Data Warehouse

Dr. Shalini Gambhir

Why we need a separate Data Warehouse?

- Data Warehouse queries are complex because they involve the computation of large groups of data at summarized levels.
- It may require the use of distinctive data organization, access, and implementation method based on multidimensional views.
- Performing OLAP queries in operational database degrade the performance of functional tasks.
- Data Warehouse is used for analysis and decision making in which extensive database is required, including historical data, which operational database does not typically maintain.
- The separation of an operational database from data warehouses is based on the different structures and uses of data in these systems.
- Because the two systems provide different functionalities and require different kinds of data, it is necessary to maintain separate databases.

Difference between Database and Data Warehouse

Database	Data Warehouse
1. It is used for Online Transactional Processing (OLTP) but can be used for other objectives such as Data Warehousing. This records the data from the clients for history.	1. It is used for Online Analytical Processing (OLAP). This reads the historical information for the customers for business decisions.
2. The tables and joins are complicated since they are normalized for RDBMS. This is done to reduce redundant files and to save storage space.	2. The tables and joins are accessible since they are de-normalized. This is done to minimize the response time for analytical queries.
3. Data is dynamic	3. Data is largely static
4. **Entity:** Relational modeling procedures are used for RDBMS database design.	4. **Data:** Modeling approach are used for the Data Warehouse design.
5. Optimized for write operations.	5. Optimized for read operations.
6. Performance is low for analysis queries.	6. High performance for analytical queries.
7. The database is the place where the data is taken as a base and managed to get available fast and efficient access.	7. Data Warehouse is the place where the application data is handled for analysis and reporting objectives.

Difference between Operational Database and Data Warehouse

Operational Database	Data Warehouse
Operational systems are designed to support high-volume transaction processing.	Data warehousing systems are typically designed to support high-volume analytical processing (i.e., OLAP).
Operational systems are usually concerned with current data.	Data warehousing systems are usually concerned with historical data.
Data within operational systems are mainly updated regularly according to need.	Non-volatile, new data may be added regularly. Once Added rarely changed.
It is designed for real-time business dealing and processes.	It is designed for analysis of business measures by subject area, categories, and attributes.
It is optimized for a simple set of transactions, generally adding or retrieving a single row at a time per table.	It is optimized for extent loads and high, complex, unpredictable queries that access many rows per table.
It is optimized for validation of incoming information during transactions, uses validation data tables.	Loaded with consistent, valid information, requires no real-time validation.
It supports thousands of concurrent clients.	It supports a few concurrent clients relative to OLTP.
Operational systems are widely process-oriented.	Data warehousing systems are widely subject-oriented
Operational systems are usually optimized to perform fast inserts and updates of associatively small volumes of data.	Data warehousing systems are usually optimized to perform fast retrievals of relatively high volumes of data.
Data In	Data Out
Less Number of data accessed.	Large Number of data accessed.
Relational databases are created for on-line transactional Processing (OLTP)	Data Warehouse designed for on-line Analytical Processing (OLAP)

Difference between OLTP and OLAP

Feature	OLTP	OLAP
Characteristic	It is a system which is used to manage operational Data.	It is a system which is used to manage informational Data.
Users	Clerks, clients, and information technology professionals.	Knowledge workers, including managers, executives, and analysts.
System orientation	OLTP system is a customer-oriented, transaction, and query processing are done by clerks, clients, and information technology professionals.	OLAP system is market-oriented, knowledge workers including managers, do data analysts executive and analysts.
Data contents	OLTP system manages current data that too detailed and are used for decision making.	OLAP system manages a large amount of historical data, provides facilitates for summarization and aggregation, and stores and manages data at different levels of granularity. This information makes the data more comfortable to use in informed decision making.
Database Size	100 MB-GB	100 GB-TB
Database design	OLTP system usually uses an entity-relationship (ER) data model and application-oriented database design.	OLAP system typically uses either a star or snowflake model and subject-oriented database design.

Difference between OLTP and OLAP

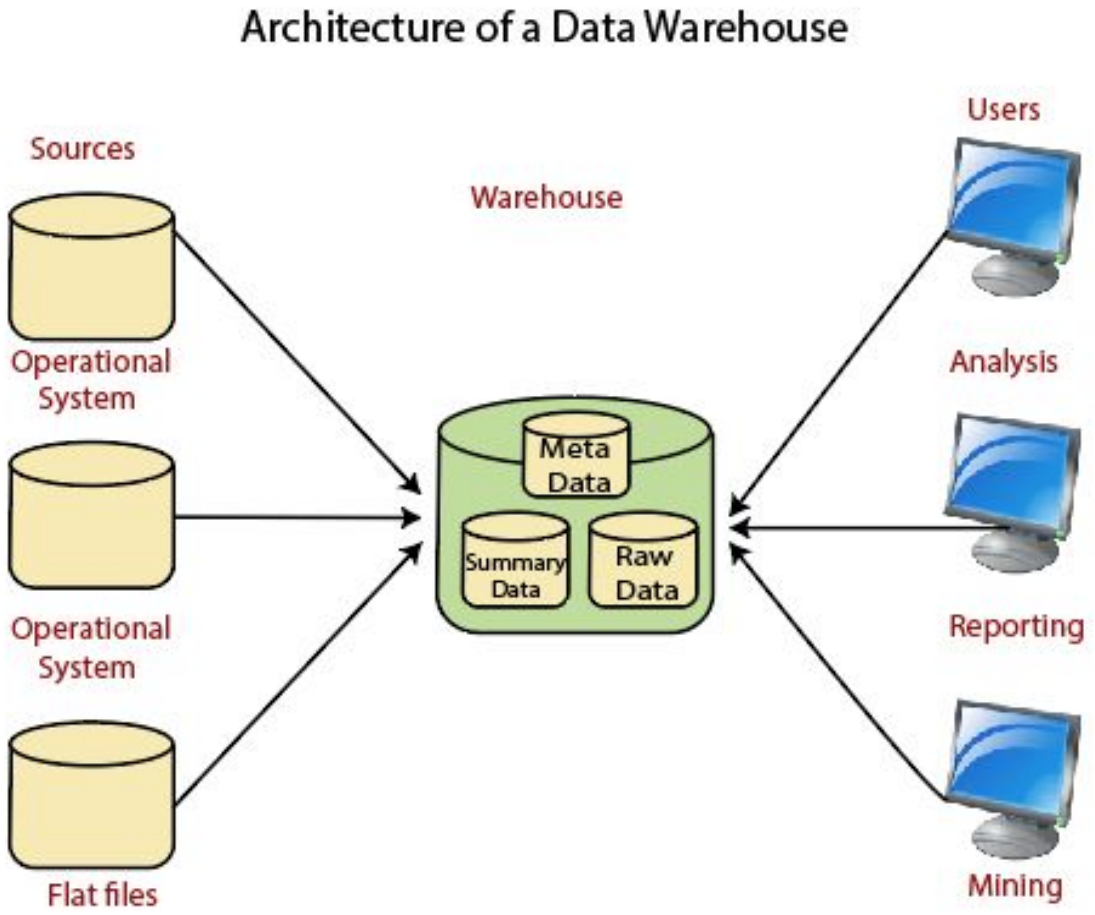
View	OLTP system focuses primarily on the current data within an enterprise or department, without referring to historical information or data in different organizations.	OLAP system often spans multiple versions of a database schema, due to the evolutionary process of an organization. OLAP systems also deal with data that originates from various organizations, integrating information from many data stores.
Volume of data	Not very large	Because of their large volume, OLAP data are stored on multiple storage media.
Access patterns	The access patterns of an OLTP system subsist mainly of short, atomic transactions. Such a system requires concurrency control and recovery techniques.	Accesses to OLAP systems are mostly read-only methods because of these data warehouses stores historical data.
Access mode	Read/write	Mostly write
Insert and Updates	Short and fast inserts and updates proposed by end-users.	Periodic long-running batch jobs refresh the data.
Number of records accessed	Tens	Millions
Normalization	Fully Normalized	Partially Normalized
Processing Speed	Very Fast	It depends on the amount of files contained, batch data refresh, and complex query may take many hours, and query speed can be upgraded by creating indexes.

