DATA WAREHOUSING AND DATA MINING

What is a Data Warehouse?

- A data warehouse (DW or DWH) is a complex system that stores historical and cumulative data used for forecasting, reporting, and data analysis.
- It involves collecting, cleansing, and transforming data from different data streams and loading it into fact/dimensional tables.
- A data warehouse represents a subject-oriented, integrated, time-variant, and non-volatile structure of data.

Data Warehouse

- A data-warehouse is a heterogeneous collection of different data sources organised under a unified schema.
- 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 Warehouse Architecture

- There are three ways we can construct a data warehouse system.
- These approaches are classified by the number of tiers in the architecture.

1. Single-tier architecture
2. Two-tier architecture
3. Three-tier architecture

Data warehouses and their architectures very **depending upon the elements** of an organization's situation.

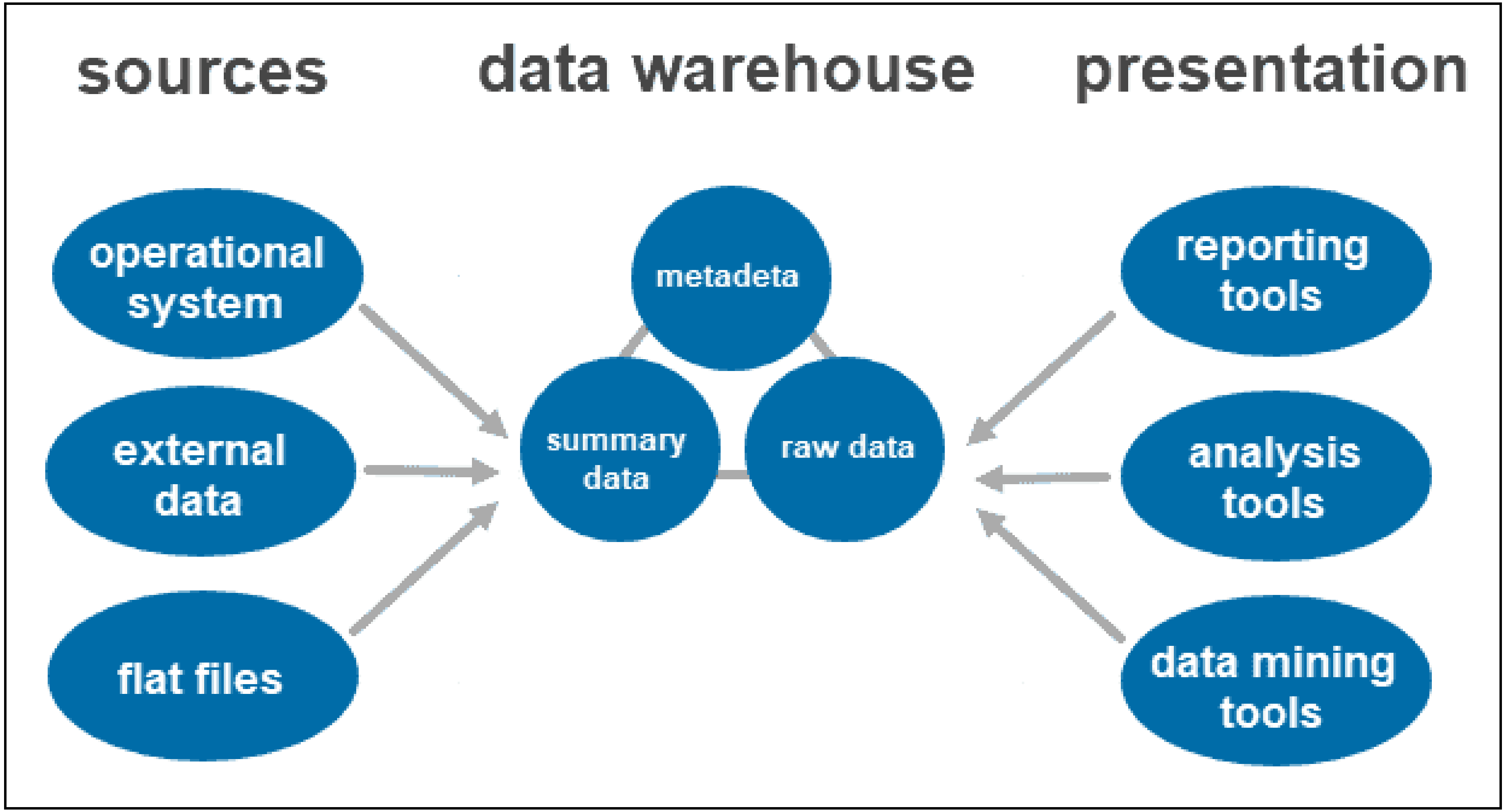
Three common architectures are:

