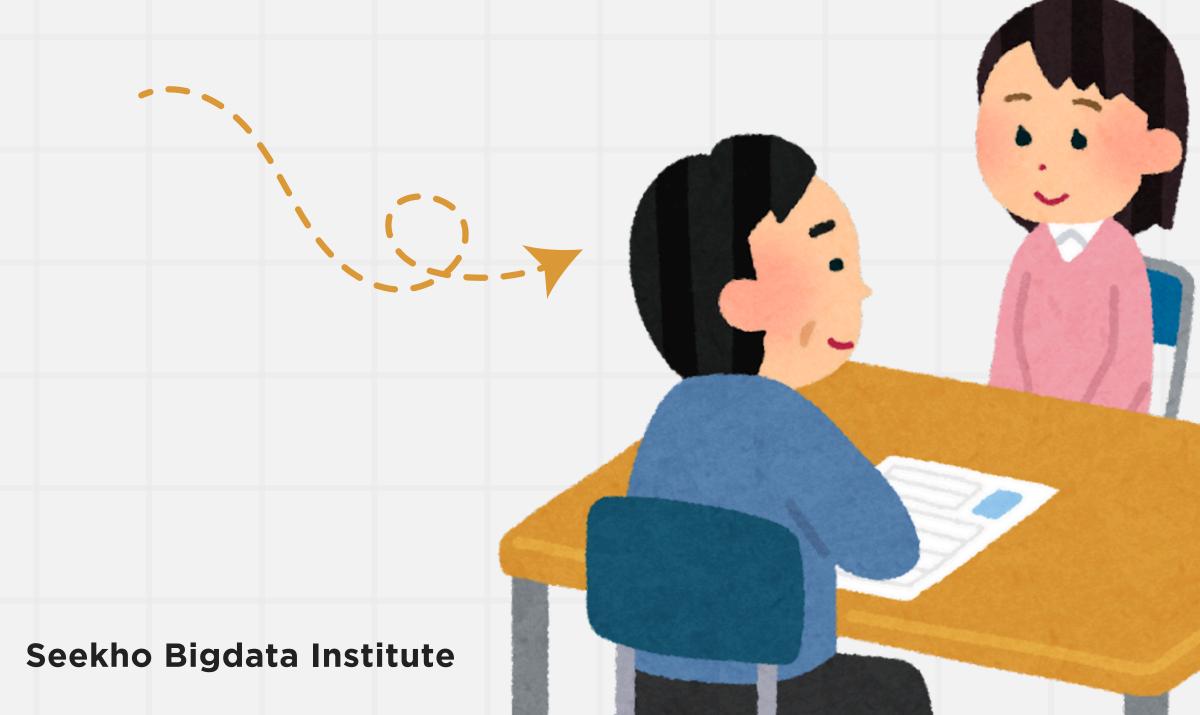


Don't Attend a Data Engineering Interview in 2025 Without These

PySpark Questions!



- 1. Why is Spark processing faster than MapReduce jobs?
- 2. Spark vs. MapReduce
- 3. Why was Spark developed?
- 4. What is Spark?
- 5. What is PySpark?
- 6. What are the characteristics of PySpark?
- 7. Features of Spark and Advantages Disadvantages of PySpark?
- 8. What is a Spark Driver?
- 9. PySpark Architecture?
- 10. PySpark Modules & Packages



- 11. Spark Components?
- 12. What is SparkContext?
- 13. What is SparkSession? Explain.
- 14. SparkContext vs. SparkSession
- 15. Repartition() vs. Coalesce()?
- 16. Difference between Cache and Persist?
- 17. What is Unpersist?
- 18. What is the difference between Broadcast Variable and Accumulator Variable?
- 19. What is Shuffling in Spark?
- 20. Difference between groupByKey() vs. reduceByKey() vs. aggregateByKey() vs. sortBy() vs. sortByKey()?



- 21. What is RDD?
- 22. How to create RDD?
- 23. Types of RDDs?
- 24. When to use RDDs?
- 25. What are RDD Operations Transformations and Actions?
- 26. map vs. flatMap vs. filter?
- 27. collect vs. collectAsList vs. select()?
- 28. Why is DataFrame (DF) faster than RDD?
- 29. RDDs vs. DataFrames vs. Datasets?
- 30. How to Pivot and Unpivot a Spark DataFrame?



- 31. What is Spark Schema?
- 32. GroupBy clause?
- 33. What is Spark SQL DataFrame?
- 34. Why use DataFrame?
- 35. Is PySpark faster than Pandas?
- 36. What is DAG and Lineage Graph, RDD Lineage?
- 37. What is Paired RDD?
- 38. What is Skewness?
- 39. How to mitigate skewed data?
- 40. Optimization Techniques in Spark



- 41. How to read CSV file using delimiter ','?
- 42. What is Star Schema & Snowflake Schema? Differentiate Star & Snowflake.
- 43. What is Data Skewness?
- 44. What is Catalyst Optimizer?
- 45. Explain Serialization and Deserialization.
- 46. What are PySpark Serializers?
- 47. What are Salting Techniques?
- 48. Explain mapPartitions in Spark.
- 49. How to track failed jobs in Spark?
- 50. What is Broadcast Join?



- 51. Deployment Mode Cluster Mode and Client Mode
- 52. Spark Submit Command?
- 53. ORC vs. Parquet vs. CSV vs. JSON
- 54. How to deal with bad data?
- 55. Why does Out Of Memory issue occur?
- 56. How to remove duplicate rows?
- 57. How to create SparkContext & SparkSession?
- 58. How to create RDD?
- 59. Create (Read) Spark DataFrame from CSV, TXT, JSON, XML



If Vou find this helpful like and share it with your friends



