

Tech Mahindra GCP Data Engineer Interview Guide – Experienced 3+

Technical Round 1

1. Python Code to Reverse a String

```
def reverse_string(s):
    return s[::-1]

print(reverse_string("TechM"))
```

2. Python Code Using Constructors in a Class

```
class Employee:
    def __init__(self, name, emp_id):
        self.name = name
        self.emp_id = emp_id

    def display(self):
        print(f"Name: {self.name}, ID: {self.emp_id}")

emp = Employee("Atul", 101)
emp.display()
```

3. Rank, Dense Rank, and Row Number

Differences:

- **RANK()**: Skips ranks for duplicates.
- **DENSE_RANK()**: Does not skip ranks.
- **ROW_NUMBER()**: Assigns unique ranks.

4. Setting Dependencies for Tasks in DAG

Using `set_upstream` and `set_downstream`:

```
task2.set_upstream(task1)
```

```
task3.set_downstream(task2)
```

5. Running Tasks in Parallel

- **Solution:** Define independent tasks and avoid setting dependencies between them.

6. Limiting Parallel Tasks

Set max_active_tasks:

DAG(max_active_tasks=3)

7. Query for 2nd Latest Joining Per Department

SQL Query:

```
SELECT dept_name, empid, name, salary, joining_date
FROM (
    SELECT *, ROW_NUMBER() OVER (PARTITION BY dept_name ORDER BY joining_date DESC) AS rnk
    FROM Employee
) subquery
WHERE rnk = 2;
```

19. Inner, Left, Right Joins

Table 1:

1
1
0
null

Table 2:

1
1
0
0
null
null

Joins:

- **Inner Join:** Matches common rows.
- **Left Join:** Includes all rows from Table 1 and matching rows from Table 2.
- **Right Join:** Includes all rows from Table 2 and matching rows from Table 1.

Technical Round 2

1. Provide Data Pipeline for GCP Data Engineering

- **Typical Pipeline:**

Data Ingestion: Use Pub/Sub or Dataflow to ingest data from various sources.

Data Storage: Store raw data in GCS (Google Cloud Storage).

Data Processing: Use Dataflow (Apache Beam) or Dataproc (Spark) for processing.

Data Warehousing: Load processed data into BigQuery for analytics.

Visualization: Use Looker Studio for creating dashboards.

2. Partitioning vs Clustering in BigQuery

- **Partitioning:** Splits data into segments based on column values (e.g., date).
 - Improves query performance for specific ranges.
- **Clustering:** Organizes data within partitions based on specified columns.
 - Optimizes queries with multiple filtering conditions.

3. Load JSON Files Stored in GCS to BigQuery

Steps:

1. Upload JSON files to GCS.
2. Use BigQuery console or CLI:

```
bq load --source_format=NEWLINE_DELIMITED_JSON \
[DATASET].[TABLE] gs://[BUCKET]/[FILE.json] [SCHEMA_FILE]
```

4. Difference Between Internal and External Tables in BigQuery

- **Internal Table:** Data is stored directly in BigQuery.
- **External Table:** Data resides in GCS but is queried through BigQuery without loading.

5. Remove Duplicate Rows in BigQuery

Query:

```
SELECT DISTINCT *
FROM `project.dataset.table`;
```

Or for deduplication based on specific columns:

```
SELECT *, ROW_NUMBER() OVER (PARTITION BY column_name ORDER BY column_name) AS row_num
FROM `project.dataset.table`
WHERE row_num = 1;
```

6. How to Use Dataflow with BigQuery

- Dataflow pipelines process data and write it to BigQuery.
- Example: Use Apache Beam SDK with Python or Java for writing data to BigQuery.

7. What is ParDo and Map?

- **ParDo**: Beam's transformation for parallel processing of elements in a PCollection.
- **Map**: Applies a function to each element in a dataset, similar to Python's map().

8. What is PCollection?

- **Definition**: Apache Beam's abstraction for distributed datasets.
- Immutable and can hold both bounded and unbounded data.

9. Are You Aware of Beam?

- **Answer**: Yes, Apache Beam is a unified programming model for batch and stream processing. It supports runners like Dataflow, Spark, and Flink.

10. Airflow Operators

- **Types**:

Action Operators: BashOperator, PythonOperator.

Transfer Operators: S3ToGCSOperator, GCSToBigQueryOperator.

Sensor Operators: ExternalTaskSensor, TimeDeltaSensor.

11. What is XCom in Airflow?

- **Cross-Communication**: Mechanism to share data between tasks in a DAG.

Glassdoor Tech Mahindra Review –

<https://www.glassdoor.co.in/Reviews/Tech-Mahindra-Reviews-E135932.htm>

Tech Mahindra Careers –

<https://www.techmahindra.com/careers/>

Subscribe to my YouTube Channel for Free Data Engineering Content –

<https://www.youtube.com/@shubhamwadekar27>

Connect with me here –

<https://bento.me/shubhamwadekar>

Checkout more Interview Preparation Material on –

https://topmate.io/shubham_wadekar