Three-Tier Data Warehouse Architecture
The bottom tier is a warehouse database server that is almost always a relational database system.

Back-end tools and utilities are used to feed data into the bottom tier from operational databases or other external sources (such as customer profile information provided by external consultants).

These tools and utilities perform data extraction, cleaning, and transformation.

The data are extracted using application program interfaces known as gateways.
• **The middle tier** is an OLAP server that is typically implemented using either

• (i) **A relational OLAP (ROLAP) model**, that is, an extended relational DBMS that maps operations on multidimensional data to standard relational operations.

• (ii) **A multidimensional OLAP (MOLAP) model**, that is, a special-purpose server that directly implements multidimensional data and operations.
• **The top tier** is a front-end client layer, which contains query and reporting tools, analysis tools, and/or data mining tools (e.g., trend analysis, prediction, and so on).
Metadata Repository

- Meta data is the data defining warehouse objects. It has the following kinds:
  - Description of the structure of the warehouse
    - schema, view, dimensions, hierarchies, derived data defn, data mart locations and contents
  - Operational meta-data
    - data lineage (history of migrated data and transformation path), currency of data (active, archived, or purged), monitoring information (warehouse usage statistics, error reports, audit trails)
  - The algorithms used for summarization
  - The mapping from operational environment to the data warehouse
  - Data related to system performance
    - warehouse schema, view and derived data definitions
  - Business data
    - business terms and definitions, ownership of data, charging policies
Data Warehouse Back-End Tools and Utilities

- **Data extraction:**
  - get data from multiple, heterogeneous, and external sources

- **Data cleaning:**
  - detect errors in the data and rectify them when possible

- **Data transformation:**
  - convert data from legacy or host format to warehouse format

- **Load:**
  - sort, summarize, consolidate, compute views, check integrity, and build indices and partitions

- **Refresh**
  - propagate the updates from the data sources to the warehouse
• From the architecture point of view, there are three data warehouse models:
  
  I. The Enterprise Warehouse
  II. The Data Mart
  III. The Virtual Warehouse.
Enterprise warehouse

- An enterprise warehouse collects all of the information about subjects spanning the entire organization. It provides corporate-wide data integration, usually from one or more operational systems or external information providers, and is cross-functional in scope.
- It typically contains detailed data as well as summarized data, and can range in size from a few gigabytes to hundreds of gigabytes, terabytes, or beyond.
- An enterprise data warehouse may be implemented on traditional mainframes, computer super servers, or parallel architecture platforms.
Data Mart

- A data mart contains a subset of corporate-wide data that is of value to a specific group of users.
- Data marts are usually implemented on low-cost departmental servers that are UNIX/LINUX- or Windows-based.
- The implementation cycle of a data mart is more likely to be measured in weeks rather than months or years.
- Depending on the source of data, data marts can be categorized as independent or dependent.
  - Independent data marts are sourced from data captured from one or more operational systems or external information providers, or from data generated locally within a particular department or geographic area.
  - Dependent data marts are sourced directly from enterprise data warehouses.
Virtual Data Warehouse:

- A virtual warehouse is a set of views over operational databases. For efficient query processing, only some of the possible summary views may be materialized.
- A virtual warehouse is easy to build but requires excess capacity on operational database servers.
- It is popular because it enables business to access & analyze data from operational system
Distributed Data Warehouse

- Distributed data warehouses are those in which certain components of the data warehouse are distributed across a number of different physical databases.
- It usually involves redundant data & as a consequence, most complex loading and updating process.
Data Warehouse Manager

• The warehouse manager is the system component that perform all the operations necessary to support the warehouse management process.

• Operations performed by warehouse manager:
  
  I. Analyze the data to perform consistency.
  II. Create indexes, Business view, Partition view against the base data.
  III. Generate new aggregations that may be required.
  IV. Update all existing aggregations.
  V. Transform into a star flake schema.
  VI. Generate the summaries.