# Big Data Analytics and Visualization (CDSC 715)

**Addisu G. Semie, PhD**
**Asst. Prof.,Computational Science Program,**
**Addis Ababa University**
**Email:** addisu.semie@aau.edu.et

# Big Data Analytics

- What is big data analytics?
- Data analytics: Key concepts
- Descriptive, Diagnostic, Predictive and Prescriptive Analytics
- Data Warehouse Architecture
- Technologies used in Big Data analytics

# Why Big Data Analytics?
## Making Smarter and More Efficient Organization

New York Police Department is utilizing data patterns, scientific analysis, and technological tools to prevent the occurrence of crime


Analyse historical arrest patterns and then maps them with events such as federal holidays, paydays, traffic flows, rainfall etc

Big Data analytics strategy helps NYPD to identify crime locations, through which they deploy their officers to these locations. Thus by reaching these locations before the crimes were committed, they prevent the occurrence of crime.

# Why Big Data Analytics?

## Optimize Business Operations by Analysing Customer Behavior

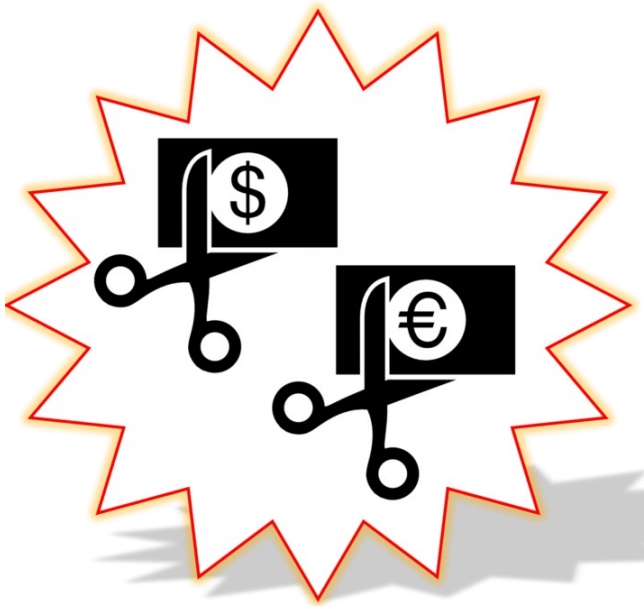
- Analysing all the clicks of every visitor on a website
- Studying the paths leading them to buy products
- Customer Satisfaction

Amazon uses customer click-stream data and historical purchase data of more than 300 million customers and each user is shown customized results on customized web pages.

All this information helps Amazon to improve their user experience, thereby improving their sales and marketing.

# Why Big Data Analytics?

## Cost Reduction




Parkland Hospital uses analytics and predictive modelling to identify high-risk patients and predict likely outcomes once patients are sent home. As a result, Parkland reduced 30-day readmissions for patients with heart failure, by 31 percent, saving $500,000 annually.

Cloud computing bring significant cost advantages when it comes to store and process Big Data.

Patients nowadays are using new sensor devices when at home or outside, which send constant streams of data that can be monitored and analysed in real-time to help patients avoid hospitalization by self-managing their conditions.

# Why Big Data Analytics?

## New Generation Products

Big Data tools are used to operate Google's Self Driving Cars. The Toyota Prius is fitted with cameras, GPS as well as powerful computers and sensors to safely drive on the road without the intervention of human beings.

Netflix launched the seasons of its TV show House of Cards based on the user reviews, ratings and viewership.

NETFLIX

A smart yoga mat has sensors embedded in the mat will be able to provide feedback on your postures, score your practice, and even guide you through an at-home practice.

# What is big data analytics?

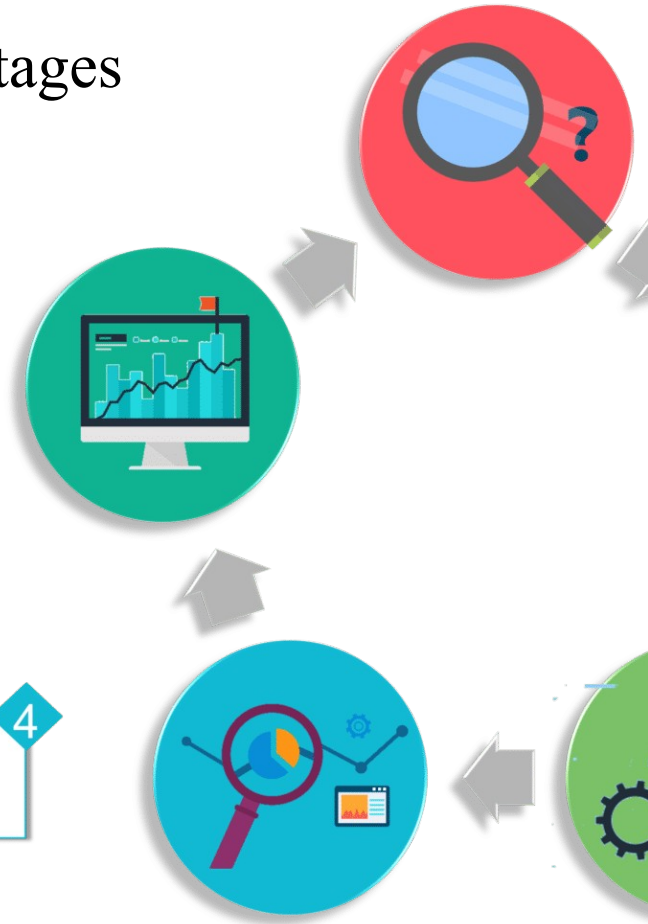
- Big data analytics is the often complex process of examining big data to uncover information such as:
  - hidden patterns
  - correlations
  - market trends
  - customer preferences

that can help organizations make informed business decisions.

Analyzing Big Data is largely used by companies to facilitate their growth and development.

