IS5102
Database Management Systems
Lecture 15: Data Warehouses

Alexander Konovalov
alexander.konovalov@st-andrews.ac.uk

(with thanks to Susmit Sarkar)

2021

Objectives

- Relational Model and Analytics
- Data Warehouse and Dimensionality Modeling
- The rise of the term NoSQL
- Different non-relational data management systems
- Graph databases
- Document-oriented databases

The Need for Rethinking the Relational Model

- Growing amounts of data in **operational databases**
  (RDBMS has been around since late 1970s)
- Not directly suited for **decision support**
- Typically: numerous operational systems with overlapping and sometimes contradictory definitions

Decision making require access to data from multiple sources

- Multiple queries to individual sources?
- Do they have historic data?

Need a **data warehouse** to store data acquired from multiple sources, under a unified schema, in a single repository

Definition

A subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management decision-making process (Inmon, 1993).

Subject-oriented Data

Organized around the major **subjects** of the enterprise (e.g. customers, products, and sales)

Rather than major **application areas** (e.g. customer invoicing, stock control, and product sales).

Integrated application-oriented data from different source systems,

. . . often including **inconsistent** data

Must be made consistent to present a unified view of the data

Time-variant Data

Data in the warehouse is **only** accurate and valid at **some point in time** or over some time interval.

Time-variance is also shown in the extended time that the data is held, the implicit or explicit association of time with all data, and the fact that the data represents a series of snapshots

Data in the warehouse is not normally updated in real-time (RT) but is refreshed from operational systems on a regular basis. (However, emerging trend is towards RT or near RT DWs).

New data is always added as a **supplement** to the database, rather than a replacement.

- DBMS
- Data loaders

Data loaders retrieve information from various data sources

DBMS use used by various tools to perform queries and analyse information

When and how to gather data?

- **source-driven architecture**: data sources submit new information, continuously or periodically
- **destination-driven architecture**: data periodically requested by the warehouse

Which schema to use?

- integrate data to a warehouse schema
- may involve **data cleansing**, **deduplication** (also known as **merge & purge**), **householding**, changing units of measurements, etc.
- warehouse as materialised view

Data Warehouse Design (2)

- How to propagate updates (view maintenance)
- Which data to summarise?

ETL: extract – transform – load model

ELT: extract – load – transform model

For decision-support queries:

Query profile generated for each user, group of users, or the data warehouse

. . . based on information that describes the **characteristics** of the queries such as

- frequency,
- target table(s),
- and size of results set

Dimensionality modeling

A logical design technique that aims to present the data in a standard, intuitive form that allows for high-performance access

Every dimensional model (DM) is composed of one table with a composite primary key, called the **fact table**, and a set of smaller tables called **dimension tables**

Attributes in the fact table classified as **measure attributes** or **dimension attributes**

Multidimensional data: data that can be modelled using dimension and measure attributes

Each dimension table has a simple (non-composite) primary key that corresponds exactly to one of the components of the composite key in the fact table.

Forms “star-like” structure, which is called a star schema or star join.

Dimensionality modeling – Star Schemas

Star schema is a logical structure that has a **fact table** (containing factual data) in the center, surrounded by **denormalized dimension tables** (containing reference data).

Facts are generated by events that occurred in the past, and are unlikely to change, regardless of how they are analysed

Star schemas can be used to speed up query performance by denormalizing reference information into a single dimension table.

Complex data warehouse designs may have more than one fact table, and also multiple levels of dimension tables, leading to **snowflake schemas**

Dimensionality modeling – Surrogate Keys

All natural keys are replaced with **surrogate keys**

Means that every join between fact and dimension tables is based on surrogate keys, not natural keys

Surrogate keys allows the data in the warehouse to have some **independence** from the data used and produced by the OLTP (Online Transactional Processing) systems.

Data lake - a repository where data can be stored in multiple formats

No up-front efforts to preprocess data ...

... at a cost of more efforts and flexibility needed when creating queries

Chapter 11, *Database System Concepts*, Silbershatz, Korth and Sudarshan

Chapter 33, *Database Systems*, Connolly and Begg