Top 200 PySpark Interview Questions for 2 Years Experience

Basic PySpark Concepts (1-30)

1. What is PySpark and how does it differ from Apache Spark?
2. Explain the architecture of Apache Spark
3. What are RDDs in PySpark?
4. What is the difference between RDD, DataFrame, and Dataset?
5. What are transformations and actions in PySpark?
6. Explain lazy evaluation in Spark
7. What is a DAG (Directed Acyclic Graph) in Spark?
8. What are the advantages of using DataFrames over RDDs?
9. How do you create a SparkSession?
10. What is SparkContext and how is it different from SparkSession?
11. Explain the difference between map() and flatMap()
12. What is the difference between reduce() and reduceByKey()?
13. What are broadcast variables in PySpark?
14. What are accumulators in PySpark?
15. Explain partitioning in PySpark
16. What is the difference between coalesce() and repartition()?
17. What are narrow and wide transformations?
18. Explain shuffling in Spark
19. What is the difference between cache() and persist()?
20. What are the different storage levels in PySpark?
21. How do you read a CSV file in PySpark?
22. How do you read a JSON file in PySpark?
23. What is schema inference?
24. How do you define a custom schema in PySpark?
25. What is the difference between collect() and take()?
26. What is the use of select() in DataFrames?
27. How do you filter data in PySpark DataFrame?
28. What is withColumn() used for?

29. Explain the difference between withColumn() and withColumnRenamed()
30. What is the use of drop() function?

Intermediate PySpark Operations (31-80)

31. How do you handle null values in PySpark?
32. What is fillna() and dropna()?
33. How do you perform joins in PySpark?
34. What are the different types of joins in PySpark?
35. What is the difference between join() and union()?
36. Explain broadcast join in PySpark
37. What is a shuffle join?
38. How do you perform aggregations in PySpark?
39. What is groupBy() in PySpark?
40. Explain the difference between groupBy() and agg()
41. What are window functions in PySpark?
42. How do you use rank() and dense_rank()?
43. What is the difference between rank(), dense_rank(), and row_number()?
44. How do you use lead() and lag() functions?
45. What is partitionBy() in window functions?
46. How do you perform sorting in PySpark?
47. What is the difference between sort() and orderBy()?
48. How do you remove duplicates in PySpark?
49. What is distinct() vs dropDuplicates()?
50. How do you perform string operations in PySpark?
51. What are UDFs (User Defined Functions)?
52. How do you create and register a UDF?
53. What is the difference between UDF and Pandas UDF?
54. What are the performance implications of UDFs?
55. How do you optimize UDF performance?
56. What is explode() function in PySpark?
57. How do you work with arrays in PySpark?
58. How do you work with maps/dictionaries in PySpark?
59. What is struct in PySpark?
60. How do you flatten nested JSON in PySpark?

61. What is pivot() and unpivot() in PySpark?
62. How do you perform cross joins?
63. What is the use of when() and otherwise()?
64. How do you use case statements in PySpark?
65. What is lit() function?
66. How do you concatenate columns in PySpark?
67. What is concat() vs concat_ws()?
68. How do you extract date parts from timestamp?
69. What are the common date functions in PySpark?
70. How do you convert data types in PySpark?
71. What is cast() function?
72. How do you handle complex data types?
73. What is createOrReplaceTempView()?
74. How do you run SQL queries in PySpark?
75. What is the difference between temp view and global temp view?
76. How do you write data to Parquet format?
77. What are the advantages of Parquet over CSV?
78. How do you write data with partitioning?
79. What is bucketing in Spark?
80. How do you read from multiple files?

Performance Optimization (81–120)

81. What are the common performance optimization techniques in PySpark?
82. How do you avoid shuffling in Spark?
83. What is data skewness and how do you handle it?
84. How do you optimize joins in PySpark?
85. What is predicate pushdown?
86. What is column pruning?
87. How does caching improve performance?
88. When should you use cache() vs persist()?
89. What is the difference between MEMORY_ONLY and MEMORY_AND_DISK?
90. How do you monitor Spark applications?
91. What is the Spark UI and what information does it provide?

92. How do you identify shuffle operations in Spark UI?
93. What are stages and tasks in Spark?
94. How do you tune the number of partitions?
95. What is the ideal partition size?
96. How does spark.sql.shuffle.partitions affect performance?
97. What is adaptive query execution (AQE)?
98. How do you enable AQE in Spark?
99. What are the benefits of AQE?
100. How do you handle small files problem?
101. What is salting technique for skewed data?
102. How do you use broadcast variables for optimization?
103. What is the benefit of using DataFrame API over RDD?
104. How does Catalyst optimizer work?
105. What is Tungsten execution engine?
106. How do you optimize memory usage in Spark?
107. What is garbage collection tuning in Spark?
108. How do you configure executor memory and cores?
109. What is the difference between executor memory and driver memory?
110. How do you calculate the right number of executors?
111. What is dynamic allocation in Spark?
112. How do you enable dynamic resource allocation?
113. What are the trade-offs of using persist()?
114. How do you unpersist cached data?
115. What is speculative execution in Spark?
116. How do you handle OOM (Out of Memory) errors?
117. What causes data spill to disk?
118. How do you optimize SQL queries in Spark?
119. What is cost-based optimization (CBO)?
120. How do you collect statistics for optimization?