# Data Analytics: Key concepts

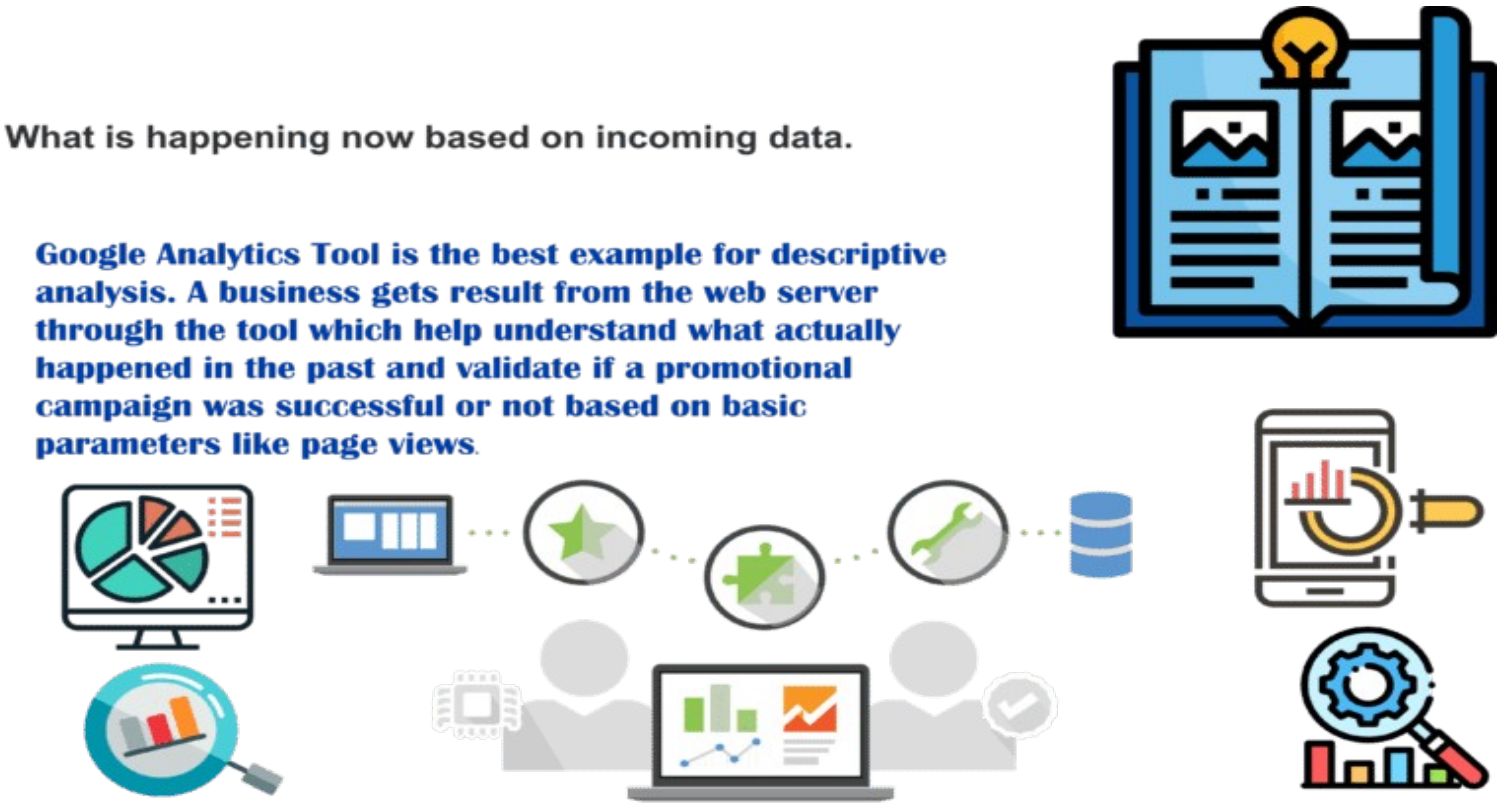
Stages



# Types of Big Data Analytics

**Descriptive Analytics**: It uses data aggregation and data mining to provide insight into the past and answer: "What has happened?"

The descriptive analytics does exactly what the name implies they "describe" or summarize raw data and make it interpretable by humans.



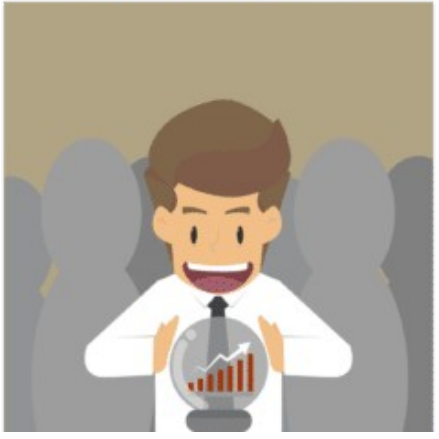
# Types of Big Data Analytics

**Predictive Analytics**: It uses statistical models and forecasts techniques to understand the future and answer: “What could happen?”

Predictive analytics provides companies with actionable insights based on data. It provides estimates about the likelihood of a future outcome.

**What might happen in the future**

**For example, Southwest Airlines analyses sensor data on their planes in order to identify patterns that indicate a potential malfunction, thus allowing the airlines to the necessary repairs before its schedule.**



# Types of Big Data Analytics

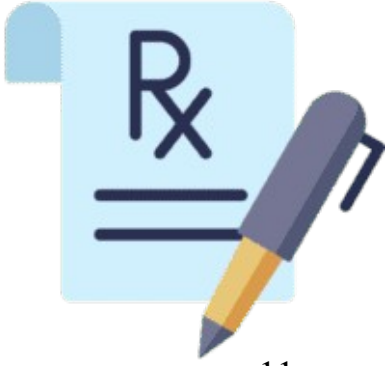
**Prescriptive Analytics**: It uses optimization and simulation algorithms to advice on possible outcomes and answers: "What should we do?"

It allows users to "prescribe" a number of different possible actions and guide them towards a solution.

In a nutshell, this analytics is all about providing advice.

**What action should be taken.**

**Google's self-driving car is a perfect example of prescriptive analytics. It analyses the environment and decides the direction to take based on data.**



# Types of Big Data Analytics

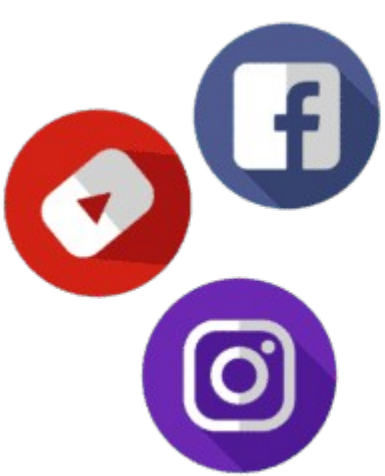
**Diagnostic Analytics:** It is used to determine why something happened in the past.

It is characterized by techniques such as drill-down, data discovery, data mining and correlations.

Diagnostic analytics takes a deeper look at data to understand the root causes of the events.

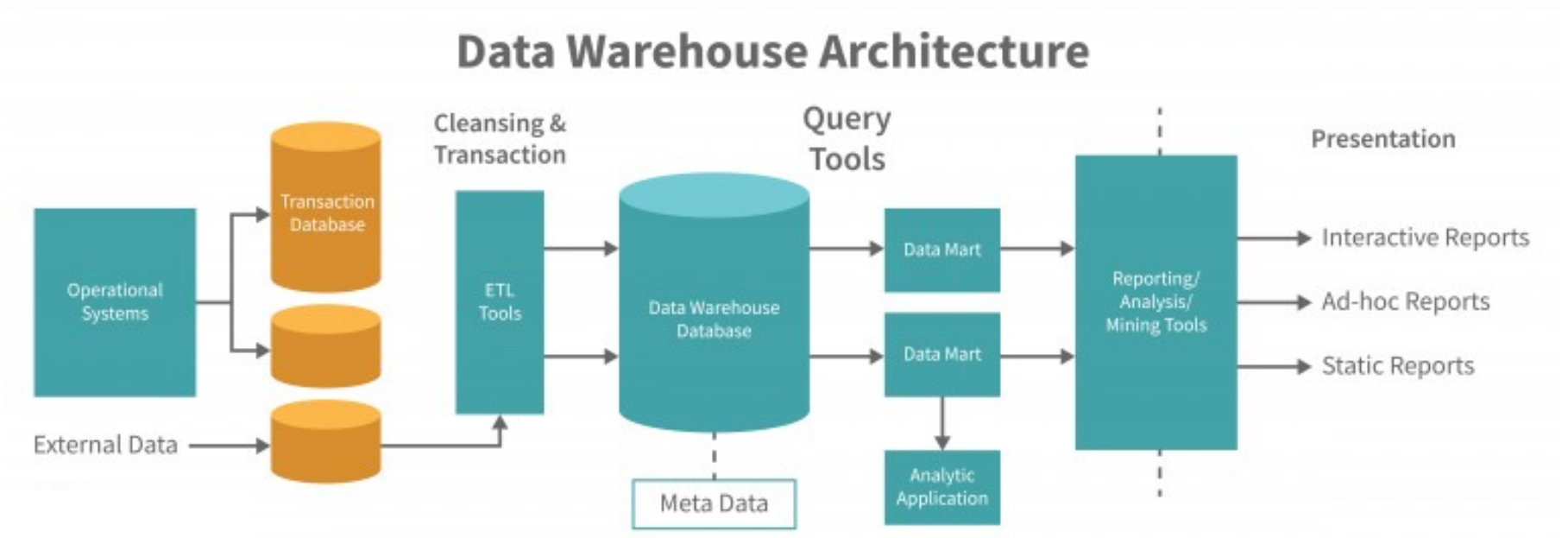
Why did it happen

For a Social Media marketing campaign, you can use diagnostic analytics to assess the number of posts, mentions, followers, fans, page views, reviews, pins, etc. and analyse the failure and success rate of the campaign at a fundamental level.

