Spark Interview Questions

1) What is Resilient Distributed Dataset (RDD) in Apache Spark? How it provides abstraction in Spark and make spark operator rich?
2) What is RDD lineage graph or linage operation in Apache Spark? Explain lineage graph operator in Apache Spark, how it enables fault-tolerance in Spark?
3) What is the difference between rdd and dataframes ?
4) Transformation vs Action, Examples for transformations and actions
5) What is the use of Spark driver, where it gets executed on the cluster ?
6) What is a Broadcast Variables?
7) What is difference between Caching and Persistence?
8) What is a Dstream?
9) How can you minimize data transfers when working with Spark?
10) What is the significance of Sliding Window operation?
11) What do you understand by Executor Memory in a Spark application?
12) Difference between groupByKey() and reduceByKey() ?
13) Does Apache Spark provide check pointing? Types of checkpointing ?
14) Hadoop uses replication to achieve fault tolerance. How is this achieved in Apache Spark?
15) What do you understand by Lazy Evaluation and how is it useful ?
16) What are the different ways of creating an RDD ?
17) What happens to RDD when one of the nodes on which it is distributed goes down?
18) What is SparkContext ?
19) While processing data from HDFS does it execute code near data?
20) How SparkSQL is different from HQL and SQL?
21) What is Catalyst framework ?
22) How does Spark code gets executed over the cluster ?
23) What is RDD partitions ? How Spark partition the data ?
24) How fault tolerance is achieved in Spark Streaming applications ?