

ETL stands for Extract, Transform, Load. It is a data integration process that moves data from one or more sources into a destination system, usually a data warehouse.

Extract:- Gather data from various sources (databases, APIs, file, etc.)

Transform:- Clean, format & structure the data for analysis (e.g. filtering, joining, aggregation).

Load:- Move the transformed data into a target system (data warehouse, database, or data lake).

ETL Pipeline is an automated sequence of processes that perform the ETL steps. It ensures data flows reliably from source to destination & can be built using tools like Apache Airflow, Talend, AWS Glue or custom scripts.

Types of ETL pipeline:

1) Batch ETL pipelines:-

- Run at scheduled intervals.
- Good for large volumes of data processed periodically.

2) Real-Time or Streaming ETL pipelines:-

- Process data as it arrives.
- Useful for real-time dashboards, fraud detection or IoT data.

3) Cloud Native ETL pipeline:-

- Build on cloud platform (AWS Glue, ADF)
- Scalable & integrated with other cloud services.



4> Open-Source ETL pipelines :-

- Use open-source tools like Apache Nifi, Airflow, Talend open studio.
- Cost-Effective with high Flexibility.

5> <sup>ELT</sup> ~~ETL~~ Pipelines (Extract, ~~Trans~~ Load, Transform):

- Data is loaded into destination first, then transform there.
- Common with modern cloud data warehouse like snowflake or bigquery.