What’s the major change in Hadoop 2 vs Hadoop 1?

* Hadoop 2 can map tens of thousands of nodes whereas Hadoop 1 can only do 4000 at its maximum that the engine can support (due to job tracker and Namenode).
* Hadoop 2 accept master node failure because of the switchable secondary master node if this secondary master node not fail simultaneously. Hadoop 1 has only one master node thus all calculation will stop if master node shuts down.
* Hadoop 2 has Yarn (resource manager) which was originally split from MapReduce. Hadoop 1 has HDFS and MapReduce.

What is secondary namenode?

Firstly, NameNode stores meta data

* While secondary name node is standby, it keeps merging edits log to fsimage. If active namenode was down, secondary namenode immediately be switched to take the responsibility of an active node.
* Secondary share same storage with active namenode such that secondary name node solves the problem of single point of failure.

Where are Hadoop config files? What’s the file names?

* Hadoop config files are under “conf” directory of Hadoop.
* File names:

1. Hadoop-env.sh
2. Core-site.xml
3. Hdfs-site.xml
4. Mapred-site.xml

Which Hadoop component own scheduler?

* Resource manager

How to tell MapReduce not to use a combiner?

* Can be set through “*min.num.spills.for.combine*” (The number of spills for which a combiner need to run).