All Interview questions Topic wise

---

1. BigQuery

- What is BigQuery?
- BigQuery Architecture.
- Internal vs External tables.
- View vs Materialized View.
- Authorized View vs Materialized View.
- Federated Queries in BigQuery.
- BQ Storage format (Columnar).
- What are partitions and clustering in BigQuery?
- Partition by vs Cluster by in BigQuery.
- How many columns can be used as partitions?
- What is the limit of clustering columns?
- Can we create index in BigQuery?
- BigQuery Slot Mechanism.
- Slot Reservation.
- Slot usage calculation.
- Cost difference between SELECT * and SELECT with Partition filter.
- Time Travel in BigQuery.
- Snapshot tables.
- Uniqueness about BigQuery.
- BigQuery Failover.
- BigQuery execution flow.
- BigQuery pricing (Computation and Storage).
- DML support in BigQuery.
- BigQuery query optimization techniques.
- How does BigQuery process SELECT * query internally?
- Group by vs Distinct – Which is more efficient?
- Row Level and Column Level Security.
- Policy Tags in BigQuery.
- How to get DDL for all tables in a dataset?

All Interview questions Topic wise

- How to get row count for 1000 tables at once?
- How to remove duplicates in BigQuery?
- Recovering deleted data (Time Travel).
- Updating schema of struct column.
- Data Quality Checks in BigQuery.
- How to load data from GCS to BigQuery?
- GCS to BigQuery transfer using Cloud Functions.
- Large dataset handling in BigQuery.
- Native vs External Tables.
- Partitioning types (3 types).
- Max partitions in BigQuery (4000).
- Data Warehouse in GCP.
- Handling schema changes in BigQuery.
- Query for 3 years continuous attendance.
- BigQuery vs SQL differences.
- BigQuery CLI, UI, Libraries (Ways to create tables).
- Columnar Storage in BigQuery.
- Removing duplicates and keeping latest records.
- Managing multi-group access.
- BQ Command line to create tables.
- Supported file formats for loading data.
- Real-time scenario-based questions.

---

2. SQL Queries & Concepts

- Second highest salary.
- Third highest salary (SQL & PySpark).
- Nth highest salary.
- Max salary in each department.
- Running total.
- Moving average (3-day and 10-day).
- Remove duplicates using Window functions.

All Interview questions Topic wise

- Rank(), Dense Rank(), Row Number() differences.
- Window functions vs Aggregate functions.
- Self Join vs Cross Join.
- Window function syntax.
- Joins (Inner, Left, Right, Full Outer).
- Count rows after different joins.
- Coalesce in SQL.
- Truncate vs Delete vs Drop.
- Filtering null values in Joins.
- Delete duplicates using ROW_NUMBER().
- Select temperature higher than the previous day.
- Lead and Lag functions.
- Accumulated sum of salary.
- Team-wise concatenation of members.
- Customers' first transaction analysis.
- Defaulters for 3 consecutive months.
- Routes with unique pairs.
- Difference between SQL and PostgreSQL.
- Types of Indexes in PostgreSQL.
- Partitioning in PostgreSQL.
- SQL query to handle prime numbers.
- Palindrome check in SQL.
- Query for max-min salary difference.
- Query to select salary rank from 10 to 45.
- Dynamic sum query based on column selection.
- Normalize data using SQL.
- Normalization in SQL.
- Handling large datasets in SQL.
- SQL routine.
- Removing junk rows from SQL output.

All Interview questions Topic wise

3. Airflow / Composer

- What is Airflow?
- DAG in Airflow.
- Airflow Operators (Python, Bash, Branch, Sensor, BQToGCS, GCSToBQ, TriggerRun, HTTP, etc.).
- Task Dependencies (Parallel vs Sequential tasks).
- XCom and methods.
- ExternalTaskSensor for DAG dependencies.
- Task Instance.
- Dummy Operator.
- BranchPythonOperator.
- Trigger Rules.
- Executors (Local, Sequential, Celery, Kubernetes).
- Control Executor.
- Sample DAG Code.
- How Airflow handles task failures.
- Version Control in Airflow.
- Reusable DAGs.
- Dynamic Workflows.
- DAG scheduling and reruns.
- Monitoring and troubleshooting in Airflow.
- Handling configurations in Airflow.
- Airflow vs Cloud Composer.
- Composer connection to Azure.
- DAG-to-DAG triggering.
- DAG code structure (Python/BashOperator example).
- Using Airflow for BigQuery-GCS transfer.
- Real-time scenario-based questions.

