Data mining

Presented by: Tek Narayan Adhikari

What is Data Mining?

- "The process of discovering meaningful patterns and trends often previously unknown by using some mathematical algorithm on huge amount of stored data"
- "Extraction of interesting, non-trivial, implicit, previously unknown and potentially useful information or patterns from data in large database."
- Data mining is basically concerned with the analysis of data and the use of software techniques for finding patterns and regularities in sets of data.

Two Approaches are:

Descriptive Data Mining

Predictive Data Mining

Descriptive Data Mining:

- It characterizes the general properties of data in the database.
- It finds patterns in data the user determinants which ones are important.
- Mostly used during data exploration.
- Typical questions answered by descriptive data mining are:
  - What is in the data?
  - What doesn't look like?
  - Are there any unusual patterns?
  - What does the data suggest for customer segmentation?
  - User may have no idea on which kind of patterns are interesting?
- Functionalities of descriptive data mining are: Clustering, Summarization, Visualization, and Association.

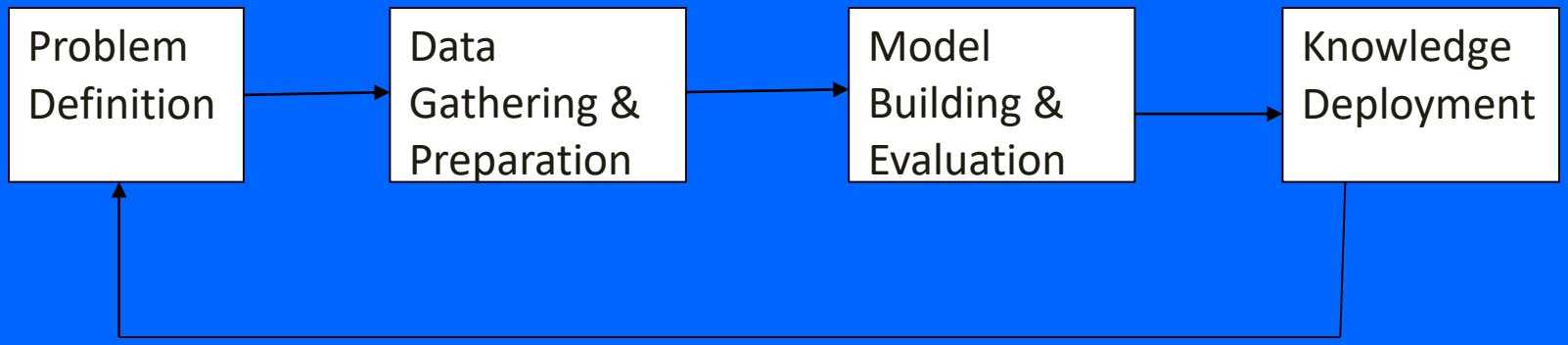
Predictive Data Mining:

- X: Vectors of independent variables.
- Y: Dependent variables
- Y = f(X)
- Users don't care about the model, they simply interested in accuracy of predictions.
- Using unknown examples the model is trained and the unknown function is learned from data.
- The more data with known outcomes is available the better is the predictive power of model.

Predictive Data Mining:

- Used to predict outcomes whose inputs are known but the output values are not realized yet.
- Never 100% accurate.
- The performance of a model on past data is not predicting the known outcomes.
- Suitable for unknown data set.
- Typical questions answered by predictive models are:
  - Who is likely to respond to next product?
  - Which customers are likely to leave in the next six months?

Data Mining Process:



Problem Definition:

- Focuses on Understanding the project objectives and requirements in terms of business perspective.
- Eg: How can I sell more of my product to customer? Which customers are most likely to purchase the product?

Data Gathering and Preparation:

- Data Collection & Exploration.
- Identify data quality, patterns in data.
- Data preparation phase covers all the tasks involved to build the model.
- Data preparation tasks are likely to be performed multiple and not in any prescribed order.

Data Mining Process cont..

Model Building and Evaluation:

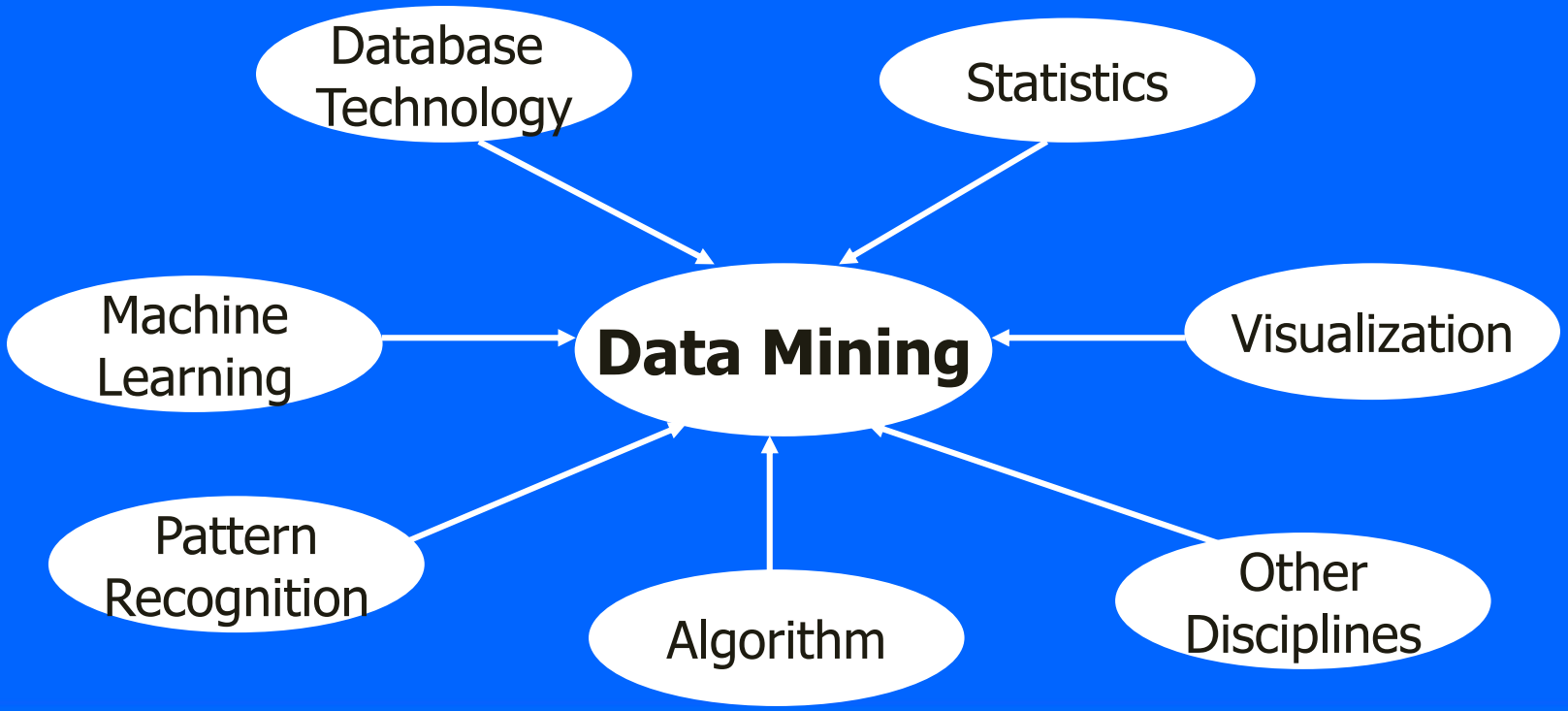
- Various modeling techniques are applied and calibrated the parameters to optimal values.
- Evaluate how well the model satisfies the originally stated business goal.

Knowledge Deployment:

- Use data mining within a target environment.
- Insight and actionable information can be derived from data.

Why Data Mining?

- Data mining is a combination of multidisciplinary field. It can be applied in many fields and can be done using many algorithm and techniques.



Data Mining Vs. Query Tools

- SQL can find normal queries from the database such as what is an average turnover? Whereas data mining tools find interesting patterns and facts such as what are the important trends in sells?
- Data mining is much more faster than SQL in trend and pattern analysis since it uses algorithm like machine learning, genetic algorithm.
- If we know exactly what we are looking for, we use SQL nut if we know only vaguely what we are looking for we use data mining.
- Hybrid information can't be easily be traced using SQL.

Data Warehouse

- In most of the organization, there occur large databases in operation for normal daily transactions called operational database.
- A data warehouse is a large database built from the operational database.
- In computing, a data warehouse (DW or DWH), also known as an enterprise data warehouse (EDW), is a system used for reporting and data analysis, and is considered a core component of business intelligence. DWs are central repositories of integrated data from one or more disparate sources.

Data Warehouse

- A data warehouse is a database, which is kept separate from the organization's operational database.
- There is no frequent updating done in a data warehouse.
- It possesses consolidated historical data, which helps the organization to analyze its business.
- A data warehouse helps executives to organize, understand, and use their data to take strategic decisions.
- Data warehouse systems help in the integration of diversity of application systems.
- A data warehouse system helps in consolidated historical data analysis.

Data Warehouse

A data warehouse should be:

- Time – dependent
  - There must be a connection between the information in the warehouse and the time when it was entered.
  - One of the most important aspect of the warehouse as it relates to data mining, because information can then be sourced according to period.
- Non-Volatile
  - Data in a warehouse is never updated, but used only for queries.
  - End-users who want to update data must use operational database.
  - A data warehouse will always be filled with historical data.