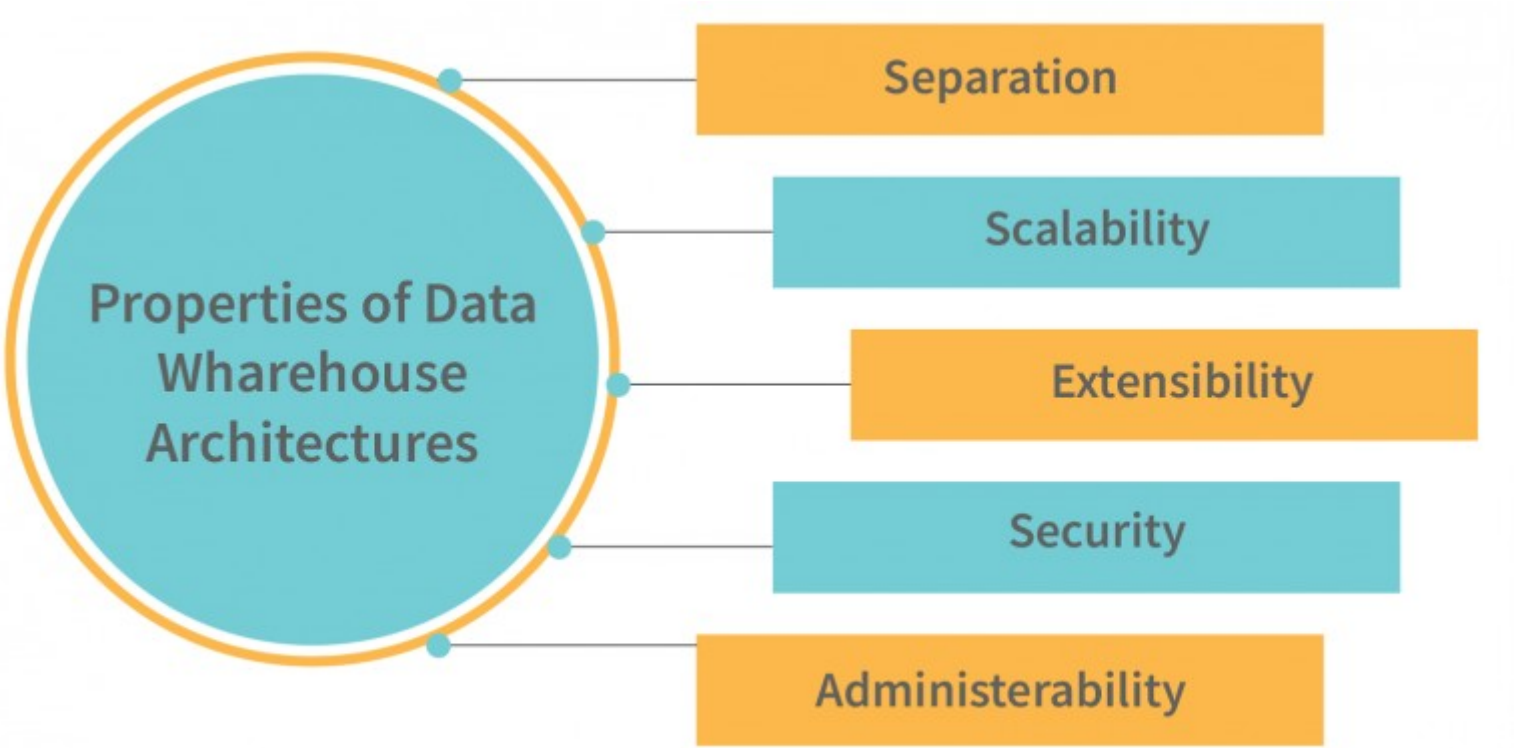



# Data Warehouse Architecture

- Is a method of defining the overall architecture of data communication processing and presentation that exist for end-clients computing within the enterprise

