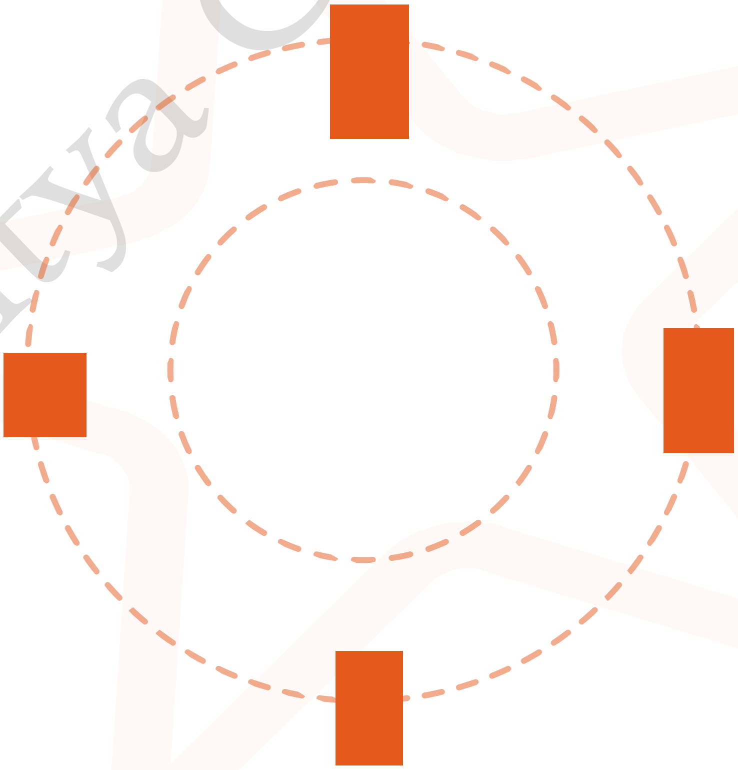
Nitya CloudTech Pvt Ltd.

PySpark Scenario-Based Interview Questions & Answers



📌 ROUND 1: Technical Interview (60 Minutes)

1. Tell me about your project and challenges you faced?
Candidate:
I explained the details of my recent project where I developed a **Data Pipeline using PySpark**. One of the challenges I faced was handling large volumes of data with multiple sources and ensuring optimal data loading times. I tackled this by implementing **incremental data loads** and optimizing **Spark queries** using caching and partitioning.

2. Difference between Incremental load and Full load?
Candidate:
In **full load**, the entire dataset is loaded every time. In **incremental load**, only the new or modified data since the last load is processed, which reduces load time and resource usage.

3. How much data are you loading daily? Any scenario-based questions related to the project?
Candidate:
I load around **X TB of data daily**, primarily from transactional systems. I explained the **ETL pipeline architecture** I used, and how we ensured **data consistency** during large-scale loads.

4. What is the difference between PySpark and Pandas?
Candidate:
While **Pandas** is optimized for **single-node** operations on smaller datasets, **PySpark** is built for **distributed computing**, which allows it to handle larger datasets across multiple nodes in a cluster.

5. Spark Optimization Techniques?
Candidate:
Some techniques include **caching**, **partitioning**, **broadcast joins**, **repartitioning**, and **increasing shuffle partitions** to avoid bottlenecks and reduce processing time.

6.
Spark architecture?
Candidate:
The **Spark architecture** consists of **Driver**, **Executors**, and **Cluster Manager**. The **Driver** controls the execution, while **Executors** execute tasks, and the **Cluster Manager** allocates resources to executors.

7. Top 3 salaries in the department?
Candidate:
I shared insights about how **compensation trends** are set based on market benchmarks and internal company policies, focusing on roles in **data engineering**, **machine learning**, and **data science**.

8. OLTP vs OLAP?
Candidate:
OLTP (Online Transaction Processing) systems handle high-volume transactional data, while **OLAP (Online Analytical Processing)** systems are designed for complex queries and data analysis, typically used in business intelligence.

9. What is SCD? Types?
Candidate:
SCD (Slowly Changing Dimensions) refers to handling changes in data over time. The main types are:

- **SCD Type 1**: Overwrite old data with new.
- **SCD Type 2**: Add new records to maintain history.
- **SCD Type 3**: Store limited historical data in the same record.

📌 ROUND 2: Technical + Managerial Interview (60 Minutes)

1. Project details? More scenario questions related to the project?
Candidate:
I shared detailed insights about my **ETL project**, the challenges we faced, and how I optimized data transformation using **PySpark**. I also

walked through
scalability and **performance improvements** we implemented.

2. Row() to column() change problem in both SQL and Python?
Candidate:
I explained the process of converting **rows to columns** using **pivoting** in SQL, and using **Pandas' pivot_table** function in Python to achieve the same.

3. What is salting?
Candidate:
Salting is a technique used to handle **data skew** in Spark by adding a random value (a “salt”) to partition keys, which ensures that data is evenly distributed across partitions.

4. Spark optimizations?
Candidate:
I discussed various **Spark optimizations**, such as using **broadcast joins** for small datasets, **caching** data for repeated access, **repartitioning** data for more balanced partitions, and configuring **spark.sql.shuffle.partitions** for optimal performance.

5. File formats in Spark and its use cases?
Candidate:
I explained the benefits of **Parquet** (optimized for columnar storage and performance), **ORC** (used in Hive and also columnar), and **CSV/JSON** (more flexible but slower for large datasets) depending on the use case.

6. How do you handle Data skewness?
Candidate:
I explained how I handle **data skew** by using **salting** for key columns, **broadcast joins** for small tables, and using **Spark’s skew hints** to manage skewed data more efficiently.

7. Why are you changing companies? Why are you asking for a 100% hike?
Candidate:

I
emphasized my growth aspirations and the desire to work with a **leading organization** like DBS, where I can expand my skills and contribute to **cutting-edge data projects**. The 100% hike request was based on industry standards and the value I can bring to the team.

8. Suppose if you joined DBS Bank and the person you have to take K.T. (Knowledge Transfer) from has a difficult personality. How would you handle it?
Candidate:
I would approach the situation with **empathy** and **professionalism**, ensuring clear communication and patience, while trying to understand their perspective. If required, I would escalate to a **manager** to ensure smooth knowledge transfer.

If you like the content, please consider supporting us by sharing it with others who may benefit. Your support helps us continue creating valuable resources!

Scan any QR using PhonePe App

Aditya chandak