Carnegie Mellon University in Qatar

Distributed Systems

15-440 - Fall 2018

Recitation 10 Handout

1 Intended Learning Outcome (ILO)

The ILO of this recitation is:

- Apply MapReduce to a real problem.

2 Objectives

- Understand the MapReduce data flow at a high level.
- Develop and run a simple MapReduce program.

3 High-Level MapReduce Data-Flow

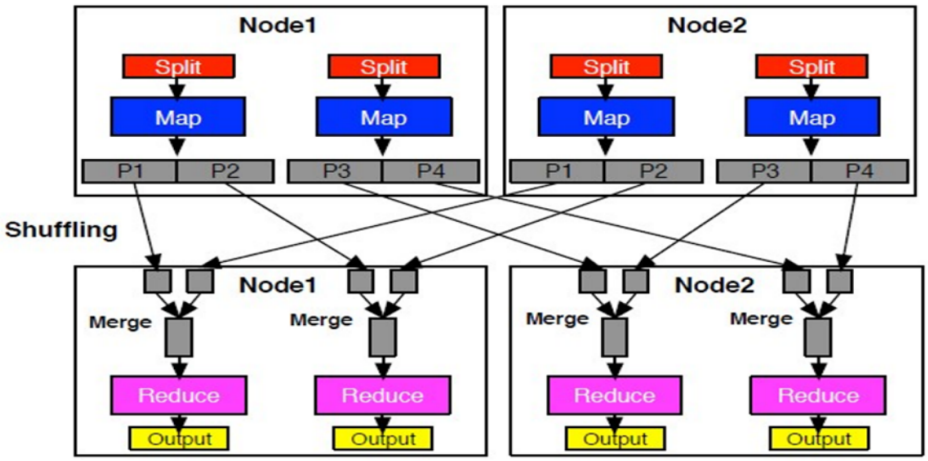

Figure 1

4 The WordCount Problem

Write a MapReduce application, referred to as WordCount, that computes the occurrence frequency of each word in text files. After you write your application, follow the following steps:

1. Create a folder for the .class files of your application using the following command

```
$ mkdir WordCount_Classes
```

2. Compile your WordCount program using the following command:

```
$ javac -classpath $(hadoop classpath)
            -d WordCount_Classes WordCount.java
```

where ***WordCount.java*** *is the program's name and that HADOOP_HOME is the root of the Hadoop installation.*

3. Create the jar file required by Hadoop to run your application using the following command:

```
$ jar -cvf WordCount.jar -C WordCount_Classes/ .
```

where "-C ***WordCount_Classes****" part of this command directs the Jar tool to go to the* ***WordCount_Classes*** *directory, and the "." following* ***WordCount_Classes/*** *directs the Jar tool to archive all the contents of in the current directory.*

4. Create two simple sample text files, `file01` and `file02`. For instance you can have them as follows:

```
echo "Welcome to MapReduce"              >> file01
echo "Welcome to MapReduce in 15440"     >> file02
```

5. Create an input directory in HDFS using the following command:

```
$ hadoop dfs -mkdir /user/hadoop/wordcount/input
```

6. Copy `file01` and `file02` to your HDFS input directory using the following commands:

```
$ hadoop dfs -copyFromLocal file01 /user/hadoop/wordcount/input
$ hadoop dfs -copyFromLocal file02 /user/hadoop/wordcount/input
```

7. Check that `file01` and `file02` now exist at `user/hadoop/wordcount/input` using the following command:

```
$ hadoop dfs -ls /user/hadoop/wordcount/input
```

8. Run your WordCount application using the following command:

```
$ hadoop jar WordCount.jar WordCount
    /user/hadoop/wordcount/input  /user/hadoop/wordcount/output
```

9. List your outputfile

```
$ hadoop dfs -ls /outputfile
```

10. Check the parts in your output

```
$ hadoop dfs -cat /outputfile/part-0000x
```