Now let us look at how to configure Hadoop.

Hadoop is configured using a number of XML files. And each file controls a number of parameters. There are three main configuration files with which you will work. core-site.xml is used to configure the parameters that are common to both HDFS and MapReduce. hdfs-site.xml contains parameters that are for the HDFS daemons, like the NameNode and DataNodes. mapred-site.xml controls the settings for MapReduce daemons, JobTracker and TaskTrackers. We are not going to spend the time covering all of the configuration files. That would just take too much time. However, you do have the option of pausing this video if you would like to read the descriptions of the other configuration files.

The hadoop-env.sh is a script that sets a number of environment variables. Normally, with Hadoop, these variables are not set but with BigInsights, they are. There is one that must always be set and that is the JAVA_HOME environment variable.

Here are some of the settings found in core-site.xml. We are not going to spend time on these nor those on this page as well. If you want to pause the video to read their description, feel free to do so.

Next we have some setting in hdfs-site.xml. If you want to set a different value for the default block size, then you would modify dfs.block.size. Likcwise, if you want to change the default replication factor, then you would modify dfs.replication. Once again, I am not going to cover all the parameters.

To change MapReduce settings, you modify mapred-site.xml. You can control which nodes can connect to the JobTracker. mapred.reduce.tasks lets you set the number of reduce tasks per job. mapred.map.tasks.speculative. execution allows the JobTracker, when having determined that there might be a problem with one map task, to start another map task running in parallel. Both map tasks process the same data and, upon successful completion of one of the tasks, the other is terminated. mapred.tasktracker.map.tasks.maximum and mapred.tasktracker.reduce.tasks.maximum lets you define the number of slots on a TaskTracker that can run map and reduce task. mapred.jobtracker.taskScheduler points to the scheduler that is to be used for MapReduce jobs.

So how do you set these parameters? First of all, you must stop the appropriate service or services before making the change. You are making changes to value element for the appropriate property element. The configuration files are in the hadoop-conf directory. The changes must be made to the configuration files on all nodes in the cluster.

Let me spend a few minutes and focus on BigInsights. With BigInsights the hadoop-conf directory is under $BIGINSIGHTS_HOME. But, and this is very important, you do not make changes to the configuration files in that directory. BigInsights has a staging directory which is $BIGINSIGHTS_HOME/hdm/hadoop-conf-staging that has copies of the configuration files.

You make changes to the files in this staging directory and then execute a script that distributes the changes to all nodes in the cluster.

Finally, let us talk about setting up rack topology.

To make Hadoop aware of the cluster's topology, you code a script that receives as arguments, one or more ip addresses of nodes in the cluster. The script returns on stdout, a list of rack names, one for each input value. Then you update core-site.xml and modify the topology.script.file.name property to point to your script. The good news is that there are examples available for you to review.

This ends this unit. Thank you for attending.

And here are some trademarks that may have been referenced in this presentation.