Data Warehousing

What is a Data Warehouse?

Data warehouse refers to a data repository that is maintained separately from the organization's operational database. It supports information processing by providing a solid platform of consolidated, historical data for analysis.

According to William H. Inmon, data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process".

Data Warehouse is the process of constructing and using data warehouses.

Subject-Oriented Data Warehouse

- Organized around major subjects, such as customer, product, sales
- Focused on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.
- Provides a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

Integrated Data Warehouse

- Constructed by integrating multiple, heterogeneous data sources such as relational databases, flat files, on-line transaction records
- Applies data cleaning and data integration techniques
- Ensures consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources.
- Converts data when moved to the warehouse

Time-variant Data Warehouse

- Data are stored to provide information from a historic perspective (e.g. the past 5 – years)
- Every key structure in the data warehouse contains either implicitly or explicitly, a time element.

Non-Volatile Data Warehouse

- A physically separate store of data transformed from the operational environment
- Operational update of data does not occur in the data warehouse environment
- Does not require transaction processing, recovery, and concurrency control mechanisms
- Usually requires only two operations in data accessing: initial loading and access of data.

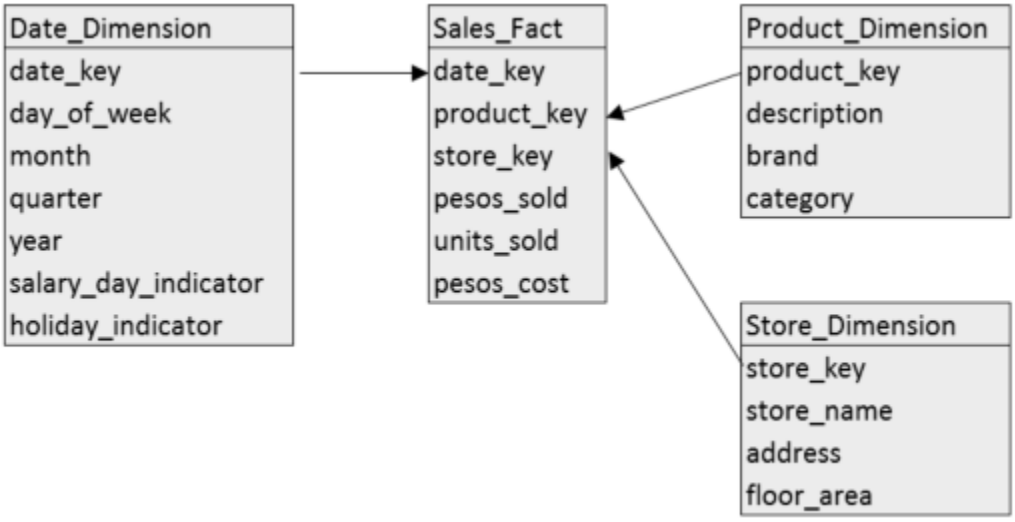
Dimensional Modelling

Dimensional modeling is a logical design technique for structuring data so such that it is intuitive for business users and delivers fast query performance. It is widely accepted as the preferred approach for Data warehousing presentation.

Dimensional Modelling divides world into measurements and context. Measurements are numeric values called facts. Context are intuitively divided into clumps called dimensions. Dimensions describe the "who, what, where, when, why, and how" of the facts.

Dimensional Model

A dimensional model consists of a fact table containing measurements surrounded by a halo of dimension tables containing textual context. It is commonly known as star join and as a star schema when stored in a relational database (RDBMS). The figure shows a typical dimensional model.



