By - Pravin Ghadage

300+ Data Engineering Real Time Interview Questions Asked in KPMG, Coforge, EPAM Systems, Deloitte, IQVIA, Synechron, EY, KPI Partners, Tredence, NTT Data, LTI Mindtree, Publicis Sapient, Infovision, Cognizant, Recro, IBM, Accenture, Persistent Systems

1) What are the best practices for managing large datasets in data bricks?

2) What is Surrogate Key? In SCD 2 can we use Surrogate Key?

3) Explain the use of Delta Lake for data versioning?

4) What are the data sources you used in your project?

5) How did you get data from your source? How you implement your notebook along with incoming data?

6) How to transform the data from on prem to cloud?

7) Explain the pipeline you have worked with?

8) Diff between Azure Data Lake Storage (ADLS) and Blob Storage?

9) How did you optimize the storage cost in ADLS?

10) Diff between server less pool and dedicated SQL pool?

11) How did you manage the pipeline from failed activity?

12) Have did you manage the audit logging for data compliance?

13) How to optimize the ADF pipeline?

14) How to implement parallel processing in ADF?

15) What is medallion architecture?

16) Diff between group by key and reduce by key?

17) Diff between partitioning and bucketing?

18) What is repartition and coalesce? What is its use? Diff between repartition and coalesce?

19) What is SCD? Explain SCD1, SCD2, SCD3?

20) How did you implement SCD2 in your project?

21) Explain the Spark Architecture?

22) What are the azure functions you have majorly worked on?

23) What are the azure services you worked on?

24) Explain azure synapse analytics?

25) What is the best way to load data into Azure Synapse?

26) What is the advantage of using Azure Synapse over Apache Spark or Hadoop?

27) What is incremental loading? How to implement it?

28) What is auto loader?

29) What are the encryption techniques?

30) What is Spark? Advantages of using Spark?

31) Diff between streaming and batch processing in Spark?

32) What is Databricks? Advantages of the Databricks?

33) How to optimize data bricks performance?

34) Explain your project in detail with day-to-day activity?

35) How would you optimize the table in Databricks?

36) What are the steps you perform for the pipeline optimization, if it is taking longer time?

37) How to use nested JSON with data bricks?

38) How do you monitor and troubleshoot Spark jobs?

39) Explain the data bricks architecture?

40) Explain Spark optimization techniques?

41) What is DAG in Spark?

42) Diff between DAG and LINEAGE?

43) What are the different modes in spark?

44) What is Partition? How spark partitions the data?

45) What is Spark driver and driver program?

46) What is an executor?

47) What is worker node in spark?

48) Diff between persist and cache in PYSPARK? How it is implemented in PYSPARK?

49) What is broadcast variable and broadcast joins in spark?

50) What is checkpointing in spark?

51) What is Spark Context? Explain the role of Spark Session in PYSPARK? Diff between Spark context and Spark session?

52) What is shuffling, why should we minimize it?

53) How do you implement disaster recovery for ADLS?

54) How to estimate the number of resources required for your spark job?

55) What is Spark SQL?

56) Explain the lazy evaluation in PYSPARK? How does it impact the execution of spark jobs?

57) How do you handle skewed data issues in PYSPARK?

58) What is serialization and deserialization?

59) What is Common Data Model (CDM)?

60) Explain wide and narrow Transformations? Diff between narrow transformation and wide transformation?

61) What is mount point in data bricks?

62) Explain the various transformations in PYSPARK?

63) What is the significance of Catalyst Optimizer in PYSPARK?

64) What is schema on read and schema on write?

65) Explain the types of Integration Runtime (IR)? Explain the use of Integration Runtime in ADF?

66) What are the types of the triggers in Azure? Explain each type in detail? Describe the role of triggers in ADF pipelines?

67) Explain the out of memory issue?

68) Explain the types of the clusters in data bricks?

69) What is the cluster size you have worked on?

70) How much data you deal with on daily basis?

71) What is the most challenging thing you faced in your project?

72) Why MapReduce is not widely used now? Similarities between Spark and MapReduce?

73) Diff between Spark and MapReduce?

74) Diff between map and flat map?

75) Diff between managed tables and external tables?

76) Explain the joins in PYSPARK?

77) What are the deployment modes you used in your project?