Data Warehouse Architecture

- A data warehouse architecture is a method of defining the overall architecture of data communication processing and presentation that exist for end-clients computing within the enterprise. Each data warehouse is different, but all are characterized by standard vital components.

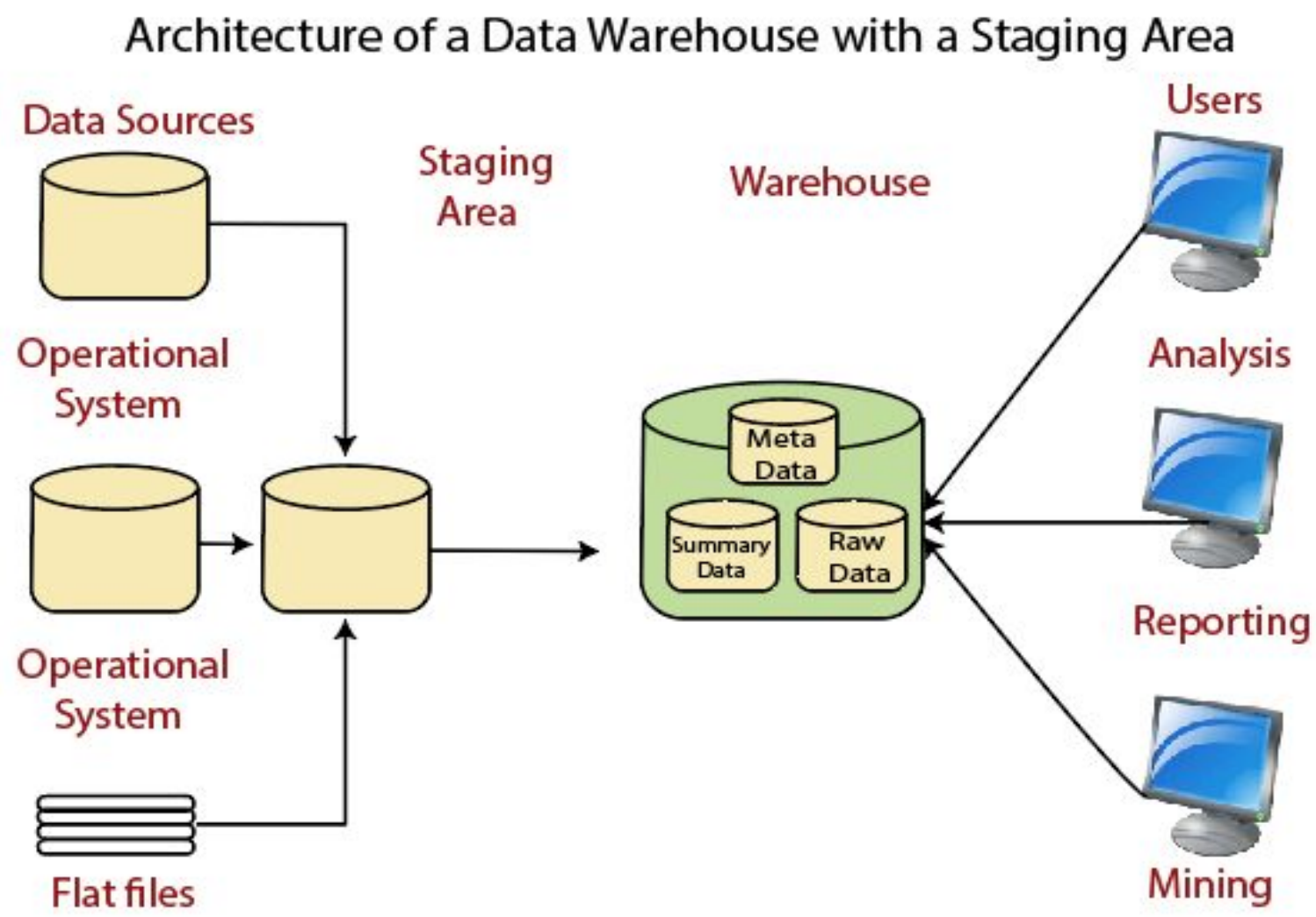
- Data warehouses and their architectures very depending upon the elements of an organization's situation.
- Three common architectures are:
- Data Warehouse Architecture: Basic
- Data Warehouse Architecture: With Staging Area
- Data Warehouse Architecture: With Staging Area and Data Marts

