Infosys Data Engineer Interview Guide – Experienced 3+

Round 1: Technical Interview

1. **Tell me about your project**: Explain your project, its goals, and the technologies you used.

2. **Hadoop Commands for Get Merge**: What Hadoop command would you use to merge multiple files into one?

3. **Hadoop Architecture**: Can you explain the architecture of Hadoop and its components?

4. **Spark Session vs Spark Context**: What is the difference between SparkSession and SparkContext in Spark?

5. **Handling Null Values in Spark**: How would you handle null values in a dataset, especially in a single column (not all columns)?

6. **YARN**: What is YARN, and how does it manage resources in a Hadoop ecosystem?

7. **Map vs FlatMap**: What is the difference between map and flatMap in Spark, and when would you use each?

8. **Sqoop Incremental and Job**: Can you explain the concept of incremental loading in Sqoop and how to use it for job processing?

9. **Performance Tuning in Sqoop and Spark**: What performance tuning techniques do you apply in both Sqoop and Spark to optimize their execution?

10. **Left Anti Join Scenario**: Explain a scenario where you would use a LeftAntiJoin in Spark SQL.

Round 2: Technical Interview

1. **Web API Read**: How would you read data from a web API? What steps would you follow after reading the data?

2. **Do you know OOzi**: Have you worked with OOzi? If yes, can you explain what it is and how it's used in data pipelines?

3. **Employee Count by Department**: Given the data with id, name, and department, how would you calculate how many employees are in each department?

4. **Mappers in Spark**: Can you explain the concept of mappers in Spark, and how are they used in data transformations?

5. **Reading RDBMS Data Using Spark**: How would you read data from an RDBMS using Spark? Provide the syntax.

   Answer: spark.read.format("jdbc").option("url", "jdbc:mysql://localhost:3306/database_name").option("dbtable", "table_name").option("user", "username").option("password", "password").load()

6. **Executor Memory in Spark**: What work is done by the executor memory in Spark?

7. **Broadcasting in Spark**: What is broadcasting in Spark, and why is it used? Can you give an example of its use?

8. **Scala Traits**: What are traits in Scala? How are they different from classes?

9. **DBUtils Function in Databricks**: Explain the use of the dbutils function in Databricks.

10. **Moving Files in DBFS**: How would you move a file to another path in Databricks File System (DBFS)?

11. **Creating a Job Cluster in Databricks**: How do you create a job cluster in Databricks?

12. **Lazy Evaluation in Spark**: What is lazy evaluation in Spark, and how does it improve performance?

13. **Managed vs External Tables**: What is the difference between managed and external tables in Hive or Spark SQL?

14. **Delta Lakehouse**: Explain the concept of a Delta Lakehouse and its architecture.

15. **Bronze/Silver/Gold Layer in Data Pipeline**: What is the purpose of the Bronze, Silver, and Gold layers in a data pipeline?

16. **Deploying from Dev to QA/Prod**: How do you deploy from a development environment to QA and production?

17. **Job Creation in Databricks**: What are the steps to create and schedule a job in Databricks?

Summary

- Round 1 (Technical 1): Focuses on Hadoop, Spark basics, Sqoop, and performance optimization techniques. It also covers data handling scenarios like null value management and join operations.
- Round 2 (Technical 2): Moves into Databricks, web API integration, advanced Spark
- concepts, and Scala traits. It also dives into deployment, Delta Lakehouse, and job scheduling in Databricks.

Glassdoor Infosys Review –

https://www.glassdoor.co.in/Reviews/Infosys-Reviews-E7927.htm

Infosys Careers –

https://www.infosys.com/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