78) What are the steps you have taken for data security in your project?

79) What are the parameters and variables?

80) Diff between procedure and functions?

81) Write a query to delete the duplicate records? Explain all other types to delete duplicate records in SQL?

82) What is CTE? What is temporary view?

83) What is normalization and denormalization? What are its uses?

84) Write a query to find the max salary without max Function?

85) Write a query to display nth highest salary employees from emp table?

86) Write a query to display first five highest salary employees from emp table?

87) Write a query to display junior most employee details from emp table?

88) Write a query to display last two rows from emp table?

89) Write a query to display first row and last row from emp table?

90) Write a query to display employee details who are getting max salary in each department from emp table?

91) Write a query to display odd number of records from emp table?

92) Write a query to display 4th highest salary employee from emp table?

93) Write a query to display the employees who are getting more salary than their manager salary from emp table?

94) Write a query to find duplicate records in table?

95) Write a query to find employees with salaries greater than the department average?

96) Diff between rank and dense rank?

97) What are the query optimization techniques in SQL?

98) What are the global temporary tables?

99) Explain candidate key, primary key and super key?

100) Write a query to calculate the top 5 products with highest sales in each region?

101) What are the union, union all, intersect, minus operators?

102) Explain the window functions in SQL?

103) Write a syntax for creating temporary table in SQL?

104) What are the materialised views?

105) Explain the different types of indexes?

106) What is order of execution of a SQL query?

107) Diff between where and having clause?

108) Diff between temporary view and global view?

109) Diff between CTE and subquery? Which is better to use?

110) Diff between CTE and temp view?

111) Diff between count (*) and count(column-name)?

112) Explain the join types in SQL?

113) Diff between DISTINCT and group by clause?

114) What is the process of normalization and why is it required?

115) Diff between DELETE and TRUNCATE?

116) Explain in detail Primary Key, Unique Key, Foreign Key, Candidate Key?

117) How do you monitor and troubleshoot issues in Azure SQL Database?

118) Why do we use CASE statement in SQL?

119) What are the prerequisites to use the UNION operator in SQL?

120) Write a SQL query to convert row-level data to column-level data using pivot?

121) What are subqueries and where can we use them?

122) Write a SQL query to find employees with no manager assigned?

123) Write a query to find the Cumulative Sum in a New Column?

124) Diff between Azure SQL Database and Azure SQL Managed Instance?

125) Find the total sales amount for each month and calculate the month-over-month sales growth?

126) Explain the concept of check pointing in PYSPARK? Why it is important in streaming applications?

127) How to call one notebook in another notebook?

128) How do you manage data lifecycle policies in ADLS?

129) How do you handle exceptions and errors in Python?

130) How will you join two bigger tables?

131) How will you extract only different data from two different tables?

132) What is the diff between SQL and NOSQL?

133) Explain various activities in azure?

134) What are the data flows in azure?

135) What is delta lake? What is the use of delta lake? How does it support ACID transactions? How to create the delta table?

136) How do you handle schema evolution in the delta lake?

137) Diff between data lake and delta lake and data warehouse?

138) Write a SQL query to list all employees who joined in the last 6 months?

139) What is dimensional modelling? What is fact and dimension table?

140) What is time travel? how did you use time travel in your project?

141) How does Z Ordering works? What is the use of Z ordering?

142) How to apply indexing on Databricks table?

143) What is serverless computing?

144) If a job is failed then how did you debug it?

145) What is azure key vault? What are its uses?

146) If two dataflows are running in parallel and we want to run a pipeline after this, how would you do that?

147) How did you get the data from Rest API?

148) What is data skewness in spark?

149) What is fault tolerance? Its use in real time application?

150) Why data bricks is good than dataflow?

151) Merge statement in data bricks?

152) What are the prerequisites before the migration?

153) What is azure DEVOPS?

154) Explain CI/CD process?

155) What is azure logic apps? How did you used azure logic apps in your project?

156) How do you handle your PYSPARK code deployment?

157) What is git? How are you integrating ADF, ADB with Git?

158) Diff between coalesce and is null?

159) Diff between dataflows and pipeline?

160) What is table and view? What will happen if we drop the view?

161) What is Microsoft fabric?

162) How to implement parallel copies in ADF using partitioning?