The records in the dimensional model can be retrieved using SQL, like in the following example

```
SELECT p.brand, sum(f.pesos_sold),
sum(f.units_sold)
FROM sales_fact f, product_dim p, date_dim d
WHERE f.productkey = p.productkey
and f.datekey = d.datekey
and d.quarter = '1 Q 2015'
GROUP BY p.brand
ORDER BY p.brand
```

Example of Dimension Attribute and Fact Table Metrics

Brand	Pesos Sales	Unit Sales
Axon	780	263
Framis	1044	509
Widget	213	444
Zapper	95	39

Dimension Attribute

Fact Table Metrics

Dimensional Modelling Paradigms

1. Relational Modelling
2. Dimensional Modelling

Normalized Models

Normalized data models are widely used in most databases. It is designed to limit redundancies. Other than keys, each attribute may appear in only one table. Third Normal Form (3NF) model is the design objective of normalized models. However, it can also be extended to 4NF or more. Modelling business processes results in numerous data entities or tables and a spaghetti-like interweaving of relationships among them. Most ERP systems have tens of thousands of tables.

Normalized models are essential to operational systems because of its speed when processing individual transactions.

Normalization

Normalization is a database design technique that reduces data redundancy and eliminates undesirable characteristics like insertion, update and deletion anomalies. Normalization rules divides larger tables into smaller tables and links them using relationships. The inventor of the relational model Edgar Codd proposed the theory of normalization with the introduction of the First and Second Normal Form and later joined Raymond F. Boyce to develop the theory of Boyce-Codd Normal

---

Normalized models are not good for data warehouse systems due to the following reasons:

- Not usable by end-users because of its complex structure
- Not usable for data warehouse queries because of slow performance due to the many table joins

---

Benefits of Dimensional Modeling According to Kimball

- Understandability.
    - Model must be easily understood by business users although represent complexities of the business
- Performance.
    - Fast response to queries that summarize millions of rows is essential
    - Limiting models to single level joins rather than multi-level joins
    - Denormalization has a significant impact on performance

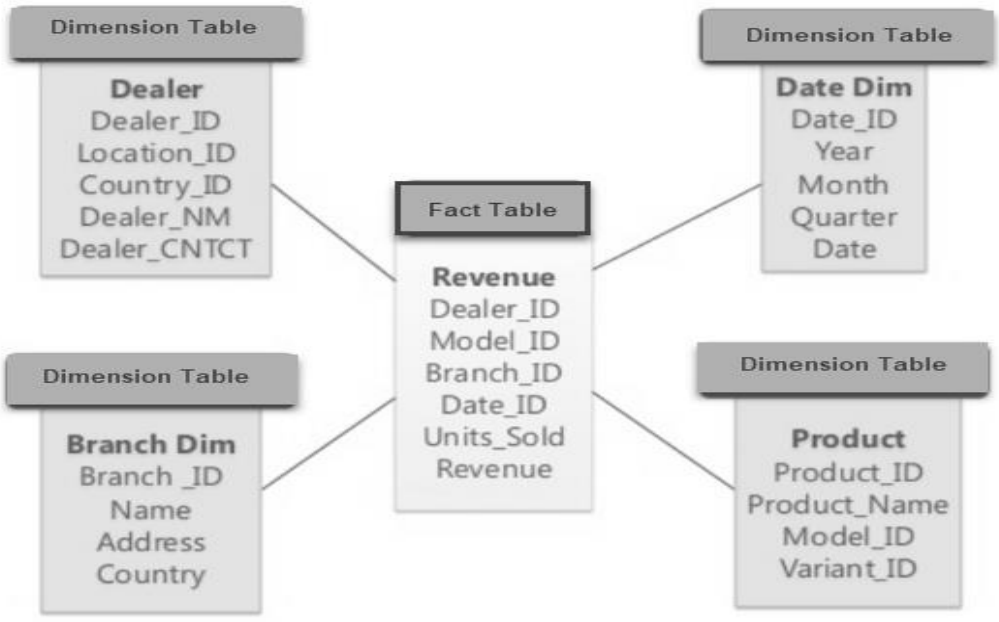
---

Examples of Dimensional Modelling

1. Star Schema

In the Star schema, the center of the star can have one fact table and a number of associated dimension tables. It is known as star schema as its structure resembles a star. The star schema is the simplest type of Data Warehouse schema. It is also known as Star Join Schema and is optimized for querying large data sets.

