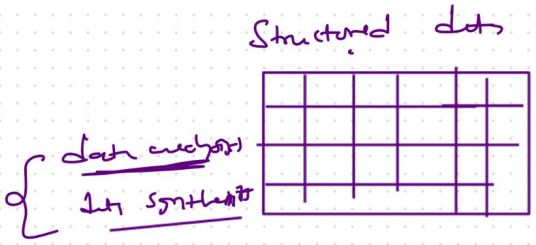
ETL vs ELT [key differences in Data Processing]

ETL [Extract, Transform, Load] ?

✓ Extract : Collect data from multiple sources [MySQL, JSON, XML, log file etc]

✓ Transform : Clean, filter, and format the data

✓ Load : Store cleaned data into a data warehouse — Data engineers

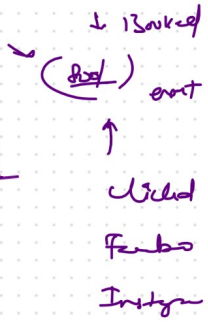


ETL Workflow Example :

1) Collect data from multiple sources

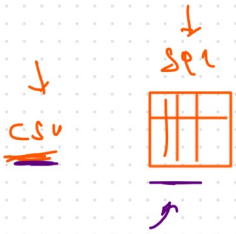
ex: At OYO Rooms data came from

- customer data in MySQL
- Booking data from PostgreSQL
- user activity data
- marketing data from social platforms
- Some data from XML format



2) Transform the data :

- Clean, filter & standardized data format
- make different data sources compatible



3) load into data warehouse

- store in optimized structure for analysis
- Serves as the foundation for business intelligence

Ans2 Food Delivery App

Collect data: Restaurants menus, Customer orders, tracking info

Transform: remove duplicate orders, invalid address

Load: Store clean data in data warehouse

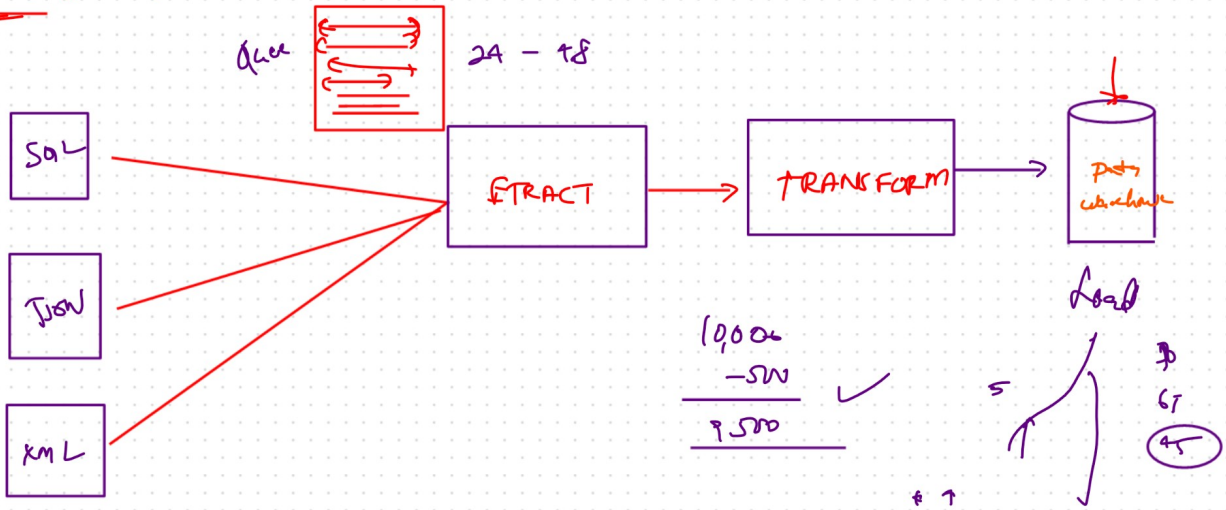
ETL Characteristics

- Processing location: Data transformed before loading

Infrastructure: Traditional data warehouses with compute power

Data types: Best for structured data

Speed: Slower as transformation occurs beforehand



Use cases: Bank transactions, order history, conversion cycles

Popular Tools

- Informatica
- Talend
- Pentaho

ELT (Extract, load, Transform)

