3.5 Data Warehousing

3.5.1 Data warehouse

- Data warehouse is data management and data analysis
- Goal: is to integrate enterprise wide corporate data into a single repository from which users can easily run queries

3.5.2 Benefits

- The major benefit of data warehousing are high returns on investment.
- Increased productivity of corporate decision-makers

3.5.3 Problems

- Underestimation of resources for data loading
- Hidden problems with source systems
- Required data not captured
- Increased end-user demands
- Data homogenization
- High demand for resources
- Data ownership
- High maintenance
- Long-duration projects
- Complexity of integration

3.5.4 Main components

1. Operational data sources for the DW is supplied from mainframe operational data held in first generation hierarchical and network databases, departmental data held in proprietary file systems, private data held on workstaions and private serves and

external systems such as the Internet, commercially available DB, or DB assoicated with and organization ‘s suppliers or customers.

2. Operational datastore (ODS)is a repository of current and integrated operational data used for analysis. It is often structured and supplied with data in the same way as the data warehouse, but may in fact simply act as a staging area for data to be moved into the warehouse.
3. query manager also called backend component, it performs all the operations associated with the management of user queries. The operations performed by this component include directing queries to the appropriate tables and scheduling the execution of queries.
4. End-user access tools can be categorized into five main groups: data reporting and query tools, application development tools, executive information system (EIS) tools, online analytical processing (OLAP) tools, and data mining tools.

3.5.5 Data flow

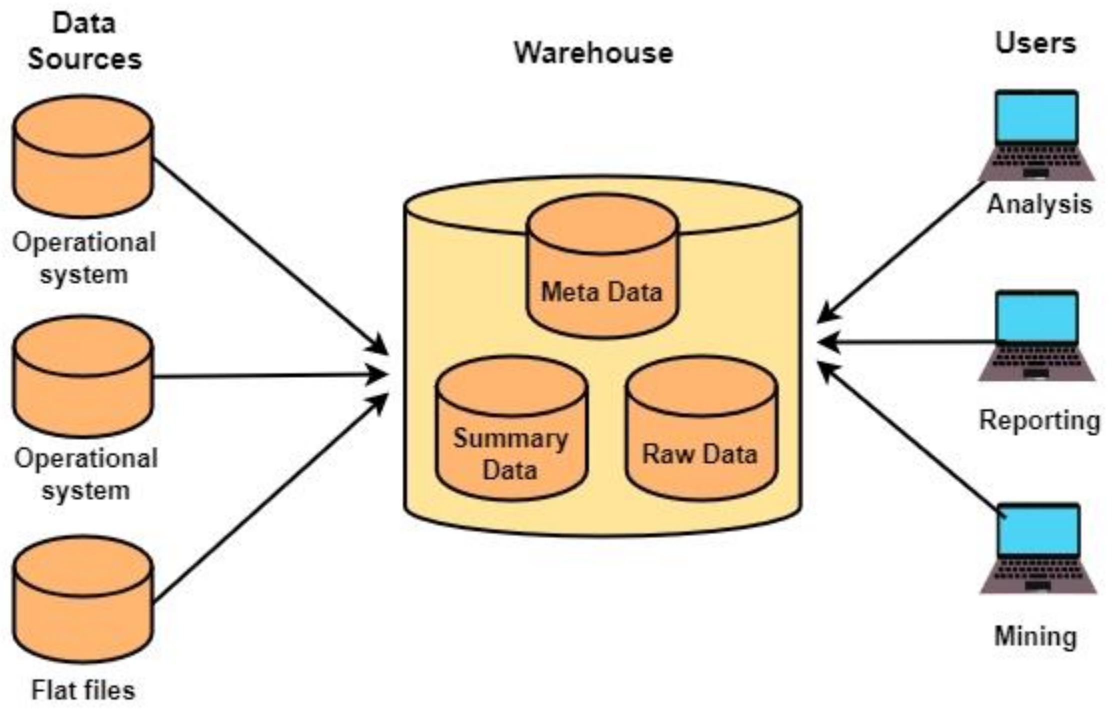
- Inflow- The processes associated with the extraction, cleansing, and loading of the data from the source systems into the data warehouse.
- upflow- The process associated with adding value to the data in the warehouse through summarizing, packaging, and distribution of the data.
- downflow- The processes associated with archiving and backing-up of data in the warehouse.