1. Data Warehouse Architecture: **Basic**
2. Data Warehouse Architecture: **With Staging Area**
3. Data Warehouse Architecture: **With Staging Area and Data Marts**

Single-tier Data Warehouse Architecture:

- The single-tier architecture is not a frequently practiced approach.
- The main goal of having such an architecture is to remove redundancy by minimizing the amount of data stored.

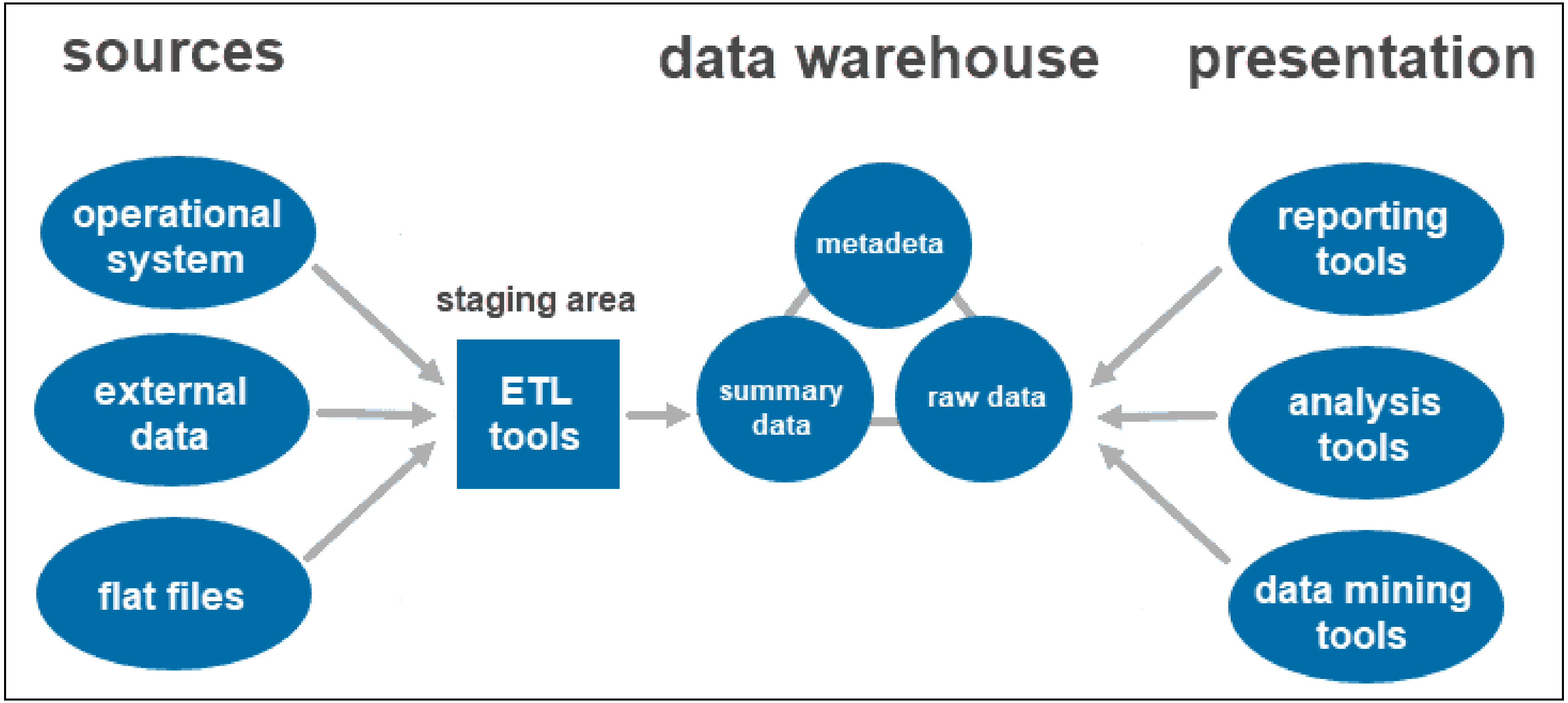
Single-tier


sources
data warehouse
presentation
operational system
external data
flat files
metadeta
summary data
raw data
reporting tools
analysis tools
data mining tools