A data warehouse should be:

- Subject Oriented
  - Not all the information in the operational database is useful for a data warehouse.
  - A data warehouse should be designed especially for decision support and expert system with specific related data.
- Integrated
  - In an operational data, many types of information being used with different names for same entity.
  - In a data warehouse, all entities should be integrated and consistent i.e. only one name must exist to describe each individual entity.

Data Warehouse

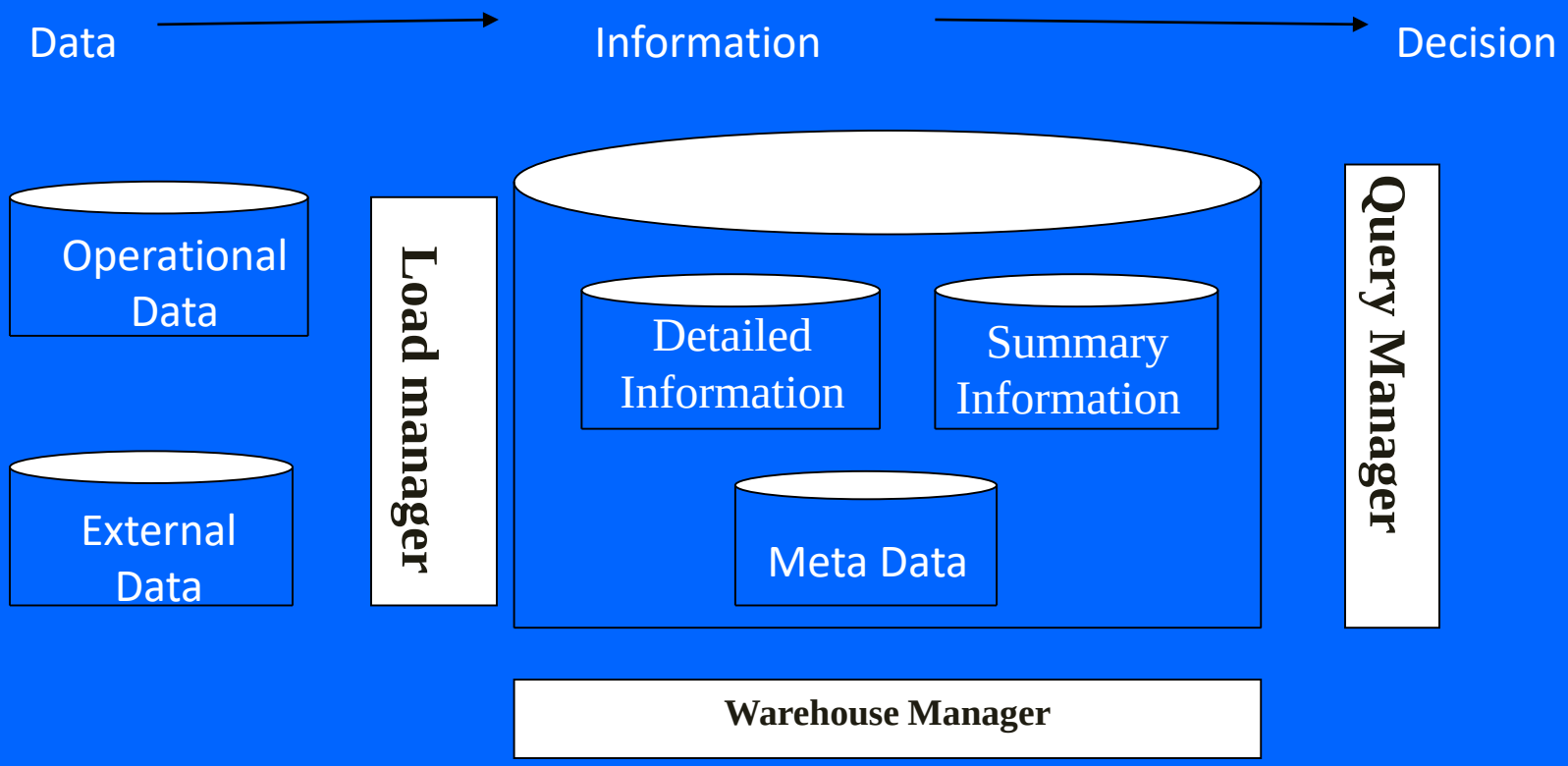
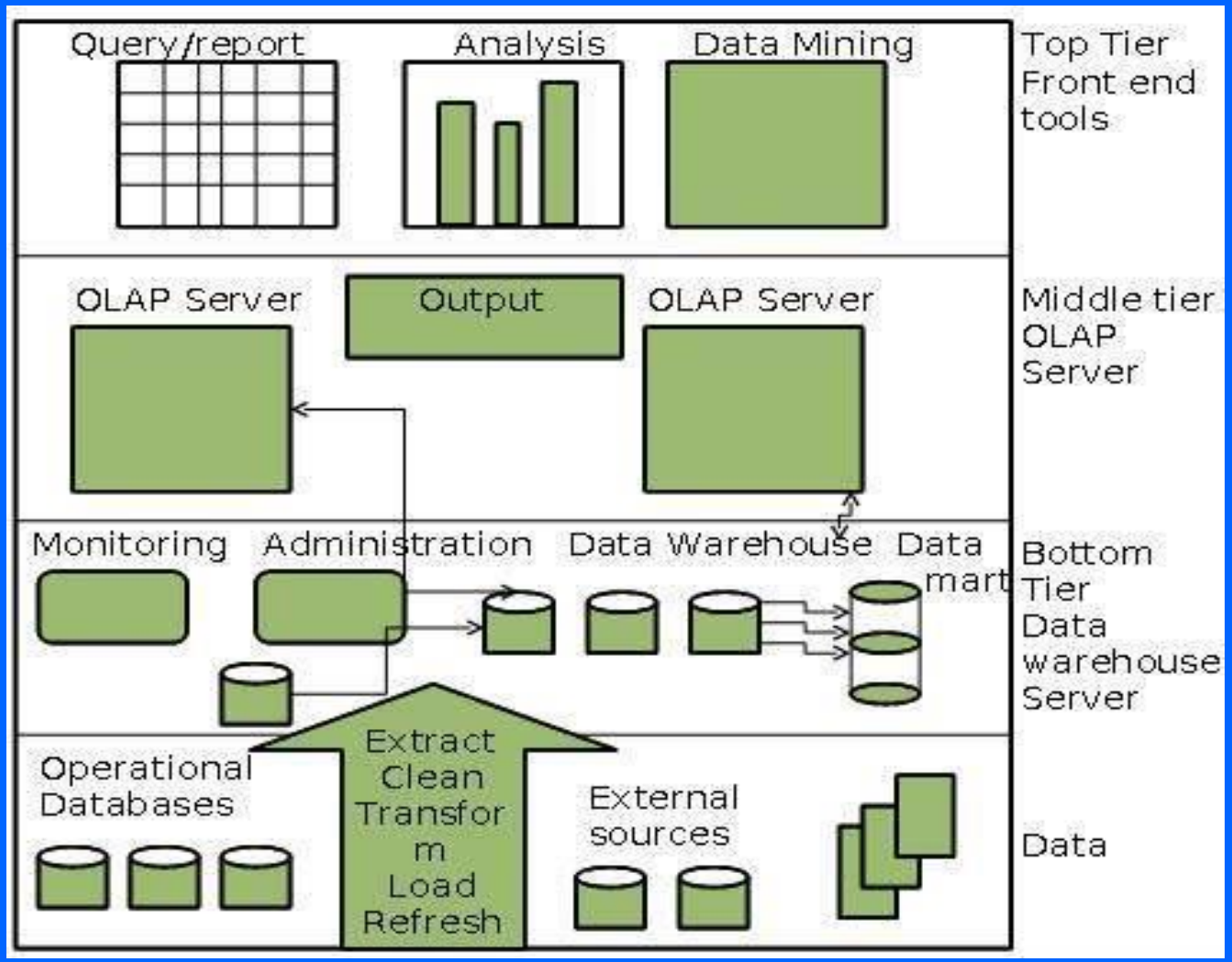


