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
- Lazily evaluated
- Catchable