Data warehousing and mining

Granularity

- Charge of a patient per day per ward.

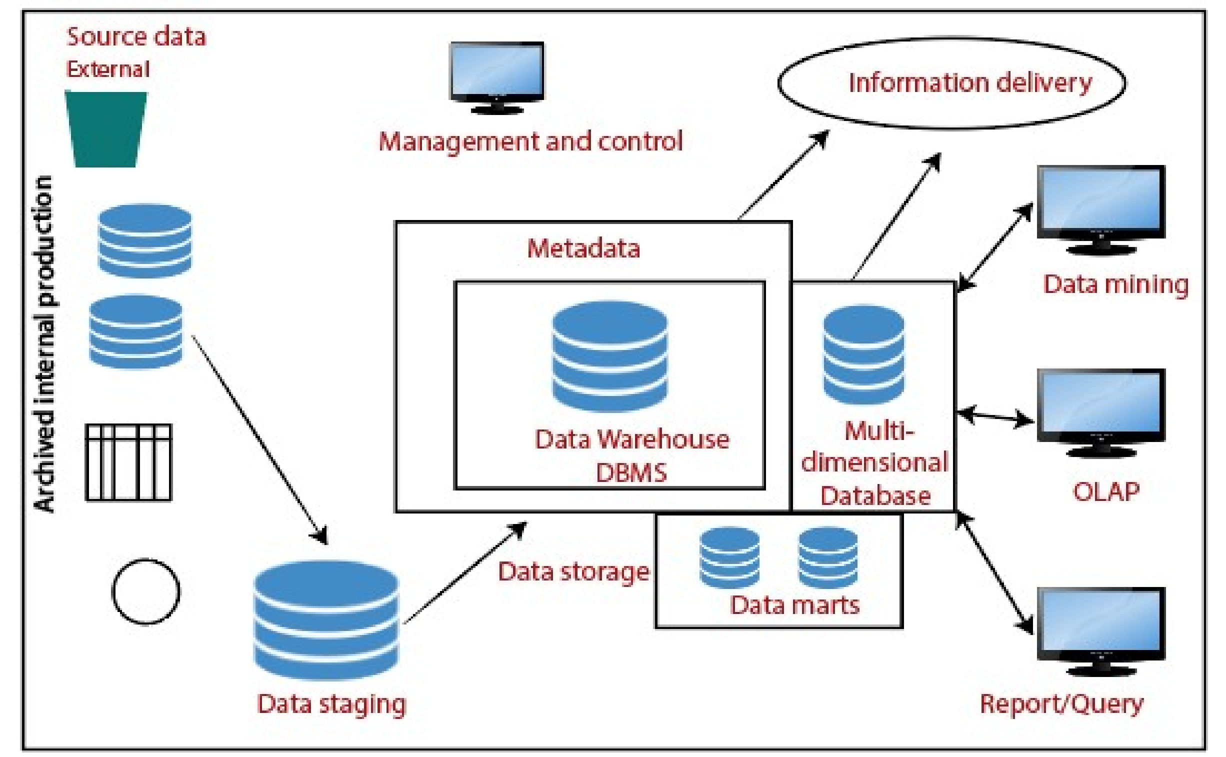
Top down approach

- Top down
  - Advantages:
    - Enterprise view of data.
    - Single central storage of data about the content.
    - Quick results with iterative implementations
  - Disadvantages:
    - Takes longer to build even with an iterative method.
    - High exposure/risk to failure.
    - High outlay without proof of concept.

Bottom up Approach

- Advantages:
  - Faster and easier to implement
  - Favourable return on investment and proof of concept.
  - Less risk of failure
  - Learn and grow
- Disadvantages:
  - Each data mart has its own narrow view of data.
  - Redundant data
  - Inconsistent and irreconcilable data
  - Unmanageable interfaces

Data warehouse architecture



Components or Building Blocks of Data Warehouse

Source Data Component

- **Production Data:** This type of data comes from the different operating systems of the enterprise. Based on the data requirements in the data warehouse, we choose segments of the data from the various operational modes.
- **Internal Data:** In each organization, the client keeps their "**private**" spreadsheets, reports, customer profiles, and sometimes even department databases. This is the internal data, part of which could be useful in a data warehouse.
- **Archived Data:** Operational systems are mainly intended to run the current business. In every operational system, we periodically take the old data and store it in archived files.
- **External Data:** Most executives depend on information from external sources for a large percentage of the information they use. They use statistics associating to their industry produced by the external department.

Data Staging Component

- A staging area, or landing zone, is an intermediate storage area used for data processing during the extract, transform and load (ETL) process. The data staging area sits between the data source(s) and the data target(s), which are often data warehouses, data marts, or other data repositories.
- Data staging areas are often transient in nature, with their contents being erased prior to running an ETL process or immediately following successful completion of an ETL process. There are staging area architectures, however, which are designed to hold data for extended periods of time for archival or troubleshooting purposes.

ETL

- Extract Transform Load

Extraction

- **Data Extraction:** This method has to deal with numerous data sources. We have to employ the appropriate techniques for each data source.

Transformation

- Data Transformation: As we know, data for a data warehouse comes from many different sources. If data extraction for a data warehouse posture big challenges, data transformation present even significant challenges. We perform several individual tasks as part of data transformation.
- First, we clean the data extracted from each source. Cleaning may be the correction of misspellings or may deal with providing default values for missing data elements, or elimination of duplicates when we bring in the same data from various source systems.
- Standardization of data components forms a large part of data transformation. Data transformation contains many forms of combining pieces of data from different sources. We combine data from single source record or related data parts from many source records.
- On the other hand, data transformation also contains purging source data that is not useful and separating outsource records into new combinations. Sorting and merging of data take place on a large scale in the data staging area. When the data transformation function ends, we have a collection of integrated data that is cleaned, standardized, and summarized.

Loading

- Data Loading: Two distinct categories of tasks form data loading functions. When we complete the structure and construction of the data warehouse and go live for the first time, we do the initial loading of the information into the data warehouse storage. The initial load moves high volumes of data using up a substantial amount of time.