weeks):
  - Enterprise-class Solutions for Large-scale Data Analysis
- Second half of the course (2 Weeks):
  - Internet and Web-scale Solutions for Large-scale Data Analysis

SPRING'2011: CMPSC 274 4/4/11

Course Grading

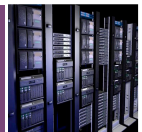
6

- First half of the course (assignment/ assessment based):
  - Home-works
  - Mid-term Exams
  - Text book
- Second half of the course:
  - Project based
  - Large programming/implementation project (2 person)

SPRING'2011: CMPSC 274 4/4/11

+

Data Management Overview



+

Historical Perspective

8

- Advent of computer technology:
  - Persistent storage of data and information
  - Value of data/information realized very early especially in the context of business entities
- Early efforts in the industry:
  - Effective data management solutions
  - Based on storing data:
    - Files: logical abstraction
    - Tapes: physical realization
- File based data management
  - Problems in accessing data
  - Problems in processing data
  - In general, problems in effective usage of data

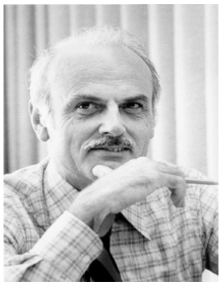
+

Historical Perspective

9

- Emergence of alternative storage models:
  - Departure from file-based storage
  - New data models to enable data access based on its attributes
  - New language models to enable effective manipulation of data
- Data models (circa 1965):
  - Network model: essentially to model business entities using the information paradigm
  - Hierarchical model: another variant
- Standardization efforts:
  - Economies-of-scale/minimize duplication
  - CODASYL

E. F. Codd



Codd'69: Relational Data Model

- Relational Data Model:
  - Tabular framework for data representation
  - Intuitive and easy to comprehend
  - Complete theoretical framework
- Relational algebra (operational framework):
  - An algebraic framework for operating on relational data
  - Well-defined algebraic operators
- Relational calculus (theoretical framework):
  - First-order logic
  - Declarative querying framework
  - Equivalent to relational algebra

Relational Data Model

- RDBMS model became extremely successful.
- Logical Data Model:
  - Intuitive
  - Well-defined
  - Design time considerations need not focus on physical issues
- Physical storage independence:
  - Run-time system maintains the access methods
  - Dynamic mapping from logical to physical level
- Declarative Query Interface:
  - Users did not need to be expert programmers

Data Management Evolution

13

- RDBMS became highly successful:
  - Widely adopted by both large and small business entities
- Enterprises became increasingly reliant on databases
- Primarily used for day-to-day operations:
  - Banking operations
  - Retail operations
  - Travel industry

Data Management Evolution

14

- Typically:
  - Database modeled the state of the enterprise
  - Client operations were applied to update the state.

Batched Transaction Processing

15

- Nomenclature:
  - Transaction Processing Systems
- Typical usage:
  - Spool client transactions during the day
  - During the night, spooled transactions applied to the database state of previous night
  - New database state becomes available for the next working day
- Advantages:
  - Almost up-to-date information on the finger-tips
  - Failure-recovery is in-built in the paradigm
  - No issues of concurrency

Batched Transaction Processing

16

- Problem:
  - Database state did not reflect up-to-date information
- Impact on daily operations:
  - Some amount of guess-work in formulating the application state of client transactions:
    - seat availability on a flight
    - funds availability in a bank account
    - inventory information for re-order
- Impact on batch update:
  - Transaction **failures** if the "guess" is wrong

SPRING'2011: CMPSC 274 4/4/11

The OLTP Paradigm

17

- On-line Transaction Processing:
  - Database state is up-to-date at all times
- Significant challenges:
  - Multiple users/clients need to be supported
  - Handle hardware and software failures
- Emergence of what is now commonly referred to as:
  - The Transaction Concept:
    - Concurrency
    - Failures

SPRING'2011: CMPSC 274 4/4/11

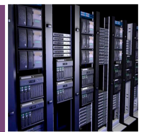
Concurrency & Failures: A Quick Preview

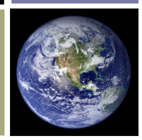
18

- List maintenance:
  - lookup/find operations in O(log N) time.
  - Read-only operations:
    - Concurrency does not cause any difficulties.
- List updates:
  - Inserts/deletes also in O(log N) time if executed sequentially.
  - What if I specify that operations are arbitrarily interleaved?
  - Worse yet: what happens if the updaters can fail?
  - Can you do it safely? Do you have the necessary tools to solve this problem?

SPRING'2011: CMPSC 274 4/4/11

+ Data Analysis Overview





Data Analysis → Business Intelligence

+ Hans Peter Luhn

21

50 Years of Business Intelligence

22

- Vision of Business Intelligence:
  - Hans Peter Luhn in a 1958 article.
  - Predates the notions of Databases and Data Management.
- A pioneer in Information Sciences:
  - New use of the term ***thesaurus***
  - Automatic creation of literature abstracts
  - 16 digit Luhn's number widely used for credit cards and other banking instruments
  - …

SPRING'2011: CMPSC 274 4/4/11

Luhn's Vision

23

- Defined BI as:

  "… provides means for selective dissemination to each of its action points in accordance with their current requirements or desires."
- Key technologies:
  - Auto-abstracting of documents,
  - Auto-encoding of documents, and
  - Auto creation and updating of profiles
- Breadth of the vision:

  "… *business* is a collection of activities carried on … be it **science, technology, commerce, industry, law, government, defense**, et cetra."