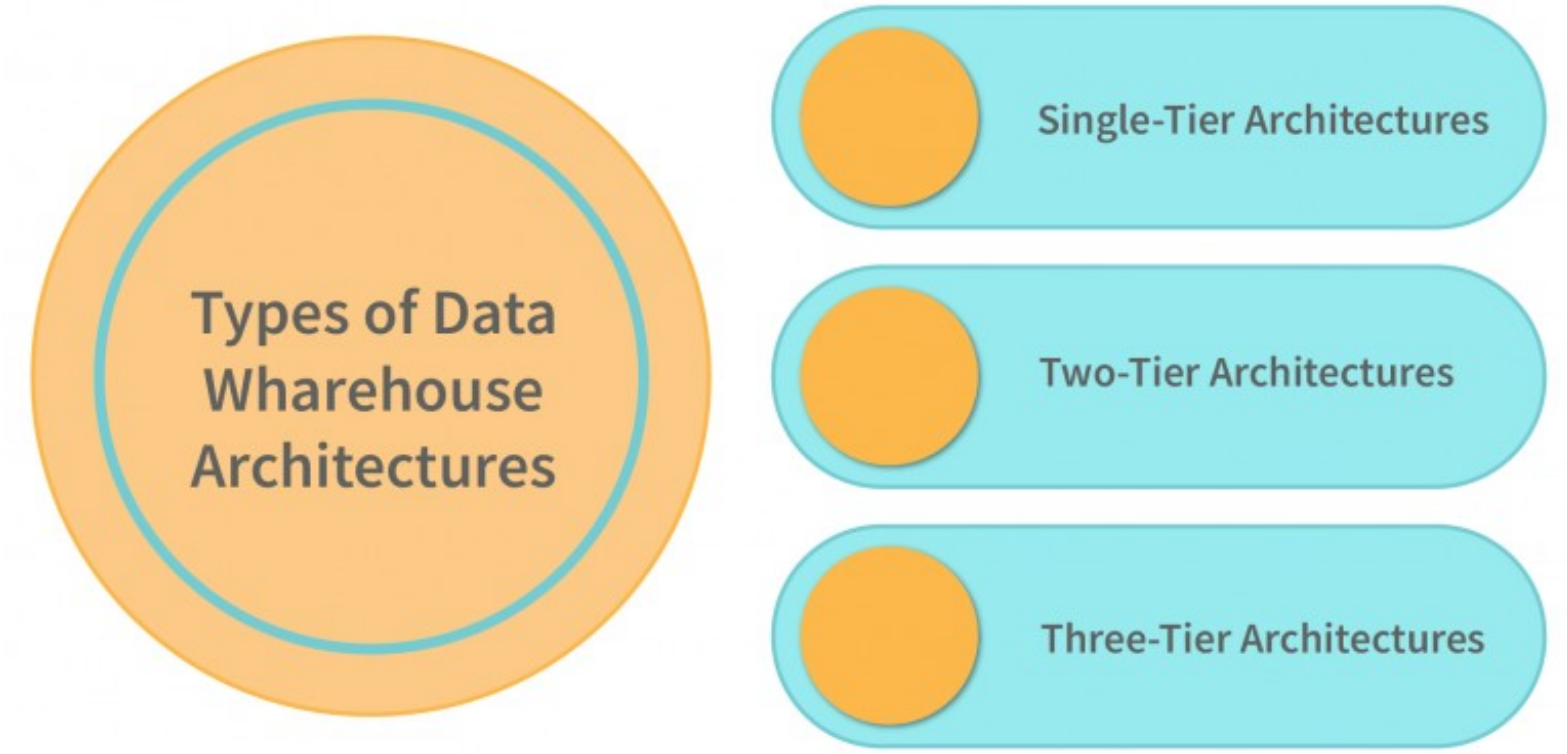


# Data Warehouse Architecture Properties



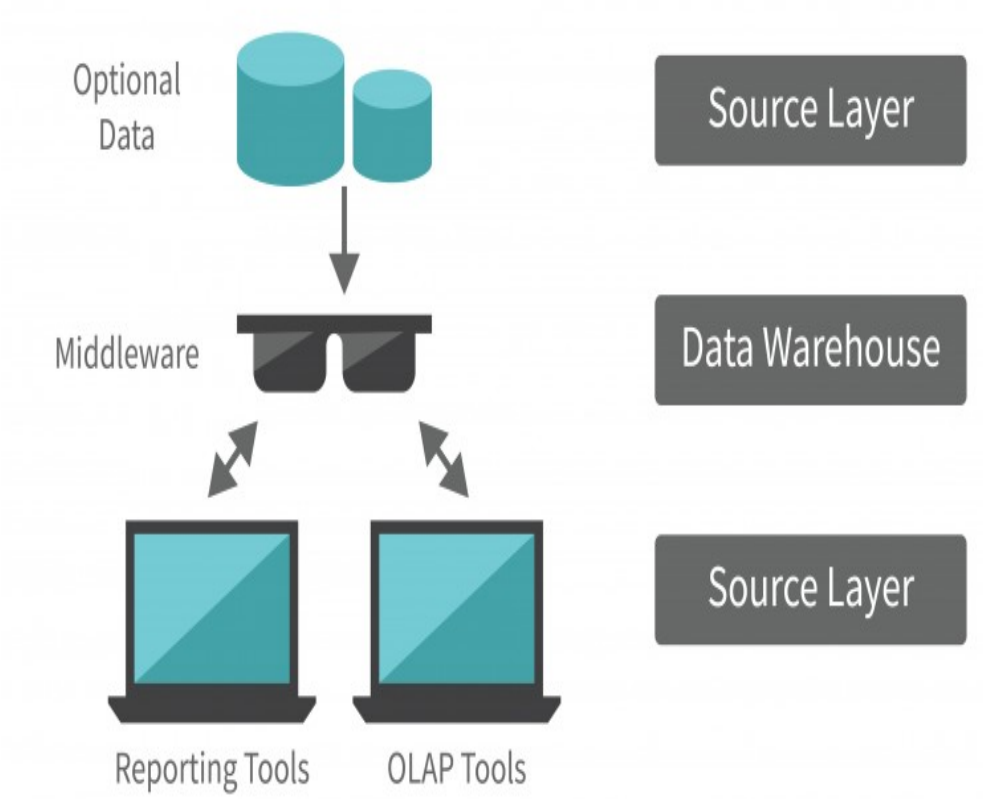
# Types of Data Warehouse Architectures

There are mainly three types of Data warehouse Architectures

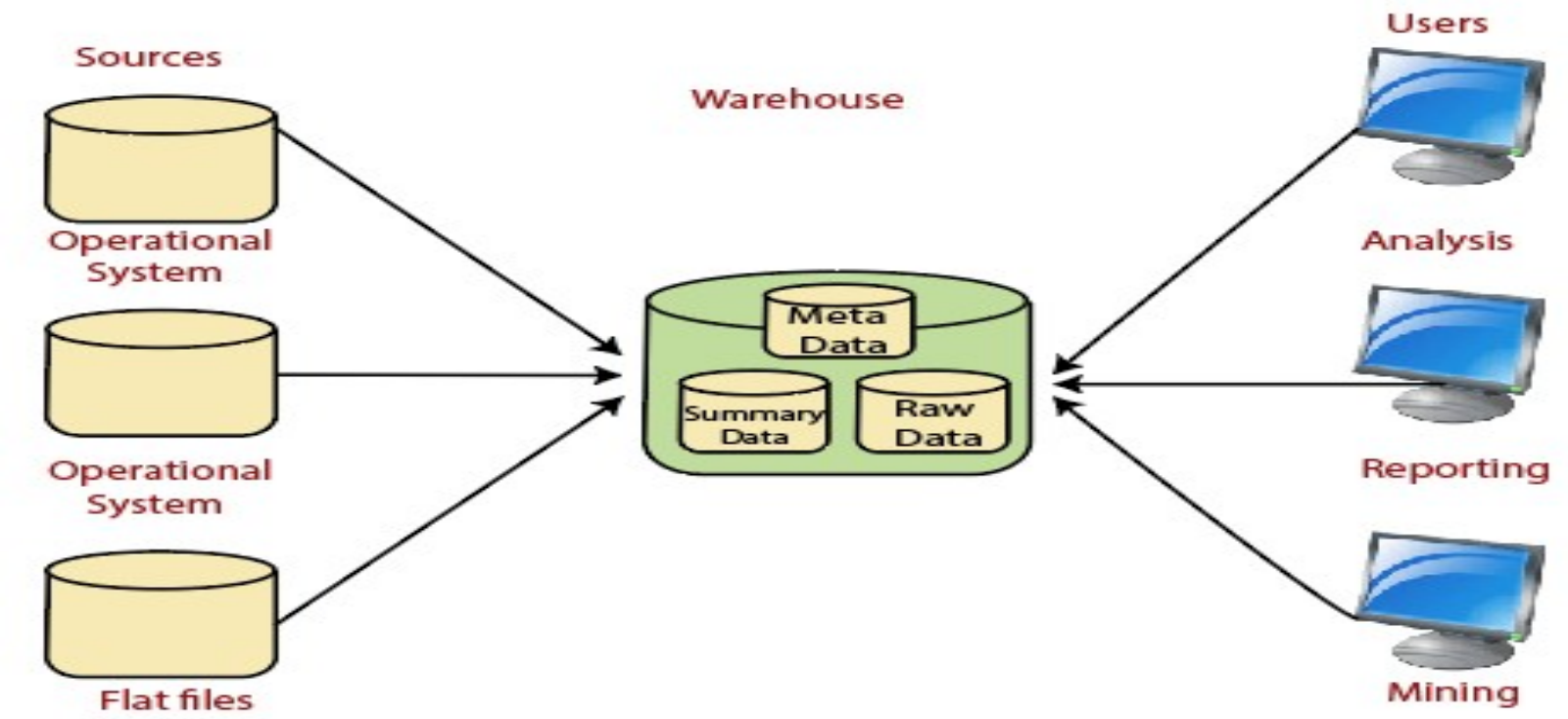


# Single-Tier Data Warehouse Architectures

- Aimed at keeping data space minimal.
- Used for batch and real-time processing.
- Are currently the most preferred way to process operational data.
- Are not implemented in real time systems.
- The quality of the data determined by the the data storage and processing middleware



