Transcript name: HDFS command line interface

English
Welcome to the unit of Hadoop Fundamentals on the HDFS command line interface. In this presentation, I will cover the general usage of the HDFS command line interface and commands specific to HDFS. Other commands should be familiar to anyone with UNIX experience and will not be covered. I will follow the presentation of most commands with an example run on an IBM BigInsights installation of the Hadoop technology stack. The HDFS can be manipulated through a Java API or through a command line interface. All commands for manipulating HDFS through Hadoop's command line interface begin with "hadoop", a space, and "fs". This is the file system shell. This is followed by the command name as an argument to "hadoop fs". These commands start with a dash. For example, the "ls" command for listing a directory is a common UNIX command and is preceded with a dash. As on UNIX systems, ls can take a path as an argument. In this example, the path is the current directory, represented by a single dot. Let's begin looking at these HDFS commands by starting up Hadoop. We'll run the start.sh script to bring up each of the Hadoop nodes: first the NameNode, the Secondary NameNode, a DataNode, the JobTracker, and a TaskTracker. Now we'll run the -ls command to give us the current directory. As we saw for the "ls" command, the file system shell commands can take paths as arguments. These paths can be expressed in the form of uniform resource identifiers or URIs. The URI format consists of a scheme, an authority, and path. There are multiple schemes supported. The local file system has a scheme of "file". HDFS has a scheme called "hdfs". For example, let us say you wish to copy a file called "myfile.txt" from your local filesystem to an HDFS file system on the localhost. You can do this by issuing the command shown. The copyFromLocal command takes a URI for the source and a URI for the destination. The scheme and the authority do not always need to be specified. Instead you may rely on their default values. These defaults can be overridden by specifying them in a file named core-site.xml in the conf directory of your Hadoop installation. Now let's examine the copyFromLocal command. We will copy a file named myfile.txt to the HDFS. Now we can just examine the HDFS with the -ls command and see that our file has been copied. And there it is. HDFS supports many POSIX-like commands. HDFS is not a fully POSIX compliant file system, but it supports many of the commands. The HDFS commands are mostly easily-recognized UNIX commands like cat and chmod. There are also a few commands that are specific to HDFS such as copyFromLocal. We'll examine a few

of these.

copyFromLocal and put are two HDFS-specific commands that do the same thing - copy files from the local filesystem to a location on another filesystem.

Their opposite is the copyToLocal command which can also be referred to as get. This command copies files out of the filesystem you specify and into the local filesystem.

getmerge is an enhanced form of get that can merge the files from multiple locations into a single local file.

Now let's try out the getmerge command. First we will copy the myfile.txt into a second copy on the HDFS called myfile2.txt. Then we will use the getmerge command to combine the two and write them as one file in the local filesystem called myfiles.txt. Now if we cat out the value, we see the text twice.

setrep lets you override the default level of replication to a level you specify. You can do this for one file or, with the -R option, to an entire tree. This command returns immediately after requesting the new replication level. If you want the command to block until the job is done, pass the -w option.

This concludes the presentation. Thank you for watching.