Globant Data Engineer Interview Guide – Experienced 3+

Technical round 1 and round 2 combined

1. Introduce Yourself

Prepare a brief and engaging self-introduction:

- Mention your experience, key skills (Spark, Hive, SQL, Python, etc.), and impactful projects.
- Highlight relevant certifications or achievements.

2. Difference Between head() and take() in PySpark

- head(n) returns the first n rows of a DataFrame as a single list of Row objects.
- take(n) returns the first n rows but as a list of rows, similar to collect().
  Usage: take() is preferred for retrieving specific rows because it avoids overhead compared to collect().

3. Convert Array Column to Multiple Columns – PySpark

Use selectExpr with posexplode to split array elements into separate columns.

df = spark.createDataFrame([(1, [10, 20, 30])], ["id", "array_column"])

df.selectExpr("id", "array_column[0] as col1", "array_column[1] as col2", "array_column[2] as col3").show()

4. Drop Columns with Null Values – PySpark

Use dropna with subset.

df.dropna(how='all', subset=['column_name']).drop('column_name').show()

5. Dynamic Partition Pruning Error

- Dynamic Partition Pruning (DPP) is used in Spark for runtime filtering.
- Common errors: Occur due to unsupported join conditions or improper configuration.
  Fix: Enable with spark.sql.optimizer.dynamicPartitionPruning.enabled=true.

6. Read and Write Modes in Spark

- Read Modes: PERMISSIVE (default), DROPMALFORMED, FAILFAST.
- Write Modes: overwrite, append, ignore, error (default).

7. Keep a Specific Column on Top (SQL)

Use **CASE** for ordering. Example:

SELECT country FROM table

ORDER BY CASE WHEN country = 'US' THEN 0 ELSE 1 END;

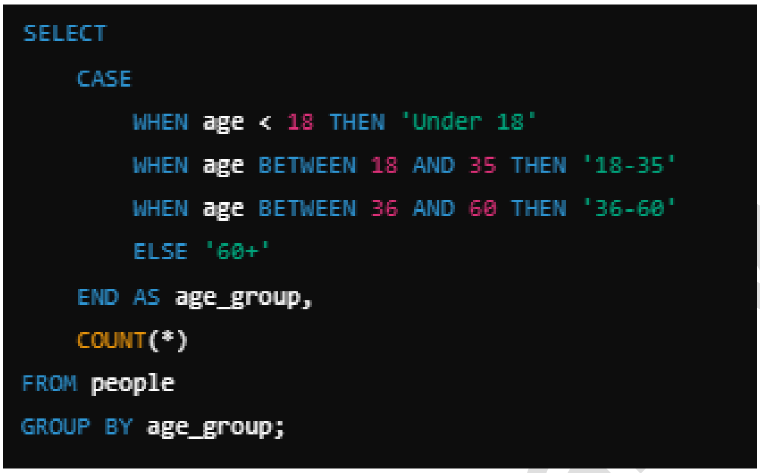
8. Count Occurrences in a Column (SQL)

SELECT column_name, COUNT(*)

FROM table

GROUP BY column_name;

9. Age Bracket Division (SQL)



```
SELECT
    CASE
        WHEN age < 18 THEN 'Under 18'
        WHEN age BETWEEN 18 AND 35 THEN '18-35'
        WHEN age BETWEEN 36 AND 60 THEN '36-60'
        ELSE '60+'
    END AS age_group,
    COUNT(*)
FROM people
GROUP BY age_group;
```

10. How Adaptive Query Execution (AQE) Works

AQE optimizes Spark queries at **runtime** by:

- Dynamically choosing join strategies.
- Dynamically optimizing partition sizes.
- Handling skewed joins automatically.

11. Difference Between MapReduce and Spark

- MapReduce is disk-based and processes in stages.
- Spark is in-memory, allowing faster execution and more complex transformations.

12. Checkpointing in Spark

Checkpointing saves the RDD/Dataset state to persistent storage to handle failures.

- Types: Metadata checkpointing and Data checkpointing.

13. Serializer in Spark

- Serializers reduce the cost of object serialization in distributed computing.
- JavaSerializer (default) and KryoSerializer (more efficient).

14. Convert 3 Rows into One Column (SQL)

```
SELECT GROUP_CONCAT(column_name SEPARATOR ', ')
FROM table;
```

15. Check if Two Strings are Anagrams – Python Example

```
def are_anagrams(str1, str2):
    return sorted(str1) == sorted(str2)
```

Glassdoor Globant Review –

https://www.glassdoor.co.in/Reviews/Globant-Reviews-E150678.htm

Globant Careers –

https://career.globant.com/

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