Relational Model for Data Warehouse

Tanzim Hossain
Lecturer
Department of Software Engineering
Daffodil International University

Star Schema

Star Schema

- When using a relational database for a data warehouse, a new data modeling technique is needed to represent multidimensional data.
- A **star schema** is a data modeling representation of multidimensional data cubes.

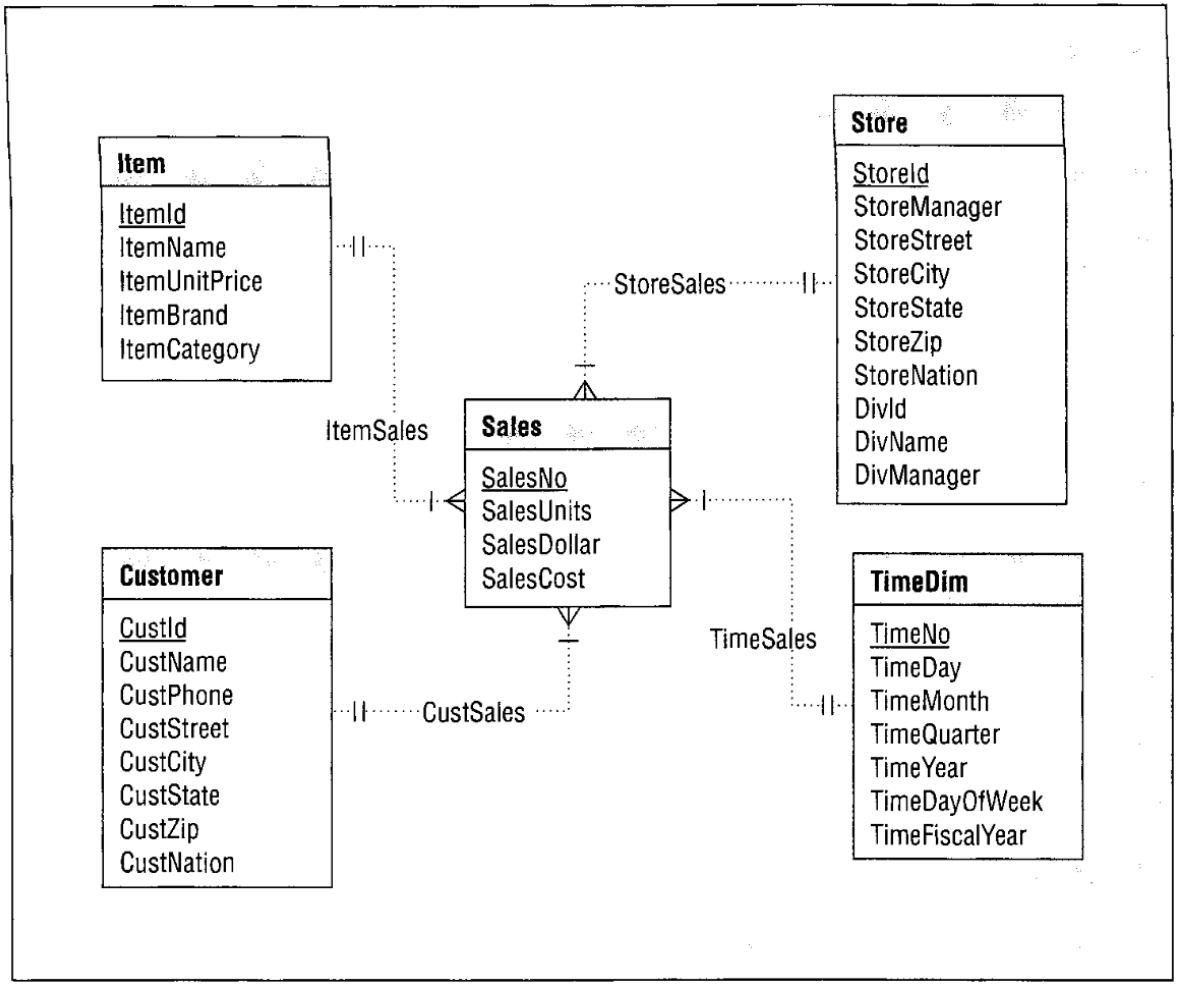
Star Schema

- In a relational database, a star schema diagram looks like a star with one large central table, called the **fact table**, at the center of the star that is linked to multiple **dimension tables** in a radial manner.