Fig: "Architecture of a Data Warehouse"

Three Tire Architecture of Data Warehouse



Data Warehouse

- ***Load Manager:*** The system components that perform all the operations necessary to support the extract and load process. It fast loads the extracted data into a temporary data store and performs simple transformations into a structure similar to the one in the data warehouse.
- ***Warehouse Manager:*** Performs all the necessary operations to support the warehouse management process. It analyzes the data to perform consistency and referential checks. It also transforms and merges the source data in the temporary data store into the published data warehouse with creating indexes and business views. Update all existing aggregations and back up data in the data warehouse.
- ***Query Manager:*** Performs all the operations necessary to support the query management process by directing queries to the appropriate tables. In some cases it also stores query profiles to allow the warehouse manager to determine which indexes and aggregations are appropriate.

Data Warehouse

- ***Detailed Information:*** Stores all the detailed information to determine the business requirements to analyze the level at which to retain detailed information in the data warehouse.
- ***Summary Information:*** Stores all the predefined aggregations generated by the warehouse manager. It is a transient area which will change on an ongoing basis in order to respond to changing query profiles. It is essentially a replication to detailed information.
- ***Meta Data:*** Meta data is data about data which describes how information is structured within a data warehouse. It maps data stores to common view of information with the data warehouse.

Data Warehouse Models

- From the perspective of data warehouse architecture, we have the following data warehouse models:
- Virtual Warehouse
- Data mart
- Enterprise Warehouse

Virtual Warehouse

- The view over an operational data warehouse is known as a virtual warehouse.
- A virtual data warehouse provides a compact view of the data inventory.
- It contains Meta data.
- It uses middleware to build connections to different data sources.
- They can be fast as they allow users to filter the most important pieces of data from different legacy applications.
- Easy to build a virtual warehouse.
- Building a virtual warehouse requires excess capacity on operational database servers.

Data Mart:

- Data Mart is a subset of the information content of a data warehouse that is stored in its own database.
- Data mart may or may not be sourced from an enterprise data warehouse i.e. it could have been directly populated from source data.
- Data mart can improve query performance simply by reducing the volume of data that needs to be scanned to satisfy the query.
- Data marts are created along functional level to reduce the likelihood of queries requiring data outside the mart.
- Data marts may help in multiple queries or tools to access data by creating their own internal database structures.
- Eg: Departmental Store, Banking System.

Enterprise Warehouse

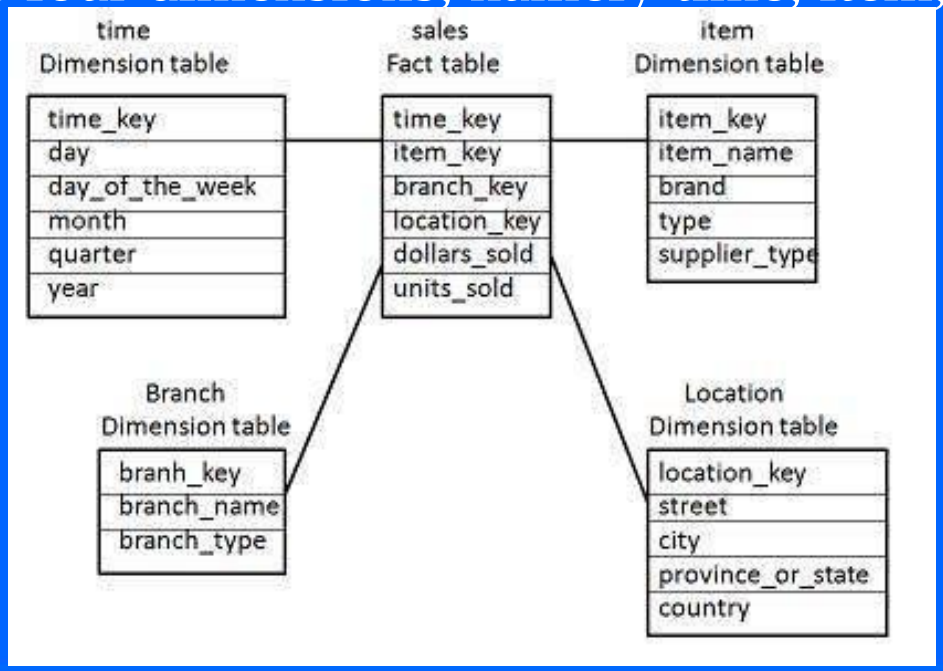
- An enterprise warehouse collects all the information and the subjects spanning an entire organization
- It provides us enterprise-wide data integration.
- The data is integrated from operational systems and external information providers.
- This information can vary from a few gigabytes to hundreds of gigabytes, terabytes or beyond.

Data Warehousing - Schemas

- Schema is a logical description of the entire database.
- It includes the name and description of records of all record types including all associated data-items and aggregates.
- Much like a database, a data warehouse also requires to maintain a schema.
- A database uses relational model, while a data warehouse uses Star, Snowflake, and Fact Constellation schema.

