Data Extraction in ETL

QUESTION 1 **Describe different types of data sources used in ETL with suitable examples**

ANSWER

In **ETL,** data sources are places from where data is extracted before transforming and loading it into a data warehouse

1) APIs / Web Services

Data fetched from external applications using APIs.

Examples:
REST API
SOAP API

Example: **Getting weather or payment data**

2) Data Warehouses

Central repositories that store historical and analytical data.

Examples:
Amazon Redshift
Snowflake
Azure Synapse
Google BigQuery

Example:
Sales data loaded daily into Snowflake for reporting.

3) Relational Databases (RDBMS)

These store data in tables (rows & columns).

Examples:
MySQL
Oracle
SQL Server
PostgreSQL

Example use case:
A company stores customer and sales data in MySQL.

QUESTION 2	**What is data extraction? Explain its role in the ETL pipeline.**
ANSWER	Data Extraction is the first step of the ETL process.
	It means collecting or pulling data from different source systems so it can be processed further.
	Data extraction - getting data from where it is stored
	Data extraction is the process of retrieving data from various data sources such as databases, files, APIs, or applications and moving it to a staging area for transformation.
	Role of Data Extraction in the ETL Pipeline
	EXTRACT
	TRANSFORM
	LOADING
	Connects to multiple data sources
	Reads data without changing it
	Moves data to a staging area
	Ensures data is complete and accurate
	Minimizes impact on source systems
	Data extraction is the first step in ETL where data is collected from various source systems and moved to a staging area for further transformation and loading.

QUESTION 3	**Explain the difference between CSV and Excel in terms of extraction and ETL usage**
ANSWER	CSV and Excel are both flat file data sources, but they behave very differently in ETL.
	CSV (Comma-Separated Values)
	What is CSV
	Plain text file
	Data separated by commas
	No formatting, no formulas
	Very easy and fast to extract
	Lightweight file
	Same structure always
	Supported by all ETL tools
	No dependency on Excel software
	Excel File (.xls / .xlsx)
	What is Excel?
	Binary / structured file
	Supports multiple sheets
	Can contain formulas, charts, formatting

QUESTION 4	**Explain the steps involved in extracting data from a relational database**
ANSWER	**Identify the Data Source**
	First, decide from which database and tables data is required.
	Example:
	Database: sales_db
	Tables: customers, orders
	Purpose: Avoid extracting unnecessary data.
	Database Connection
	Create a secure connection using:
	Host name
	Port number
	Username & password
	Database name
	Select Required Data (Write SQL Query)
	Use SELECT queries to extract required columns and rows.

QUESTION 5		**Explain three common challenges faced during data extraction.**
ANSWER		Data extraction is the first step of ETL, but it comes with several challenges that can affect the entire pipeline.
		1) Data Quality Issues
		Missing values (NULLs)
		Duplicate records
		Incorrect or inconsistent data
	Example:	Customer age is NULL
		2) Performance & Source System Load
		Extracting large volumes of data
		Heavy queries slow down source systems
		3)Data Format & Schema Changes
		Source schema changes frequently
		Columns added, removed, or datatype changed
	Example:	New column discount added

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

ANSWER | **API stands for Application Programming Interface.**

An API is a set of rules and endpoints that allows one application to communicate with another application and exchange data.

1) Real-Time Data Access

APIs allow data to be fetched immediately when an event occurs.

example- New order placed → API sends order data instantly

2) Event-Driven Data Extraction

APIs support:
Webhooks
Streaming endpoints

Example: Payment completed → API triggers data extraction

3) Access to Live Systems

APIs connect directly to:
Cloud applications
SaaS platforms

Examples: Salesforce API
Google Analytics API

QUESTION 7	**Why are databases preferred for enterprise-level data extraction?**
ANSWER	Enterprises handle large, critical, and continuously growing data, so databases are the most reliable source for extraction.
	1) Structured and Consistent Data
	Databases store data in tables with fixed schema
	Enforced by constraints (PK, FK, NOT NULL)
	2) Handles Large Volumes of Data Efficiently
	Designed for millions of records
	Optimized with indexes and partitions
	3) Supports Incremental Data Extracti
	Uses timestamps, IDs, or CDC (Change Data Capture)
	4) High Performance and Reliability
	Databases support transactions
	Ensure ACID properties

QUESTION 8	**What steps should an ETL developer take when extracting data from large CSV files (1GB+)?**
ANSWER	Large CSV files need careful handling to avoid memory issues, slow performance, and failures.
	1) Understand the File Structure First
	Before extraction:
	Check column count & order
	Verify delimiter (, | ;)
	Identify header rows
	2) Avoid Loading the Entire File into Memory
	Don’t read the full CSV at once.
	3) Split the File (If Possible)
	For very large files:
	Split into smaller chunks (e.g., 200MB each)
	4) Use Staging Tables
	Load raw CSV data into:
	Staging tables
	Temporary tables