Data Mining and Warehousing

DATA 604

Leanne Wu

lewu@ucalgary.ca

Department of Computer Science

UNIVERSITY OF CALGARY

Data mining

- Recall that **knowledge** is different from data or information
  - information which is merged with conceptual models of the world

- The goal of data mining is to take data or information, and transform it into knowledge
  - What are patterns or rules in your data which might actually indicate knowledge you've collected about the world?
  - **Inductive**: Patterns and rules are based upon the data you have (does not create new facts)
- Broader terms: Data analytics, business intelligence, knowledge Discovery

Common techniques

- Clustering (unsupervised learning)
  - For example: k-means
  - Partition a population into groups such that each group shares common characteristics
- Classification (supervised learning)
  - Given a population and a set of events involving that population, divide that population into a hierarchy of classes
  - Examples: Twitter topics, Google ads settings
- Sequential or time-series based patterns
  - Look for events which always appear in sequence (for example, transit delays)
  - Look for events which always appear over a period of time (stocks, climate patterns)

Association Rule Mining

Techniques drawn from artificial intelligence

- Regression
  - specific kind of classification
  - tries to identify classification rules of a specific format (instead of identifying a class, attempts to predict a specific value for a variable)
- Neural Networks
  - Used to deduce a value for a new set of inputs given a set of pre-existing inputs
  - Inspired by how neurons organize computation in the brain
- Genetic algorithms
  - Used to search large, varied spaces by representing a problem as an smaller encoding, then trying different variations of the encoding
  - Trials are done by evaluating which variations did best, then trying variations of the variations

Data Warehousing

- How do we organize data so that it is easy to apply analytical techniques to it?
- Consider:
  - Performance
  - Accuracy
  - Relatively few users
  - Read-only (or read-heavy)
  - Queries are known in advance (and performed on a regular basis)

OLTP vs OLAP

OLAP (Online Analytical Processing)

- Read-only (modifications are rare)
- Complex queries performed on a regular basis
- Interested in information and knowledge

OLTP (Online Transactional Processing)

- Frequent modifications
- Simple queries and transactions
- Records data

Typical operations for data warehousing

- Selecting
- Sorting
- Aggregation (Also called rolling-up or drilling-up)
- Disaggregation (Also called drilling-down)
- Deriving attributes (applying formulae)
- Pivoting (performing cross-tabulations)

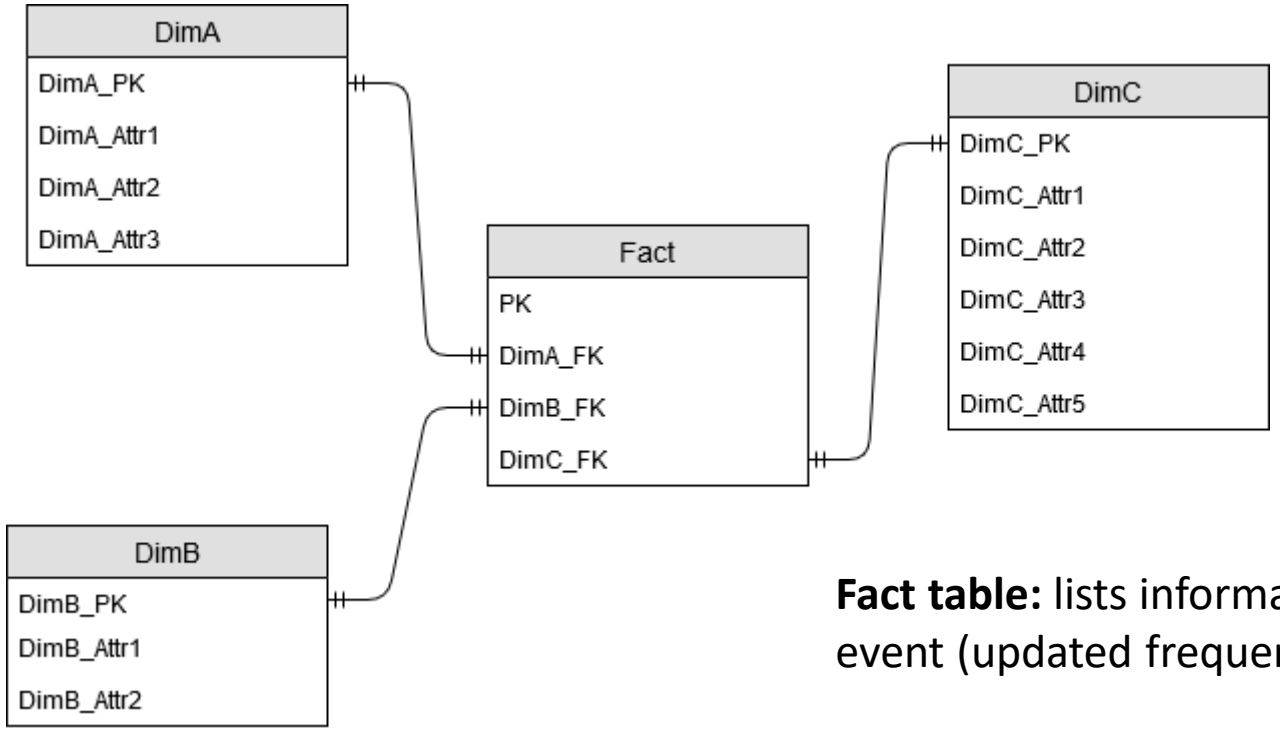
Creating a warehouse: Extract, Transform, Load