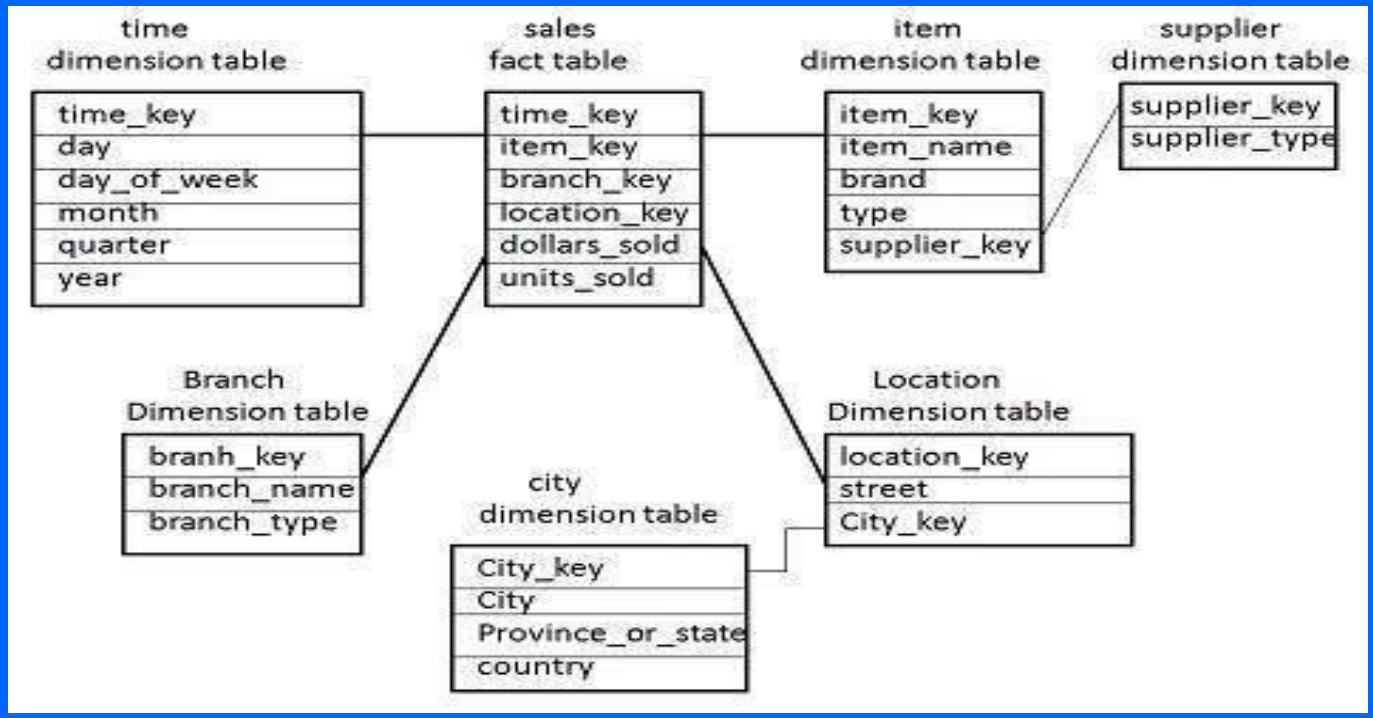
Star Schema

- Each dimension in a star schema is represented with only one-dimension table.
- This dimension table contains the set of attributes.
- The following diagram shows the sales data of a company with respect to the four dimensions, namely time, item, branch, and location.