In the following example, the fact table is at the center which contains keys to every dimension table like Dealer_ID, Model ID, Date_ID, Product_ID, Branch_ID & other attributes like Units sold and revenue.



Example of Star Schema

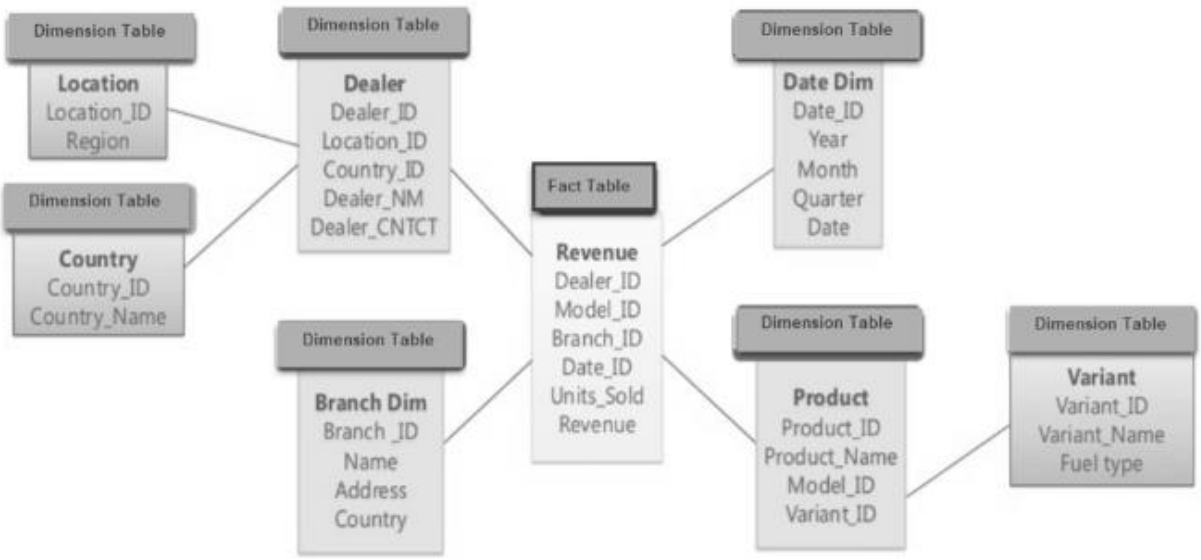
Characteristics of Star Schema

- Every dimension in a star schema is represented with the only one-dimension table.
- The dimension table should contain the set of attributes.
- The dimension table is joined to the fact table using a foreign key
- The dimension table are not joined to each other
- Fact table would contain key and measure
- The Star schema is easy to understand and provides optimal disk usage.
- The dimension tables are not normalized.
- The schema is widely accepted by business intelligence tools.

---

2. Snowflake Schema

Snowflake schema is a logical arrangement of tables in a multidimensional database such that the entity relationship diagram resembles a snowflake shape. A Snowflake Schema is an extension of a Star Schema, and it adds additional dimensions. The dimension tables are normalized which splits data into additional tables.