Logical design for OLAP

- Consider the workload
  - Fewer reads (SELECTs)
  - Considerably fewer writes (INSERT/UPDATE/DELETEs)
- Databases will likely be robust even when denormalized
- What is the best way to link together as much information as possible about a given event?

Star Schema

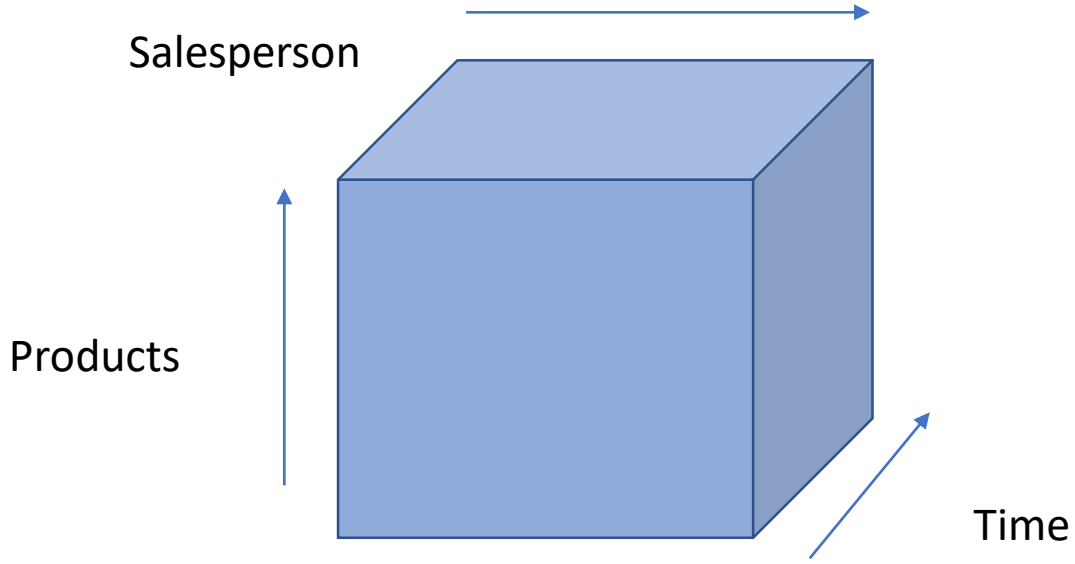
Dimension table: lists information which are relevant to data stored in the fact table but which rarely change



Fact table: lists information about a recorded event (updated frequently)

Other possible logical designs

- Snowflake schema: Similar to star schema, but each dimension table may be further normalized
- Constellations: Multiple fact tables share dimensional tables
- Cubes: Tables which are organized along multiple dimensions



Adapting Data Warehouses for Modern Organizations

Federated Databases

- Incorporate databases from many sources (potentially many organizations)
- Focus is on how you exchange data between these sources
  - Keeping semantics consistent is challenging!

Data Marts

- Focus on a small component part of an organization
- Allows you to minimize the dimensionality of the data

Master Data Management

- Deals with the governance, processes, etc. of the most important pieces of data within an organization
- Master data is typically extremely high quality with rigorous guards around how it can be modified or used

Data Warehousing: Challenges

- Data quality
- Tracking the evolution of schemata
  - For example:
    - format of information (7 digit phone numbers to 10 digit phone numbers)
    - gender vs sex
    - Impact of integrating new data sources, mergers, acquisitions, shutdown of organizations
- Understanding where your data came from (provenance)
- Data retention (when is data no longer useful?)
- Ad-hoc and real-time reporting