Two-tier Data Warehouse Architecture:

- A two-tier architecture includes a staging area for all data sources, before the data warehouse layer.
- By adding a staging area between the sources and the storage repository, you ensure all data loaded into the warehouse is cleansed and in the appropriate format.

Two-tier


sources
data warehouse
presentation
operational system
external data
flat files
staging area
ETL tools
metadeta
summary data
raw data
reporting tools
analysis tools
data mining tools

Three-tier Data Warehouse Architecture:

The three-tier approach is the most widely used architecture for data warehouse systems.

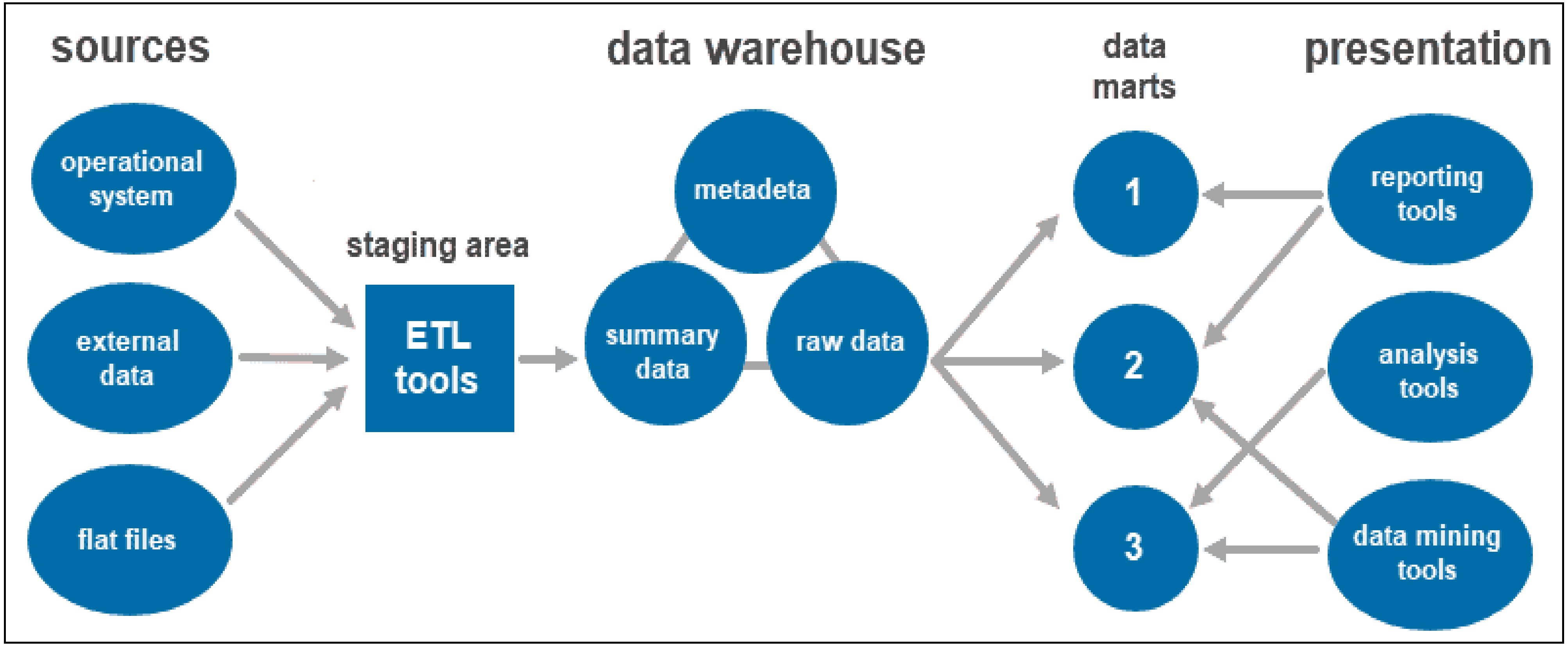
Essentially, it consists of **three tiers**:

The **bottom tier** is the database of the warehouse, where the cleansed and transformed data is loaded.