Fact Table & Dimensions

- The fact table stores numeric data (facts), such as sales results, while the dimension tables store descriptive data corresponding to individual dimensions of the data cube such as product, location, and time.
- There is a **1-M relationship** from each dimension table to the fact table.

Star Schema Example



Star Schema Example

- This ERD consists of four **dimensions** entity types, **Item**, **Customer**, **Store**, and **Time**, along with one **fact** entity type called **Sales**.
- When converted to a table design, the Sales table has foreign keys to each dimension table (Item, Customer, Store, and TimeDim).

Time Rperesentation

- For dimension tables, time representation involves the level of historical integrity, an issue for updates to dimension tables.
- When a dimension row is updated, related fact table rows are no longer historically accurate.
- For example, if the city column of a customer row changes, the related sales rows are no longer historically accurate.
- To preserve historical integrity, the related sales rows should point to an older version of the customer row.
- Kimball (April 1996) presents three alternatives for historical integrity:

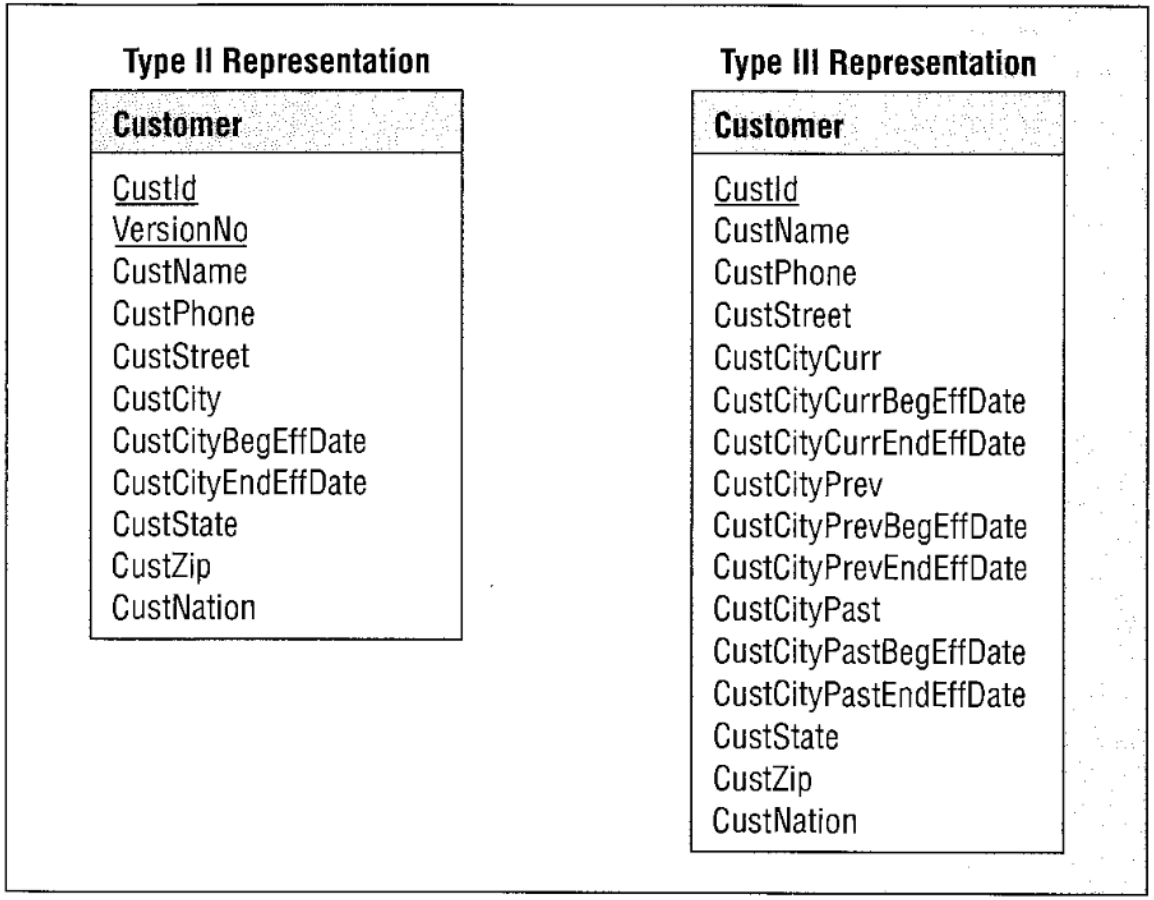
Type I & Type II

- **Type I:** overwrite old values with the changed data. This method provides no historical integrity.
- **Type II:** use a version number to augment the primary key of a dimension table.
- For each change to a dimension row, insert a row in the dimension table with a larger version number.
- For example, to handle the change to the city column, there is a new row in the Customer table with the same customer number but a larger version number than the previous row.