Advanced Topics (121–170)

121. What is Structured Streaming in PySpark?
122. How is Structured Streaming different from DStreams?
123. What are streaming sources in PySpark?

124. What are streaming sinks?
125. How do you handle late data in streaming?
126. What is watermarking in Structured Streaming?
127. How do you perform windowed aggregations in streaming?
128. What are trigger types in Structured Streaming?
129. What is checkpointing in streaming?
130. How do you handle stateful operations in streaming?
131. What is mapGroupsWithState()?
132. What is flatMapGroupsWithState()?
133. How do you read from Kafka in PySpark?
134. How do you write to Kafka in PySpark?
135. What is Delta Lake?
136. What are the advantages of Delta Lake over Parquet?
137. How do you perform ACID transactions with Delta Lake?
138. What is time travel in Delta Lake?
139. How do you optimize Delta tables?
140. What is Z-ordering in Delta Lake?
141. How do you handle schema evolution in Delta Lake?
142. What is the difference between merge and upsert?
143. How do you implement SCD Type 2 in PySpark?
144. What are the different file formats supported by Spark?
145. How do you work with Avro files?
146. How do you work with ORC files?
147. What is the difference between Parquet and ORC?
148. How do you connect to databases using JDBC?
149. How do you optimize JDBC reads?
150. What is partitionColumn in JDBC reads?
151. How do you handle incremental loads?
152. What is change data capture (CDC)?
153. How do you implement CDC in PySpark?
154. What are the best practices for partition keys?
155. How do you handle data quality checks?
156. What is Great Expectations with PySpark?
157. How do you implement data validation?
158. What are MLlib capabilities in PySpark?

159. How do you prepare data for machine learning?
160. What is feature engineering in PySpark?
161. How do you handle categorical variables?
162. What is VectorAssembler?
163. How do you split data into train and test?
164. What are transformers and estimators in MLlib?
165. How do you build a pipeline in MLlib?
166. What is cross-validation in PySpark?
167. How do you save and load models?
168. What are the common issues with PySpark serialization?
169. How do you handle lambda serialization issues?
170. What is Py4J and how does it work?

Databricks Specific (171–200)

171. What is Databricks and how is it different from Apache Spark?
172. What are Databricks notebooks?
173. How do you create a cluster in Databricks?
174. What are the different cluster modes?
175. What is Unity Catalog?
176. How do you manage permissions in Databricks?
177. What are Databricks workflows?
178. How do you schedule jobs in Databricks?
179. What is Delta Live Tables (DLT)?
180. What are medallion architecture layers?
181. What is the bronze, silver, gold pattern?
182. How do you use dbutils in Databricks?
183. What are widgets in Databricks notebooks?
184. How do you pass parameters between notebooks?
185. What is %run command in Databricks?
186. How do you mount storage in Databricks?
187. What are secrets in Databricks?
188. How do you use secret scopes?
189. What is Databricks Connect?
190. How do you optimize costs in Databricks?

191. What are spot instances in Databricks?
192. What is photon engine?
193. How do you enable photon?
194. What is Auto Loader in Databricks?
195. How is Auto Loader different from traditional streaming?
196. What are the advantages of Unity Catalog?
197. How do you implement data governance?
198. What is Databricks SQL?
199. How do you create dashboards in Databricks?
200. What are the best practices for production deployments in Databricks?

Most Frequently Asked Questions (Top 20)

1. **What is the difference between RDD, DataFrame, and Dataset?**
2. **Explain transformations vs actions**
3. **What is lazy evaluation?**
4. **How do you optimize joins?**
5. **What are broadcast variables?**
6. **How do you handle null values?**
7. **What are UDFs and their performance implications?**
8. **Explain partitioning and bucketing**
9. **What is data skewness and how to handle it?**
10. **Difference between cache() and persist()**
11. **What are window functions?**
12. **How do you handle small files problem?**
13. **Explain Structured Streaming**
14. **What is Delta Lake?**
15. **How do you read/write from different file formats?**
16. **What is shuffle and how to minimize it?**
17. **Explain narrow vs wide transformations**
18. **How do you optimize Spark configuration?**
19. **What is AQE (Adaptive Query Execution)?**
20. **How do you debug performance issues in Spark?**

These questions cover the essential topics for a 2-year experienced PySpark

developer role in Databricks!