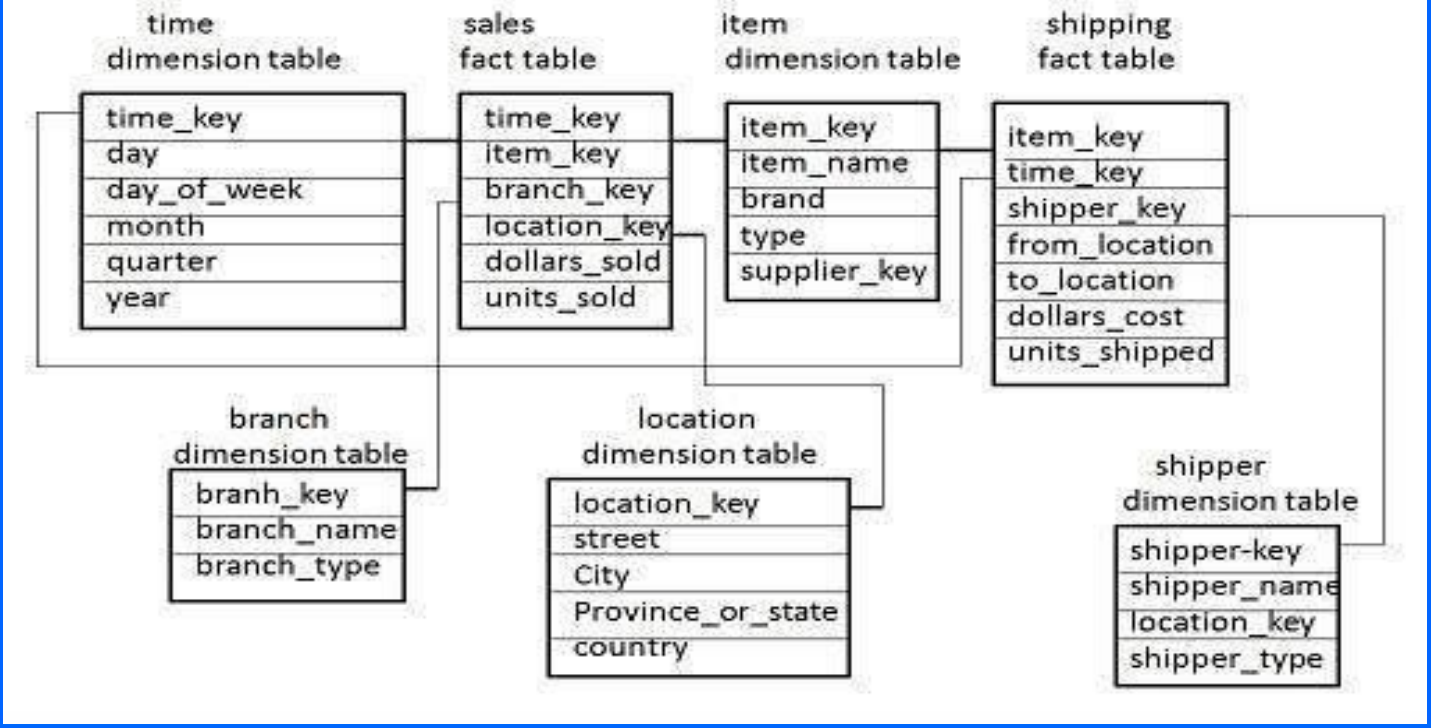
Snowflake Schema

- Some dimension tables in the Snowflake schema are normalized.
- The normalization splits up the data into additional tables.
- Unlike Star schema, the dimensions table in a snowflake schema are normalized. For example, the item dimension table in star schema is normalized and split into two dimension tables, namely item and supplier table.