The **middle tier** is the application layer giving an abstracted view of the database. It arranges the data to make it more suitable for analysis. This is done with an OLAP server, implemented using the ROLAP or MOLAP model.

The **top-tier** is where the user accesses and interacts with the data. It represents the front-end client layer. You can use reporting tools, query, analysis or data mining tools.

Three-tier



sources
operational system
external data
flat files
staging area
ETL tools
data warehouse
metadeta
summary data
raw data
data marts
1
2
3
presentation
reporting tools
analysis tools
data mining tools

Data Warehouse Components:

1. ETL Tools
2. The Database
3. Data
4. Access Tools
5. Data Marts

ETL Tools:

- ETL stands for **E**xtract, **T**ransform, and **L**oad.
- The staging layer uses ETL tools to extract the needed data from various formats and checks the quality before loading it into the data warehouse.
- The data coming from the data source layer can come in a variety of formats. Before merging all the data collected from multiple sources into a single database, the system must clean and organize the information.

The Database

- The most crucial component and the heart of each architecture is the database.
- The warehouse is where the data is stored and accessed.
- When creating the data warehouse system, you first need to decide what kind of database you want to use.

There are **four types** of databases you can choose from:

1. **Relational** databases (row-centered databases).
2. **Analytics** databases (developed to sustain and manage analytics).
3. Data **warehouse** applications (software for data management and hardware for storing data offered by third-party dealers).
4. **Cloud-based** databases (hosted on the cloud).

Data

- Once the system cleans and organizes the data, it stores it in the data warehouse.
- The data warehouse represents the central repository that stores metadata, summary data, and raw data coming from each source.

- **Metadata** is the information that defines the data. Its primary role is to simplify working with data instances. It allows data analysts to classify, locate, and direct queries to the required data.

- **Summary data** is generated by the warehouse manager. It updates as new data loads into the warehouse. This component can include lightly or highly summarized data. Its main role is to speed up query performance.

- **Raw data** is the actual data loading into the repository, which has not been processed. Having the data in its raw form makes it accessible for further processing and analysis.

Access Tools

- Users interact with the gathered information through different tools and technologies.
- They can analyze the data, gather insight, and create reports.

