47 Databricks interview questions to ask your applicants

Questions

1. Can you explain the main components of the Databricks platform and how they interact?

2. How do you handle data ingestion in Databricks? Can you describe the process?

3. What are some best practices for optimizing Spark jobs in Databricks?

4. How would you manage version control for notebooks in Databricks?

5. Can you discuss how Delta Lake improves data management in Databricks?

6. What is your experience with using Databricks for machine learning workflows?

7. How do you set up and manage clusters in Databricks?

8. Can you explain the difference between Databricks SQL and Databricks notebooks?

9. How do you monitor and troubleshoot performance issues in Databricks?

10. What strategies do you employ for ensuring data governance and security in Databricks?

11. How do you approach troubleshooting issues in a Databricks environment?

12. What steps would you take to ensure data quality in Databricks?

13. Can you describe your experience with integrating Databricks with other data tools?

14. How do you handle permissions and access controls in Databricks?

15. What strategies would you use to optimize data storage in Databricks?

16. How do you ensure your Databricks workflows are scalable?

17. Can you describe a challenging project you worked on in Databricks and how you overcame obstacles?

18. How do you stay updated with the latest features and updates in Databricks?

19. What are the differences between Apache Spark and Databricks, and why would you choose one over the other?

20. Can you explain the role of the Databricks workspace and how it facilitates collaboration among data teams?

21. How do you implement and manage job scheduling in Databricks, specifically for ETL processes?

22. What are the different ways to share notebooks and dashboards in Databricks, and how would you ensure they are up to date?

23. How do you handle library dependencies in your Databricks projects, and what are some common pitfalls?

24. Can you describe how to use Databricks REST APIs for automating tasks?

25. What steps would you take to optimize data pipeline performance in Databricks?

26. How do you manage and monitor resource utilization within your Databricks workspace?

27. Describe your experience with using Databricks Delta for streaming data applications.

28. What is your approach to testing and validating code in Databricks notebooks?

29. Can you explain the concept of lazy evaluation in Apache Spark and why it is beneficial?

30. How do you handle data skewness in a distributed computing environment like Databricks?

31. What are some common strategies for error handling and recovery in Databricks workflows?

32. How do you approach optimizing data read and write operations in Databricks?

33. Can you explain the significance of caching in Databricks and when you would use it?

34. How do you ensure data quality in your Databricks data processing pipelines?

35. What are your strategies for managing and organizing large datasets in Databricks?

36. Can you explain how data partitioning works in Spark and its impact on performance?

37. What are the differences between DataFrames and RDDs in Spark, and when would you use one over the other?

38. How do you optimize joins in Spark, particularly when dealing with large datasets?

39. Can you discuss how Spark SQL can be utilized for querying structured data?

40. What is the role of broadcast variables in Spark, and how do they improve performance?

41. How do you implement error handling in Spark jobs, and what strategies do you use for recovery?

42. Can you describe the difference between transformations and actions in Spark?

43. How would you approach handling late data in a Spark streaming application?

44. What methods can you use to ensure data consistency when using Spark with multiple writers?

45. Can you explain the concept of shuffling in Spark and its implications on performance?