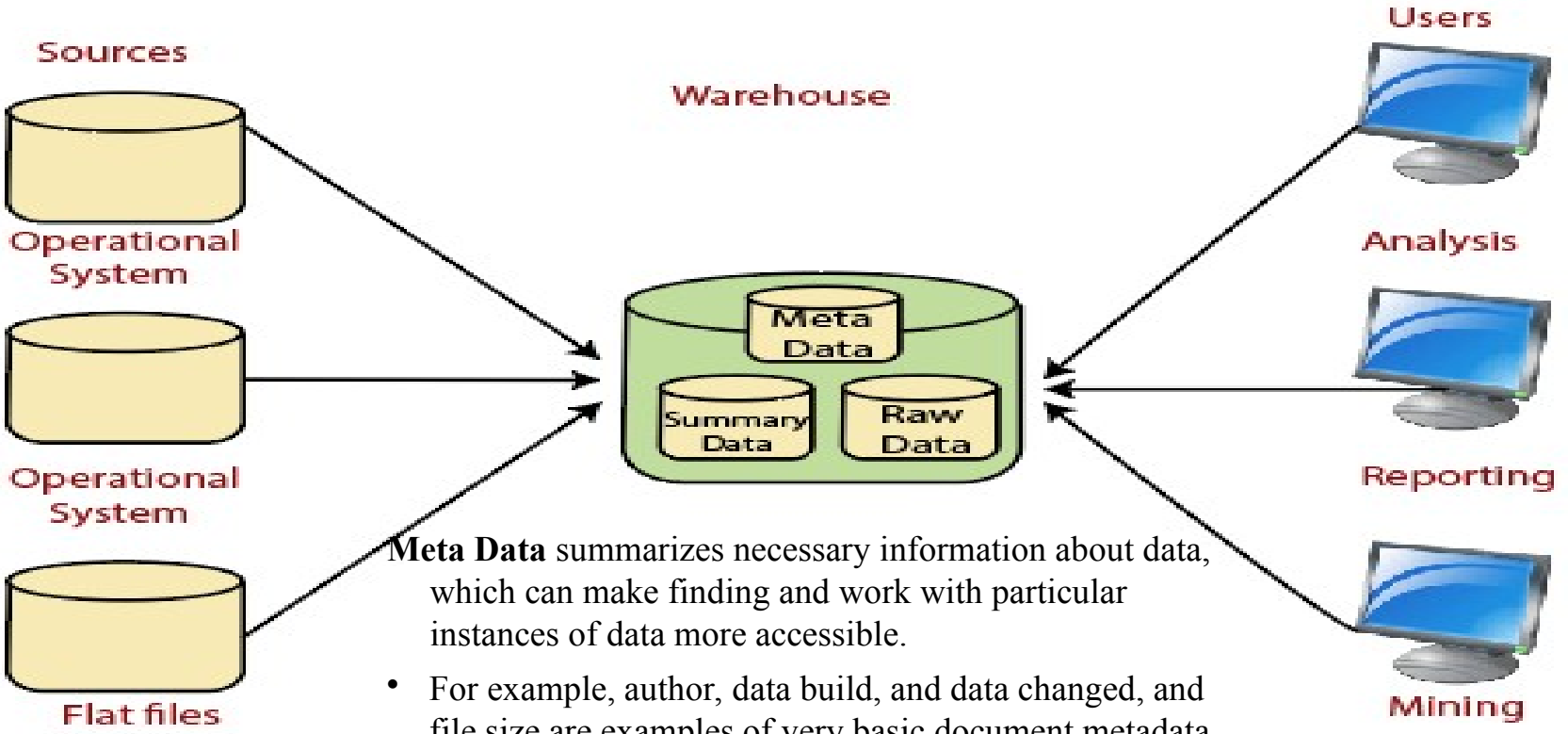
# Single-Tier Data Warehouse Architectures




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.

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.

# Single-Tier Data Warehouse Architectures




**Meta Data** summarizes necessary information about data, which can make finding and work with particular instances of data more accessible.

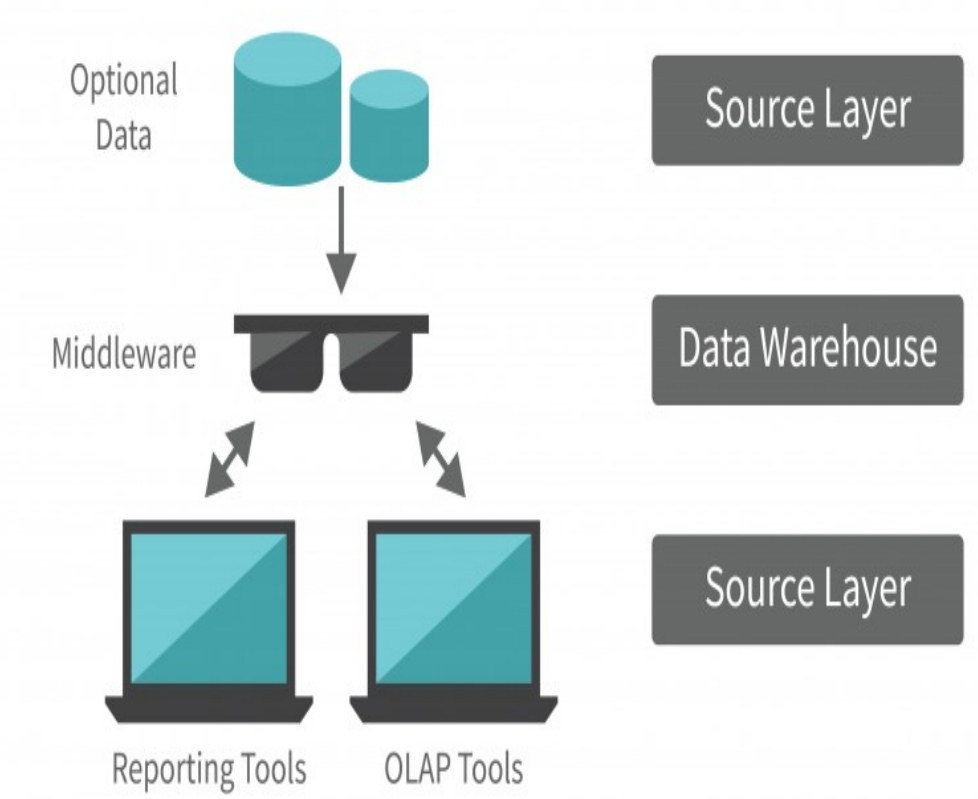
- For example, author, data build, and data changed, and file size are examples of very basic document metadata.

Metadata is used to direct a query to the most appropriate data source.