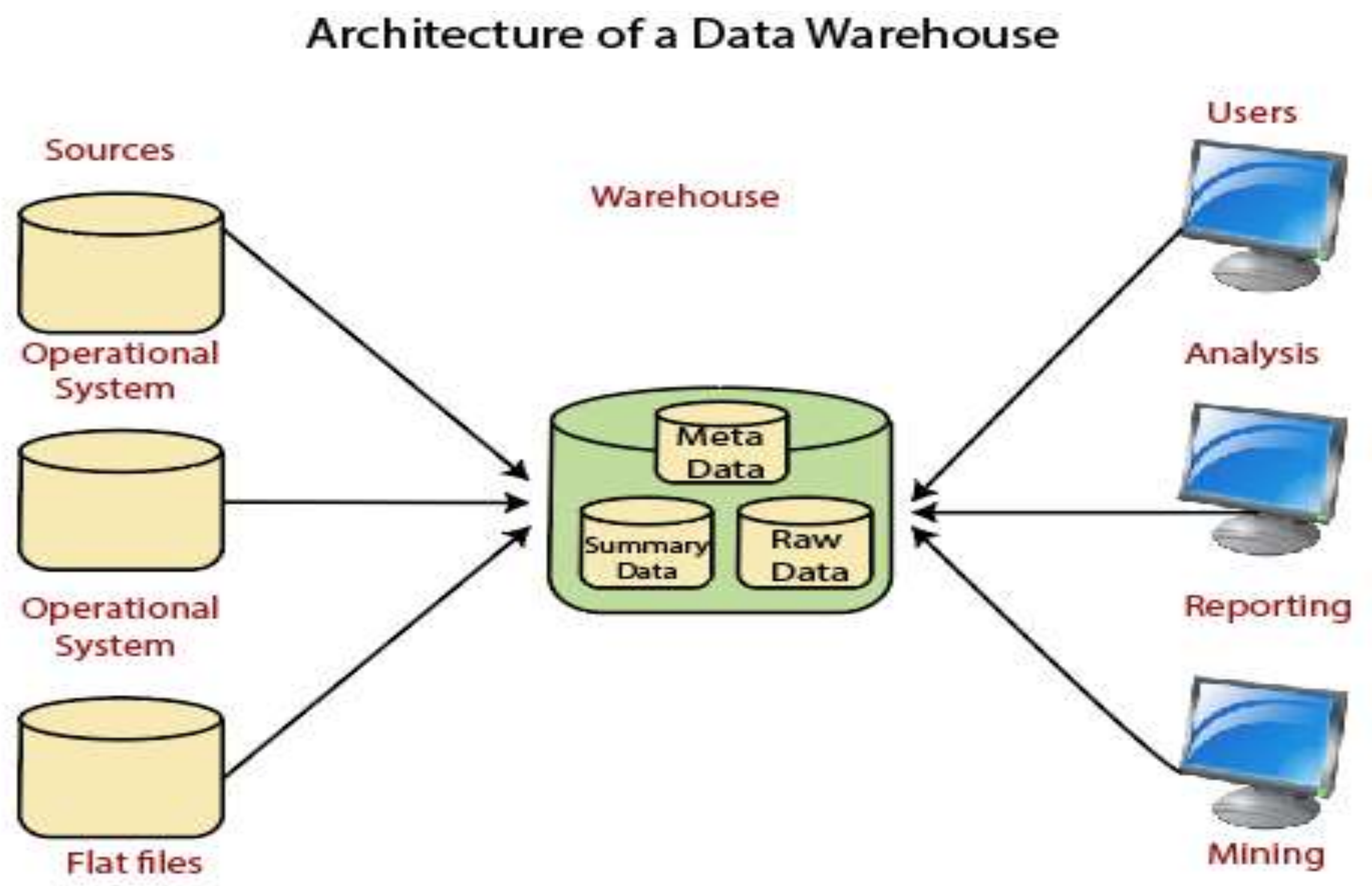
Some of the **tools used include**:

- **Reporting tools.** They play a crucial role in understanding how your business is doing and what should be done next. Reporting tools include visualizations such as graphs and charts showing how data changes over time.
- **OLAP tools.** Online analytical processing tools which allow users to analyze multidimensional data from multiple perspectives. These tools provide fast processing and valuable analysis. They extract data from numerous relational data sets and reorganize it into a multidimensional format.
- **Data mining tools.** Examine data sets to find patterns within the warehouse and the correlation between them. Data mining also helps establish relationships when analyzing multidimensional data.

Data Marts

- Data marts allow you to have multiple groups within the system by segmenting the data in the warehouse into categories.
- It partitions data, producing it for a particular user group.