Data Warehouse Architecture: Basic

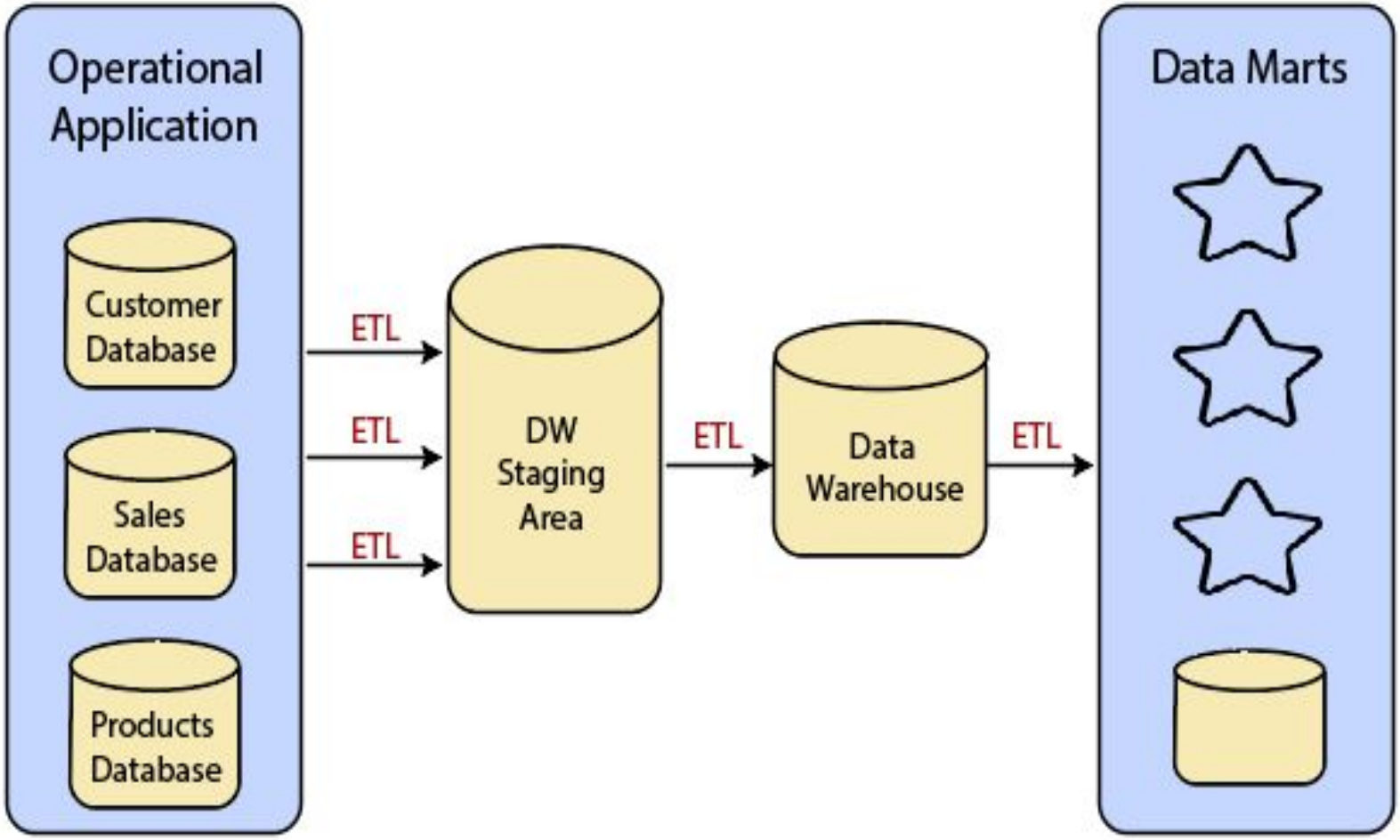


Operational System

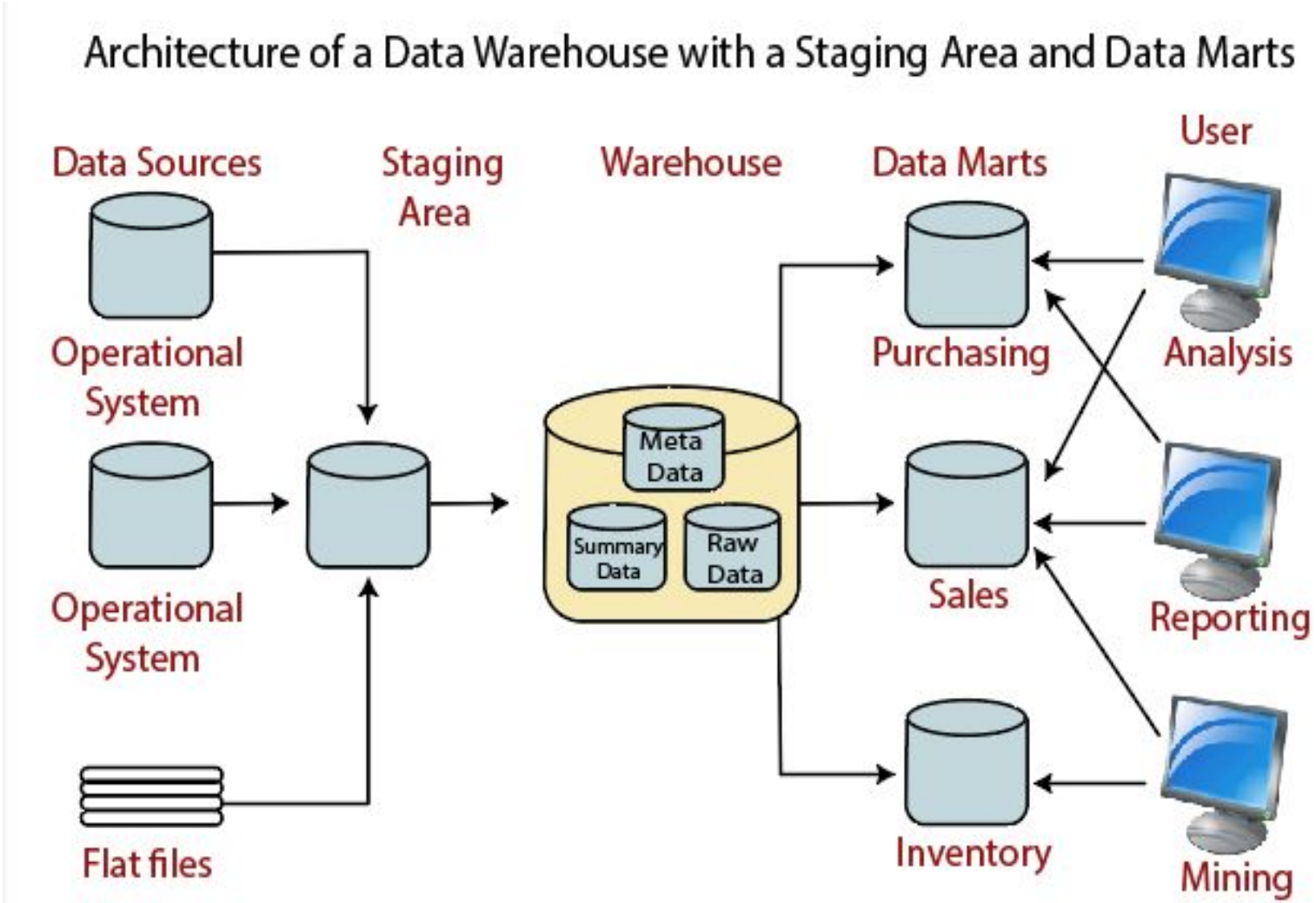
Data Warehouse Architecture: With Staging Area



