100+

Azure Data Engineer

Interview Questions

You Must Prepare

Spark • Databricks • ADF • Delta Lake • SQL

Based on Real Project Experience

Helpful for 3–7 Years Azure Data Engineers

www.linkedin.com/in/yalla-manideep-reddy

APACHE SPARK / PYSPARK

1. What is the difference between a DataFrame and a Dataset in Spark? Which one do you prefer in real-time projects and why?

2. What is a DAG in Spark?

3. What is a lineage graph and why is it important in Spark?

4. What are transformations in Spark? Explain narrow and wide transformations and how they impact Spark performance.

5. What happens internally once you submit a Spark job?

6. Explain Spark architecture, including Spark job, stage, and task, and explain the role of Driver, Executor, and Cluster Manager.

7. If a Spark job has two wide transformations and one action on 2GB of data, how many jobs, stages, and tasks will be created?

8. What is a partition in Spark and based on what factors are Spark partitions created?

9. What is the data locality principle in Spark and why is it important?

10. What is the difference between client mode and cluster mode? What real-time challenges have you faced using them?

11. What is on-heap and off-heap memory in Spark and how does Spark manage memory internally?

12. What transformations have you commonly used in PySpark?

13. What Spark optimization techniques do you use in real-time projects?

14. What is repartition and coalesce in Spark and when would you use each?

15. What types of joins are available in Spark and which join is most efficient in terms of performance?

16. What is a broadcast join?

17. When both tables are large, which join strategy will you use instead of broadcast join?

18. What is the salting technique and why is it used?

19. What is predicat e pushdown and how does it improve performance?

20. What is the explode function and when do you use it?

21. How would you resolve an out-of-memory issue in Spark step by step?

22. What is bucketing in Spark and when should you use partitioning vs bucketing?

23. Explain data skewness and how you handled it.

24. What is Z-Ordering and OPTIMIZE in Databricks?

25. What are StructType and StructField in Spark?

26. How do you read data from a CSV file in Spark?

27. What is SparkContext and what is its role in Spark?

28. You have millions of records with four states in a column where one state has very less data. How does partitioning work in this case?

29. If you have one Spark action, how many stages and tasks will be created?

DELTA LAKE / STORAGE

30. What is the difference between Delta Lake and a traditional Data Lake?

31. Why do we use Delta format in Databricks?

32. What are ACID transactions in Delta Lake?

33. What is the difference between Parquet and Delta format?

34. What is Delta cache?

35. What is the difference between Delta cache and normal Spark cache?

36. What is the VACUUM command in Delta Lake and when will you use it?

37. What is Delta Lake time travel and how does it work internally?

38. How do you handle historical and incremental data?

39. What is SCD Type 2 and how do you implement it?

40. What is a surrogate key and why is it required?

41. What is the difference between ADLS Gen1 and ADLS Gen2?

42. What is the difference between Blob Storage and ADLS Gen2?

DATABRICKS

43. What is Databricks?
44. Why do we use Databricks when Apache Spark already exists?
45. What are the key differences between Databricks and open-source Spark?
46. What is Unity Catalog in Databricks?
47. What is the purpose of Unity Catalog in real-time projects?
48. What are Delta Live Tables?
49. How do consumers connect to the Gold layer and consume data?
50. What are Databricks Utilities (dbutils) and how do you use them?
51. What are Databricks notebooks?
52. What notebook languages are supported in Databricks and their use cases?
53. What cluster configuration did you use in your Databricks project?
54. Based on what factors do you decide the Databricks cluster configuration?

AZURE DATA FACTORY (ADF)

55. What is Azure Data Factory?
56. What is Integration Runtime in ADF and what are the different types?
57. What is the Copy Data activity in ADF and how does batch count impact performance?
58. What is the ForEach activity in ADF and how does batch count work?
59. What is the Lookup activity in ADF?
60. How do you implement incremental load in ADF?
61. How do you update the watermark column during incremental load?
62. How do you send notifications from ADF pipelines?
63. How do you handle a pipeline failure that occurs in the middle?

64. How would you optimize a 100GB data pipeline in ADF?

65. What are Mapping Data Flow and Wrangling Data Flow and how do they differ?

66. What are Web activity and Webhook activity in ADF?

67. What are triggers in ADF and how many types of triggers are there?

68. How do you design an ADF pipeline when the source file is a ZIP file?

69. How do you connect Azure Data Factory to Databricks?

70. How do you handle errors in ADF and where do you store logs?

71. How do you confirm the file format in an ADF pipeline?

72. You have JSON and CSV files in Blob Storage. How do you move data from Blob to ADLS using ADF?

DATA GOVERNANCE / ETL / MODELING

73. What is data profiling and why is it important?

74. What is data governance and how do you implement it?

75. What is OLTP and OLAP and what are the differences?

76. What is a fact table and what is a dimension table?

77. What is a primary key and what is a foreign key?

78. What is normalization and why is it used?

79. What are facts in ETL?

80. What is the difference between Star schema and Snowflake schema?

81. What are Slowly Changing Dimensions (SCD) and explain all types.

82. Explain SCD Type 2 with a real-time example.

83. How much data are you handling in your current project?

84. What is the most challenging problem you have faced in your project?

85. What is the difference between deep copy and shallow copy?

SQL (PRACTICAL)

86. How do you find the 2nd highest salary in SQL?

87. How do you find the 3rd highest salary using window functions?

88. How do you identify duplicate records and their count?

89. How do you remove duplicate records from a table?

90. How do you find employees who earn more than their manager or joined before their manager?

91. How do you calculate a 7-day rolling average?

92. How do you calculate a running total in SQL?

93. How do you find departments with no employees?

94. How do you find employees who work in more than one department?

95. How do you calculate the average salary of each department?

96. How do you calculate a cumulative monthly average?

97. How do you find customers who have not placed any orders in the last 6 months?

98. How do you find employees with high tenure?

99. How do you rank departments based on average salary?

100. Given a table with duplicate values, how do ROW_NUMBER(), RANK(), and DENSE_RANK() work differently?

SQL THEORY

101. What is indexing in SQL, why is it required, what are the different types of indexes, and how do clustered and non-clustered indexes differ?

102. What is the difference between UNION and UNION ALL and when would you use each?

103. What are views in SQL, what are the different types of views, and why do we use them?

104. What are the different types of joins in SQL and when should each be used?

105. What is a stored procedure, how is it used in real-time projects, and how does it differ from a function?

106. What are window functions in SQL and why are they used?

107. Explain ROW_NUMBER(), RANK(), and DENSE_RANK() with real-time use cases.

108. What are SQL constraints and why are they important?

109. Explain PRIMARY KEY, FOREIGN KEY, UNIQUE, CHECK, and NOT NULL constraints.

110. What is the difference between DELETE, TRUNCATE, and DROP?

111. What is normalization and denormalization in SQL and when would you prefer denormalization?

112. What is the difference between WHERE and HAVING clauses and when should each be used?

ALL THE BEST