# Single-Tier Data Warehouse Architectures

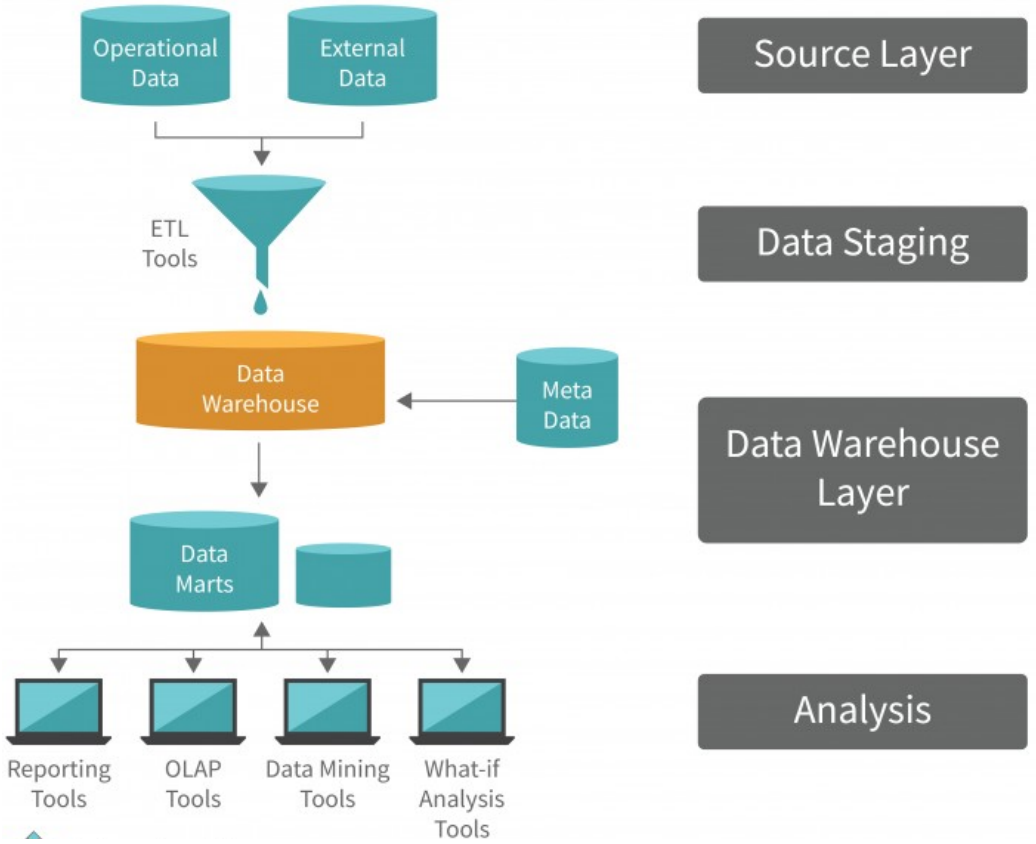
The examples of some of the end-user access tools can be:

- Reporting and Query Tools
- Application Development Tools
- Executive Information Systems Tools
- Online Analytical Processing Tools
- Data Mining Tools



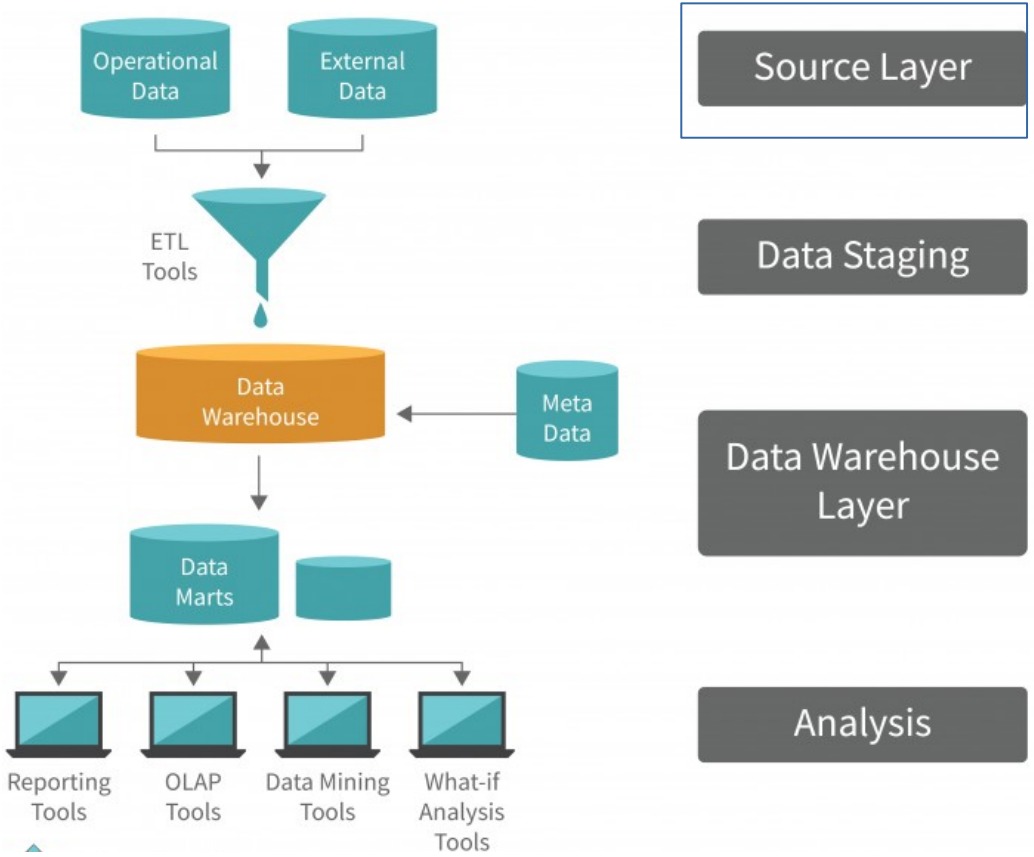
# Two-Tier Data Warehouse Architectures

- In a two-tier data warehouse, physical sources are separated from data warehouses.
- Provides a better understanding of the data and allows for more informed decisions



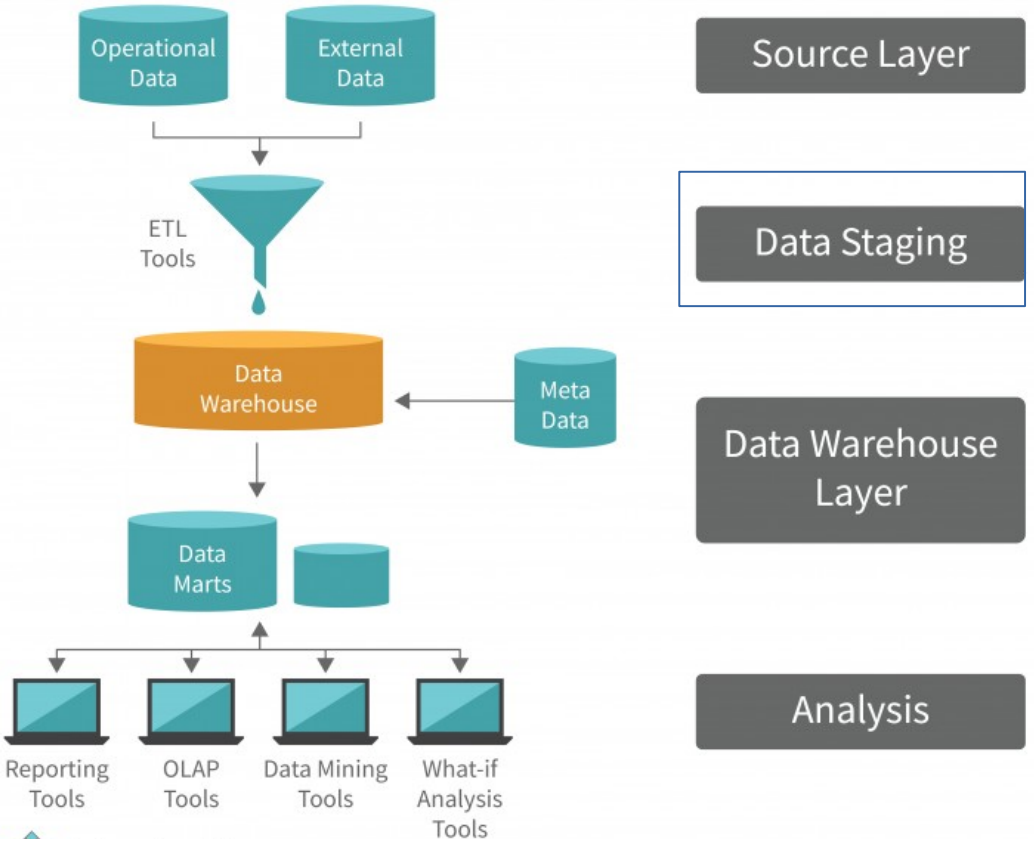
# Two-Tier Data Warehouse Architectures

- 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.