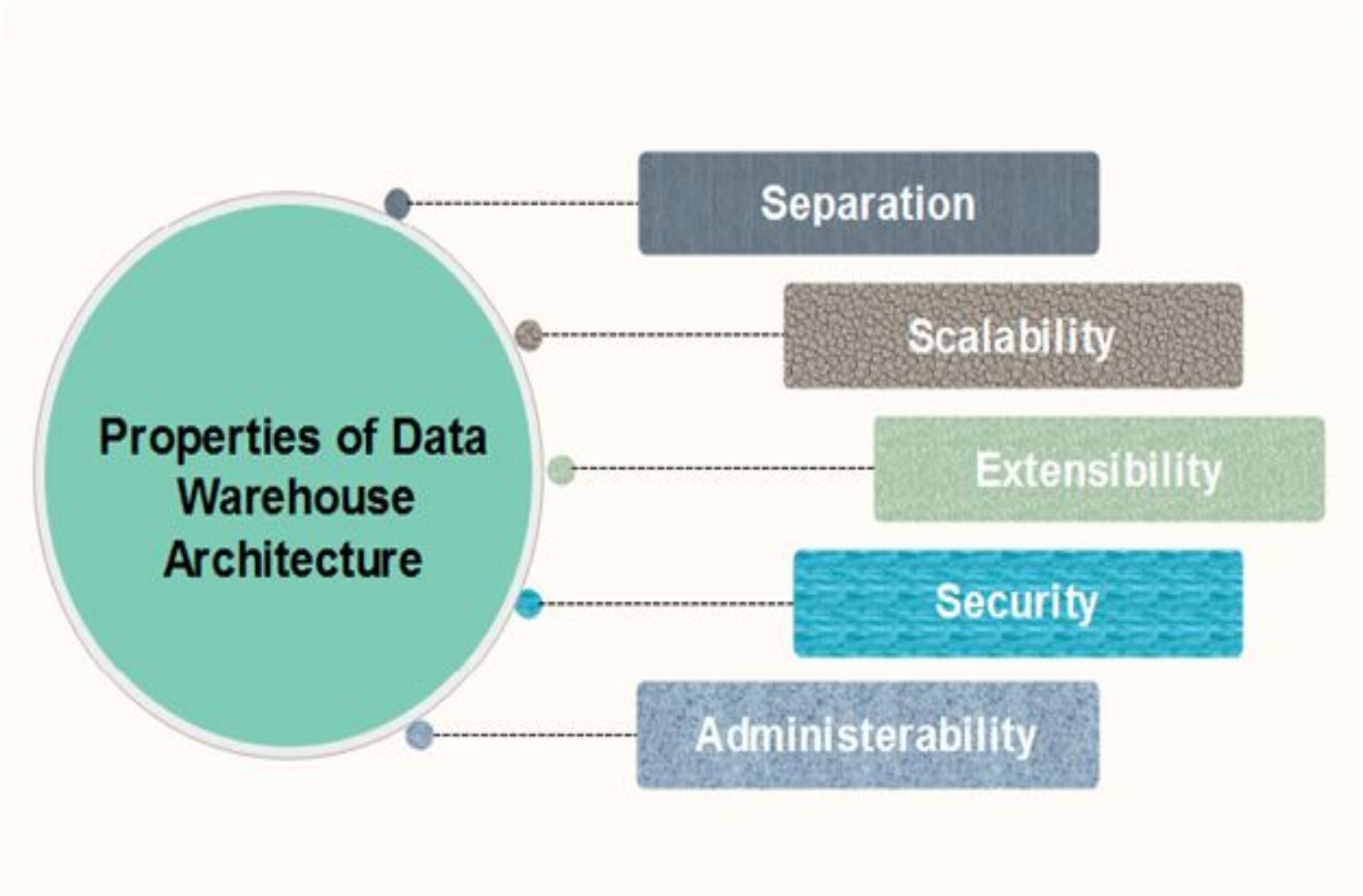
Data Warehouse Staging Area is a temporary location where a record from source systems is copied.