Type III

- **Type III:** use additional columns to maintain a fixed history.
- For example, to maintain a history of the current city and the two previous city changes, three city columns (Cust-CityCurr, CustCityPrev, CustCityPast) can be stored in the Customer table along with associated six date columns (two date columns per historical value column) to record the effective dates.

Time Rperesentation



Type II Vs. Type III

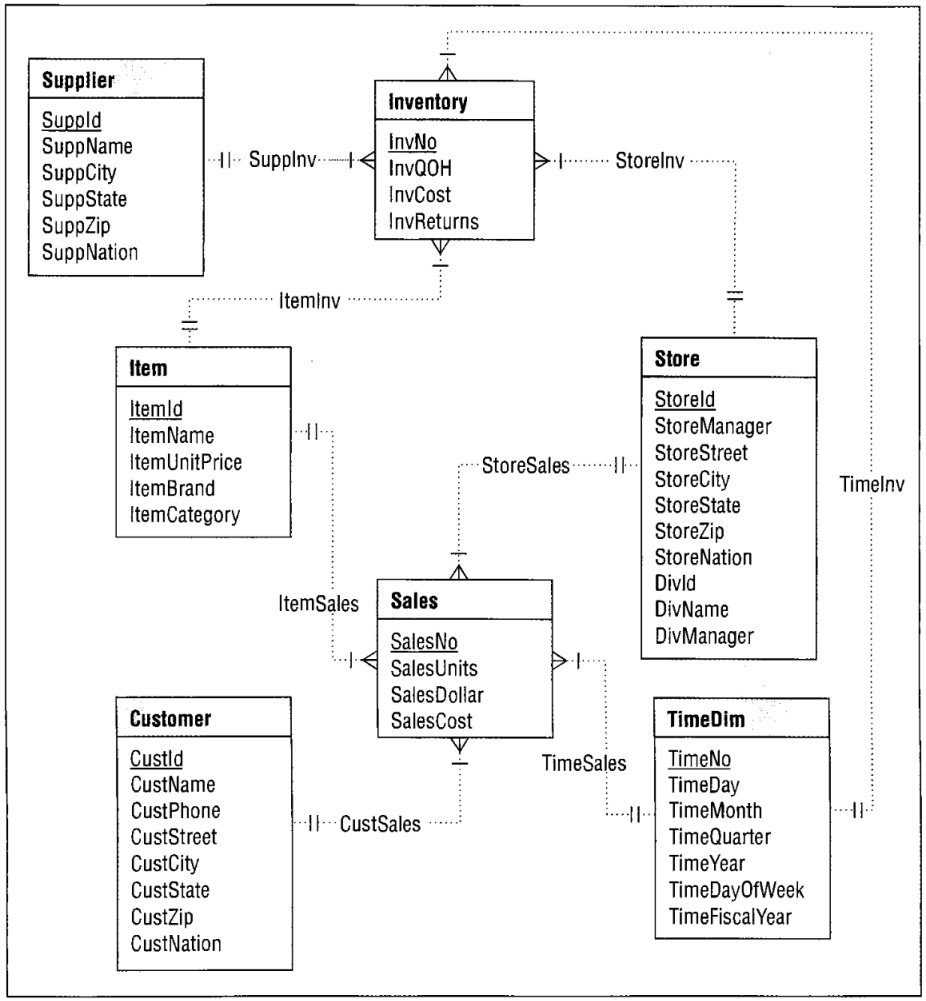
- Type II alternative involves multiple rows for the same customer, but the entire history is represented.
- The Type III alternative involves just a single row for each customer, but only a limited history can be represented.