- Data Warehouse Architecture: Basic

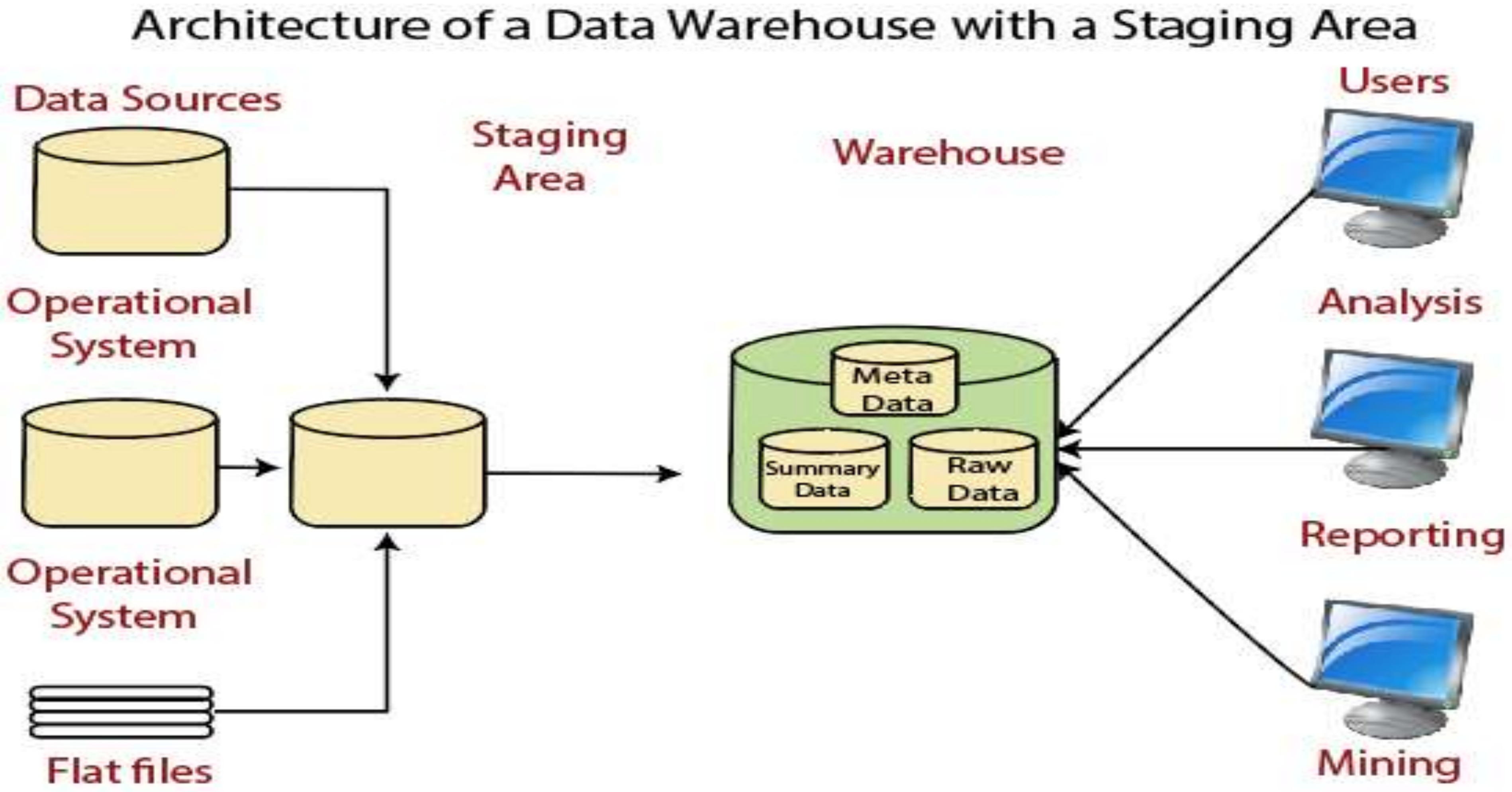


- **Operational System:** An operational system is a method used in data warehousing to refer to a system that is used to process the day-to-day transactions of an organization.
- **Flat Files:** A Flat file system is a system of files in which transactional data is stored, and every file in the system must have a different name.
- **Meta Data:** A set of data that defines and gives information about other data.

Data Warehouse Architecture: With Staging Area

- A staging area simplifies data cleansing and consolidation for operational method coming from multiple source systems, especially for enterprise data warehouses where all relevant data of an enterprise is consolidated.