Data Warehouse Architecture: With Staging Area and Data Marts

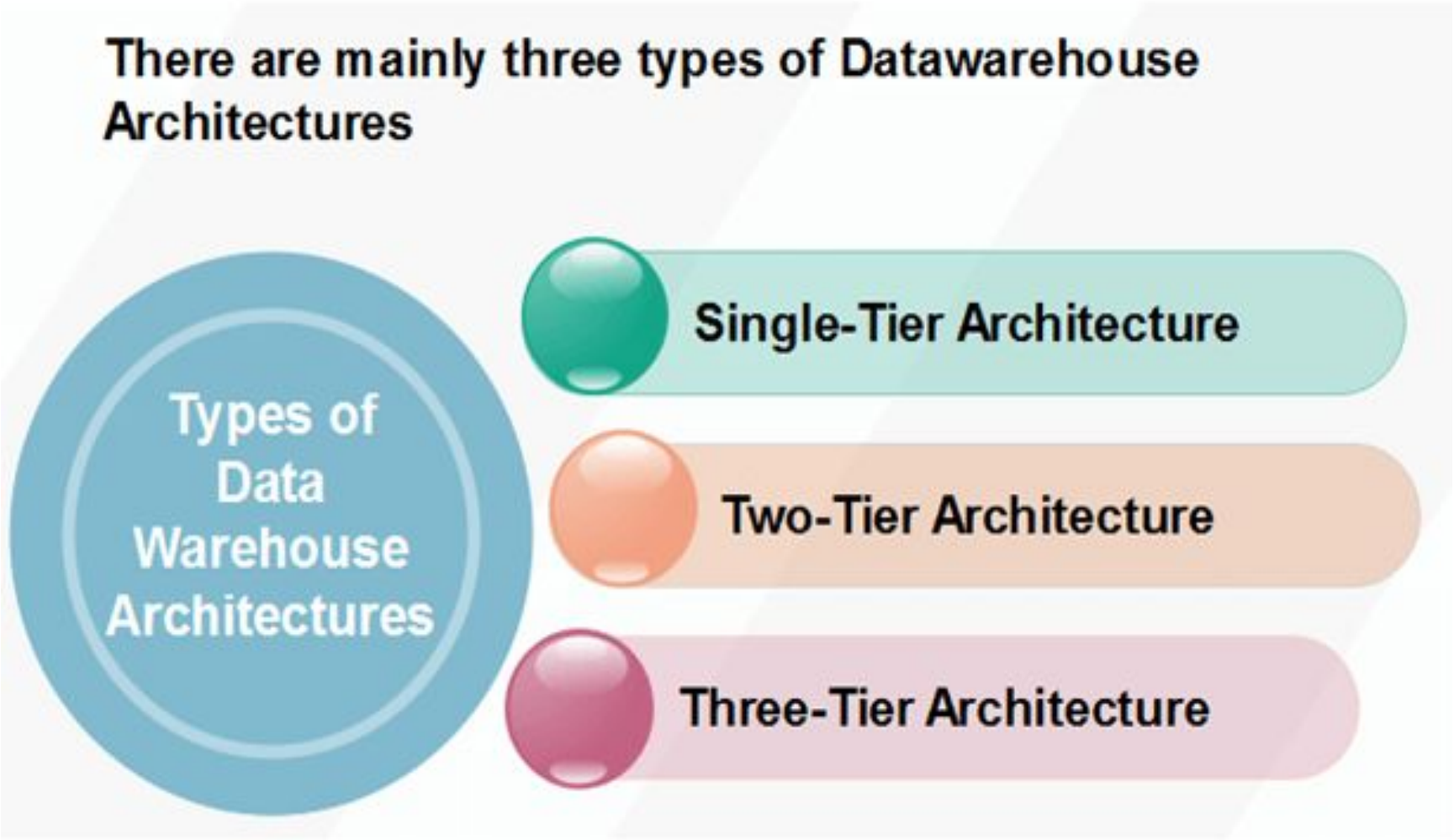


Properties of Data Warehouse Architectures

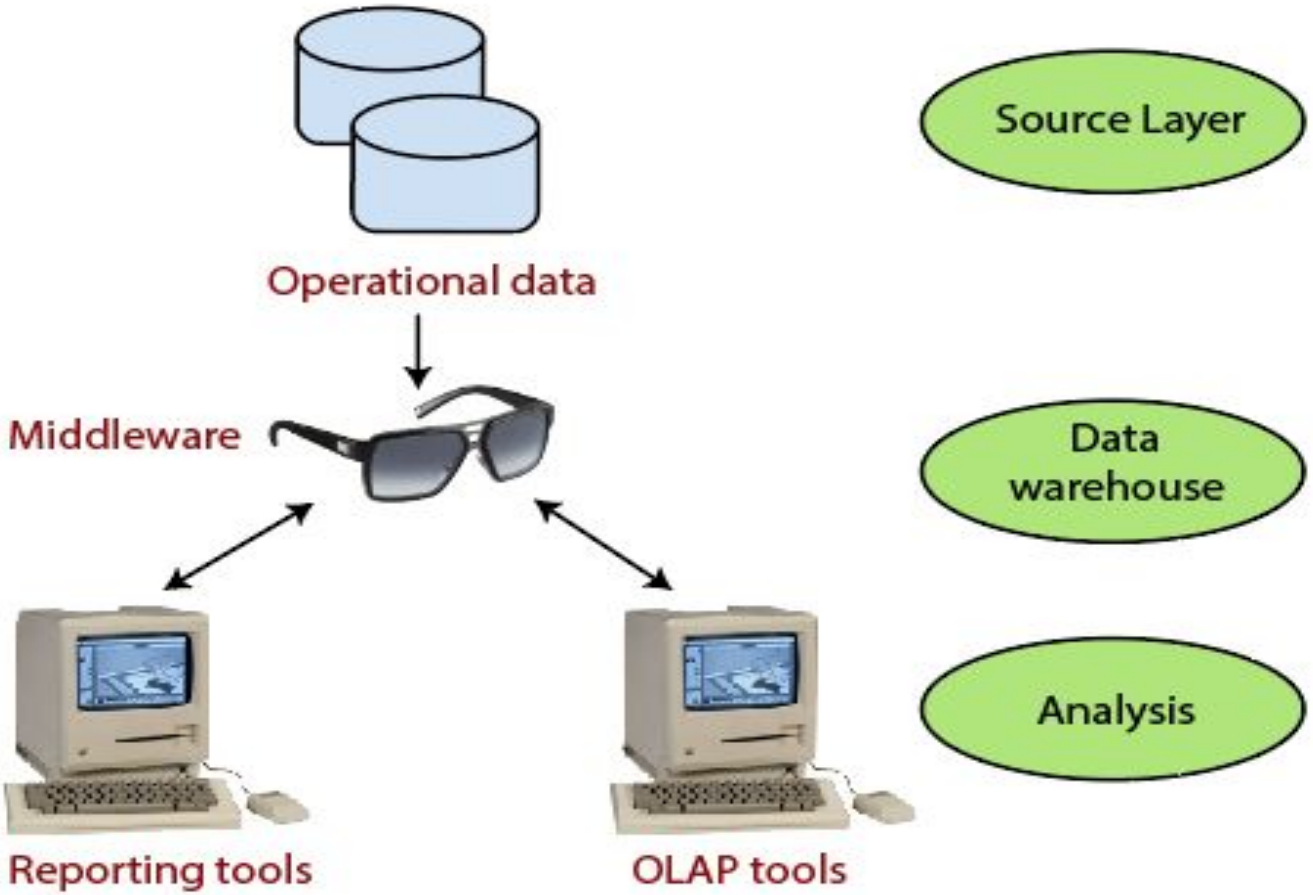


- **1. Separation:** Analytical and transactional processing should be keep apart as much as possible.
- **2. Scalability:** Hardware and software architectures should be simple to upgrade the data volume, which has to be managed and processed, and the number of user's requirements, which have to be met, progressively increase.
- **3. Extensibility:** The architecture should be able to perform new operations and technologies without redesigning the whole system.
- **4. Security:** Monitoring accesses are necessary because of the strategic data stored in the data warehouses.
- **5. Administerability:** Data Warehouse management should not be complicated.

