**1. What is ETL in data engineering, and why is it important?**

**ETL stands for:**

* **Extract:** Get data from different sources (like databases, files, APIs)
* **Transform:** Clean, filter, format, or enrich the data
* **Load:** Save the transformed data into a storage system (like a data warehouse)

**Why it’s important:**

It helps businesses make sense of **raw data** by converting it into **structured data** for analysis and decision-making.

**Real-world example:**

Let’s say you run an e-commerce website:

* **Extract:** Pull order data from MySQL and customer feedback from Google Sheets.
* **Transform:** Clean up duplicate orders, convert all prices to USD, and tag negative reviews.
* **Load:** Push the cleaned data to BigQuery for analysis.

**2. What are the key differences between cloud storage and traditional on-premise storage?**

| **Feature** | **On-Premise Storage** | **Cloud Storage (e.g., GCS)** |
| --- | --- | --- |
| Location | In your own server room | Hosted by Google in data centers |
| Cost | High upfront hardware cost | Pay-as-you-go model |
| Scalability | Limited, manual upgrades | Auto-scalable |
| Maintenance | You manage everything | Google maintains it |
| Access | Local or VPN-based access | Accessible from anywhere (internet) |

**Simple example:**

* If you use **on-premise**, you buy a hard disk and install it in your office.
* With **cloud storage**, you upload your files to GCS, and Google keeps them safe and available online.

**3. What is Google Cloud Storage (GCS), and what are some common use cases?**

🔹 **Google Cloud Storage (GCS)** is an object storage service for storing any amount of data (like images, videos, CSVs, backups).

**Key features:**

* Secure
* Scalable
* Cheap (has multiple storage classes like Standard, Nearline, Coldline)
* Supports public/private access

**Common use cases:**

* Hosting website images or videos
* Storing logs and backups
* Holding CSV or JSON data for ETL pipelines
* Keeping training datasets for machine learning

**4. What is the role of BigQuery in Google Cloud, and how is it different from a traditional database?**

**BigQuery** is Google’s serverless data warehouse. It’s made for **big data analysis** using **SQL**.

**Difference from traditional databases:**

| **Feature** | **Traditional DB (like MySQL)** | **BigQuery** |
| --- | --- | --- |
| Scaling | Manual | Auto-scaling |
| Performance | Slower for large datasets | Super fast for large queries |
| Maintenance | You manage | Google manages |
| Cost | Based on instance size | Pay only per query/GB |

**Real-world use:**

Imagine analyzing **millions of sales records**. BigQuery can scan **GBs of data in seconds** with simple SQL like:

SELECT category, SUM(amount)

FROM `myproject.sales.orders`

GROUP BY category;

**5. What are the main types of cloud service models (IaaS, PaaS, SaaS), and which Google Cloud services fall under each?**

**Cloud Service Models:**

| **Model** | **Full Form** | **You Manage** | **Example GCP Services** |
| --- | --- | --- | --- |
| IaaS | Infrastructure as a Service | Networking, VM | Compute Engine, VPC |
| PaaS | Platform as a Service | Apps, Data | App Engine, Cloud Functions |
| SaaS | Software as a Service | Just use the app | Google Docs, Gmail |

**Examples:**

* IaaS → You get a **virtual machine** on GCP (Compute Engine) and install whatever you want.
* PaaS → You just deploy code, and Google handles everything else (App Engine).
* SaaS → You just log in and use the app like **Google Sheets**.