PySpark Most Asked Interview Questions 2️⃣ 0️⃣ 2️⃣ 5️⃣   
  
  
1. Basic & Conceptual Questions  
What is PySpark? How is it different from Apache Spark?  
What are the main components of PySpark architecture?  
What is RDD, DataFrame, and Dataset? Which one is better?  
Explain lazy evaluation in PySpark.  
What is the difference between narrow and wide transformations?  
  
2. DataFrame Operations (Very Common)  
How do you read and write files in PySpark?  
CSV, Parquet, JSON — with schema and without schema.  
Explain how to handle missing or null values in PySpark.  
How do you filter rows and select specific columns in a DataFrame?  
How to add or rename a column in PySpark?  
What is the difference between select, selectExpr, and withColumn?  
  
3. Aggregation & Grouping (High Frequency)  
Explain groupBy() vs groupByKey() in PySpark.  
How do you calculate aggregates like sum, avg, count using PySpark DataFrame API?  
How to apply multiple aggregations on different columns?  
  
4. Window Functions (Top Trend in 2025)  
What is a Window function in PySpark?  
Difference between rank(), dense\_rank(), and row\_number()?  
Explain use of lead() and lag() functions with example.  
What does rowsBetween and rangeBetween mean in window spec?  
  
5. Joins & Performance Optimization  
Explain different types of joins in PySpark.  
How to optimize join operations in PySpark (e.g., broadcast join)?  
What is broadcast join? When should we use it?  
How to identify and resolve skewed joins?  
  
6. PySpark SQL  
How do you register a DataFrame as a temporary table?  
Write a PySpark SQL query to get top N records by group.  
What is the use of spark.sql() vs DataFrame API?  
  
7. Real-time Scenarios & Coding  
Remove duplicates from a DataFrame based on one or more columns.  
Get the latest record per group (e.g., latest transaction per customer).  
Explode nested arrays or structs.  
Flatten a deeply nested JSON structure.  
Read data from S3, transform, and write to Parquet with partitioning.  
  
8. Performance Tuning & Optimization (Highly Asked)  
What is the difference between cache() and persist()?  
When to use repartition() vs coalesce()?  
How do you tune Spark configurations for better performance?  
What is Tungsten and Catalyst optimizer?  
  
9. UDFs and Error Handling  
How do you write and register a UDF in PySpark?  
When should you avoid UDFs?  
How to handle errors and exceptions in a PySpark pipeline?  
  
10. Integration & Advanced  
How do you integrate PySpark with Hive or AWS S3?  
What are broadcast variables and accumulators?  
How do you work with time-series data in PySpark?  
How do you write unit tests for PySpark code?