Lab 7 Map/Reduce

Hadoop Map/Reduce là một khung nền (software framework) mã nguồn mở, hỗ trợ người lập trình viết các ứng dụng theo mô hình Map/Reduce. Để hiện thực một ứng dụng theo mô hình Map/Reduce, sinh viên cần sử dụng các interface lập trình do Hadoop cung cấp như: Mapper, Reducer, JobConf, JobClient, Partitioner, OutputCollector, Reporter, InputFormat, OutputFormat, v.v..

Yêu cầu sinh viên thực thi ứng dụng WordCount trên hai mô hình: Pseudo-Distributed Operation và Fully-Distributed Operation để hiểu rõ hoạt động của mô hình Map/Reduce và kiến trúc HDFS (Hadoop Distributed FileSystem).

1 Cài đặt và sử dụng Map/Reduce

SV có thể cài đặt mô hình Pseudo-Distributed Operation trên một máy đơn. Yêu cầu cấu hình 3 tập tin trong thư mục conf của hadoop như sau:

conf/core-site.xml:

```
<configuration>
  <property>
    <name>fs.default.name</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>
```

conf/hdfs-site.xml:

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
</configuration>
```

conf/mapred-site.xml:

```
<configuration>
  <property>
    <name>mapred.job.tracker</name>
    <value>localhost:9001</value>
  </property>
</configuration>
```

Để hiện thực mô hình Fully-Distributed Operation, SV thay đổi nội dung của 3 tập tin trên như sau:

conf/core-site.xml:

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>fs.default.name</name>
    <value>hdfs://172.28.10.182:9000</value>          // SV thay doi dia chi IP tuong ung
  </property>
</configuration>
```

conf/hdfs-site.xml:

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<!-- Put site-specific property overrides in this file. -->

<configuration>
  <property>
    <name>dfs.replication</name>
    <value>2</value>
  </property>
</configuration>
```

conf/mapred-site.xml:

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
```

<!-- Put site-specific property overrides in this file. -->

<configuration>

<property>
  <name>mapred.job.tracker </name>
  <value>http://172.28.10.45:9001</value> ***// SV thay doi dia chi IP tuong ung***
</property>

<property>
   <name>mapred.local.dir</name>
   <value>/home/phuong/local</value> ***// SV thay doi thu muc tuong ung***
   <description> The local directory where MapReduce stores intermediate, data files. May be a comma-separated list of directories on different devices in order to spread disk i/o. Directories that do not exist are ignored.
   </description>
</property>

<property>
  <name>mapred.map.tasks</name>
  <value>20</value> ***// SV thay so luong map task***
</property>

<property>
  <name>mapred.reduce.tasks</name>
  <value>4</value> ***// SV thay doi so luong reduce task***
</property>

<property>
  <name>mapred.task.tracker.http.address</name>
  <value>http://172.28.10.45:50060</value> ***// SV thay doi dia chi IP tuong ung***
</property>

</configuration>

Sau khi cấu hình các tập tin .xml, SV tạo hai tập tin masters và slaves trong thư mục conf với nội dung như sau:

masters
    172.28.10.182 ***// SV thay doi dia chi IP tuong ung***
slaves
    172.28.10.182 ***// SV thay doi dia chi IP tuong ung***
    172.28.10.45

Để sử dụng nhiều hơn 2 máy, SV điền thêm địa chỉ IP các máy vào tập tin slaves.

2 Thực thi ứng dụng WordCount

SV có thể sử dụng mã nguồn WordCount.java của Google như bên dưới hoặc tự viết.

```
package org.myorg;
import java.io.IOException;
import java.util.*;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapred.*;
import org.apache.hadoop.util.*;
public class WordCount {
  public static class Map extends MapReduceBase implements
Mapper<LongWritable, Text, Text, IntWritable> {
   private final static IntWritable one = new IntWritable(1);
   private Text word = new Text();
   public void map(LongWritable key, Text value,
OutputCollector<Text, IntWritable> output, Reporter reporter)
throws IOException {
    String line = value.toString();
    StringTokenizer tokenizer = new StringTokenizer(line);
    while (tokenizer.hasMoreTokens()) {
     word.set(tokenizer.nextToken());
     output.collect(word, one);
    }
   }
  }
  public static class Reduce extends MapReduceBase implements
Reducer<Text, IntWritable, Text, IntWritable> {
   public void reduce(Text key, Iterator<IntWritable> values,
OutputCollector<Text, IntWritable> output, Reporter reporter)
throws IOException {
    int sum = 0;
    while (values.hasNext()) {
```

```
      sum += values.next().get();
     }
     output.collect(key, new IntWritable(sum));
    }
   }
   public static void main(String[] args) throws Exception {
    JobConf conf = new JobConf(WordCount.class);
    conf.setJobName("wordcount");
    conf.setOutputKeyClass(Text.class);
    conf.setOutputValueClass(IntWritable.class);
    conf.setMapperClass(Map.class);
    conf.setCombinerClass(Reduce.class);
    conf.setReducerClass(Reduce.class);
    conf.setInputFormat(TextInputFormat.class);
    conf.setOutputFormat(TextOutputFormat.class);
    FileInputFormat.setInputPaths(conf, new Path(args[0]));
    FileOutputFormat.setOutputPath(conf, new Path(args[1]));
    JobClient.runJob(conf);
   }
}
```

Để thực thi ứng dụng, sv cần thực hiện:

2.1 Chuyển chương trình WordCount.java thành file .jar: vd WordCount.jar

$ mkdir wordcount_class
$ javac -classpath {HADOOP_HOME}/hadoop-${HADOOP_VERSION}-core.jar -d wordcount_classes WordCount.java
$ jar –cvf wordcount.jar –C wordcount_classes/ .

2.2 Thực thi chương trình

SV tạo hai tập tin file1, file2 có nội dung tùy ý, chuyển chúng vào thư mục input và thực thi lệnh bên dưới:

$ bin/hadoop jar WordCount.jar org.myorg.WordCount input output

Chú ý: Một số lệnh thao tác trên HDFS

$ bin/hadoop dfs –put <source> <dest>   : cung cấp input cho chương trình
$ bin/hadoop dfs –get <dest>   <source> : lấy về output của chương trình.
$ bin/hadoop dfs –rmr <dir>   : xóa thư mục.
$ bin/hadoop dfs –rm <file>   : xóa tập tin

3 Bài tập

1 SV thực thi chương trình WordCount ver2 của Google.

2 SV viết chương trình tính PI theo mô hình Map/Reduce.