Types of Data Warehouse Architectures

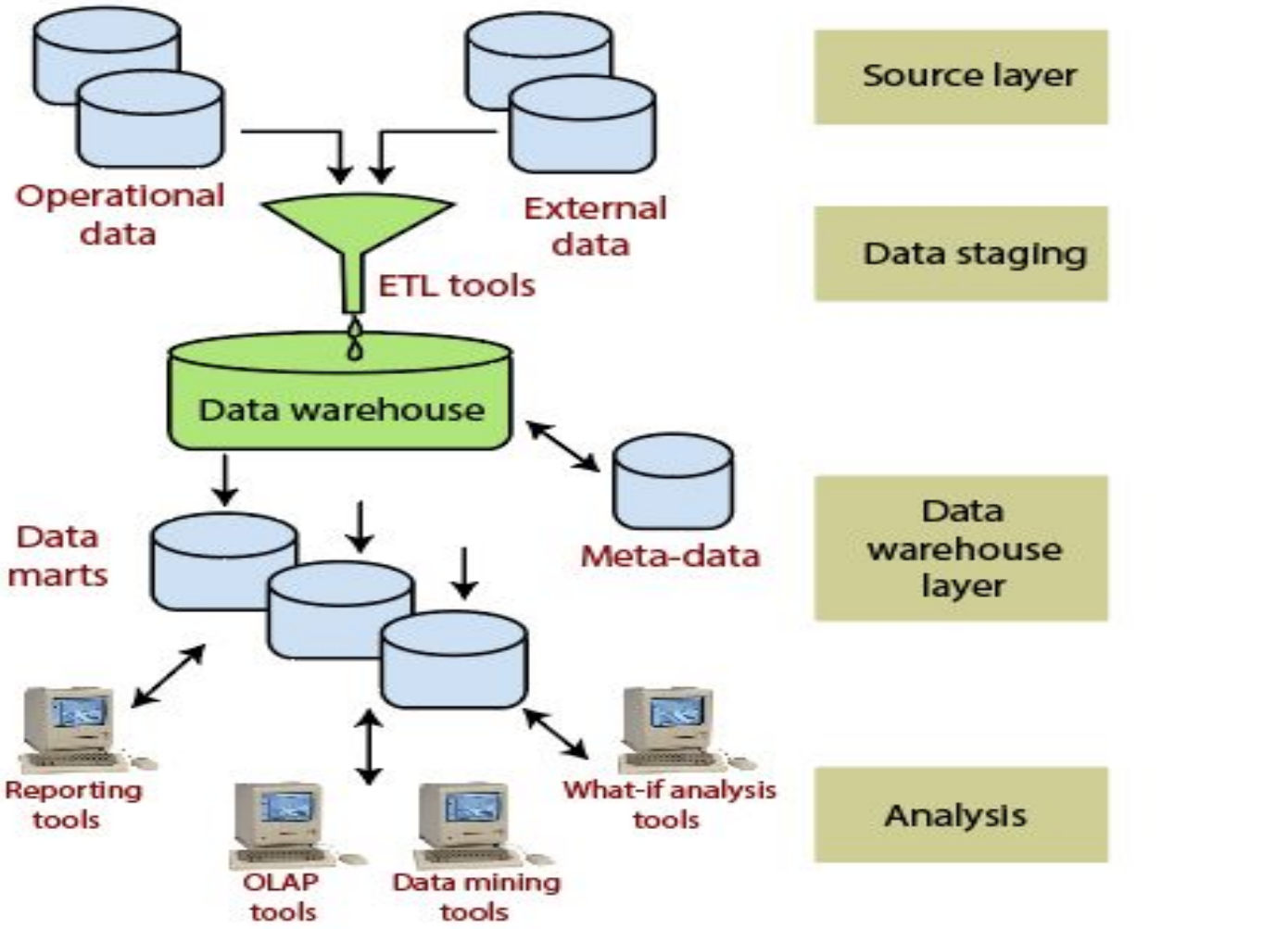


Single-Tier Architecture



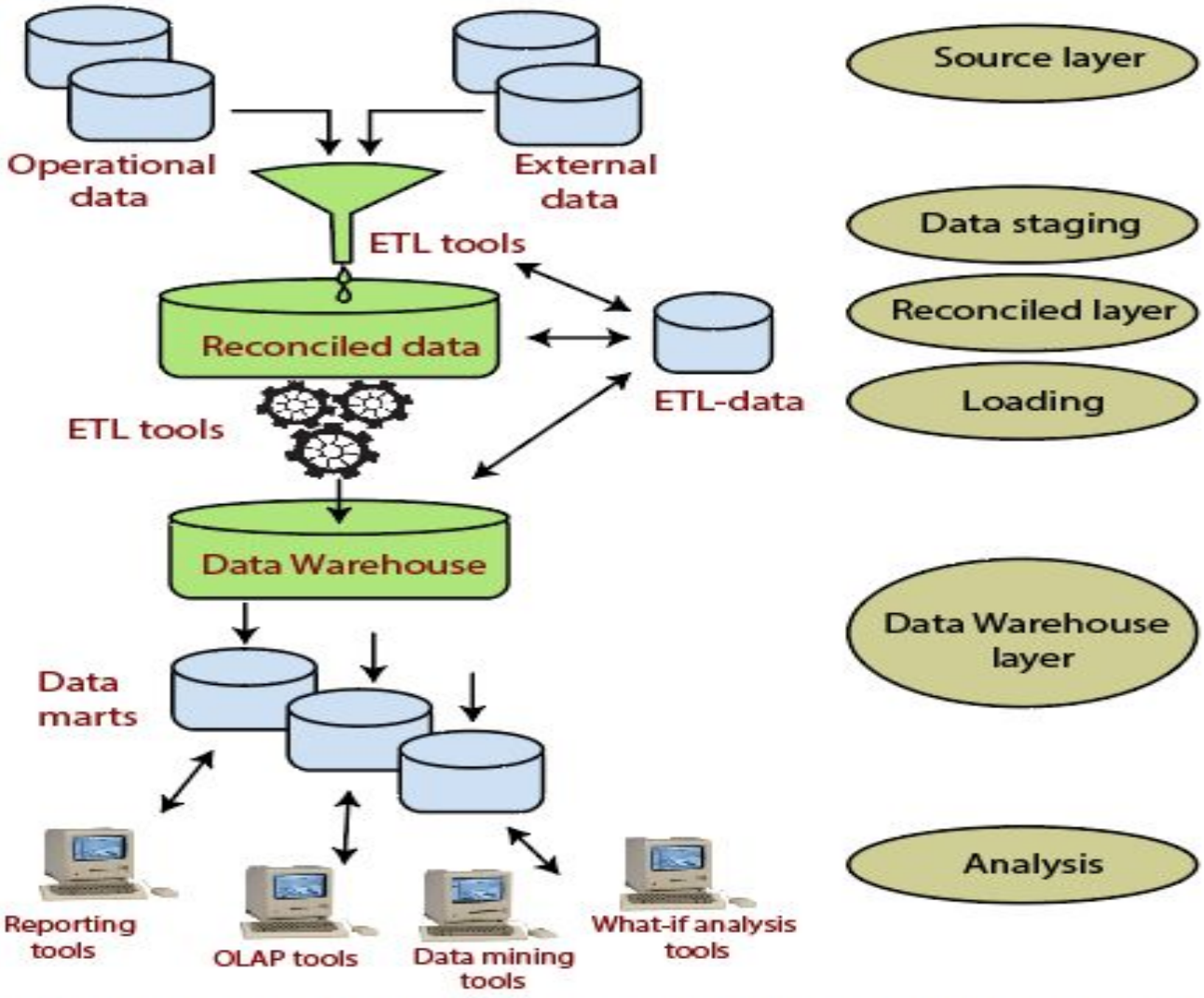
Single-Tier Data Warehouse Architecture