---

4. Dataflow

- What is Dataflow?

All Interview questions Topic wise

- Dataflow vs Dataproc.
- Predefined vs Custom Templates.
- PCollections.
- PaDo vs DoFn.
- ParallelDo in Dataflow.
- Windowing in Dataflow.
- Fault tolerance in Dataflow.
- Batch data processing in Dataflow.
- Real-time scenario (GCS to Dataflow to BigQuery).
- Data pipeline from GCS to Dataflow to BQ.
- Dataflow vs Datafusion.
- Transferring JSON from GCS to BigQuery via Dataflow.
- Apache Beam pipeline sample.
- Dynamic work rebalancing.
- Reading GCS files in Dataflow.
- Real-time scenario-based questions.

---

5. Dataproc

- What is Dataproc?
- Dataproc vs Dataflow.
- Spark and PySpark in Dataproc.
- RDD vs DataFrame.
- Spark Architecture.
- Spark Session.
- Transformations and Actions in Spark.
- Repartition vs Coalesce.
- Narrow vs Wide Transformations.
- Broadcast Join.
- Map vs FlatMap.
- GroupByKey vs ReduceByKey.
- Reading/Writing CSV, JSON, Parquet files.

All Interview questions Topic wise

- Handling 2TB data distribution.
- Fault tolerance in Spark (DAG, RDD).
- Spark optimization techniques.
- Dataproc cluster resizing.
- Different machine types in Dataproc.
- Real-time scenario-based questions.

---

6. GCS (Google Cloud Storage)

- Storage Classes.
- Object Lifecycle Management.
- Versioning in GCS.
- Bucket Versioning configuration.
- GCS Bucket Setup (Multi-region).
- Bucket lifecycle hooks.
- GCS to BigQuery integration.
- gsutil commands (List objects).
- Real-time scenario questions.
- Moving files from GCS to Archive.
- Detect file arrival in GCS.
- Bucketing in GCS.
- Limitations of GCS Buckets.

---

7. Pub/Sub

- What is Pub/Sub?
- Pub/Sub vs Kafka.
- Pub/Sub Message Acknowledgement.
- Pub/Sub in Dataflow streaming pipelines.
- Pub/Sub Scenario.
- Integrating with Dataflow.

---

8. Cloud Functions

All Interview questions Topic wise

- What is Cloud Functions?
- Event triggers in Cloud Functions.
- Upload CSV from GCS to BigQuery.
- Transferring data from GCS to BQ using Cloud Functions.

---

9. Python

- List vs Tuple.
- Lambda Function.
- Map Function.
- Reverse a string.
- List Comprehension.
- Count vowels in a string.
- Count frequency of elements in a list.
- Flatten a nested list using recursion.
- ODD/EVEN check using Lambda.
- Python program to read CSV.
- Merge DataFrames using Pandas.
- Extract substring from string.
- Add column in DataFrame.
- Converting Integer to String.
- Empty string to NULL.
- Generator and Decorators.
- Handling large datasets in Pandas.
- UDF in PySpark.
- PySpark Actions & Transformations.

---

10. General GCP Services

- IAM and Roles.
- Service Accounts.
- Cloud Run.
- VPC and Peering.

All Interview questions Topic wise

- Docker image storage.
- Namespaces.
- Deployment status in GCP.
- ETL vs ELT.
- OLAP vs OLTP.
- Facts and Dimensions.
- SCD Types (Type 1, 2).
- Star vs Snowflake Schema.
- Surrogate Keys.
- Distributed Computing.
- CAP Theorem.
- Data Modeling.
- Data Cleansing.
- Cost Optimization in GCP.
- Real-time streaming.
- Handling schema changes across GCP.
- Resilient systems.

---

11. Project & Responsibilities

- Self Introduction.
- Project Pipeline.
- Roles & Responsibilities.
- Daily Activities.
- Source and Destination in Pipeline.
- Transformations applied in the project.
- Handling production issues.
- Sprint duration.
- Team Structure.
- Business Impact if the pipeline fails.

---

12. Miscellaneous

All Interview questions Topic wise

- JIRA explanation.
- Story points estimation.
- Merge conflicts in Git.
- Rebase in Git.
- Stash in Git.
