SNS COLLEGE OF ENGINEERING

Department of Information Technology

19IT601 – Data Science and Analytics
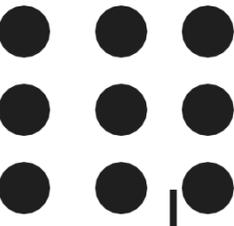
III Year / VI Semester

Unit 3 – PREDICTIVE MODELING AND MACHINE LEARNING

Topic 2: Datawarehouse Overview
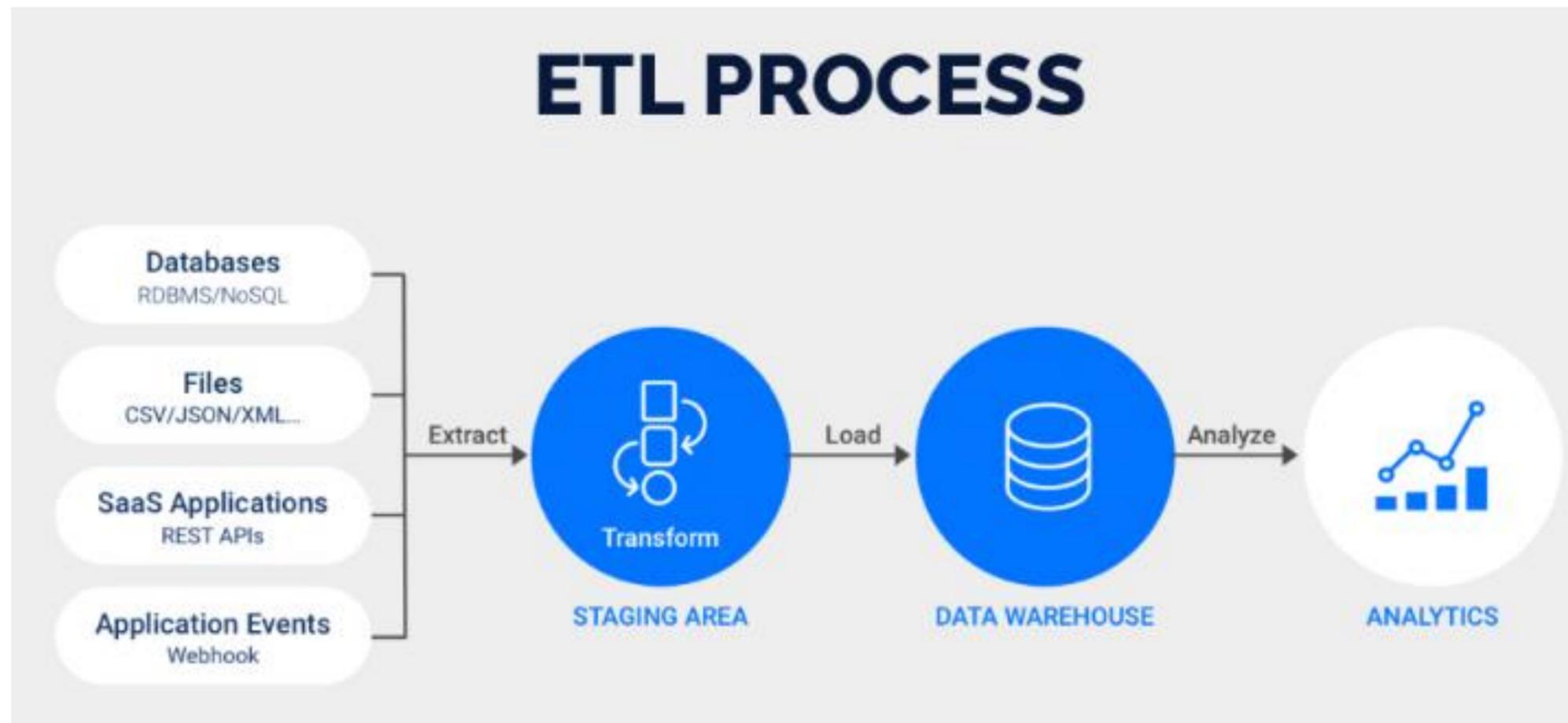
Data Warehouse

- It's basically a giant database that contains information from many different sources and ties them together.

- A Data Warehouse (DW) is a relational database that is designed for query and analysis rather than transaction processing. It includes historical data derived from transaction data from single and multiple sources.

- A Data Warehouse provides integrated, enterprise-wide, historical data and focuses on providing support for decision-makers for data modeling and analysis.

- A data warehouse has the challenge of taking data from many different sources, transforming them into some sort of schema that allows us to query these different data sources simultaneously, and it lets us make insights, through data analysis.

- Data Warehouse environment contains an extraction, transportation, and loading (ETL) solution, an online analytical processing (OLAP) engine, customer analysis tools, and other applications that handle the process of gathering information and delivering it to business users.

ETL versus ELT

ETL - Extract, Transform, and Load

- Extract, transform, and load (ETL) is a data integration methodology that extracts raw data from sources, transforms the data on a secondary processing server, and then loads the data into a target database.

- The method emerged in the 1970s, and remains prevalent amongst on-premise databases that possess finite memory and processing power.