Example of Snowflake Schema

Characteristics of Snowflake Schema

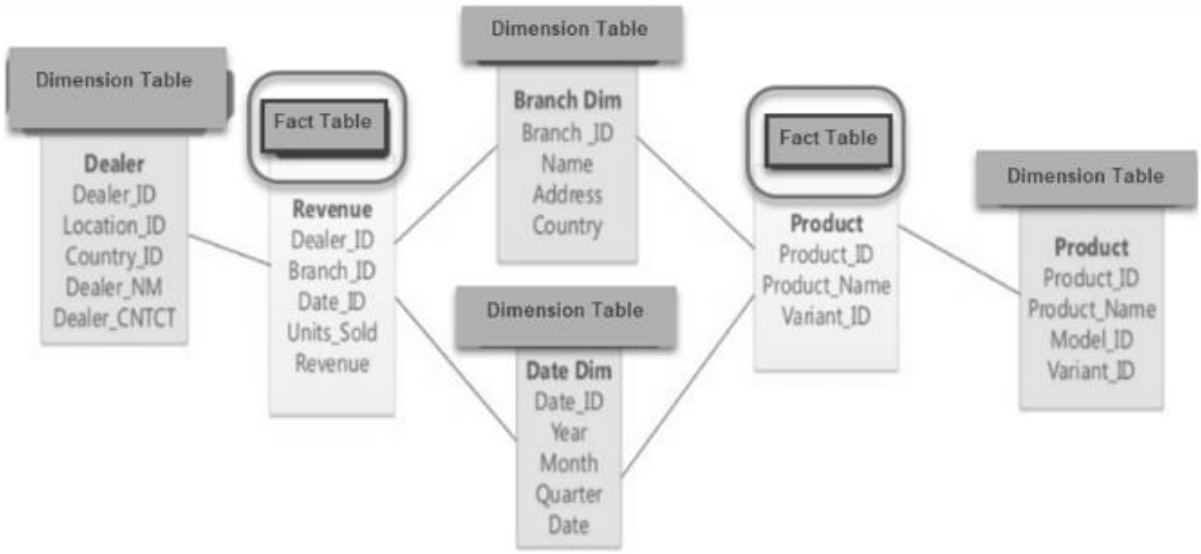
- The main benefit of the snowflake schema it uses smaller disk space.
- Easier to implement a dimension is added to the Schema
- Due to multiple tables query performance is reduced
- The primary challenge that you will face while using the snowflake Schema is that you need to perform more maintenance efforts because of the more lookup tables.

Key Difference Between Star and Schema Tables

Star Schema	Snow Flake Schema
Hierarchies for the dimensions are stored in the dimensional table.	Hierarchies are divided into separate tables.
It contains a fact table surrounded by dimension tables.	One fact table surrounded by dimension table which are in turn surrounded by dimension table
In a star schema, only single join creates the relationship between the fact table and any dimension tables.	A snowflake schema requires many joins to fetch the data.
Simple DB Design.	Very Complex DB Design.
Denormalized Data structure and query also run faster.	Normalized Data Structure.
High level of Data redundancy	Very low-level data redundancy
Single Dimension table contains aggregated data.	Data Split into different Dimension Tables.
Cube processing is faster.	Cube processing might be slow because of the complex join.
Offers higher performing queries using Star Join Query Optimization. Tables may be connected with multiple dimensions.	The Snow Flake Schema is represented by centralized fact table which unlikely connected with multiple dimensions.

3. Galaxy Schema

A galaxy schema contains two fact table that share dimension tables between them. It is also called Fact Constellation Schema. The schema is viewed as a collection of stars hence the name Galaxy Schema.



Example of Galaxy Schema

In the example, there are two facts table

1. Revenue
2. Product.

In Galaxy schema shares dimensions are called Conformed Dimensions.

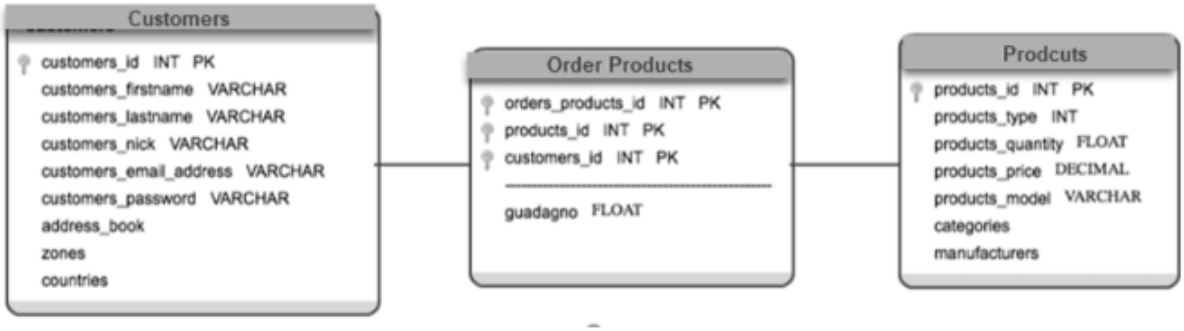
Characteristics of Galaxy Schema:

- The dimensions in this schema are separated into separate dimensions based on the various levels of hierarchy.
- For example, if geography has four levels of hierarchy like region, country, state, and city then Galaxy schema should have four dimensions.
- Moreover, it is possible to build this type of schema by splitting the one-star schema into more Star schemes.
- The dimensions are large in this schema which is needed to build based on the levels of hierarchy.
- This schema is helpful for aggregating fact tables for better understanding.

---

4. Star Cluster Schema

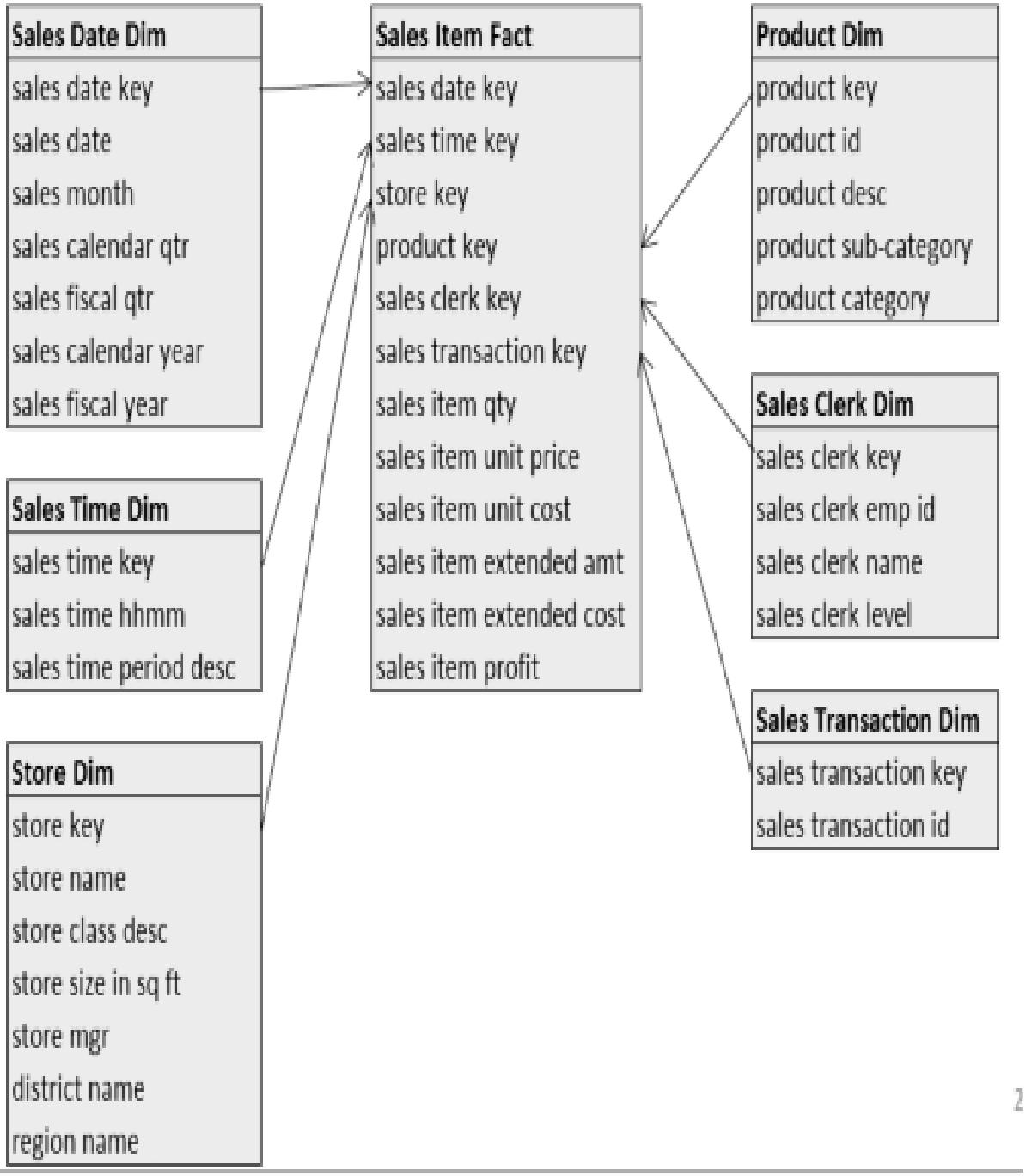
Snowflake schema contains fully expanded hierarchies. However, this can add complexity to the Schema and requires extra joins. On the other hand, star schema contains fully collapsed hierarchies, which may lead to redundancy. So, the best solution may be a balance between these two schemas which is Star Cluster Schema design.