Constellation Schema

Constellation Schema

- The star schema in represents only a single business process for sales tracking.
- Additional star schemas may be required for other processes such as shipping and purchasing.
- For related business processes that share some of the dimension tables, a **star schema** can be extended into a **constellation schema** with multiple fact entity types,

Constellation Schema



Constellation Schema

- When converted to a table design, the Inventory entity type becomes a fact table and 1-M relationships become foreign keys in the fact table.
- The Inventory entity type adds a number of measures including the quantity on hand of an item, the cost of an item, and the quantity returned.
- All dimension tables are shared among both fact tables except for the Supplier and Customer tables.

Normalization

- Fact tables are usually normalized while dimension tables are often not in third normal form.
- For example, the Store entity type is not in 3NF because DivId determines DivName and DivManager.
- Normalizing dimension tables to avoid storage anomalies is generally not necessary because they are usually stable and small.

Normalization

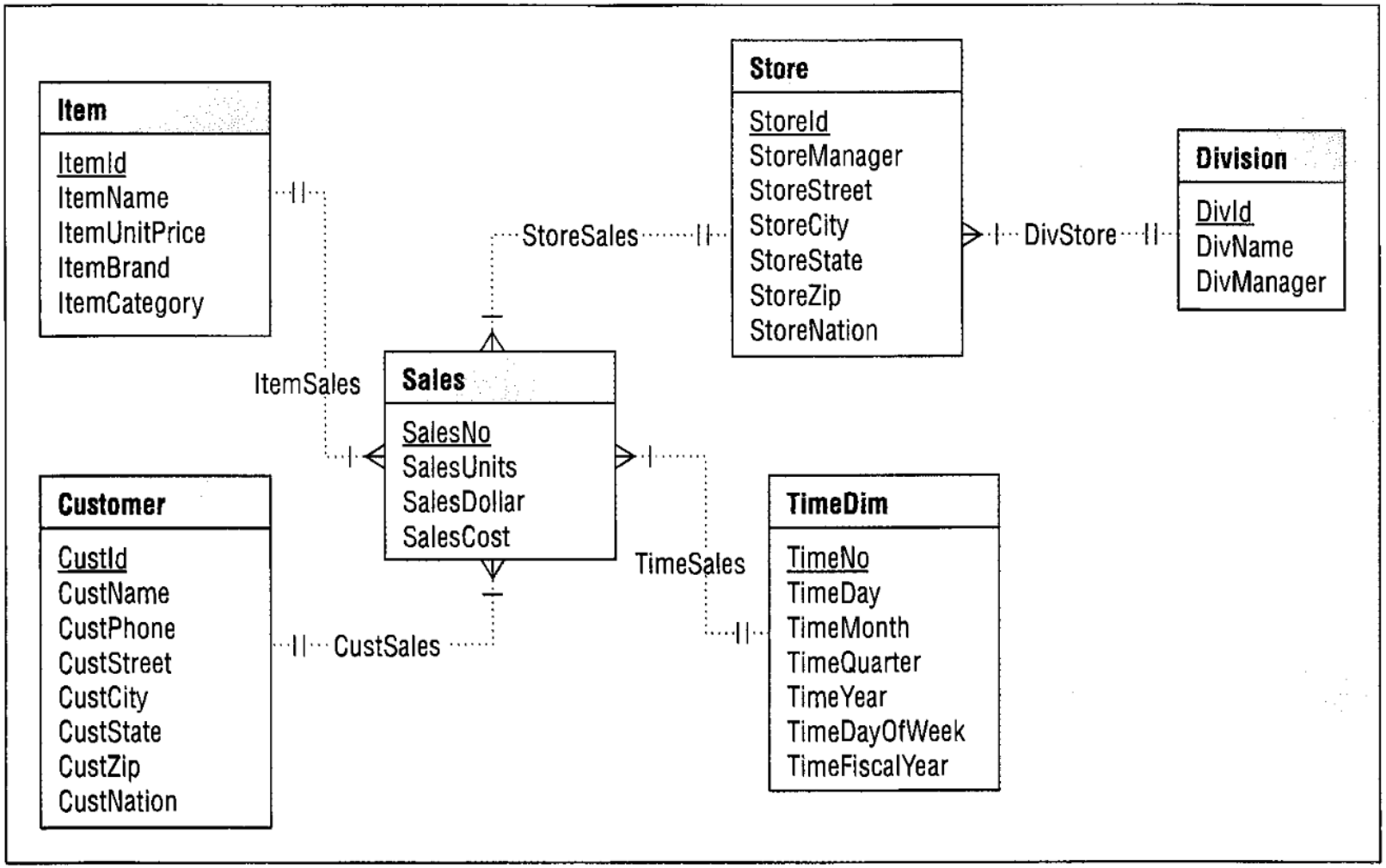
- The nature of a data warehouse indicates that dimension tables should be designed for retrieval, not update.
- Retrieval performance is improved by eliminating the join operations that would be needed to combine fully normalized dimension tables.

