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 operations

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
- Can create any number of partitions
- By default HDFS block size is partition size, possible to change partition size w/Split

How does Spark store data?

- Spark is a processing engine
- Does not store data
- Retrieve from many sources

Is it mandatory to start Hadoop before Spark?
No - can load data from local system if needed ?

What are the Components of the Spark Ecosystem?

- Spark SQL (Shark)
- Spark Streaming
- MLib
- GraphX

What is SparkCore?

- Base engine of framework
- Memory management
- Fault tolerance
- Scheduling and monitoring
- Interacting with Store systems

How is SparkSQL different from HQL and SQL ?

- Supports both SQL and HQL
- Can join tables from both SQL and HQL

When do we use Spark Streaming?

- Real time processing API
- Different resources can be used

How does Spark Streaming work?

- Spark sets a duration per batch
- Dstream
- Feeds batches to engine

- Streaming batches are generated as output

What is Spark MLib

- Machine learning library

GraphX?

- Manipulates graphs and collections
- Unifies ETL, iterative graph computation
- Fastest fault tolerant graph system

FS API?

- Reads data from a variety of file systems

Why are partitions immutable?

- Every transformation generates new partition
- Uses HDFS API so that partition is immutable, distributed, and fault tolerant
- Partition is aware of data locality

What is a transformation?

- Follows lazy operation and temporarily holds data unless called by action
- Each transformation creates new RDD
- Map, flatMap, groupbykey, reducebykey, filter, cogroup, join, sortbykey, union, distinct, sample

What is an action?

- RDD's operation - > return to Spark drivers -> execute job on cluster
- A transformations output is the *input* of an Action
- Reduce, collect, takeSample, take, first, saveAsTextFile, saveAsSequenceFile, countByKey, forEach

What is RDD lineage?

- **RDD Lineage** is a process to reconstruct lost partitions

- RDD uses lineage to rebuild lost data
- Each RDD remembers how the RDD was built from other datasets