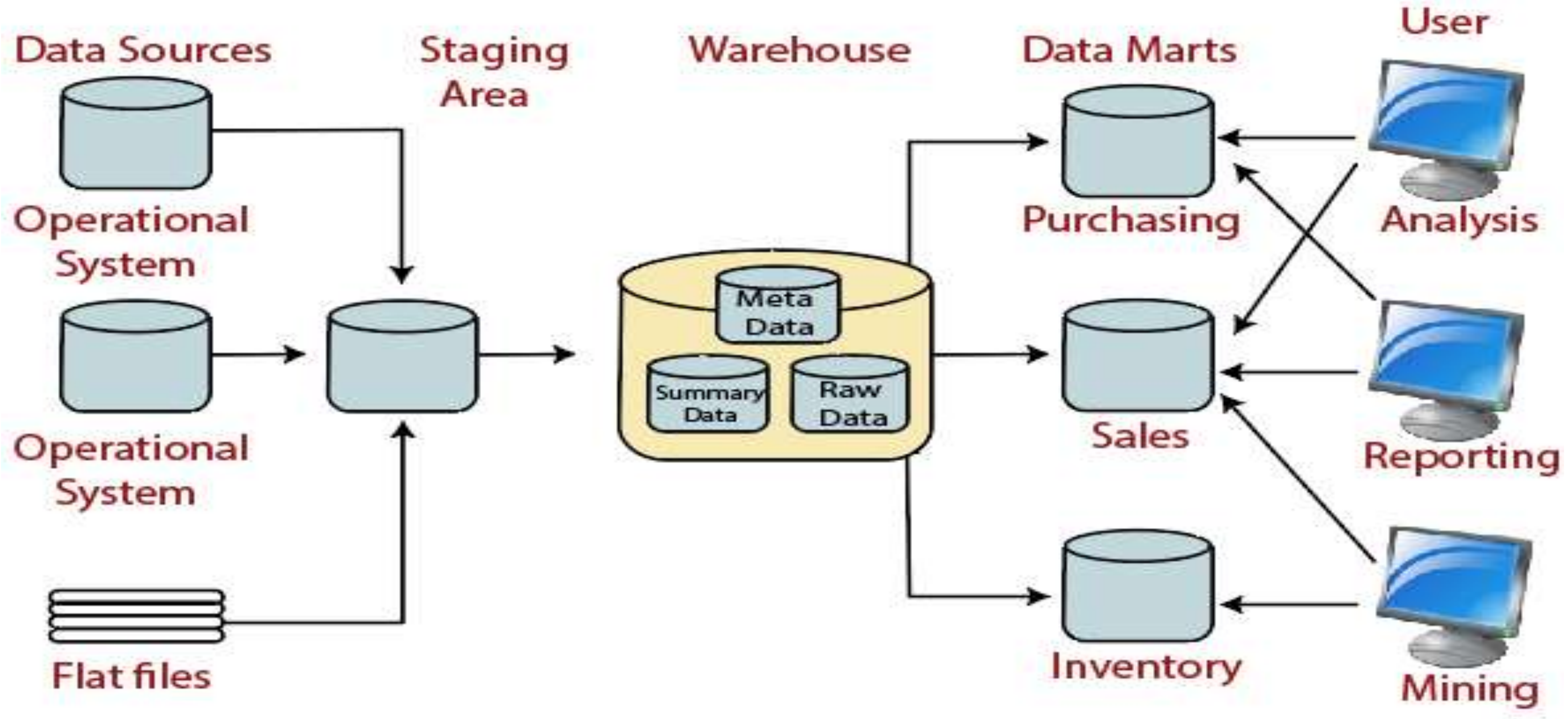
With Staging Area : Data Warehouse Staging Area is a temporary location where a record from source systems is copied.



Data Warehouse Architecture: With Staging Area and Data Marts:

A data mart is a segment of a data warehouses that can provided information for reporting and analysis on a section, unit, department or operation in the company, e.g., sales, payroll, production, etc.