163) Can we use function within the procedure? can we use procedure within function?

164) Diff between group by and partition by?

165) Out of gold, silver, bronze in which layer you have worked?

166) What is shuffling in transformations?

167) How do you move your code from one environment to another? When you move code, where is it get saved which branch?

168) If in source we have three columns and in destination we want to add one more column then how can you achieve this in ADF?

169) How to use nested JSON with data bricks?

170) How many pipelines you have created?

171) PYSPARK command to read the data from a file into a data frame?

172) How to handle nulls and duplicates in PYSPARK and SQL?

173) How to change the date format for date column?

174) What is explode function in PYSPARK?

175) Write a code to read a parquet file?

176) Code to add a column to parquet file?

177) Different approaches to create the RDD in PYSPARK?

178) What are the different ways to create the data frame?

179) What is RDD, Data Frame, Dataset in spark?

180) What is the diff between RDD and Data frame?

181) How to convert RDD to Data Frame and Data Frame to RDD?

182) How do you replace null values in DF with zero?

183) How you can query the data frame in spark SQL?

184) What is accumulator in spark?

185) What is YARN?

186) Explain the meta store concept?

187) Diff between delta vs parquet files?

188) How did you create the Temp View?

189) What is spark streaming?

190) How to remove duplicate records in data frame?

191) How to add column in data frame?

192) Diff between list and tuple?

193) Diff between def and lambda?

194) Write the Python code to count the frequency of words in a string?

195) Write the Python code for reverse the string?

196) Windows function (ranking functions) in SQL and PYSPARK?

197) Explain the different file formats that you are using in your project (CSV, JSON, PARQUET, Delta Table)?

198) Major issue faced in spark development?

199) What is type safety and how we can achieve it?

200) Why you save your data in delta form in your notebook?

201) What is the unity catalog? How it is different from hive meta store?

202) You have CSV file in ADLS and you get again same CSV file next day, how would you load that data without duplication?

203) Explain Agile methodology?

204) What are the benefits of using serverless compute options in Azure Synapse?

205) Diff between Azure Blob storage and Azure Synapse Analytics?

206) Are there any restrictions on what type of data can be loaded into Azure Synapse?

207) What is MPP? How can it benefit enterprises?

208) Diff between Azure Stream Analytics and Azure Databricks?

209) What types of files are supported by Azure Synapse?

210) Explain compression concept in ADF?

211) Explain different file formats?

212) What is User Defined Function (UDF)? how to register UDF in PYSPARK?

213) What are delta logs? How to track versioning in delta tables?

214) How do you monitor and troubleshoot ADF pipeline failures?

215) Explain the concept of managed identity in Azure and its use in data engineering?

216) How to perform aggregations in PYSPARK?

217) Write a PYSPARK code for filter the records based on a condition?

218) What is broadcast joins and broadcast variables? How they optimize the join operation? Which is more useful for optimization?

219) You have the key vaults and you need to pass the value in the notebook, how you do it?

220) Join two data frames based on composite key and filter rows where sales amount exceeds a threshold?

221) Calculate the running total of sales for each customer, partition by customer ID and order by date?

222) Handle the missing values in a Data frame and replace them with the mean or median of the respective column?

223) Find the top 10 most occurring words in a text dataset?

224) Extract and transform the nested JSON data into a structured data frame?

225) Write a PYSPARK script to detect the duplicate rows based on multiple columns?

226) Use PYSPARK to merge multiple CSV files into a single data frame and remove duplicates?

227) Diff between full load and incremental load in ADF?

228) Explain the notebook optimization strategies in data bricks?

229) Diff between star schema and snowflake schema?

230) Explain how to load the data in synapse tables?

231) Explain where and how to store the secret keys?

232) Describe the process of integrating ADF with Azure Synapse Analytics?

233) Write PYSPARK code to calculate the total sales for each product category?

234) Explain how broadcast joins improve performance in PYSPARK?

235) Diff between cluster mode and client mode?

236) How do you implement schema drift handling in ADF?

237) Write Python code to check if a number is a palindrome?

238) How do you handle incremental data load in Databricks?

239) Explain Adaptive Query Execution (AQE) in Spark?

240) How do you optimize data partitioning in ADLS?