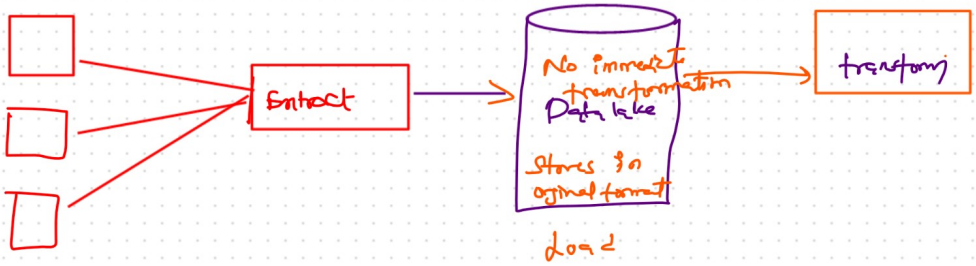
What is ELT?

Extract : Get data from multiple sources

load : Store raw data in a data lake

Transform: Process data after it's stored, as needed

Workflow

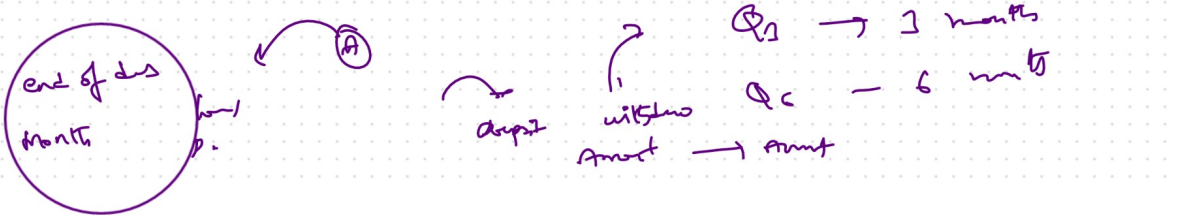


Use Case Example : Food Delivery App

* **Raw data collection :** Customer reviews (text), food photos (images), GPS data
* **Load into data lake :** Store all formats without transformation
* **Transform data :** Process as needed for specific analytics

ELT Characteristics

* **Processing location :** Data transformed after loading
* **Infrastructure :** Data lakes / cloud platforms with high compute power
* **Data types :** Handles structured, semi-structured & unstructured data
* **Speed :** faster data ingestion as transformation happens after loading
* **Use cases :** Social media analysis, storing raw data for future use



Common ELT Tools

- Azure Data Factory
- Azure Databricks
- AWS Glue
- AWS SJ

Career Options

Data Engineer vs BIS Data Engineering

Data Engineer — Core Concepts

Def: A data engineer designs, builds & maintains systems that can store process & make data accessible for analysis

Analogy: Water Pipeline Builder

Sources: Raw data from different origins (like water from river, lake, ocean)

Process: Design & build pipelines [to collect → clean → store → distribute data]

End Users: Data analysts, data scientists, management, others

Primary Responsibilities

- Data pipeline creation
- Data transformation
- ensuring data quality
- Automate

Traditional Tools & Approaches

→ SQL, Python, Airflow, ETL tools

→ ETL + Data warehouse

→ focus on structured data, manageable data

→ Storage : database based & data warehouses

ex: Monthly sales reports, Standard business analytics

Big Data Engineering

Definition: Specialize in tools & technologies for processing massive datasets that traditional systems cannot handle effectively

Analogy: DAM Engineer

- Handles "massive floods" of data
- Uses specialized systems for extreme volume & complexity
- required distributed computing & processing system

Characteristics of Big Data : (Five V's)

Volume : Extremely large datasets

Variety : Structured, semi-structured, & unstructured data

Velocity : high-speed data processing

Veracity : Data quality & reliability

Value : Extracting meaningful insights

Tools & Technologies:

* Hadoop, Spark, Hive, Kafka
* NoSQL databases
* Distributed Storage system
* Data lakes

Modern Approach

→ ELT + Big Data technologies

→ Parallel processing for large-scale data

→ eg: Analyzing twitter trends, real-time data processing