Verizon Azure Data Engineer Interview Guide – Experienced 3+

Round 1: SQL, Python, BigQuery, and Data Pipelines

1. Benefits of ETL

- Centralizes data from multiple sources.
- Enhances data quality and consistency.
- Optimizes storage and query performance.
- Enables better business insights and decision-making.

2. About Jira

- A project management tool for tracking tasks, bugs, and workflows.
- Features include sprint planning, customizable dashboards, and integration with CI/CD pipelines.

3. Integrating an API with a Database

- Steps:
    1. Fetch data from the API using HTTP requests.
    2. Parse the response (e.g., JSON or XML).
    3. Insert parsed data into the database using SQL queries.

4. SQL Query: Group Employees by Technology and Order by ID

```sql
SELECT technology, ARRAY_AGG(STRUCT(emp_id, name, salary) ORDER BY emp_id) AS employees
FROM employee_table
GROUP BY technology;
```

5. Incremental Data Load in BigQuery

- Techniques:
    - Use timestamps or version columns to identify changes.
    - Merge new data with existing tables using DML (e.g., MERGE).

6. Steps to Verify Source and Target Data Match After Load

1. Count validation (row counts in source and target).
2. Checksum/hash comparison.
3. Spot-check data samples.
4. Field-by-field data comparison.

7. Schema Changes in Source (New Column Addition)

- Impact:
    - **Landing Zone Tables:** May require schema evolution.
    - Update ETL pipelines to accommodate the new column.

8. Retrieve Schema of BigQuery Table

```
SELECT column_name, data_type
FROM `project_id.dataset.INFORMATION_SCHEMA.COLUMNS`
WHERE table_name = 'your_table';
```

9. Nested and Repeated Fields in BigQuery

- **Nested Fields:** Allow storing complex data types like structs.
- **Repeated Fields:** Allow storing arrays within a table.

10. Stored Procedures in SQL

- Encapsulates reusable logic in SQL.
- Example:

```
CREATE PROCEDURE update_salary()
BEGIN
  UPDATE employee_table SET salary = salary * 1.1;
END;
CALL update_salary();
```

Round 2: Cloud Composer, Git, GCP Authentication, and Data Engineering Concepts

11. Cloud Composer Overview

- Managed Apache Airflow service on GCP.
- Automates data workflows with DAGs.

12. Defining Tasks in DAG

- **PythonOperator:** Executes Python code.
- **BashOperator:** Runs bash scripts.
- **BigQueryOperator:** Executes SQL in BigQuery.

13. Task Dependencies in DAG

- Use >> or .set_downstream() for sequential dependencies.

task1 >> task2

14. Explain XComs

- Cross-communication between tasks.
- Parent DAG can pass variables to the child DAG using xcom_push and xcom_pull.

15. Push and Pull in Tasks

- **Push:** Store task-specific data using xcom_push.
- **Pull:** Retrieve stored data using xcom_pull.

16. Using BashOperator to Trigger Python Script with Arguments

```
BashOperator(
    task_id='run_script',
    bash_command='python script.py arg1 arg2'
)
```

17. Virtual Environment in Python

- Isolates project dependencies.
- Commands:
    - python -m venv venv_name
    - source venv_name/bin/activate

18. GCP Authentication with Jenkins

- Use service accounts with JSON key files.
- Configure authentication using gcloud auth activate-service-account.

19. Using Service Accounts in GCP

- Access control for applications.
- Example:
    - Assign roles to a service account.
    - Authenticate via a key file.

20. Connecting Local Machine to GCP Console

- Install gcloud CLI.
- Authenticate using gcloud auth login.

21. Deploying DAGs

- Copy files to Composer bucket:

gsutil cp dag_file.py gs://composer-bucket/dags/

22. Git: Copying a Branch

git checkout existing_branch

git checkout -b new_branch_name

23. Git Bash Commands

- git status: Check repository status.
- git add: Stage changes.
- git commit: Commit changes.
- git push: Push commits to remote.

24. Discarding Local Changes

git reset --hard

git clean –fd

25. Git Stash

- Temporarily saves changes without committing.
- Changes are not pushed to the remote.

26. Cloud Functions

- Event-driven serverless execution.
- Real-time use case: Trigger ETL on new file upload.

27. Connecting BigQuery with Linux

- Install bq command-line tool.
- Execute queries using:

```
bq query --use_legacy_sql=false 'SELECT * FROM dataset.table'
```

28. Real-Time Scenarios with Stored Procedures

- Automating repetitive tasks like daily updates.
- Example: Archiving old data.

29. Business Role of Data Pipeline

- Data ingestion, transformation, and analytics for decision-making.

Glassdoor Verizon Review –

https://www.glassdoor.ca/Reviews/Verizon-Reviews-E8986492.htm

Verizon Careers –

https://mycareer.verizon.com/jobs/

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