241) Write PYSPARK code to perform a left join between two Data Frames?

242) How do you design a fault-tolerant architecture for big data processing?

243) How do you handle schema evolution in Delta Lake?

244) How do you manage metadata use in data bricks?

245) Explain the difference between Spark SQL and PYSPARK Data Frame APIs?

246) How do you handle large-scale data ingestion into ADLS?

247) Write Python code to split a name column into first name and last name?

248) What are fact and dimension tables in data modelling?

249) Explain the concept of Poly-Base in Azure SQL Data Warehouse?

250) How do you manage partitioning in PYSPARK?

251) How do you read data from URL in data bricks?

252) How to connect data bricks notebook from ADF pipeline?

253) How do you optimize joins in PYSPARK for large datasets?

254) Describe the process of setting up CI/CD for Azure Data Factory?

255) Write Python code to reverse a string?

256) What are the key features of Databricks notebooks?

257) How do you handle late-arriving data in ADF?

258) Explain the concept of Data Lakehouse?

259) How will you apply indexing on a table in data bricks?

260) How do you manage metadata use in data bricks?

261) How to connect data bricks notebook from ADF pipeline?

262) Describe the process of integrating ADF with Databricks for ETL workflows?

263) Explain the concept of mounting ADLS and secret scope in Azure Databricks?

264) Can you pass the data from data bricks to ADF?

265) How do you implement window function in PYSPARK to calculate a moving average over specific time window?

266) What is the significance of partitioning in PYSPARK? How does it affect the performance of the data processing?

267) How to handle real-time streaming data using PYSPARK structured streaming? Explain the key components involved?

268) How do autoscaling clusters work in Databricks?

269) Explain the use of hierarchical namespaces in ADLS?

270) Describe the process of setting up and managing an Azure Synapse Analytics workspace?

271) Write PYSPARK code to calculate the average salary by department?

272) How do you implement streaming pipelines in Databricks?

273) Explain the purpose of Delta Lake checkpoints?

274) Write Python code to find the largest number in a list?

275) What are the challenges in integrating on-premises data with Azure services?

276) Diff between a job cluster and an interactive cluster in Databricks?

277) How to copy all tables from one source to the target using metadata-driven pipelines in ADF?

278) Write Python code to generate Fibonacci numbers?

279) What are the best practices for managing and optimizing storage costs in ADLS?

280) How do you implement security measures for data in transit and at rest in Azure?

281) How do you implement data deduplication in PYSPARK?

282) How do you monitor ADF pipeline performance?

283) Describe the role of Azure Key Vault in securing sensitive data?

284) How to track file names in the output table while performing copy operations in ADF?

285) Describe the process of setting up disaster recovery for ADLS?

286) Write Python code to calculate the factorial of a number?

287) What are the security features in Azure Synapse Analytics?

288) How do you implement CI/CD for Azure Synapse Analytics?

289) How do you monitor and optimize performance in Azure Synapse?

290) Write Python code to identify duplicates in a list and count their occurrences?

By - Pravin Ghadage

291) What are the key features of Azure DevOps?

292) How do you use Azure Logic Apps to automate data workflows in SQL databases?

293) Write a SQL query to find gaps in a sequence of numbers?

294) How do you ensure high availability and disaster recovery for Azure SQL databases?

295) Explain the role of pipelines in Azure DevOps?

296) How do you implement data masking in ADF for sensitive data?

297) How many jobs, stages, and tasks are created during a Spark job execution?

298) How do you integrate ADLS with Azure Databricks for data processing?

300) How do you implement data governance in a data lake environment?

301) How to handle incremental load in PYSPARK when the table lacks a last modified column?

302) How do you use Azure Stream Analytics for real-time data processing?

303) What are the key considerations for designing a scalable data architecture in Azure?

304) How do you integrate Azure Key Vault with other Azure services?

305) Write Python code to replace vowels in a string with spaces?

306) How do you implement data encryption at rest and in transit in ADLS?

307) Describe the use of Azure Synapse Analytics and how it integrates with other Azure services?

308) Explain the role of metadata in data modelling and data architecture?

309) How to create and deploy notebooks in Databricks?

310) What are the best practices for data archiving and retention in Azure?

311) How do you connect ADLS (Azure Data Lake Storage) to Databricks?