KPMG Data Engineer Interview Guide – Experienced 4+

Overview

In this detailed guide, we'll walk through a 3-round interview process for a Data Engineering role. This experience includes conceptual questions, hands-on coding, and behavioral discussions. By breaking down each round, we'll focus on the key expectations, example questions, and strategies to ace each part.

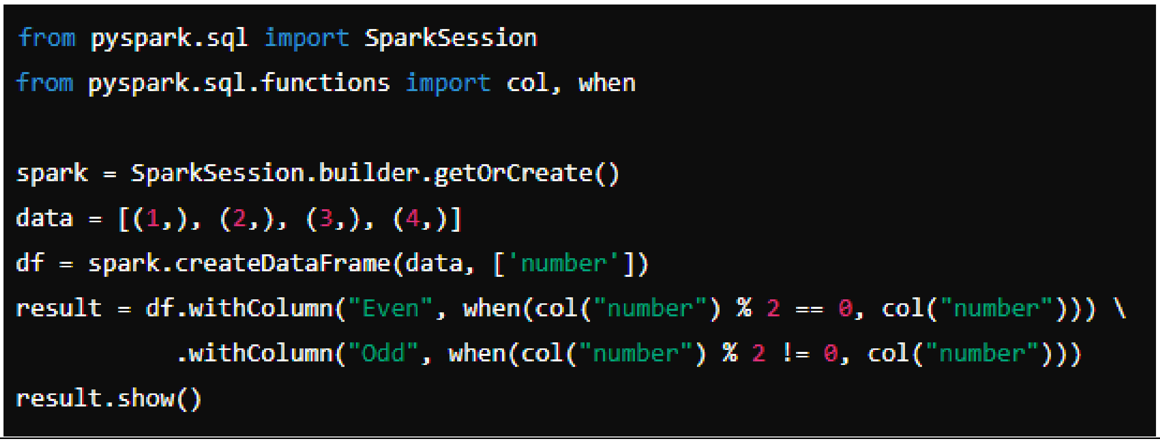
Round 1: Technical Interview

Overview

The first round assesses your technical proficiency, problem-solving abilities, and understanding of the tech stack. Be ready to discuss your experience and justify your technical decisions.

Interview Process Breakdown

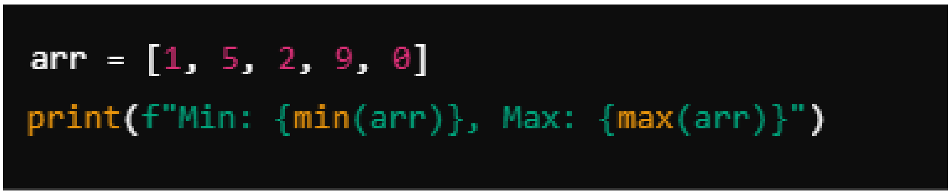
1. **Work & Conceptual Questions**
    - Introduce yourself, highlighting key projects and tech stacks.
    - Explain your day-to-day responsibilities as a Data Engineer.
    - Justify the choice of your current tech stack. Why Spark, Hadoop, or cloud platforms?
    - **Key Discussion**: Alternatives to the Medallion Architecture.
    - Demonstrate knowledge of file formats, like converting JSON to Parquet to improve efficiency.
      Tip: Talk about Parquet's columnar storage benefits like compression and faster querying.
    - Discuss the nature and volume of data you manage daily.
2. **Coding Questions**
    - ○ **PySpark Problem**:
      Split a DataFrame such that even numbers appear in one column and odd numbers in another.



```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, when

spark = SparkSession.builder.getOrCreate()
data = [(1,), (2,), (3,), (4,)]
df = spark.createDataFrame(data, ['number'])
result = df.withColumn("Even", when(col("number") % 2 == 0, col("number"))) \
           .withColumn("Odd", when(col("number") % 2 != 0, col("number")))
result.show()
```

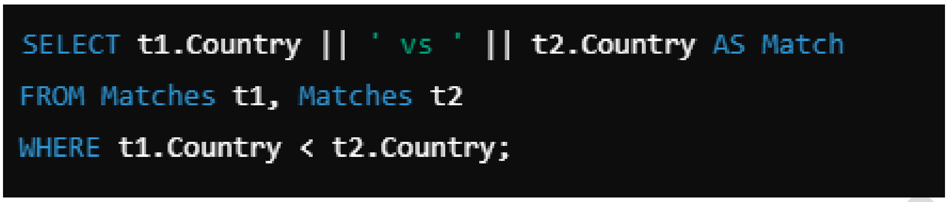
Python Problem:
Find the minimum and maximum values in an array.



```python
arr = [1, 5, 2, 9, 0]
print(f"Min: {min(arr)}, Max: {max(arr)}")
```

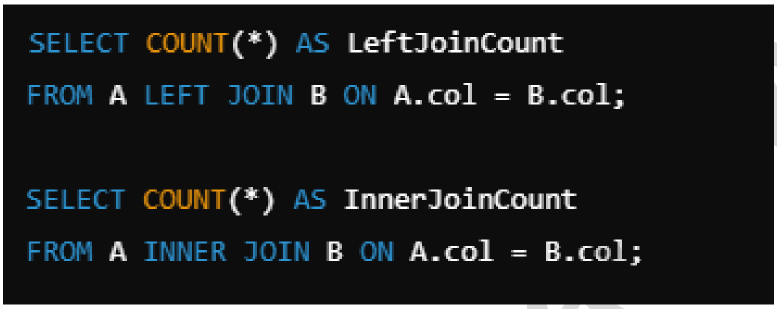
SQL Problem:

Match countries in a pairwise format:



```sql
SELECT t1.Country || ' vs ' || t2.Country AS Match
FROM Matches t1, Matches t2
WHERE t1.Country < t2.Country;
```

Find Left Outer Join and Inner Join record counts.



```sql
SELECT COUNT(*) AS LeftJoinCount
FROM A LEFT JOIN B ON A.col = B.col;

SELECT COUNT(*) AS InnerJoinCount
FROM A INNER JOIN B ON A.col = B.col;
```

Demonstrate the difference between DENSE_RANK() and RANK().