Snowflake Schema

Snowflake Schema

- When the dimension tables are small, denormalization provides only a small gain in retrieval performance.
- Thus, it is common to see small dimension tables normalized.
- This variation is known as the **snowflake schema** because multiple levels of dimension tables surround the fact table.
- For the Customer and Item tables, full normalization may not be a good idea because these tables can contain many rows.

Snowflake Schema



Storage Engines

MOLAP

- Originally, vendors of decision support software developed a storage architecture that directly manipulates data cubes.
- This storage architecture, known as MOLAP for Multidimensional OLAP, was the only choice as a storage technology for data warehouses until the mid-1990s.
- At the current time, MOLAP has been eclipsed as the primary storage architecture for data warehouses.
- But it still is an important technology for summary data cubes and small data warehouses and data marts.

MOLAP

- MOLAP storage engines directly manipulate stored data cubes.
- The storage engines of MOLAP systems are optimized for the unique characteristics of multidimensional data such as sparsity and complex aggregation across thousands of cells.
- Because data cubes are precomputed, MOLAP query performance is generally better than competing approaches that use relational database storage.
- Even with techniques to deal with sparsity, MOLAP engines can be overwhelmed by the size of data cubes.

ROLAP

- Because of the potential market size and growth of data warehouse processing, vendors of relational DBMSs have extended their products with additional features to support operations and storage structures for multidimensional data.
- These product extensions are collectively known as ROLAP for Relational OLAP.
- In the ROLAP approach, relational databases store multidimensional data using the star schema or its variations

ROLAP Vs. MOLAP

- Despite the intensive research and development on ROLAP storage and optimization techniques, MOLAP engines still provide faster query response time.
- However, MOLAP storage suffers from limitations in data cube size so that ROLAP storage is preferred for fine-grained data warehouses.
- In addition, the difference in response time has narrowed so that ROLAP storage m ay involve only a slight performance penalty if ROLAP storage and optimization techniques are properly utilized.

HOLAP

- Because of the tradeoffs between MOLAP and ROLAP, a third technology known as HOLAP for Hybrid OLAP has been developed to combine ROLAP and MOLAP.
- HOLAP allows a data warehouse to be divided between relational storage of fact and dimension tables and multidimensional storage of summary data cubes.
- When an OLAP query is submitted, the HOLAP system can combine data from the ROLAP managed data and the MOLAP managed data.

HOLAP Disadvantages

- First, HOLAP can be more complex than either ROLAP or MOLAP, especially if a DBMS vendor does not provide full HOLAP support.
- To fully support HOLAP, a DBMS vendor must provide both MOLAP and ROLAP engines as well as tools to combine both storage engines in the design and operation of a data warehouse.
- Second, there is considerable overlap in functionality between the storage and optimization techniques in ROLAP and MOLAP engines.
- It is not clear whether the ROLAP storage and optimization techniques.

HOLAP Disadvantages

- It is not clear whether the ROLAP storage and optimization techniques should be discarded or used in addition to the MOLAP techniques.
- Third, because the difference in response time has narrowed between ROLAP and MOLAP, the combination of MOLAP and ROLAP may not provide significant performance improvement to justify the added complexity.

Questions

Questions

- What is a star schema?
- What are the differences between fact tables and dimension tables?
- How does a snowflake schema differ from a star schema?
- What is a constellation schema?
- How is time represented for a fact table?
- What is an accumulating fact table?
- What is the difference between Type II and Type III representations for historical dimension integrity?

Questions

- What are the pros and cons of a MOLAP storage engine?
- What are the pros and cons of a ROLAP storage engine?
- What are the pros and cons of a HOLAP storage engine?