Two-Tier Architecture



Two-Tier Data Warehouse Architecture

Three-Tier Architecture



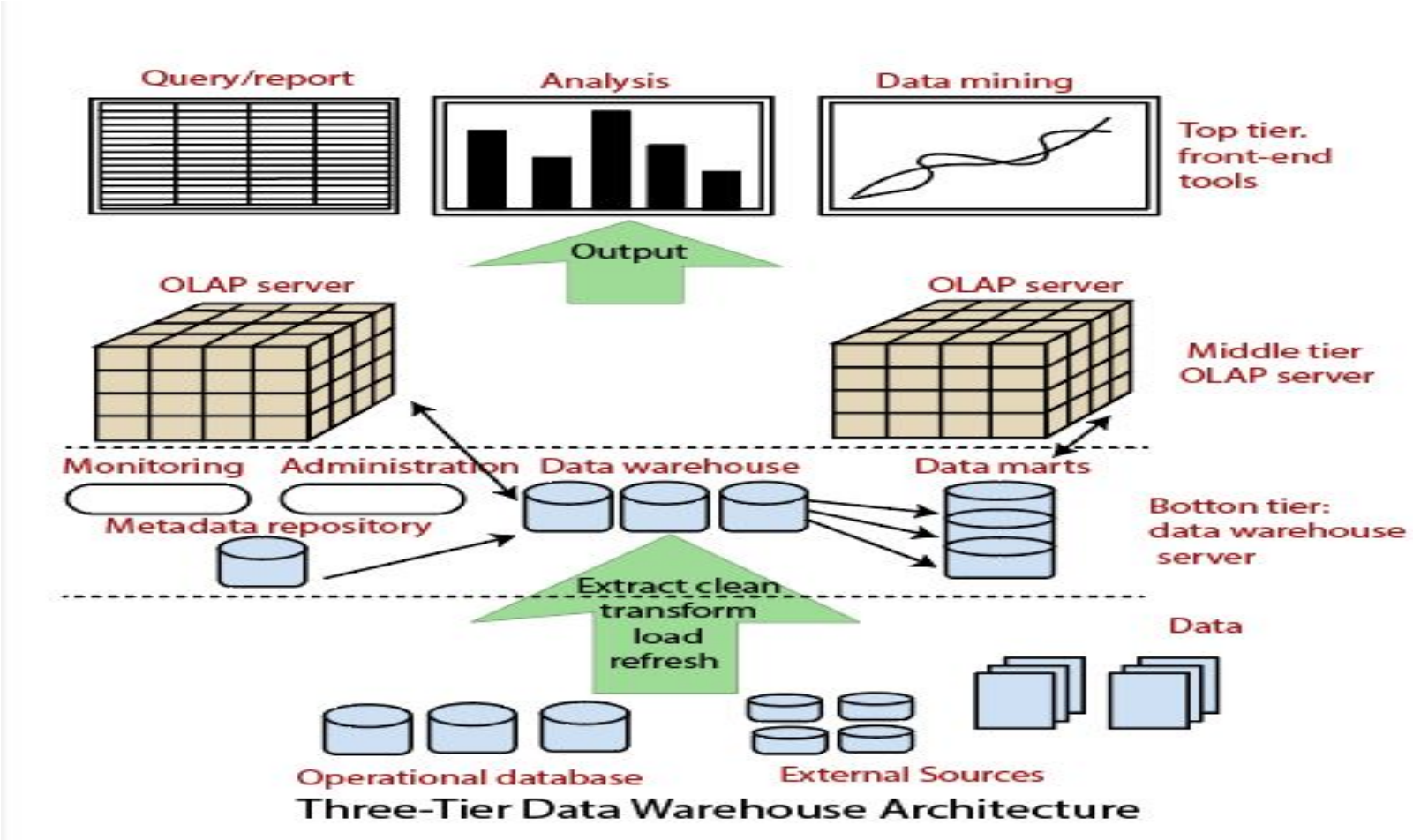
Three-Tier Architecture for a data warehouse system

