Dr. SNS RAJALAKSHMI COLLEGE OF ARTS & SCIENCE
(Autonomous)
Coimbatore -641049

Accredited by NAAC(Cycle–III) with ‘A+’ Grade
(Recognized by UGC, Approved by AICTE, New Delhi and
Affiliated to Bharathiar University, Coimbatore)

DEPARTMENT OF COMMERCE (IT)

COURSE NAME : 21UCI508 - Business Intelligence

III YEAR / V SEMESTER

Unit IV

Extract, Transform, Load (ETL)

Unit 4
Extract, Transform, Load (ETL)

ETL is one of the most commonly used techniques in data integration. It involves three steps: extracting data from various sources, transforming it into a common format, and loading it into a central repository such as a data warehouse. ETL ensures data consistency, accuracy, and readiness for analysis or reporting.

Evolution of ETL:

Businesses have been collecting the data for a long time but in the modern era the possibility of storage of data will be only with the computers and digital storage.

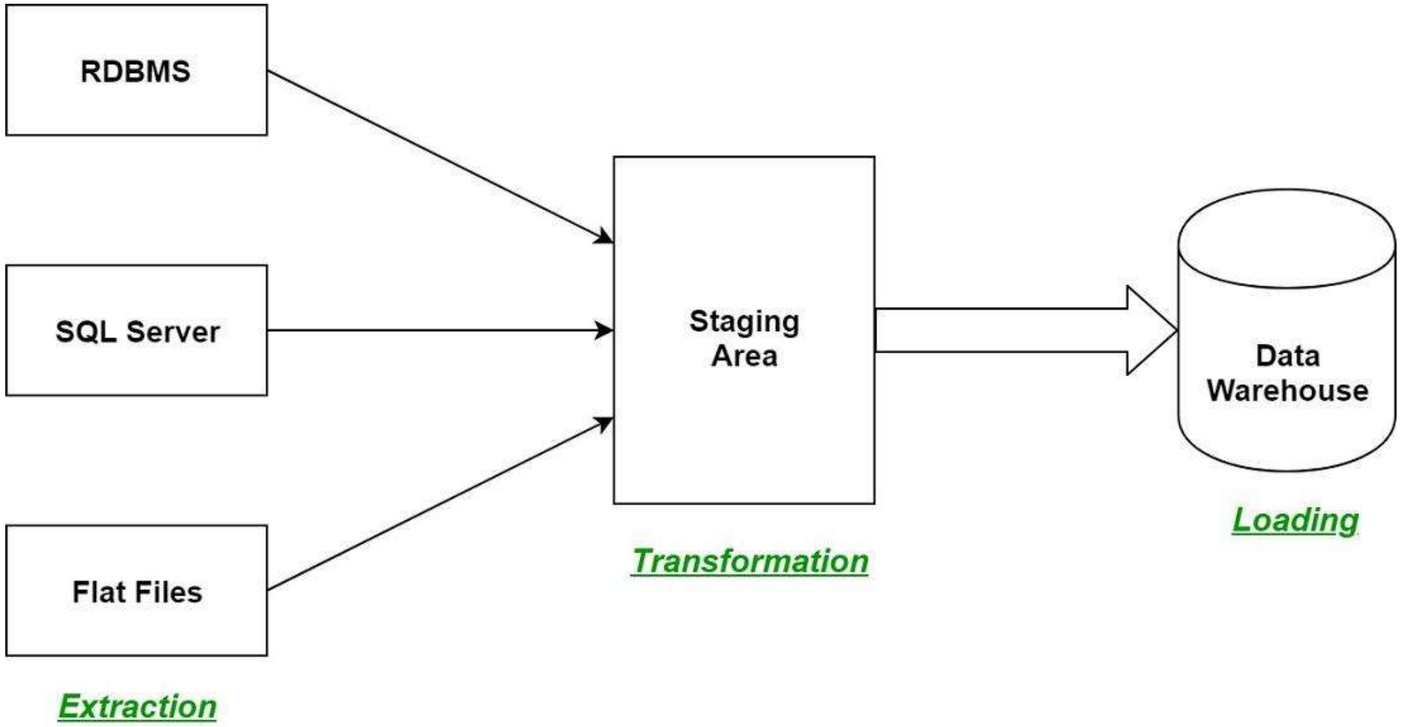
1970s - Introduction of ETL: In 1970's larger centralized databases will be invented with ETL (Extract, transform, and load) also introduced and processed and merged processed for the data analysis.

1980s - Rise of Data Warehouses and Relational Databases: In 1980's data warehouses (Storage which used for data storing purpose) and relational databases will be popular for making better analytics and decision making**1990s - Automation and Big Data**: Until the invention of the ETL software the all process of carry with manual efforts by the IT team for extracting the data from the different systems and connectors for the transformation of the data into the common well known format. Then transfer into the interconnected tables. Still the early ETL will be best option as a algorithm, addition of neural network will be making more opportunities for the analytics of the data.

1990s - ETL in the Cloud: In 1990's the Big data invented as result the computing speed and the storage capacity will increasing efficiently, where in which large amount of data will get from the different sources like social media and IOT (Internet of Things).

2000s and Beyond - Advanced Analytics and AI: In 1990's ETL and cloud computing were become more popular. With using of the data warehouses such as (AWS) Amazon Web Services, Microsoft Azure and Snowflake which making the availability of these data around the world.

ETL PROCESS:


RDBMS
SQL Server
Flat Files
Staging Area
Data Warehouse
Extraction
Transformation
Loading

Techniques Used in ETL:

1. Removing Duplicate Records

Duplicates can result from multiple entries or merging databases and may lead to incorrect results.

2. Standardization of Data:

Standardization ensures that all data follows a consistent format and style, which is crucial for data integration.

3. Handling Missing or Incomplete Data

Missing data can cause analytical distortions and needs appropriate handling.

4. Data Validation

This ensures data meets business and logical rules, maintaining integrity and reliability.

5. Data Transformation

Involves changing the structure or representation of data to meet analysis or business needs.

6. Error Correction

7. Lookup and Reference Table Usage

Reference tables help in validating and cleaning data efficiently.

8. Parsing and Reformatting

Helps in cleaning and organizing complex fields.

a. Field Parsing:

9. Outlier Detection and Handling

Outliers may indicate errors or special cases and need careful treatment.

10. Automation and Tool-Based Cleansing

Cleansing becomes faster and more accurate with tools and scripts.


Thank
you