**Coforge Data Engineer Interview Guide – Experienced 2+**

**Round 1: Technical Round**

1. **Project Explanation**: Explain your project and the technologies used so far.

2. **Traits in Scala**: What are traits in Scala, and how are they different from classes?

3. **String in Scala**: How are strings handled in Scala? How are they different from Java strings?

4. **Cache and Persist**: What are the differences between cache() and persist() in

Spark? When do you use each?

5. **Group By Key and Reduce By Key**: What is the difference between groupByKey and reduceByKey in Spark?

6. **DAG (Directed Acyclic Graph)**: Explain the DAG in Spark and how it plays a role in execution.

7. **Stages in Spark Job**: How many stages are created in a Spark job, and how are they formed?

8. **Partitions and Repartitions in Spark**: What is the difference between partitions and repartitions in Spark, and when do you use each?

9. **Map and FlatMap**: What is the difference between map and flatMap in Spark transformations?

10. **Narrow and Wide Transformations**: Explain the difference between narrow and wide transformations in Spark.

11. **Spark Architecture**: Can you explain the architecture of Apache Spark and its components?

12. **Join Operations**: Given the data below, explain the results of different types of joins:

 **Data1**:

**a b** x 1 x 1

y 2

 **Data2**:

**c d** x 2 x 1 x 1

y 3

**c d**

 Perform:

 Inner Join

 Left Join

 Right Join

 Will a schema be created?

13. **Difference Between DataFrame and Datasets**: Can you explain the difference between DataFrame and Dataset in Spark?

14. **Partitioning vs. Bucketing**: What is the difference between partitioning and bucketing in Spark, and when would you use bucketing?

15. **Disadvantage of Scala**: What is one disadvantage of using Scala for data engineering tasks?

16. **Creating External Table in Hive**: How do you create an external table in Hive?

17. **Print Prime Numbers Using Scala**: Write a Scala code to print prime numbers.

18. **Find Highest Salary**: How would you find the highest salary from a dataset in

Spark?

19. **Lazy Evaluation**: What is lazy evaluation in Spark? Explain with an example.

20. **Import Command from HDFS to Hive**: What is the command to import data from

HDFS to Hive?

21. **Handling Unstructured Data in Hive**: How would you handle unstructured data in

Hive?

22. **Memory Check in Linux**: How do you check the memory of your laptop using Linux commands?

23. **Jenkins Question**: Can you explain your experience with Jenkins in your project?

**Round 2: Technical + Managerial Round**

1. **Spark Performance Tuning**: What are the key performance tuning techniques you apply in Spark jobs to improve performance?

2. **Dynamic Resource Allocation in Spark**: Can you explain dynamic resource allocation in Spark? How does it help optimize job performance?

3. **Handling Large Datasets in Spark**: How do you handle very large datasets in Spark to ensure scalability and efficiency?

4. **Fault Tolerance in Spark**: Explain how Spark handles fault tolerance. How does it recover from node failures?

5. **UDFs (User Defined Functions)**: Have you worked with UDFs in Spark? When do you use them, and how do they differ from built-in functions?

6. **Distributed Computing**: How does Spark handle distributed computing, and what challenges have you faced while working on distributed systems?

7. **Data Modeling in Spark**: Can you walk me through a scenario where you had to design an efficient data model in Spark for a real-world application?

8. **Handling Schema Evolution**: How do you handle schema evolution in Spark, especially when reading data from sources like Parquet or Avro?

9. **Managing Data Quality**: How do you ensure data quality in a big data pipeline, and what strategies do you use for data validation?

10. **Handling Data Shuffling**: What is data shuffling in Spark, and how do you minimize its impact on job performance?