# DATA EXTRACTION IN ETL

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

**Answer:** *In ETL, data sources refer to the locations from which raw data is extracted. These sources can vary widely based on business needs.*

1. ***Databases:***

   *Relational databases like MySQL, Oracle, SQL Server, and PostgreSQL are common ETL sources.*

   *Example: Customer data stored in a MySQL database.*

2. ***Flat Files:***
   *These include CSV, TXT, JSON, and XML files.*
   *Example: Sales data stored in a CSV file generated daily.*

3. ***Spreadsheets:***
   *Excel files are often used for small-scale or manual data storage.*
   *Example: Employee attendance data in an Excel sheet.*

4. ***APIs:***
   *APIs provide data from web applications or cloud services.*
   *Example: Fetching real-time weather data using a REST API.*

5. ***Cloud Storage & Applications:***
   *Data from platforms like Google Drive, AWS S3, Salesforce, etc.*
   *Example: Customer interaction data from Salesforce CRM.*

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

**Answer:** *Data extraction is the first phase of the ETL process, where data is collected from various source systems.*

***Role in ETL Pipeline:***

*The extraction process ensures that accurate, complete, and relevant data is retrieved without affecting the performance of the source systems. It acts as the foundation for transformation and loading. If extraction is incorrect, the entire ETL process becomes unreliable.*

***Extraction can be:***

- ***Full extraction:*** *The entire dataset is extracted.*
- ***Incremental extraction:*** *Only new or updated.*

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

**Answer:** ***CVS (Common Separated Values):***

- *Plain text format*
- *Lightweight and faster to process*
- *No formatting or formulas*
- *Ideal for large-scale ETL processes*
- *Easy to automate*

***Excel Files:***

- *Binary format with sheets, formulas, formatting*
- *Slower extraction compared to CSV*
- *Not suitable for very large datasets*
- *Often requires additional parsing*

***Difference:***

*CSV is preferred for ETL due to simplicity, speed, and scalability, while Excel is better for manual analysis and reporting.*

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

Answer: ***Steps involved:***

1. ***Understand database schema:***
   *Identify tables, relationships, and required columns.*

2. ***Establish connection:***
   *Use database credentials and drivers (JDBC/ODBC).*

3. ***Write SQL queries:***
   *Select required data using "SELECT" statements, joins, and filters.*

4. ***Execute extraction:***
   *Run queries to fetch data.*

5. ***Store extracted data:***
   *Save data temporarily in the staging area or files.*

6. ***Validate extracted data:***
   *Check row counts, data types, and completeness.*

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

**Answer:** ***Three common challenges faced during data extraction.***

1. ***Data inconsistency:***
   *Different formats, missing values, or incorrect entries cause extraction issues.*

2. ***Performance impact:***
   *Large extractions can slow down source systems, especially production databases.*

3. ***Schema changes:***
   *Changes in table structure can break extraction logic and cause failures.*

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

**Answer:** ***APIs (Application Programming Interfaces) allow different software systems to communicate and exchange data.***

***Role in real-time extraction:***

*APIs provide live or near real-time data access. ETL tools can call APIs at regular intervals to fetch updated data instantly.*

***Example:***

*Extracting real-time stock prices or social media data using REST APIs.*

*APIs are essential for modern ETL pipelines that require continuous data flow.*

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

Answer: ***Databases are preferred because they:***

- *Handle large volumes of data efficiently*
- *Support indexing and optimized queries*
- *Ensure data integrity and consistency*
- *Allow secure access control*
- *Enable incremental extraction using timestamps or IDs*

*Enterprise systems rely on databases because they are scalable, reliable, and performance-optimized.*

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

Answer: ***When handling huge CSV files, an ETL developer should:***

1. ***Use chunk-based processing:***
   *Read data in smaller chunks instead of loading the entire file at once.*

2. ***Validate file structure:***
   *Check delimiters, headers, and encoding before extraction.*

3. ***Remove unnecessary columns early:***
   *Reduce memory usage by extracting only required fields.*

4. ***Use staging area:***
   *Load data into temporary storage before transformation.*

5. ***Monitor performance and errors:***
   *Track memory usage, processing time, and failures.*

*These steps prevent system crashes and improve extraction efficiency.*