Page No. :
Date. : / /

Q1) Explain How ETL Works ?

Ans. ETL (Extract, Transform, Load) is a process used to collect data from different sources, process it and store it in a structured database or data warehouse. i.e. a centralized system.

- ETL Process steps :-

1. Extract

. Data is collected from multiple sources such as databases, files, or Legacy systems.

. The Data can be in different formats (structured, semi-structured, or unstructured.)

For Eg:- Extracting sales data from different branch database.

2. Transform.

. The Extracted data is cleaned and converted into a suitable format.

. This steps include the following :-

i) Removing Duplicates.

ii) Removing errors

iii) Converting data types (e.g. text to date Format.)

iv) Applying Business rules.

. The Goal is to ensure consistency and quality of Data.

Example :- Convert all date to same Format.

3. Load

. The transformed data is loaded into a target system like data warehouse or SQL Database.

. Loading can be:

i) Full Load : Entire data is loaded at once.

ii) Incremental Load: Only new or updated data is loaded

Page No. :
Date. : / /

- Processed data is stored into:
  - Data Warehouse
  - Data Lake.

Example: Load into a centralized reporting system

ETL Work Flow : - [Extract → Transform → Load.]

Q2.) What are the Benefits and Challenges of ETL?

ANS. Benefits of ETL:

1. Data Integration
   - Combine data form multiple sources into single system, making it easier to analyze.

2. Improved Data Quality.
   - Cleans and standardizes data by removing duplicate or errors.

3. Historical Data Storage.
   - Help maintain historical records, useful for trends, Forecasting and Analysis.

4. Automation and Scalability.
   - Removes manual intervention and handles large volume of Data.

5. Better Decision Making
   - Provides accurate and consistent data

6. Supports Data Warehousing.
   - Essential for Analytics and Reporting.

→ Challenges of ETL:-

1. Complexity.
   Designing ETL pipelines can be complicated, especially with large and diverse sources.

2. Time consuming
   - Processing large volumes of data may take significant time, especially during transformation.

3. Data Quality Losses.
   - If source data is inaccurate or in-complete, it can affect final output even after transformation.

4. Maintainance
   - Requires regular updates and monitoring.

5. Performance Issues
   - Slow processing for huge datasets.

6. Cost.
   - Infrastructure and tools can be expensive.