ETL - Extract, Transform, and Load

Extract:

- Extraction refers to pulling the source data from the original database or data source. With ETL, the data goes into a temporary staging area.
- In a traditional ETL scenario, the source data is extracted to a staging area and moved into the target system

Transform:

- Transformation refers to the process of changing the structure of the information, so it integrates with the target data system and the rest of the data in that system.
- In staging area, Transformation can involve converting all data types to the same format, cleansing by removing inconsistent or inaccurate data, combining data elements from multiple data models, pulling in data from other sources, and other processes.
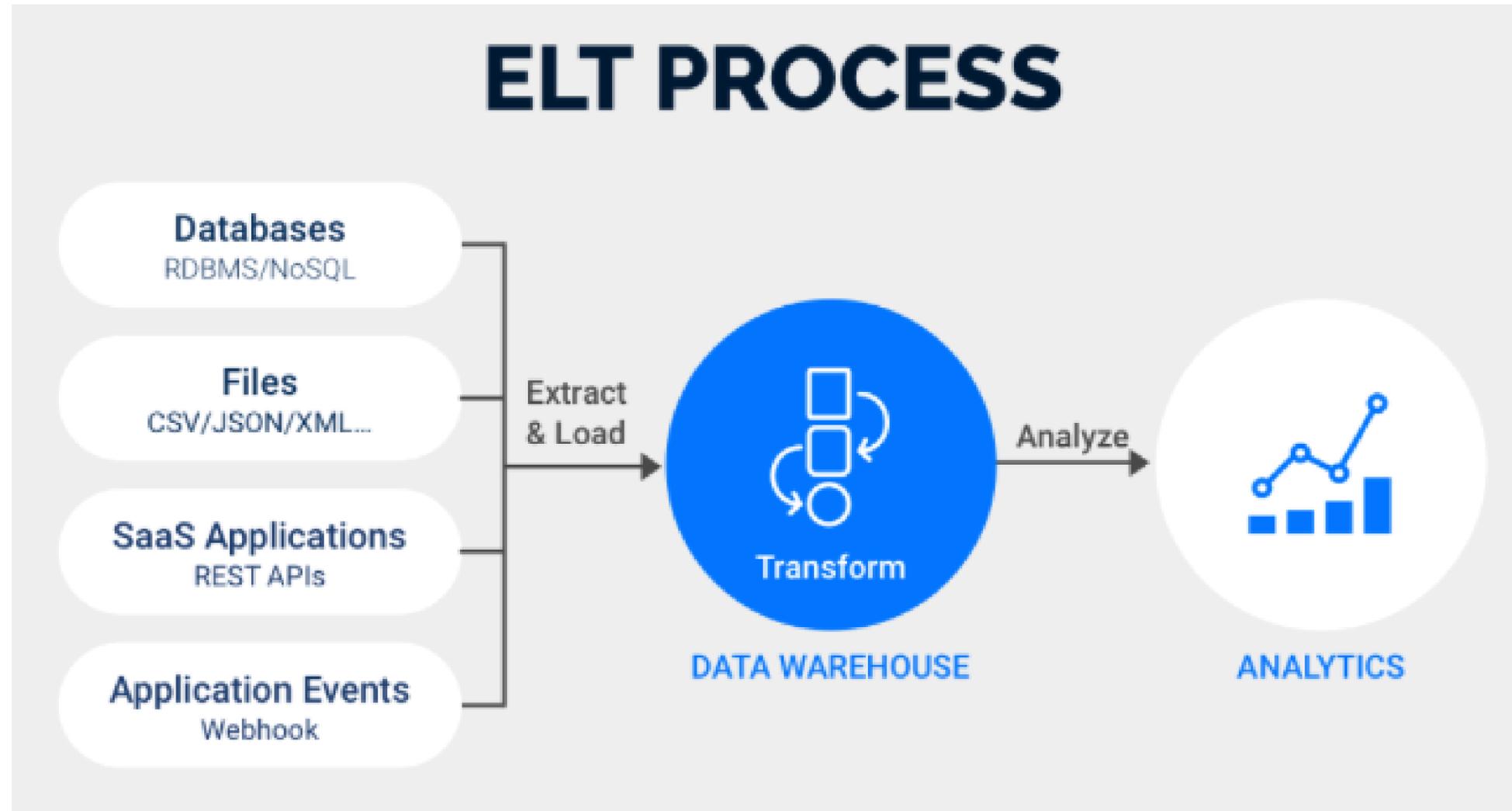
Load:

- Loading refers to the process of depositing the information into a target data storage system.
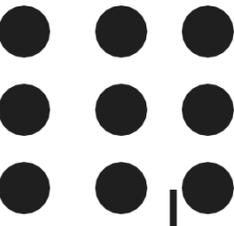
Example – OLAP Data warehouse

ETL versus ELT

ELT – Extract, Load, Transform

ETL versus ELT

ELT – Extract, Load, Transform

* Unlike ETL, extract, load, and transform (ELT) does not require data transformations to take place before the loading process.

* ELT loads raw data directly into a target data warehouse, instead of moving it to a processing server for transformation.

* With ELT, data cleansing, enrichment, and transformation all occur inside the data warehouse itself. Raw data is stored indefinitely in the data warehouse, allowing for multiple transformations.

* ELT is a relatively new development, made possible by the invention of scalable cloud-based data warehouses.

* Example – Hadoop, Data Lakes are special kinds of data stores that—unlike OLAP data warehouses—accept any kind of structured or unstructured data.

ETL versus ELT

Advantages of ELT
- ELT lets the data destination do the transformation, eliminating the need for data staging.
- Flexibility,
- Faster loading time,
- Scalability.

Advantages of ETL
- ETL can help with data privacy and compliance, cleansing sensitive data before loading into the data destination, while ELT is simpler and for companies with minor data needs.

THANK YOU