3.5.6 Tools and Technologies

The critical steps in the construction of a data warehouse:

- Extraction
- Cleansing
- Transformation

after the critical steps, loading the results into target system can be carried out either by separate products, or by a single, categories:



- code generators
- database data replication tools
- dynamic transformation engines

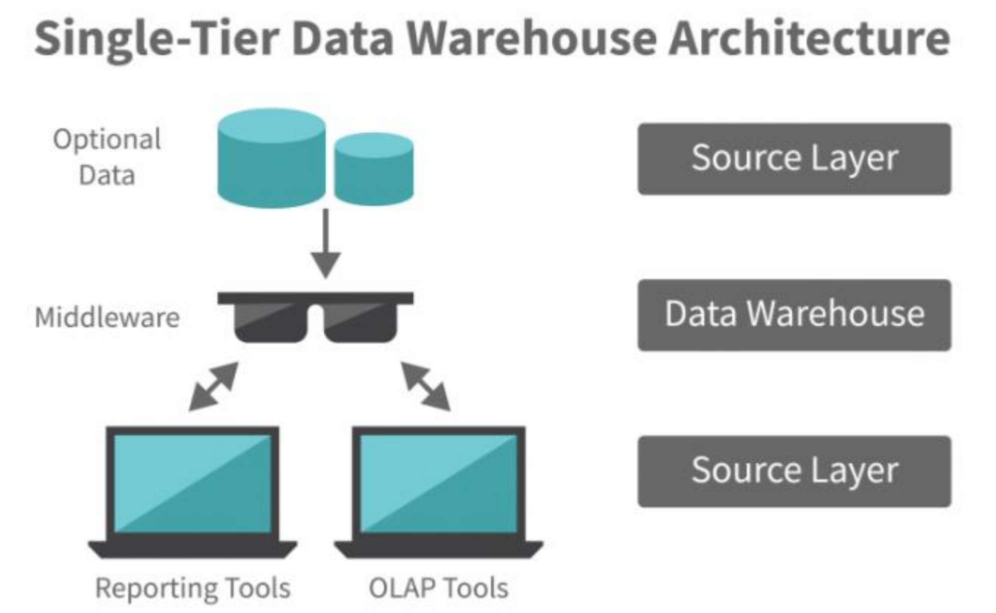
For the various types of meta-data and the day-to-day operations of the data warehouse, the administration and management tools must be capable of supporting those tasks:

- Monitoring data loading from multiple sources
- Data quality and integrity checks
- Managing and updating meta-data
- Monitoring database performance to ensure efficient query response times and resource utilization
- Auditing data warehouse usage to provide user chargeback information
- Replicating, subsetting, and distributing data
- Maintaining effcient data storage management
- Purging data;
- Archiving and backing-up data
- Implementing recovery following failure

- **Single-Tier Architecture**

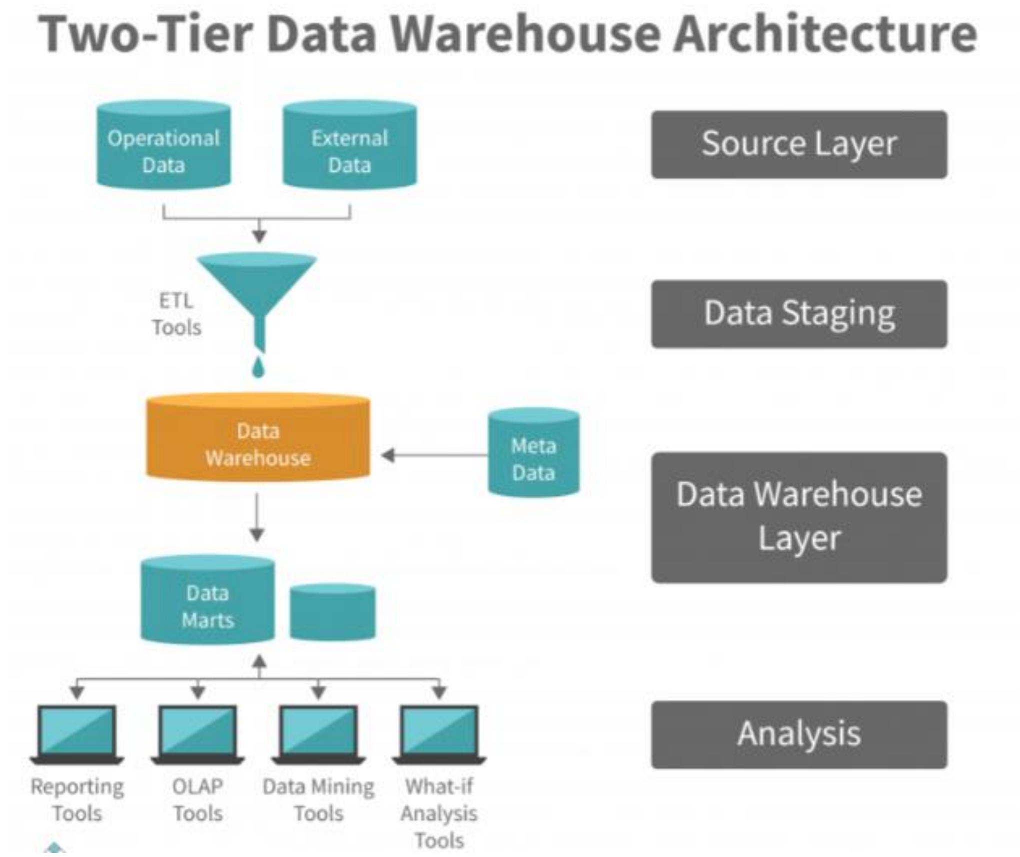
Single-tier architectures are not implemented in real-time systems. They are used for batch and real-time processing. The data is first transferred to a single-tier architecture where it is converted into a format that is suitable for real-time processing. This architecture is known as “single-threaded”. After this, the data is transferred to a real-time system. Single-tier architectures are currently the most preferred way to process operational data. It is important to note that single-tier architectures are not implemented in real time systems.

