Spark Interview Questions

Why Spark?
Difference between MapReduce and Spark
MapReduce - will try to read and write operations to hard disk linearly, network IO and serialization ops

What is Spark ?
Parallel data processing framework
Cluster manager
Each worker node has an executor and a cache?

What is an RDD?
Primary core abstraction
Resilient Distributed Dataset
RDD Properties

- Immutable
- Distributed
- Lazily evaluated
- Catchable

What is SparkContext?
Every time RDD is created, new SparkContext connects to the Cluster
Tells Spark how to access cluster
Key factor in application
All configuration details

What is a Partition?

- Logical division of data
- Specifically derived to process the data
- Small chunks support scalability and speed up process
- Everything is partitioned RDD

How does Spark partition Data?

- Uses MapReduce API
- By default HDFS block size is partition size, possible to change partition size w/Split