  "… *intelligence* is also defined … as the ability to **apprehend the interrelationships** of presented **facts** in such a way as **to guide action** towards a desired goal."

SPRING'2011: CMPSC 274 4/4/11

The intervening years

24

SPRING'2011: CMPSC 274 4/4/11

The Early Years (1970s-1980s)

25

- Contrary to Luhn's overarching vision – early efforts on business information remained focused on ***database management technology***.
- With the advent of the relational model:
  - DBMS technology became pervasive and matured.
  - Widely adapted by most enterprises.
  - Online Transaction Processing became a proven paradigm for business operations.
- Consequence:
  - Massive proliferation of OLTP systems especially within a single enterprise.
  - Data-driven decision making became a norm.
  - Disparate reporting from multiple operational data sources.

SPRING'2011: CMPSC 274 4/4/11

Notion of "Data Analytics" (1990s)

26

- Presence of multiple operational systems created a ***fractured*** view of an enterprise.
- Devlin & Murphy introduced the term ***business data warehouse*** in 1988:
  - A unified view of the enterprise primarily for integrated reporting.
- Catalysts:
  - Demand for reporting – key factors being PCs and spread-sheets.
  - Market potential – Teradata, Red-brick Systems, etc.
- Negative factors:
  - Unproven, immature, and expensive technology proposition.
  - Distinction between DBMS and DW: no clarity, ***?duplication?***
  - Fairly laborious and time-consuming data integration process
  - No clear stake-holders → ***2nd Class Entity*** often resulting in adversarial atmosphere.

SPRING'2011: CMPSC 274 4/4/11

Data Warehousing: Current State

27

- Keys to success:
  - Enormous contribution of DW evangelist Ralph Kimball
  - STAR schema & Dimensional model for DW: intuitive and scalable
  - No compromise on the autonomy of operational data sources
- Persisting head-winds:
  - Since does not directly contribute to P&L:
    - ROI question still persists.
  - Not a plug & play technology:
    - Very high consulting costs.
  - Legacy of significant time and cost over-runs of most data warehousing projects.
  - Batch-oriented DW Architecture:
    - Deemed too costly just for integrated reporting.
    - Needed intuitive analytical capabilities.

SPRING'2011: CMPSC 274 4/4/11

+ Hither "Business Intelligence" (2000-)

28

- Gray et al. [1996] introduced the CUBE operator for roll-up and drill-down analysis of multi-dimensional data (i.e., DW Model).
- DW enterprises (Hyperion, Cognos, Analysis Services, etc.) adapted the CUBE architecture and called it:
  - ***business intelligence.***
- Problem:
  - Early BI (CUBE) technology had serious issues of scaling → only accentuated the ill-repute of DW/BI technologies
  - Underlying problem: exponential explosion of data storage

SPRING'2011: CMPSC 274 4/4/11

+ Business Intelligence: Current State

29

- While the BI/Cube technology was still evolving – the spin doctors needed to undo the early damage.
- Hence, perhaps the term **Real-time Business Intelligence** – to convey the "criticality" of such technology to business leaders.
- Current debate: what exactly is meant by "real-time" in Business Intelligence?
  - In 2006, in this workshop, Donovan Schneider – gave numerous examples of "degree of timeliness" for a variety of analysis tasks.
  - My personal view is that the correct term should have been: ***Online Business Intelligence***.
- Assuming that – redefine the DW/BI architecture to support RTBI.

SPRING'2011: CMPSC 274 4/4/11

+ The present & the future

30

SPRING'2011: CMPSC 274 4/4/11

+ Real-time Business Intelligence: Required?

31

- Anecdotal evidence from Sam Walton

Airplane & Parking Lot Story

• Demonstrates the power of 10,000 feet view (from the airplane) versus the local view (from the parking lot).

• Numerous cases where "timeliness" of "intelligence" is extremely valuable.

➔ The case of RTBI is very-well justified.
➔ The question however is at what cost?

SPRING'2011: CMPSC 274 4/4/11

+ Concluding Remarks

32

- Data Management:
  - Will study the models, paradigms, theory, and algorithms needed for Enterprise Scale Data Management (& application development)
  - Will then examine the disruption that has occurred with the Internet and Web-based application:
    - underlying factors for this disruption and
    - Proposed solution

SPRING'2011: CMPSC 274 4/4/11

+ Concluding Remarks

33

- Data Analysis:
  - Will study the well-accepted principles, architecture, and solutions for enterprise class Data Analysis platforms.
  - Will then explore the disruption caused by Internet and Web-scale Applications.
- Time permitting:
  - Multi-core processors
  - GPU platforms and databases
  - Data stream processing

SPRING'2011: CMPSC 274 4/4/11