- Data Warehouses usually have a three-level (tier) architecture that includes:
- Bottom Tier (Data Warehouse Server)
- Middle Tier (OLAP Server)
- 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).

- 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.

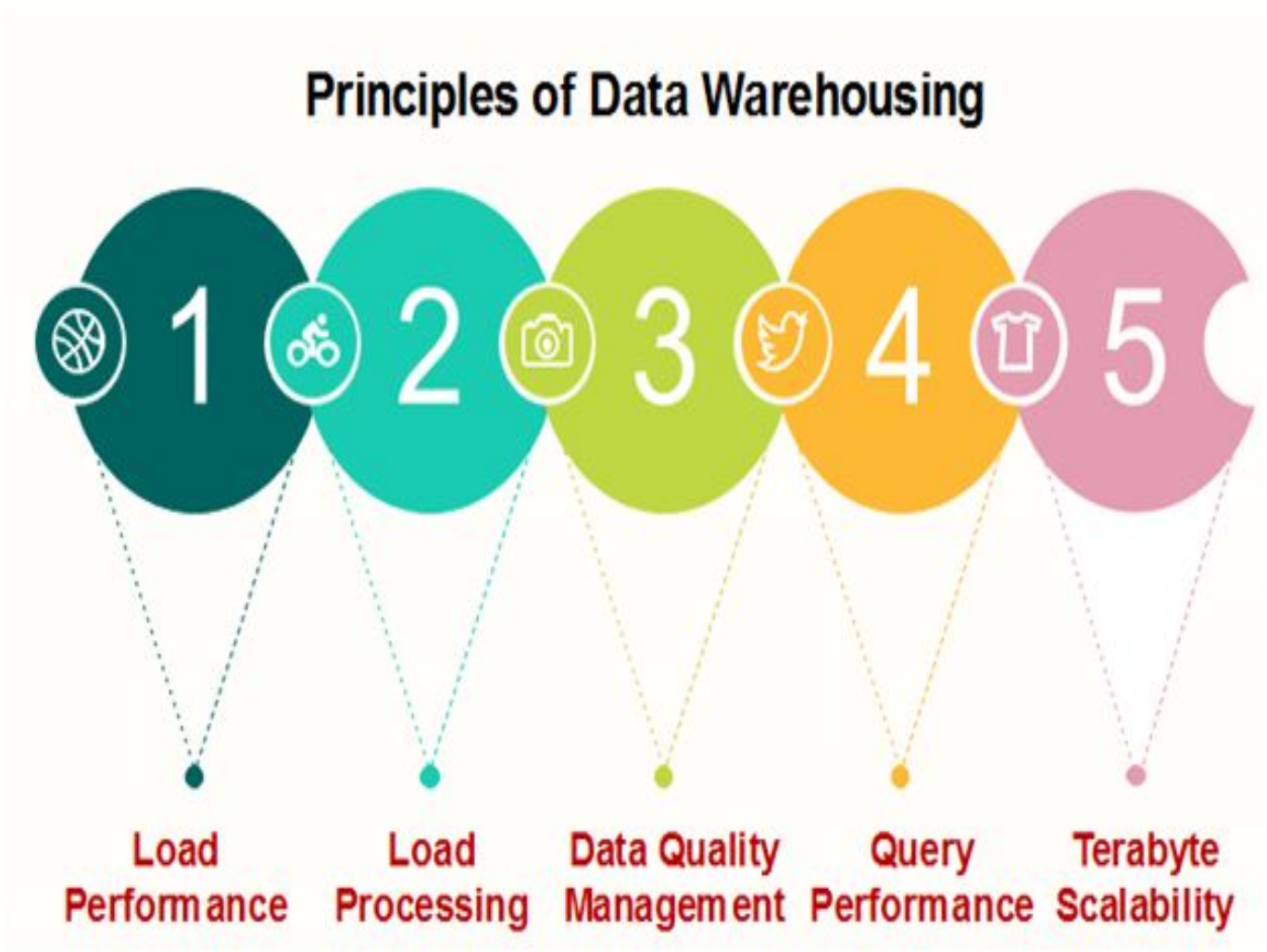
The overall Data Warehouse Architecture



Metadata in data warehouse

- The **metadata repository** stores information that defines DW objects. It includes the following parameters and information for the middle and the top-tier applications:
- A description of the DW structure, including the warehouse schema, dimension, hierarchies, data mart locations, and contents, etc.
- 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.
- System performance data, which includes indices, used to improve data access and retrieval performance.
- Information about the mapping from operational databases, which provides source **RDBMSs** and their contents, cleaning and transformation rules, etc.
- 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.