Spark Lesson 1

1.

Apache Spark was developed in order to provide solutions to shortcomings of another project, and eventually replace it. What is the name of this project?

- MapReduce
- Hadoop
- HDFS
- Pig

2.

Why is Hadoop MapReduce slow for iterative algorithms?

The Java Virtual Machine uses too much memory

Communication is a bottleneck

It needs to read off disk for every iteration

Iterative algorithms do not scale well

3.

What is the most important feature of Apache Spark to speedup iterative algorithms?

Caching datasets in memory

Caching datasets on disk

Python interface

Resiliency to data loss

4.

Which other Hadoop project can Spark rely to provision and manage the cluster of nodes?

MapReduce

YARN

HDFS

Pig

5.

When Spark reads data out of HDFS, what is the process that interfaces directly with HDFS?

Cluster Manager

YARN

Executor

Driver

6.

Under which circumstances is preferable to run Spark in Standalone mode instead of relying on YARN?

When you only plan on running Spark jobs

Never

For iterative algorithms

When we want to mix MapReduce and Spark jobs.