|  |  |
| --- | --- |
| Lab 3 – Hadoop | Name: SHUYUE JIAID: 56846018 |

1. Read and setup your Hadoop machine environment according to the lab 3 setup guide in CANVAS.
2. Login into your machine with Hadoop and open up a terminal (e.g. ctrl+alt+t)
3. Issue the shell command “jps”. What is it? You could search for its meaning on the web.

~ » jps

2265 Jps

607

Answer:

jps is the Java Virtual Machine Process Status Tool. It is the “host identifier of the host for which the process report should be generated. The hostid may include optional components that indicate the **communications protocol**, **port number**, and other **implementation specific data**.” [1]

Reference:

[1] https://docs.oracle.com/javase/7/docs/technotes/tools/share/jps.html

1. Check if you have the necessary environment for running Hadoop.
2. Fill in the following table by navigating the related information on the web.

|  |  |
| --- | --- |
| **Shell Command** | **Meaning** |
| >start-dfs.sh | Start the process of HDFS (for Storage), i.e., start NameNode, DataNode, and secondaryNameNode processes. |
| >start-yarn.sh | Start the process of YARN (for resources scheduling). |
| >mr-jobhistory-daemon.sh start historyserver | Start the process to monitor and check previous (history) work or processes. |

1. You may use the above commands for helping you setup the Hadoop environment.
2. In the terminal, change your directory to “Exercises/Ex1”.

Answer:

cd Exercises/Ex1

1. There is a java file “WordCount.java” for counting words using Hadoop.
2. Open and read the java file “WordCount.java”.

Answer:

nano WordCount.java

1. There are lots of Hadoop package imports. Fill in the following table:

|  |  |
| --- | --- |
| **Java Import Statement** | **Meaning** |
| import org.apache.hadoop.conf.Configuration; | Import the configurations of system parameters. |
| import org.apache.hadoop.fs.Path; | Import the file or directory names in a FileSystem. |
| import org.apache.hadoop.io.IntWritable;  import org.apache.hadoop.io.Text; | Import a Writable Comparable for ints. And import the class to store text using standard UTF-8 encoding |
| import org.apache.hadoop.mapreduce.Job;  import org.apache.hadoop.mapreduce.Mapper;  import org.apache.hadoop.mapreduce.Reducer; | Import MapReduce Job, Mapper, and Reducer. |

1. Explain the “map” function in less than 20 words:

Answer:

Map is to map the key-value to an intermediate key-value set.

1. Explain the “reduce” function in less than 20 words:

Answer:

Reduce is to reduce the values set of a key into a smaller values set.

1. In the same directory, there is a makefile which can help you run the java file.
2. Open and read the makefile for understanding its logical flow.
3. What is the use of the command “hadoop com.sun.tools.javac.Main $<” ?

Answer:

The command is to use Java main class to execute the WordCount.java file.

1. What is the use of the command “jar cf $@ $(basename $<)\*.class” ?

Answer:

This command is to pack the class file, i.e., the compiled file, into jar file.

1. What is the use of the command “hadoop jar $< $(basename $<) $(INPUTPATH) $(WCOUT)” ?

Answer:

This command is to execute the program through the jar file.

1. Now you can type “make WordCount” in the terminal to run the word counting program on Hadoop.
2. After you have typed “make WordCount”, you can browse the output by issuing the command:

hdfs dfs -cat wc\_out/\*

1. What can you see? Please summarize it in less than 10 words.

Answer:

The result of the word counting.

1. How would you modify this program so that it counts the word-length frequencies instead?

Answer:

The mapper can be changed to use word length as the key.

word.set(itr.nextToken().length() + “”);

context.write(word, one)

1. (Optional) If you are interested, you could also modify the original WorldCount program to compute n-gram frequencies where the n value is specified as a command line argument.
2. This is the end; please upload this sheet with your answers to the submission system.