Name: Priyanshu Sahu
Batch: Data Analytics with Ai Nov Live Batch

Assignment 17: Data Extraction in ETL

Question: 1. Describe different types of data sources used in ETL with suitable examples

Answer: Different types of data sources:

1. Databases (Structured Data)

Example: MySQL
PostgreSQL
SQL Server
Oracle

2. Files

Example: CSV
Excel
JSON
XML
Text files

Name: Priyanshu Sahu
Batch: Data Analytics with Ai Nov Live Batch

3. APIs (Real-time Data)

- Payment gateway API
- Weather API
- Social media API

4. Cloud Storage

- AWS S3
- Google Cloud Storage
- Azure Blob

Example:
Uploading log files to S3 daily.

Question: 2. What is data extraction? Explain its role in the ETL pipeline.

Answer: Data Extraction:

Data extraction means collecting data from different sources for processing.

Role in ETL:

Name: Priyanshu Sahu
Batch: Data Analytics with Ai Nov Live Batch

ETL = Extract → Transform → Load

1. Extract → Get raw data
2. Transform → Clean and process
3. Load → Store in data warehouse

Example:

Extract sales data from MySQL → Clean duplicates → Load into Power BI database.

Question: 3. Explain the difference between CSV and Excel in terms of extraction and ETL usage.

Answer :

Feature	CSV	Excel
Multiple Sheets	Single sheet only	Supports multiple sheets
Formatting	No formatting (plain text)	Supports formulas,

		colors, borders
File Size	Small, lightweight	Larger due to formatting & metadata
Parsing Speed	Fast to read/parse	Slower due to structure and formatting
ETL Usage	High (ideal for automation & large data)	Moderate (better for manual analysis & reporting)

Question: 4. Explain the steps involved in extracting data from a relational database.

Answer:

1 Connect to database (using credentials)

2 Write SQL query

Name: Priyanshu Sahu
Batch: Data Analytics with Ai Nov Live Batch

3 Fetch required data

4 Export data (CSV / Data warehouse)

5 Validate extracted data

Example:

SELECT * FROM sales WHERE salary > 20000;

Question: 5. Explain three common challenges faced during data extraction.

Answer:

1. Large Data Volume

Problem: Slow performance
Solution: Batch processing

2. Data Quality Issues

Problem: Missing values, duplicates
Solution: Cleaning during transformation

Name: Priyanshu Sahu
Batch: Data Analytics with Ai Nov Live Batch

3. Data Security

Problem: Sensitive data exposure

Solution: Encryption & Access control

Question: 6. What are APIs? Explain how APIs help in real-time data extraction.

Answer:

API = Application Programming Interface

It allows systems to communicate.

Example:

Payment app sends data to server using API.

Real-time Data Extraction:

Instead of waiting for file,
API sends data instantly.

Example:

Stock price API gives live prices.

Name: Priyanshu Sahu
Batch: Data Analytics with Ai Nov Live Batch

Question: 7. Why are databases preferred for enterprise-level data extraction?

Answer: Databases preferred for enterprise-level data extraction because:

. Structured data
. Fast querying
. Security control
. Large data handling
. Backup & Recovery

Companies don't depend on Excel for enterprise-level work.

Question 8: What steps should an ETL developer take when extracting data from large CSV files (1GB+)?

Answer:

ETL Developer should:

Name: Priyanshu Sahu
Batch: Data Analytics with Ai Nov Live Batch

. Use chunk reading (read small parts)

. Use optimized tools (Spark)

. Remove unnecessary columns

. Check data types

. Handle memory properly

. Validate row count