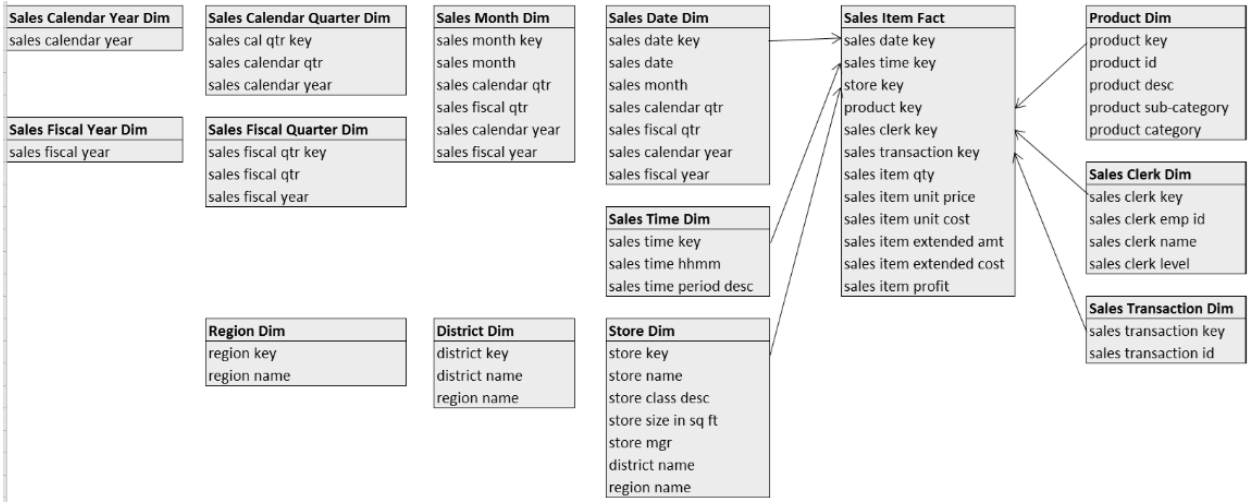
Example of Star Cluster Schema

Overlapping dimensions can be found as forks in hierarchies. A fork happens when an entity acts as a parent in two different dimensional hierarchies. Fork entities then identified as classification with one-to-many relationships.

