107 Databricks Interview Questions to Hire Top Engineers

Questions

1. Can you explain what Databricks is, like you're explaining it to a friend who's never heard of it?

2. What are the key components of the Databricks platform, and how do they work together?

3. Have you used any cloud platforms before, like AWS, Azure, or GCP? If so, how does Databricks relate to them?

4. What is Apache Spark, and why is it important in the context of Databricks?

5. Can you describe the difference between a DataFrame and a Dataset in Spark?

6. What are some common data formats that Databricks can work with, such as CSV, JSON, and Parquet? How do you load data from them?

7. What are the advantages of using the Parquet format compared to CSV?

8. Have you written any code using PySpark or Scala? Could you walk me through a simple example?

9. What is a Spark transformation, and what is a Spark action? Can you give examples of each?

10. Can you explain what a Spark job, stage, and task are? How are they related?

11. What is the difference between 'lazy evaluation' and 'eager evaluation' in Spark? How does lazy evaluation help with performance?

12. How do you handle missing or null values in a Spark DataFrame?

13. What are user-defined functions (UDFs) in Spark, and when might you use them?

14. How can you optimize the performance of a Spark job, such as by reducing shuffles?

15. What is the purpose of partitioning data in Spark? How can you control the number of partitions?

16. What is the Databricks File System (DBFS), and how is it used for storing data?

17. Have you used Databricks notebooks before? What are some of their advantages?

18. How do you collaborate with others on Databricks projects?

19. What are some ways to monitor the performance of a Spark job in Databricks?

20. How would you approach debugging a Spark job that is running slowly or failing?

21. What is Delta Lake, and what benefits does it offer compared to traditional data lakes?

22. Can you explain ACID properties and how Delta Lake ensures them?

23. What is the difference between a 'managed table' and an 'external table' in Databricks?

24. How would you implement a simple data pipeline using Databricks, from data ingestion to data transformation?

25. Imagine Databricks is a toybox. What are some toys (features) you'd find inside, and what do they do?

26. If you have a big pile of LEGO bricks (data), how would Databricks help you sort and build something cool with them?

27. What's the difference between a small and a big 'cluster' in Databricks, like having a few or many friends helping you build?

28. Can you explain, simply, what a 'notebook' is in Databricks, and why we use it?

29. If you had to teach a computer to add all the numbers from 1 to 100, how would you tell Databricks to do it?

30. What's 'Spark' in Databricks? Think of it like the engine that makes everything run fast. Why is that important?

31. Let's say a file is like a recipe. What would you do in Databricks to follow that recipe and bake a cake (get insights)?

32. Why is it useful to keep your data (like toys) organized in a special 'house' (data lake/warehouse)?

33. If you accidentally broke some LEGOs (corrupted data), what's one way Databricks could help you fix them?

34. What does it mean to 'scale' your Databricks project, like adding more tables to a restaurant?

35. Explain how you would use Databricks to find the tallest building of all the buildings in New York City.

36. How would you use Databricks to count how many times the word 'the' appears in a book?

37. If you are given a large file containing customer names and emails, how can you use Databricks to find all customers from California?

38. Explain how Databricks can help a company predict future sales based on past sales data.

39. What's an API in simple terms and how does it help Databricks connect to other things?

40. Imagine you have two lists of names. How would you find the names that are on both lists using Databricks?

41. How can Databricks help detect fraud by looking at patterns in credit card transactions?

42. What are some of the data visualization tools that you can use with Databricks to display data?

43. How would you use Databricks to analyze website traffic and find out which pages are most popular?

44. How can Databricks help improve the accuracy of machine learning models?

45. Explain how you would set up a simple Databricks job to run automatically every day.

46. If you have a file with missing data, how would you use Databricks to handle it?

47. How do you ensure that your Databricks code is easy to read and understand for others?

48. Let's say you need to process data from multiple sources (like databases and files). How can Databricks help?

49. How would you optimize a Databricks notebook that is running slower than expected?

50. Explain how you would handle skewed data in a Spark DataFrame within Databricks.

51. Describe your experience with using Delta Lake for data warehousing in Databricks. What are the benefits?

52. How do you manage dependencies (e.g., Python libraries) in a Databricks environment to ensure reproducibility?

53. What are some strategies for handling small files in Databricks to improve performance?

54. How would you set up a CI/CD pipeline for Databricks notebooks and jobs?

55. Explain how you would monitor Databricks jobs for performance and errors. What metrics are important?

56. Describe your experience with Databricks SQL Analytics. How does it differ from using Spark SQL?

57. How would you implement row-level security in Databricks to control data access?

58. Explain how you would use Databricks to build a real-time data pipeline.

59. What is the difference between using dbutils.notebook.run and calling a notebook as a job in Databricks?

60. How would you use Databricks to perform machine learning tasks? Explain your preferred workflow.

61. Describe how you've used Databricks to integrate with other data sources (e.g., cloud storage, databases).

62. How do you handle data versioning and reproducibility in Databricks projects?

63. Explain how you would use Databricks to build a data dashboard for stakeholders.

64. What are some best practices for writing efficient Spark SQL queries in Databricks?

65. How would you use Databricks to implement data governance policies?

66. Describe your experience with using Databricks for data exploration and visualization.

67. How would you use Databricks to perform A/B testing?

68. Explain how you would handle personally identifiable information (PII) in Databricks to comply with privacy regulations.

69. What is the role of the Databricks metastore, and how would you manage it?

70. How do you approach debugging complex Spark applications in Databricks?

71. Describe a time when you had to troubleshoot a performance issue in Databricks. What steps did you take?

72. How do you secure your Databricks workspace and prevent unauthorized access?

73. What are the advantages of using Databricks clusters over other cloud-based Spark solutions?

74. How would you use Databricks to build a recommendation engine?

75. Explain the difference between a Databricks job cluster and an interactive cluster. When would you use each?

76. How would you implement data lineage tracking in Databricks?

77. Describe how you would handle data quality issues in a Databricks pipeline.

78. How do you use Databricks Repos for version control and collaboration?

79. How would you optimize a Databricks notebook that's running very slowly?

80. Describe a time you had to debug a complex Spark job in Databricks.

81. What are your experiences with Delta Lake, and what advantages does it offer?

82. Explain how you would set up a CI/CD pipeline for Databricks notebooks.

83. Tell me about a time you used Databricks to solve a real-world business problem.

84. How would you implement row-level security in Databricks?

85. Describe your experience with different Databricks cluster configurations and when you would choose one over another.

86. What strategies do you use for monitoring and alerting on Databricks jobs?

87. How familiar are you with Databricks SQL Analytics, and what are its key features?

88. Explain the process of migrating data from a legacy system to Databricks.

89. How do you handle data versioning and reproducibility in Databricks?

90. Describe your experience with integrating Databricks with other cloud services.

91. What are some best practices for writing efficient Spark code in Databricks?

92. How do you approach troubleshooting performance bottlenecks in Spark SQL queries?

93. Explain your understanding of Databricks Workflows.

94. How would you implement a data quality framework within Databricks?

95. Describe your experience with using Databricks for machine learning tasks.

96. What are some strategies for cost optimization in Databricks?

97. How do you handle personally identifiable information (PII) in Databricks to comply with data privacy regulations?

98. Explain your experience with using structured streaming in Databricks.

99. How would you design a data lake using Databricks and Delta Lake?

100. Describe your experience with using Databricks for ETL (Extract, Transform, Load) processes.

101. What are some challenges you have faced when working with large datasets in Databricks, and how did you overcome them?

102. How do you ensure data consistency and integrity when performing data transformations in Databricks?

103. Explain your experience with using Databricks Repos for version control.