The data storage and processing middleware should be able to determine the quality of the data before the data is accepted by the analytical engine and transformed into relevant information. If these steps are not performed, then the middleware can be penetrated by malicious or faulty code. As an example, consider a credit score calculation. If a malicious hacker controls the middleware, then the hacker can modify the score and extract valuable data.



Two-Tier Architecture:

In a two-tier data warehouse, an analytical process is separated from a business process. This allows for greater levels of control and efficiency. A two-tier system also provides a better understanding of the data and allows for more informed decisions.



Two-layer architecture describes a four-stage data flow in which physical sources are separated from data warehouses by a two-layered architecture.

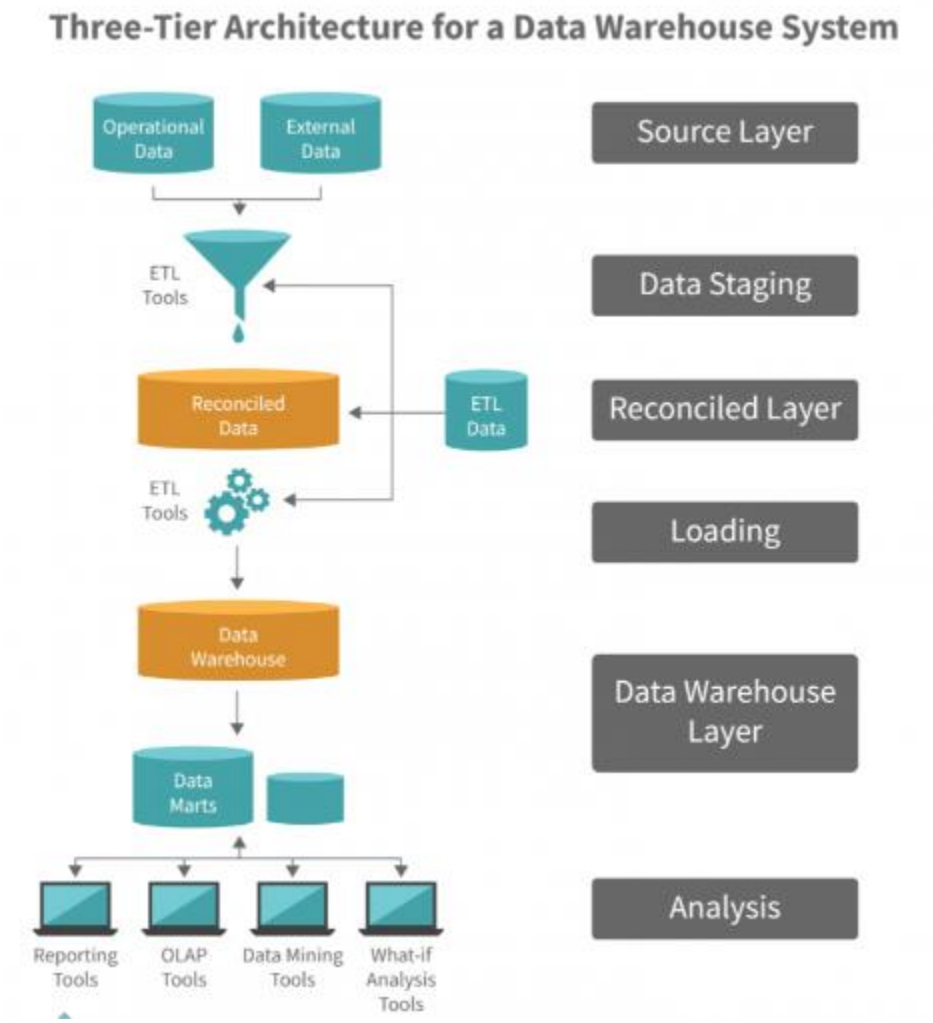
- The source of the data is critical to the data warehouse's integrity. The integrity of the data stored in the data warehouse must be guaranteed. Data integrity is the degree to which data values in a database record are true or accurate. A data warehouse is a system that stores information in a database so that it can be searched and analyzed.