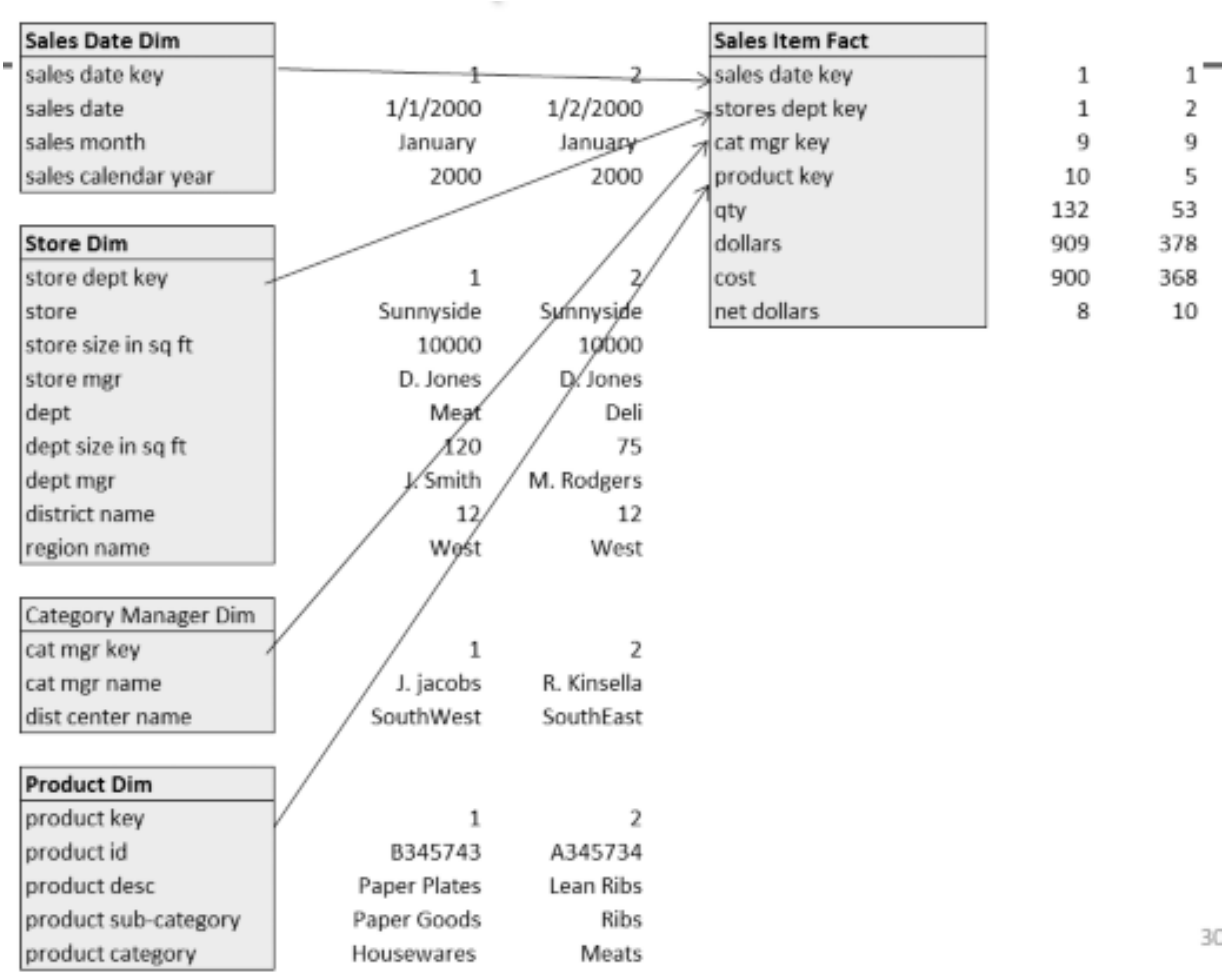
Example of Star Schema



With Dimension Families



Sample Data



References

- Data Mining and Warehousing. Jiawei Han
- Business Analytics Data Warehousing – UP National Engineering Center
- Guru99 Learning Portal