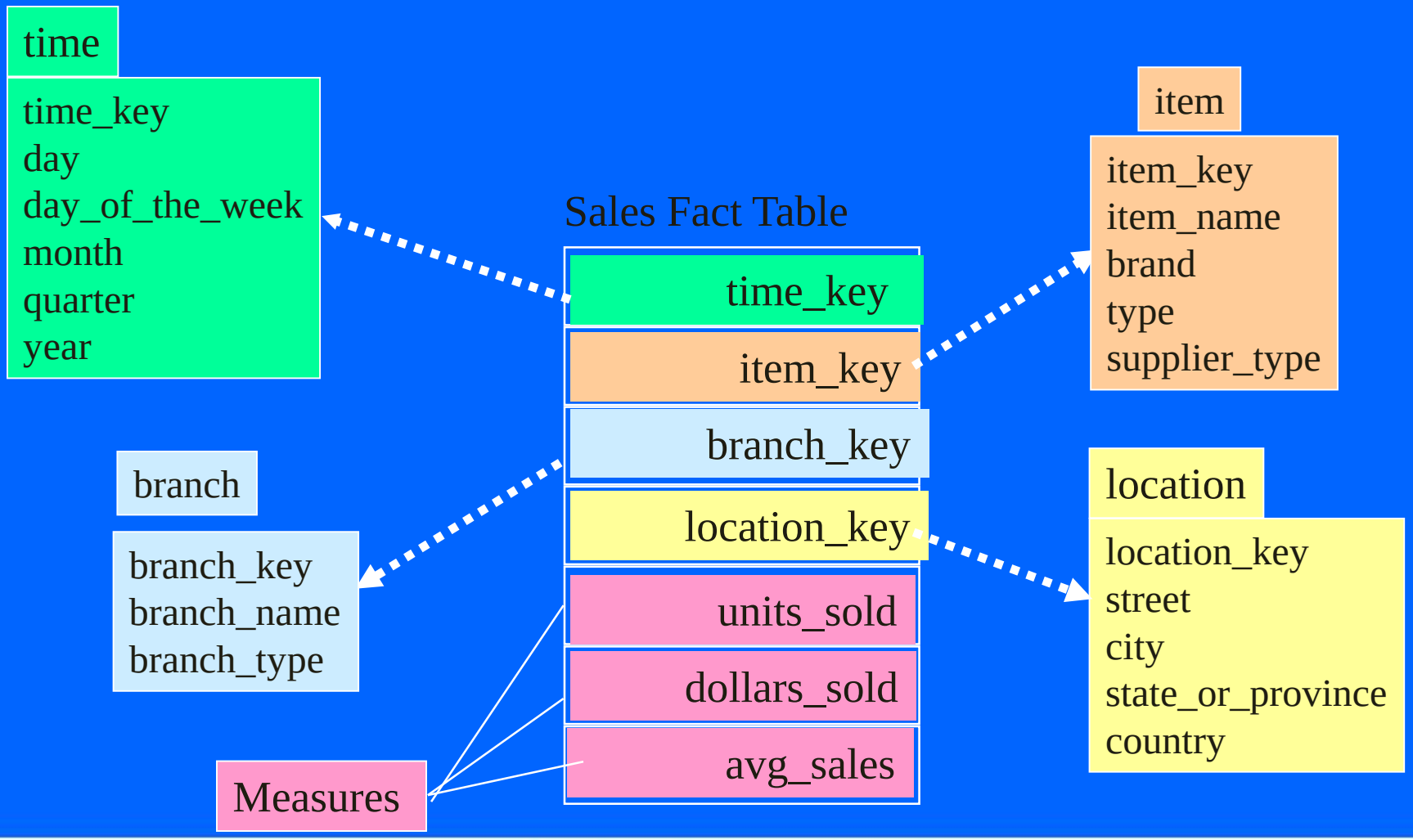


Fact Constellation Schema

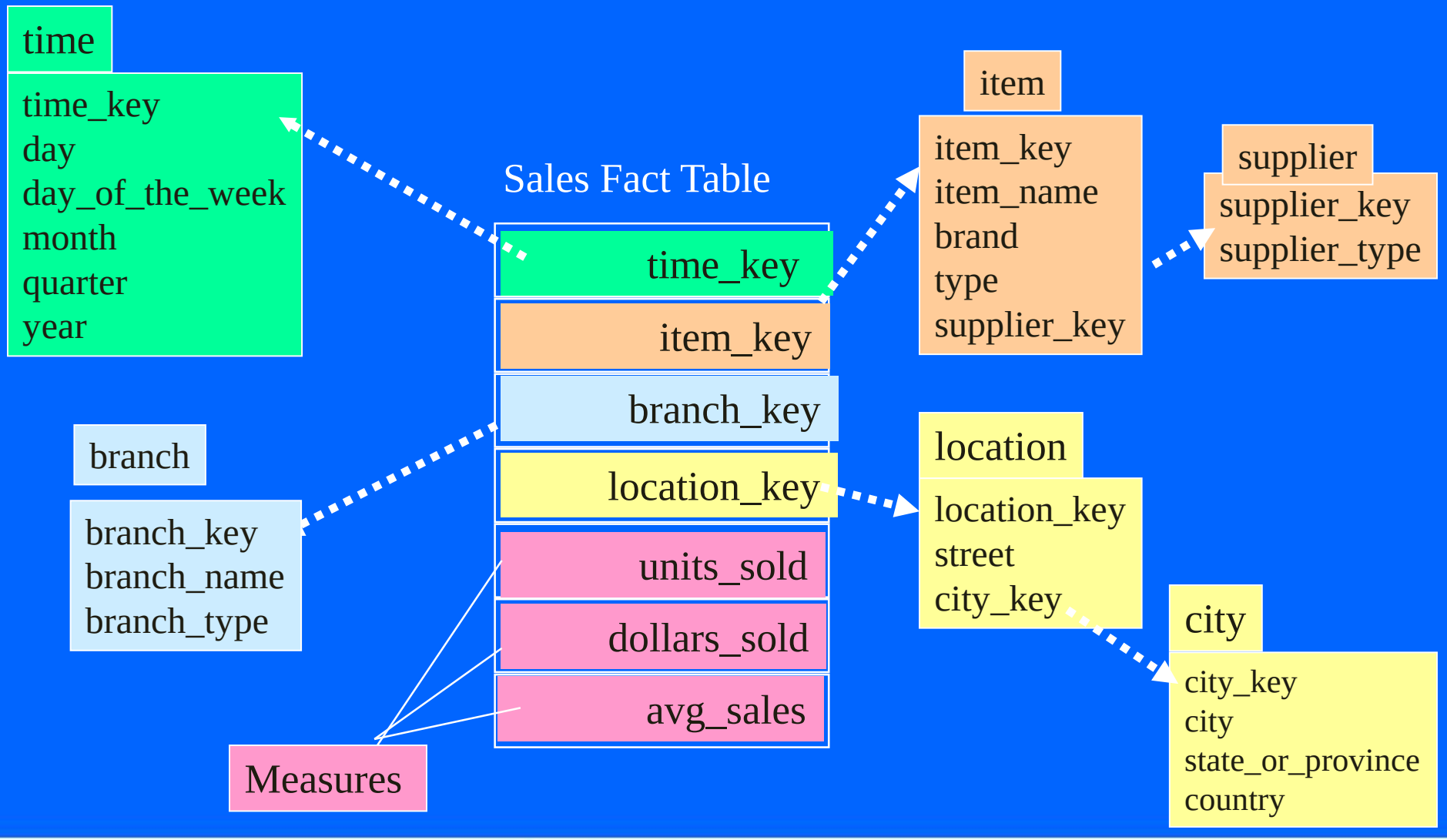
- A fact constellation has multiple fact tables. It is also known as galaxy schema.
- The following diagram shows two fact tables, namely sales and shipping.



Example of **Star Schema**



Example of **Snowflake Schema**



Example of **Fact Constellation**