- Data staging is a key process in the ETL process, and one that can significantly reduce the time it takes to extract, transform, and load (ETL) a large data set. ETL tools can extract data from various storage sources, transform the data with corporate-specific functions, and load the data into a data warehouse. Data warehouse functions such as monitoring the system, provisioning new data, and making decisions on the basis of the data are all done through data warehouse functions such as ETL. Data warehouse functions such as ETL can be implemented through a data warehouse.

- Data warehouse metadata is a critical component of the data warehouse. It is the information that helps a data warehouse administrator decide which data to delete, which data to retain, and which data to use in future reports. It is also important to maintain data warehouse consistency. Data warehouse administrators must determine which data should be updated or deleted when new data arrives, and which data should be left untouched. When data warehouse consistency is not guaranteed, application developers and users must be careful about which tables and reports they create.
- Data profiling is also very important for this level as it helps in validating data integrity and presentation standards. It also comes with advanced analytics such as real-time and batch reporting, data profiling and visualizations, and rating functions. It is important to keep in mind that this is not just a data warehouse but a live data platform that receives and analyzes massive amounts of data. This is why it is important to keep track of data changes, scalability, and performance of the system.

Three-Tier Architecture

A three-tier structure is employed in the source layer, the reconciled layer, and the data warehouse layer. The reconciled layer sits between the source data and data warehouse. The main disadvantage of the reconciled layer is the fact that it is not possible to completely ignore the problems of the data before it is reconciled. Therefore, the main focus of the reconciler should be on data integrity, accuracy, and consistency. For example, assume that the data warehouse contains a collection of company data elements that are updated frequently, such as order book information. In such a case, the best approach would be to use a web-based data warehouse refresh tool, which extracts the latest data from the data warehouse and refreshes the data in the corporate application. This architecture is appropriate for systems with a long-life cycle. Whenever a change occurs in the data, an extra layer of data review and analysis is done to ensure that no erroneous data was entered. This architecture is also known as data-driven architecture. This structure is mainly used for large-scale systems. It is important to note that the extra layers of data review and analysis created by this structure does not consume any extra space in the storage device.



Summary:

- Data warehouse is an information system that contains historical and commutative data from single or multiple sources. These sources can be traditional Data Warehouse, Cloud Data Warehouse or Virtual Data Warehouse.

- A data warehouse is subject oriented as it offers information regarding subject instead of organization's ongoing operations.
- In Data Warehouse, integration means the establishment of a common unit of measure for all similar data from the different databases
- Data warehouse is also non-volatile means the previous data is not erased when new data is entered in it.
- A Datawarehouse is Time-variant as the data in a DW has high shelf life.
- There are mainly 5 components of Data Warehouse Architecture: 1) Database 2) ETL Tools 3) Meta Data 4) Query Tools 5) DataMarts
- These are four main categories of query tools 1. Query and reporting, tools 2. Application Development tools, 3. Data mining tools 4. OLAP tools
- The data sourcing, transformation, and migration tools are used for performing all the conversions and summarizations.
- In the Data Warehouse Architecture, meta-data plays an important role as it specifies the source, usage, values, and features of data warehouse data.

****