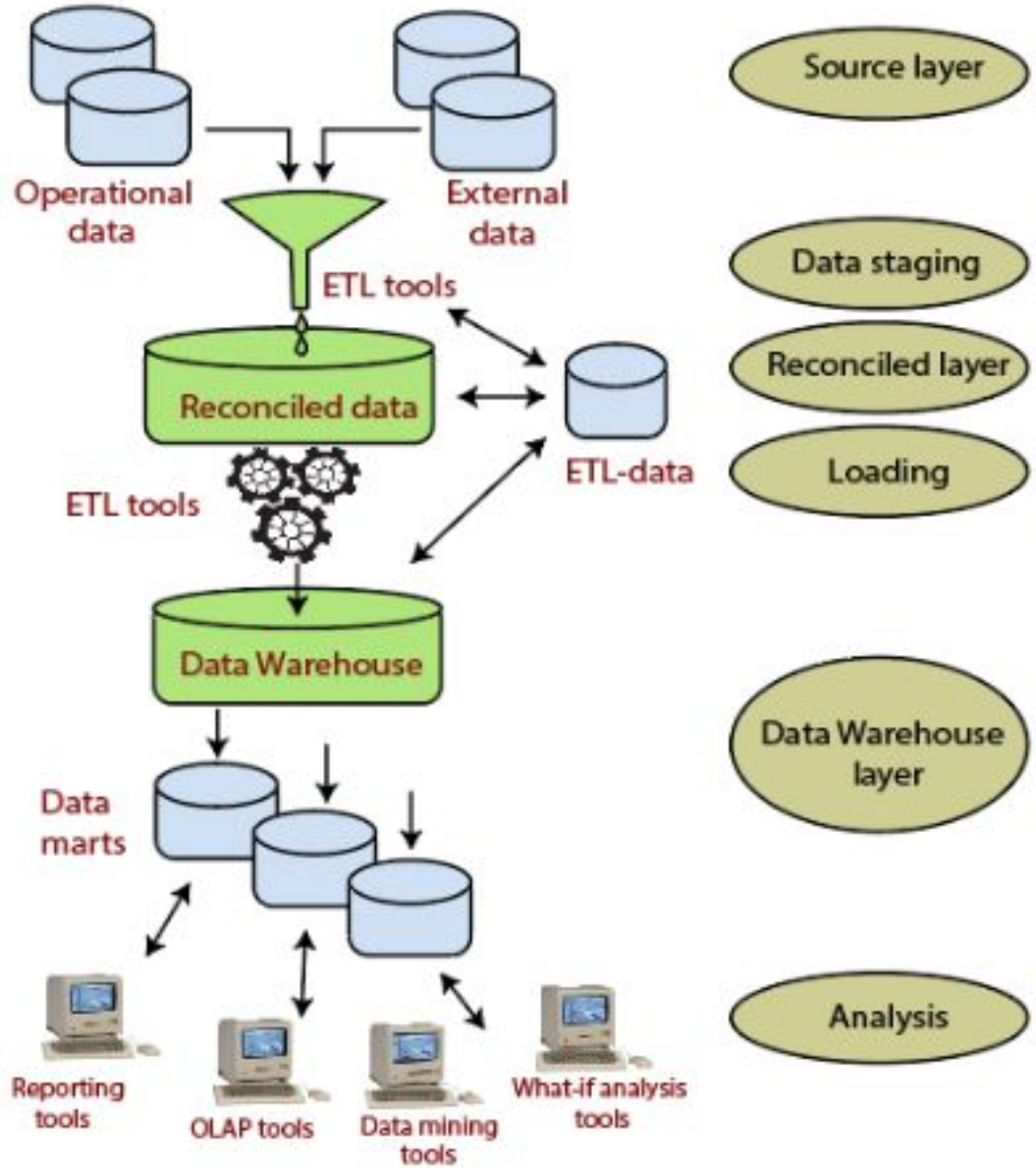
Three-Tier Data Warehouse Architecture

Data Warehouses usually have a three-level (tier) architecture that includes:

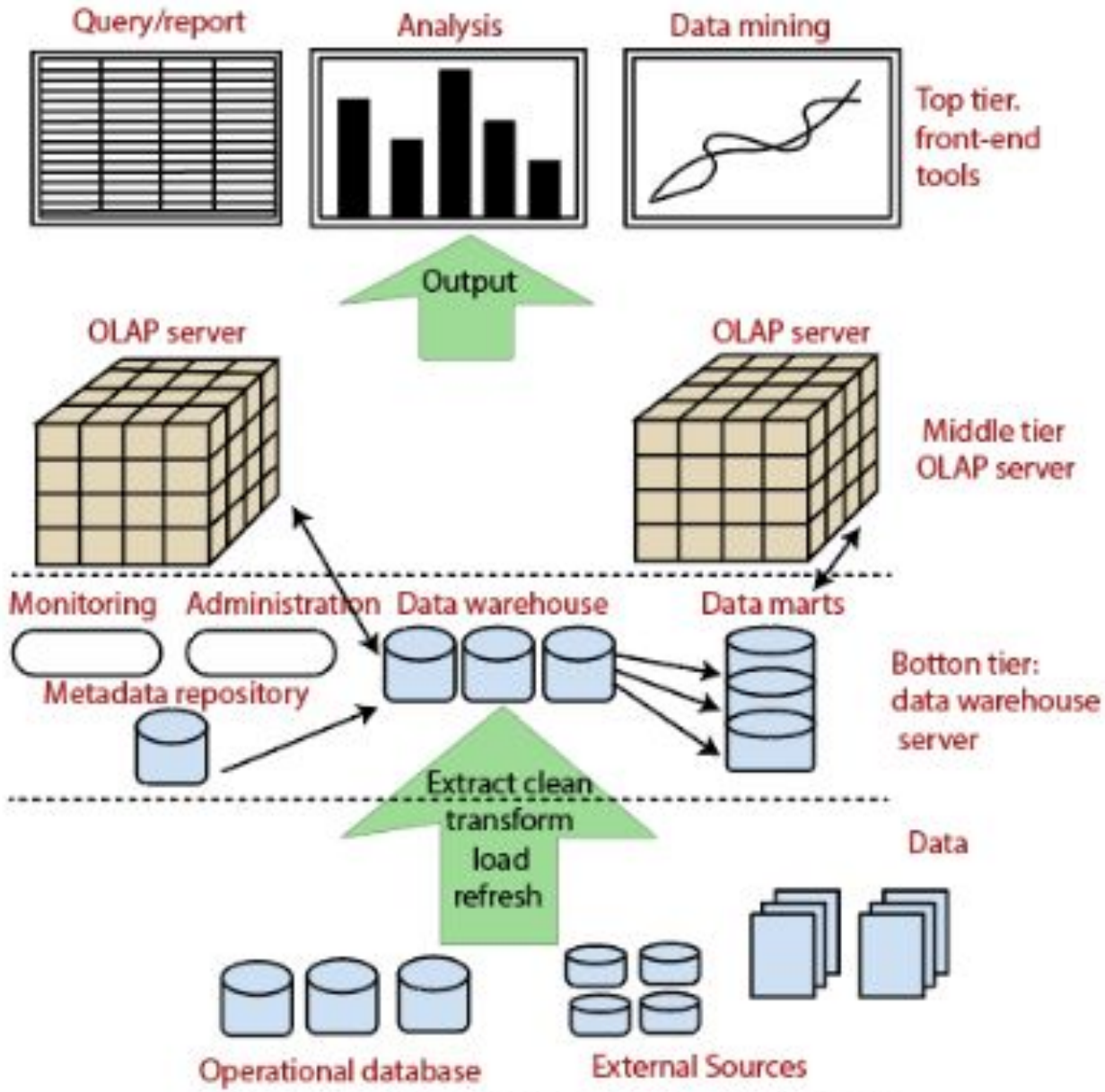
1. Bottom Tier (Data Warehouse Server)
2. Middle Tier (OLAP Server)
3. Top Tier (Front end Tools).

- A **bottom-tier** that consists of the **Data Warehouse server**, which is almost always an RDBMS. It may include several specialized data marts and a metadata repository.
- Data from operational databases and external sources (such as user profile data provided by external consultants) are extracted using application program interfaces called a gateway. A gateway is provided by the underlying DBMS and allows customer programs to generate SQL code to be executed at a server.
- **Examples** of gateways contain **ODBC** (Open Database Connection) and **OLE-DB** (Open-Linking and Embedding for Databases), by **Microsoft**, and **JDBC** (Java Database Connection).



Three-Tier Architecture for a data warehouse system

- A **middle-tier** which consists of an **OLAP server** for fast querying of the data warehouse.
- The OLAP server is implemented using either
- **(1)** A **Relational OLAP (ROLAP) model**, i.e., an extended relational DBMS that maps functions on multidimensional data to standard relational operations.
- **(2)** A **Multidimensional OLAP (MOLAP) model**, i.e., a particular purpose server that directly implements multidimensional information and operations.
- A **top-tier** that contains **front-end tools** for displaying results provided by OLAP, as well as additional tools for data mining of the OLAP-generated data.



Three-Tier Data Warehouse Architecture

- The **metadata repository** stores information that defines DW objects. It includes the following parameters and information for the middle and the top-tier applications:

1. A description of the DW structure, including the warehouse schema, dimension, hierarchies, data mart locations, and contents, etc.
2. Operational metadata, which usually describes the currency level of the stored data, i.e., active, archived or purged, and warehouse monitoring information, i.e., usage statistics, error reports, audit, etc.
3. System performance data, which includes indices, used to improve data access and retrieval performance.
4. Information about the mapping from operational databases, which provides source **RDBMSs** and their contents, cleaning and transformation rules, etc.
5. Summarization algorithms, predefined queries, and reports business data, which include business terms and definitions, ownership information, etc.

Principles of Data Warehousing

Load Performance

- Data warehouses require increase loading of new data periodically basis within narrow time windows; performance on the load process should be measured in hundreds of millions of rows and gigabytes per hour and must not artificially constrain the volume of data business.

Load Processing

- Many phases must be taken to load new or update data into the data warehouse, including data conversion, filtering, reformatting, indexing, and metadata update.

Data Quality Management

- Fact-based management demands the highest data quality. The warehouse ensures local consistency, global consistency, and referential integrity despite "dirty" sources and massive database size.

Query Performance

- Fact-based management must not be slowed by the performance of the data warehouse RDBMS; large, complex queries must be complete in seconds, not days.

Terabyte Scalability

- Data warehouse sizes are growing at astonishing rates. Today these size from a few to hundreds of gigabytes and terabyte-sized data warehouses.