Architecture of a Data Warehouse with a Staging Area and Data Marts



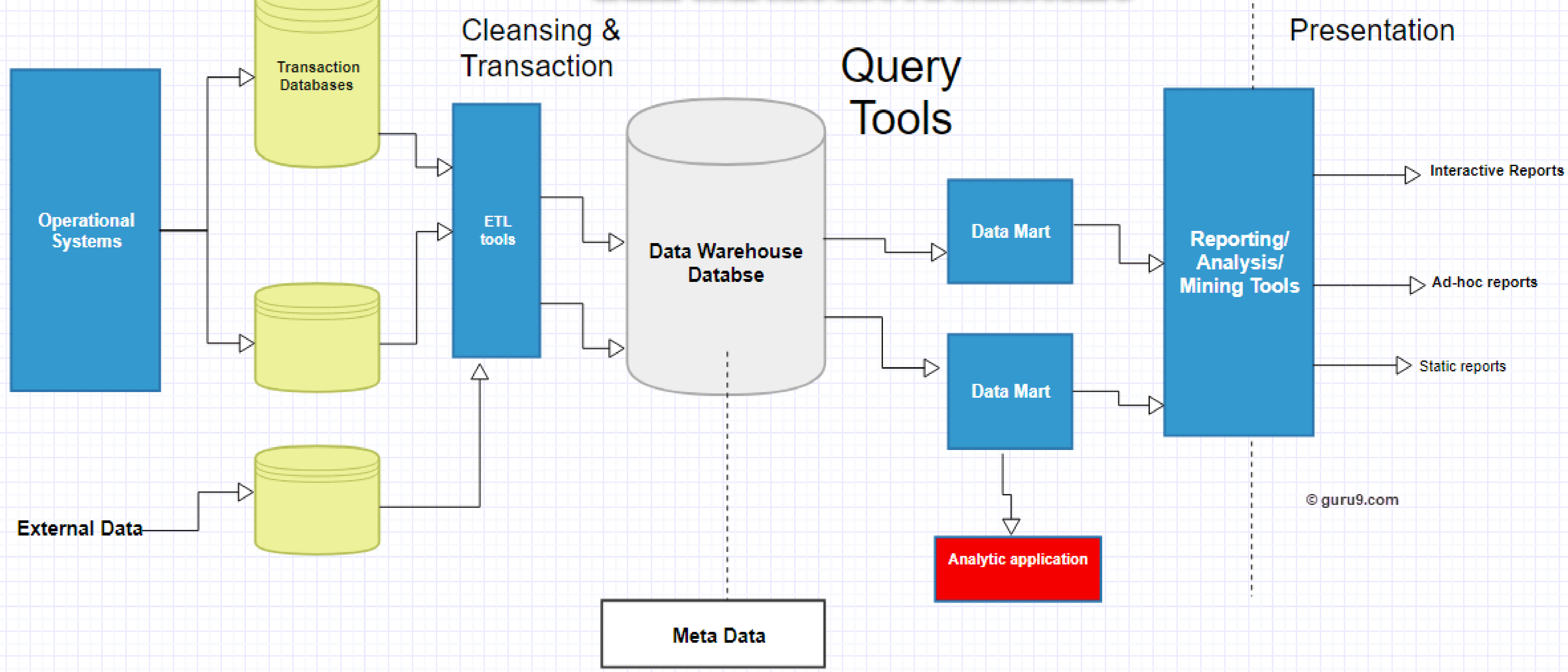
Datawarehouse Components:

- The Data Warehouse is based on an RDBMS server which is a central information repository that is surrounded by some key Data Warehousing components to make the entire environment functional, manageable and accessible.

There are mainly **five** Data Warehouse **Components**:

1. Data Warehouse Database
2. Sourcing, Acquisition, Clean-up and Transformation Tools (ETL)
3. Metadata
4. Query Tools
5. Data Marts

Data Warehouse Architecture



© guru9.com

Data Warehouse Database

- In a datawarehouse, relational databases are deployed in parallel to allow for scalability.
- Parallel relational databases also allow shared memory or shared nothing model on various multiprocessor configurations or massively parallel processors.
- New index structures are used to bypass relational table scan and improve speed.
- Use of multidimensional database (MDDBs) to overcome any limitations which are placed because of the relational Data Warehouse Models. Example: Essbase from Oracle.

ETL:

- The data sourcing, transformation, and migration tools are used for performing all the conversions, summarizations, and all the changes needed to transform data into a unified format in the Datawarehouse.
- They are also called Extract, Transform and Load (ETL) Tools.

Metadata

- The name Meta Data suggests some high-level technological Data Warehousing Concepts.
- Metadata is data about data which defines the data warehouse.
- It is used for building, maintaining and managing the data warehouse.
- In the Data Warehouse Architecture, meta-data plays an important role as it specifies the source, usage, values, and features of data warehouse data.
- It also defines how data can be changed and processed.
- It is closely connected to the data warehouse.

Query Tools

- One of the primary objects of data warehousing is to provide information to businesses to make strategic decisions.
- Query tools allow users to interact with the data warehouse system.

Data Marts

- A data mart is an access layer which is used to get data out to the users.
- In a simple word Data mart is a subsidiary of a data warehouse.
- The data mart is used for partition of data which is created for the specific group of users.
- Data marts could be created in the same database as the Datawarehouse or a physically separate Database.