Top 50 PySpark Interview Questions (Beginner & Intermediate)

1. What is PySpark? How is it different from Pandas and Spark?
2. Explain the architecture of Apache Spark.
3. What is RDD in PySpark?
4. What is DataFrame in PySpark? How is it different from RDD?
5. What are the main features of PySpark?
6. What is lazy evaluation in Spark?
7. Explain transformations and actions in PySpark.
8. What is a Spark session? How do you create it in PySpark?
9. What is partitioning in PySpark? Why is it important?
10. Explain broadcast variables and accumulators in Spark.
11. How do you read a CSV file into a DataFrame?
12. How do you write a DataFrame to a CSV or Parquet file?
13. How do you select specific columns from a DataFrame?
14. How do you filter rows using filter() or where()?
15. Explain groupBy() and aggregation functions like count(), sum(), etc.
16. How do you handle missing values in PySpark?
17. How do you add or drop columns in a DataFrame?
18. What is the use of withColumn()? Provide a coding example.
19. How do you sort or order a DataFrame?
20. How do you perform joins in PySpark? Explain inner, left, right, and outer joins.
21. What is Spark SQL? How is it integrated with PySpark?
22. How do you register a DataFrame as a temporary table?
23. Write a Spark SQL query to find the maximum value in a column.
24. How do you handle date and timestamp formats in PySpark?
25. Explain the difference between cache() and persist() methods.

26. What are some common optimizations in PySpark?

27. How do you broadcast a small dataset for optimization?

28. Explain partition tuning and shuffling in PySpark.

29. What is the Catalyst optimizer in Spark?

30. How can you monitor the Spark application performance?

31. How do you debug PySpark applications?

32. What is the common memory-related issue in Spark and how to resolve it?

33. How do you deal with skewed data in joins?

34. How do you ensure fault tolerance in Spark applications?

35. What are accumulators and how do you use them for debugging?

36. Write a PySpark program to remove duplicate rows based on a column.

37. How do you calculate the average of a numeric column grouped by another column?

38. Write a PySpark script to replace null values with the mean of the column.

39. How do you perform a self-join in PySpark?

40. Write code to convert a DataFrame column into an array.

41. How do you explode an array column into multiple rows?

42. Write code to pivot data in PySpark.

43. How do you use window functions to calculate running totals?

44. Write a PySpark script to find the top N records per group.

45. How do you apply user-defined functions (UDFs) in PySpark?

46. What is the difference between map(), flatMap(), and reduceByKey() in RDDs?

47. How do you checkpoint RDDs in Spark?

48. How do you handle schema evolution when reading from sources like JSON or Parquet?

49. Explain how PySpark handles serialization.

50. Write a program to read data from multiple files and combine them into a single DataFrame with consistent schema.