Tips

- Understand PySpark fundamentals like withColumn, transformations, and actions.
- Brush up on Python coding basics—strings, arrays, and dictionaries.
- Master SQL joins, window functions, and aggregation queries.

Round 2: Technical Interview

Overview

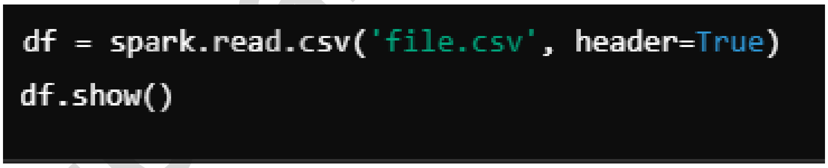
This round dives deeper into Spark internals, coding challenges, and real-world problem-solving.

Interview Process Breakdown

1. **Work & Conceptual Questions**
    - Walkthrough Spark's architecture, focusing on driver, executors, and DAGs.
    - Explain job execution in Spark:
        - Spark jobs are split into stages, tasks, and optimized using the Catalyst Optimizer.
    - Discuss Logical Plan vs Physical Plan: Logical plan is generated first and optimized into an execution-ready physical plan.
    - Compare ORC and Parquet:
        - Parquet: Best for analytics (columnar).
        - ORC: Optimized for Hive workloads (compression).
2. **Coding Challenges**

    PySpark Problems:

    - Read a CSV file into a DataFrame:

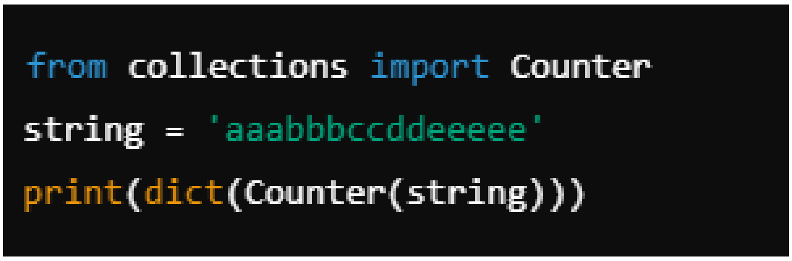


```
df = spark.read.csv('file.csv', header=True)
df.show()
```

    - Create a DataFrame with default column types.

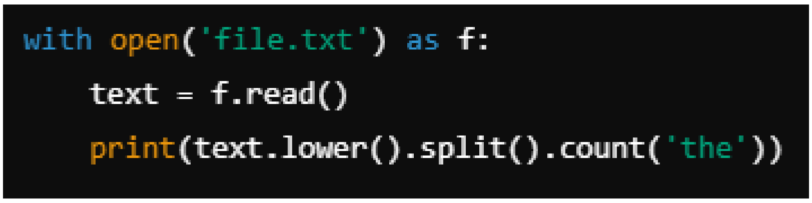
    Python Problems:

    - Count occurrences of each character in a string.



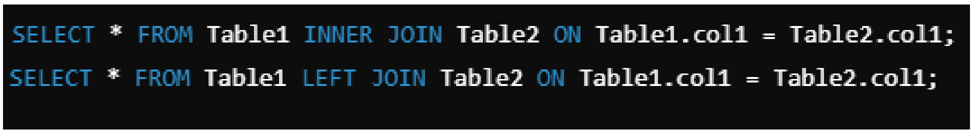
```
from collections import Counter
string = 'aaabbbccddeeeee'
print(dict(Counter(string)))
```

    - Count occurrences of a specific word in a file.



```
with open('file.txt') as f:
    text = f.read()
    print(text.lower().split().count('the'))
```

SQL Problems:
Perform **Inner**, **Left**, **Right**, and **Full** Joins.



```
SELECT * FROM Table1 INNER JOIN Table2 ON Table1.col1 = Table2.col1;
SELECT * FROM Table1 LEFT JOIN Table2 ON Table1.col1 = Table2.col1;
```

Tips

- Practice coding with PySpark transformations and Python data structures.
- Dive deep into Spark internals like the Catalyst Optimizer and logical/physical plans.
- Be proficient with advanced SQL joins, aggregations, and string manipulations.

Round 3: Hiring Manager Discussion

Overview

The final round evaluates your career goals, motivation, and fit within the company.

Interview Process Breakdown

1. **Conceptual Questions**
   - Data Lake vs Delta Lake: Delta Lake provides ACID transactions, schema enforcement, and time travel on top of Data Lakes.
2. **Behavioral Questions**
   - Why did you leave your previous job?
   - If you already have an offer, why are you exploring other roles?
   - Are you willing to relocate to Bangalore?

Glassdoor KPMG Review –

https://www.glassdoor.co.in/Reviews/KPMG-Reviews-E2867.htm

KPMG Careers –

https://kpmg.com/in/en/careers.html

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