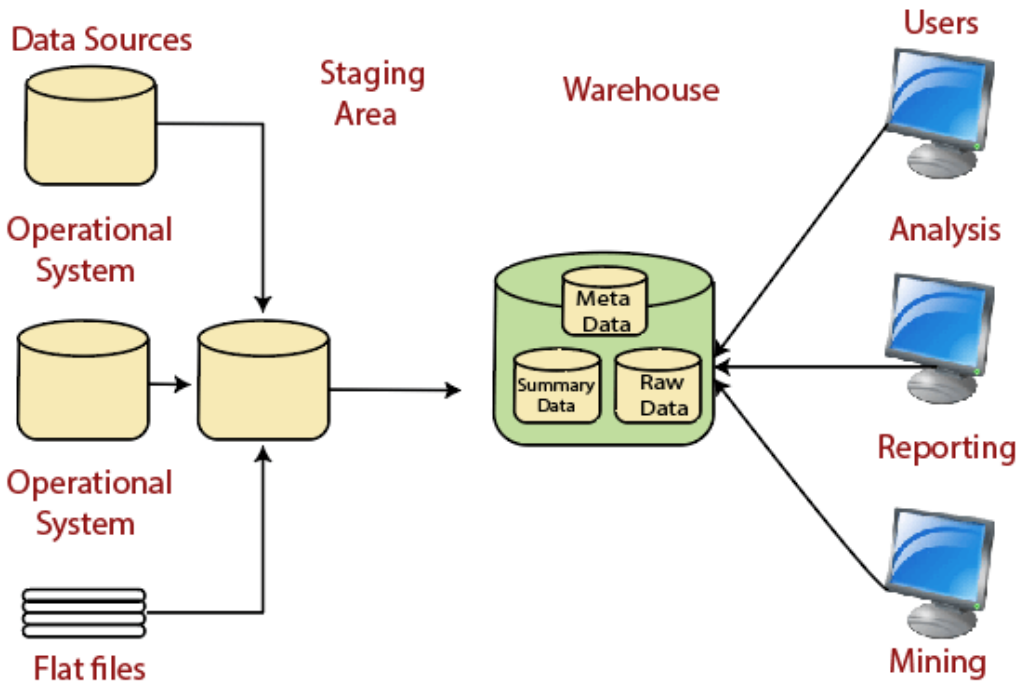
# Two-Tier Data Warehouse Architectures

- Staging area - a place where data is processed before entering the warehouse.
- It simplifies data cleansing and consolidation for operational method coming from multiple source systems
- Data Warehouse Staging Area is a temporary location where a record from source systems is copied



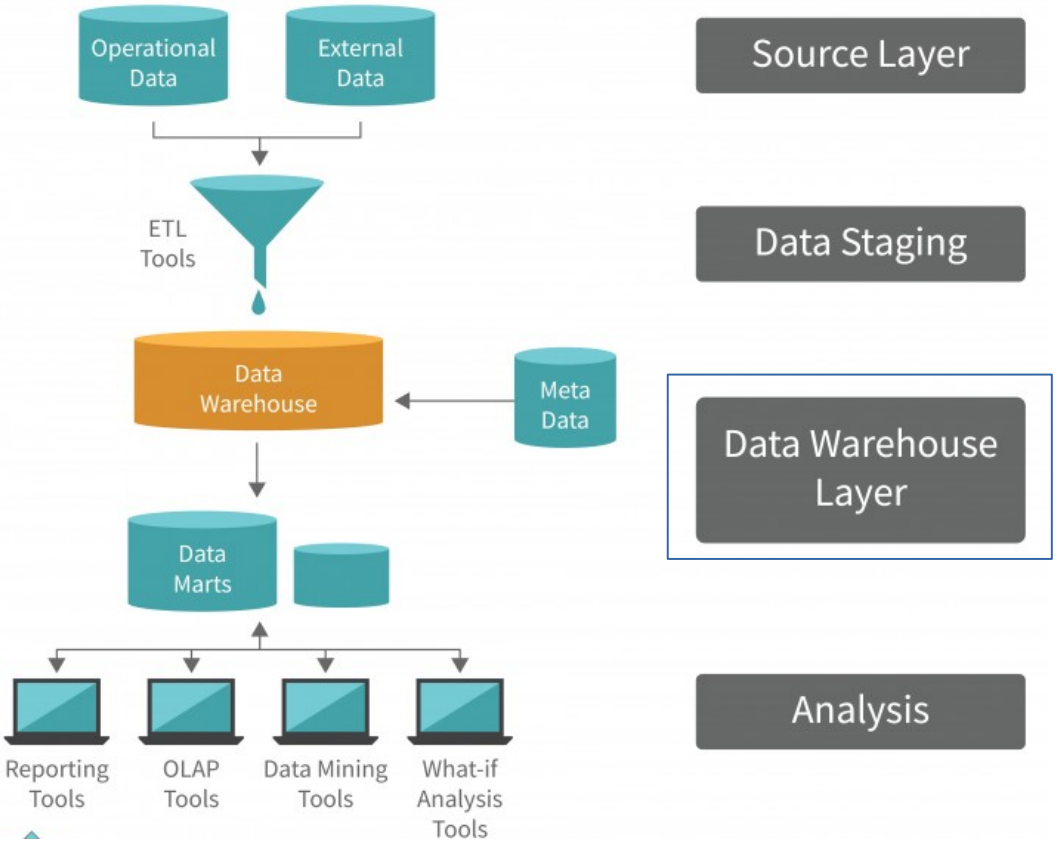
# Two-Tier Data Warehouse Architectures

- Data staging is a key process 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.



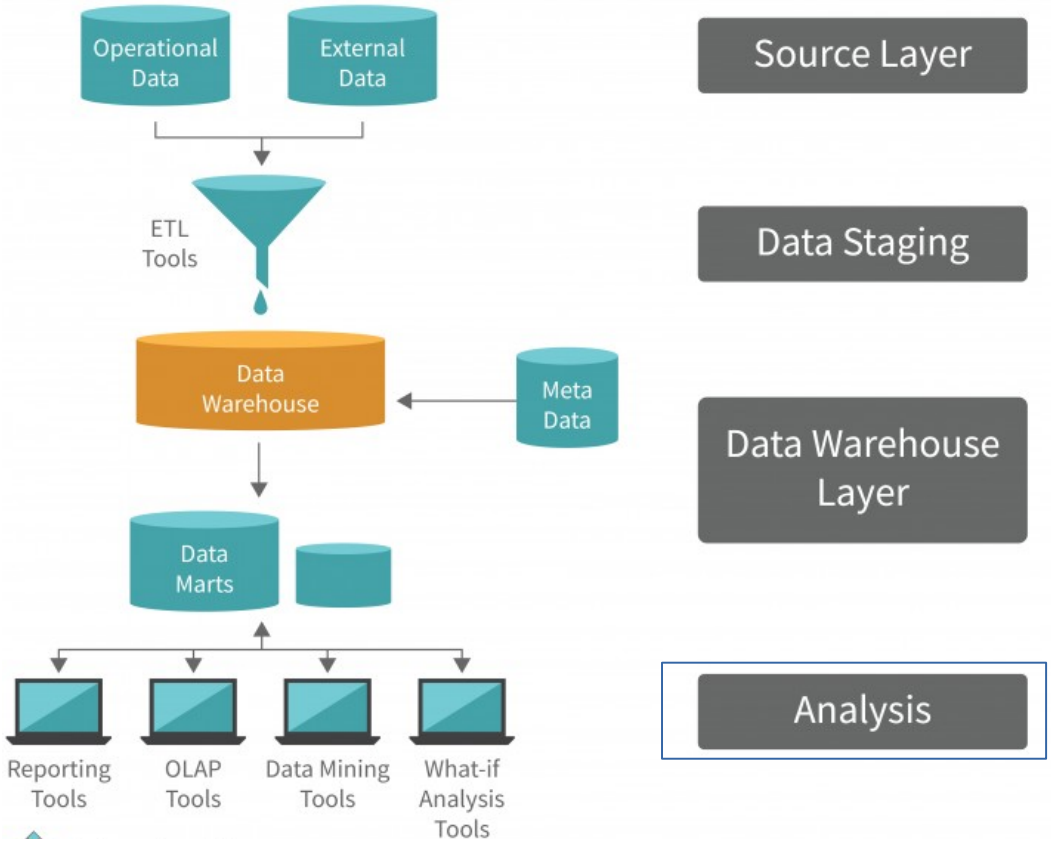
# Two-Tier Data Warehouse Architectures

- 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.



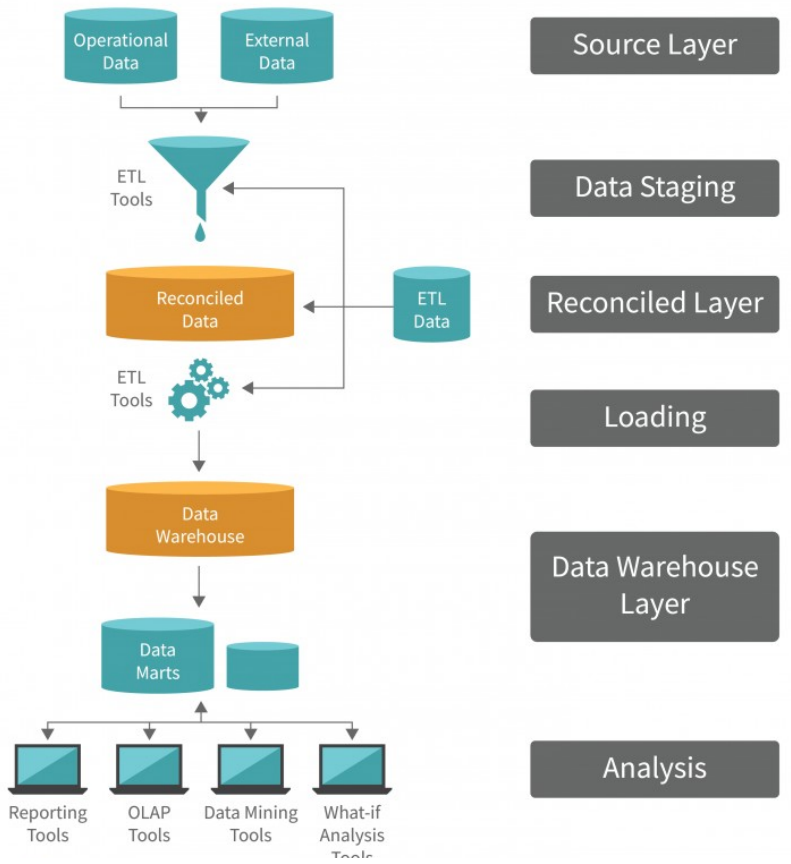
# Two-Tier Data Warehouse Architectures

- Analysis stage comes with advanced analytics such as:
  - real-time reporting
  - batch reporting
  - data profiling
  - visualizations
  - rating functions



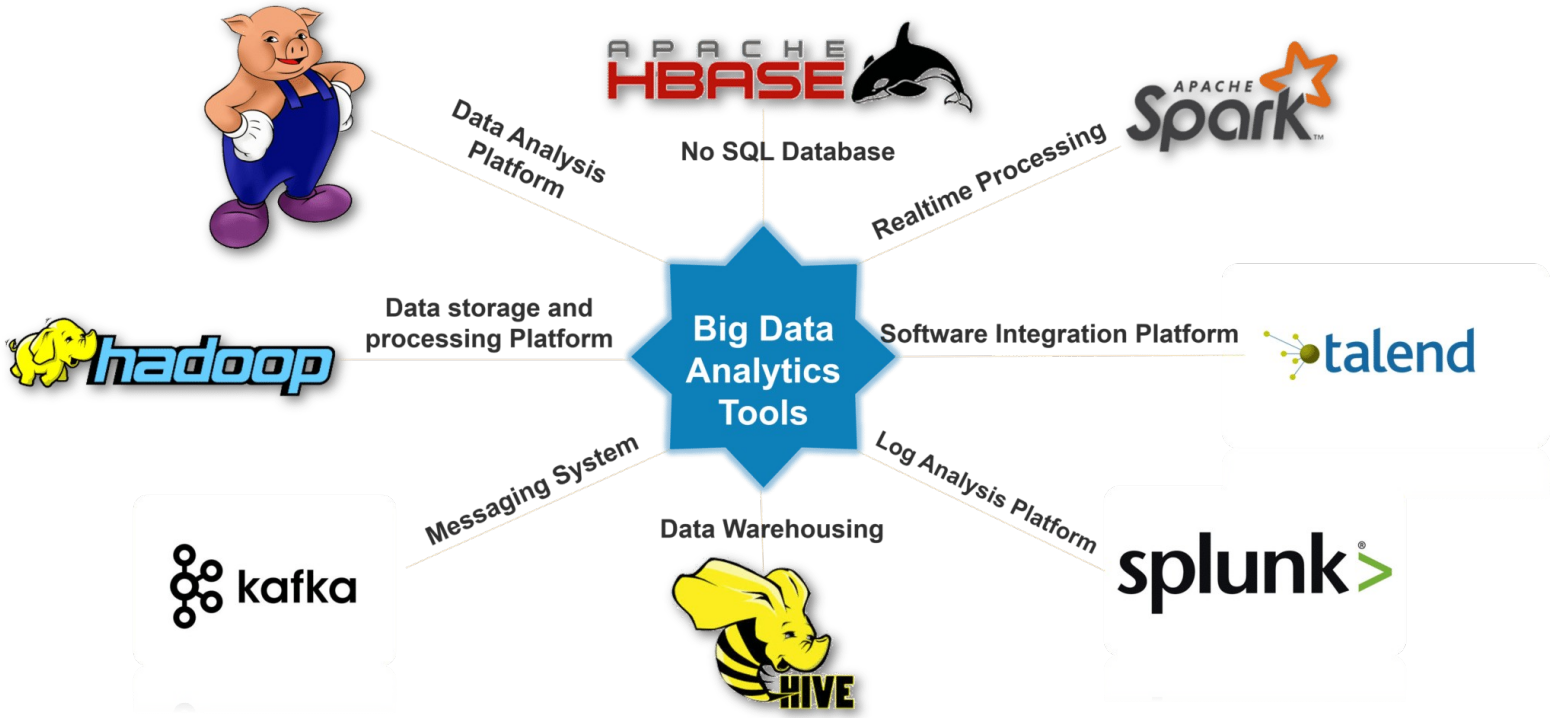
# Three-Tier Data Warehouse Architectures

- The reconciled layer sits between the source data and data warehouse
- The main focus of the reconciler should be on data integrity, accuracy, and consistency
- 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



# Big Data Tools

These are some of the following tools used for Big Data Analytics: Hadoop, Pig, Apache Hbase, Apache Spark, Talend, Splunk, Apache Hive, Kafka



# Assignment 1

Choose one of the Big Data tools and prepare a presentation that address the following points:

- Define and elaborate on its advantages and disadvantages.
- Discuss the key features of the Big Data tool
- Provide an example of how the tool can be used to perform data analytics on a large dataset.