

Thực hành MapReduce

Biên soạn: Lê Thị Minh Châu

I. Mô tả

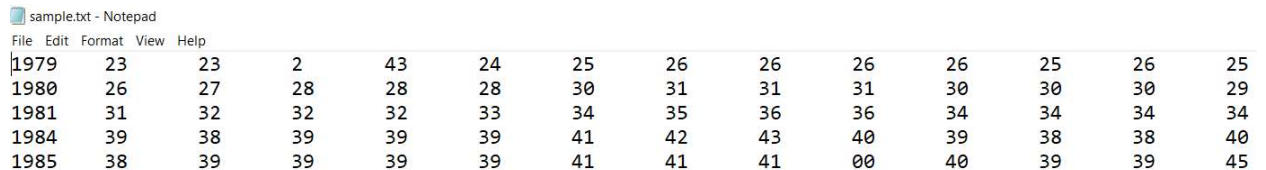
Bài toán thống kê lượng tiêu thụ điện của một tổ chức. Tập dữ liệu bên dưới chứa thông tin về mức tiêu thụ điện hàng tháng và mức trung bình hàng năm trong các năm khác nhau.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Avg
1979	23	23	2	43	24	25	26	26	26	26	25	26	25
1980	26	27	28	28	28	30	31	31	31	30	30	30	29
1981	31	32	32	32	33	34	35	36	36	34	34	34	34
1984	39	38	39	39	39	41	42	43	40	39	38	38	40
1985	38	39	39	39	39	41	41	41	00	40	39	39	45

Để xử lý tập dữ liệu trên, chúng ta phải viết các ứng dụng để xử lý và tìm ra năm sử dụng tối đa, năm sử dụng tối thiểu... Điều này chỉ khả thi khi tập dữ liệu có số lượng các bộ giới hạn. Khi tập dữ liệu phát triển, chứa thông tin của tất cả các ngành công nghiệp quy mô lớn của một tiểu bang cụ thể kể từ khi hình thành, chúng ta sẽ mất rất nhiều thời gian để thực hiện và cần một lưu lượng truy cập mạng lớn khi duy chuyển dữ liệu từ nguồn sang các máy chủ mạng. Để giải quyết các vấn đề này, chúng ta cần sử dụng framework MapReduce.

II. Tạo bộ dữ liệu

Tạo tập tin sample.txt chứa dữ liệu trên



sample.txt - Notepad

File	Edit	Format	View	Help										
1979	23	23	2	43	24	25	26	26	26	26	25	26	25	
1980	26	27	28	28	28	30	31	31	31	30	30	30	29	
1981	31	32	32	32	33	34	35	36	36	34	34	34	34	
1984	39	38	39	39	39	41	42	43	40	39	38	38	40	
1985	38	39	39	39	39	41	41	41	00	40	39	39	45	

III. Tạo chương trình MapReduce

Tạo file ProcessUnits.java chứa code sau

```
package hadoop;
```

```

import java.util.*;

import java.io.IOException;
import java.io.IOException;

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 ProcessUnits {
    //Mapper class
    public static class E_EMapper extends MapReduceBase
implements
    Mapper<LongWritable ,/*Input key Type */
    Text,                /*Input value Type*/
    Text,                /*Output key Type*/
    IntWritable>         /*Output value Type*/
    {
        //Map function
        public void map(LongWritable key, Text value,
            OutputCollector<Text, IntWritable> output,

            Reporter reporter) throws IOException {
            String line = value.toString();
            String lasttoken = null;
            StringTokenizer s = new StringTokenizer(line,"\\t");
            String year = s.nextToken();

            while(s.hasMoreTokens()) {
                lasttoken = s.nextToken();
            }
        }
    }
}

```

```

    }
    int avgprice = Integer.parseInt(lasttoken);
    output.collect(new Text(year), new
IntWritable(avgprice));
    }
}

//Reducer class
public static class E_EReduce extends MapReduceBase
implements Reducer< Text, IntWritable, Text, IntWritable > {

    //Reduce function
    public void reduce( Text key, Iterator <IntWritable>
values,
        OutputCollector<Text, IntWritable> output, Reporter
reporter) throws IOException {
        int maxavg = 30;
        int val = Integer.MIN_VALUE;

        while (values.hasNext()) {
            if((val = values.next().get())>maxavg) {
                output.collect(key, new IntWritable(val));
            }
        }
    }
}

//Main function
public static void main(String args[])throws Exception {
    JobConf conf = new JobConf(ProcessUnits.class);

    conf.setJobName("max_eletricityunits");

```

```

        conf.setOutputKeyClass(Text.class);
        conf.setOutputValueClass(IntWritable.class);
        conf.setMapperClass(E_EMapper.class);
        conf.setCombinerClass(E_EReduce.class);
        conf.setReducerClass(E_EReduce.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);
    }
}

```

```

hadoopminhchau@minhchau-master:~$ ls
apache-hive-3.1.3-bin.tar.gz  hadoop          pig              sample.txt
db-derby-10.15.2.0-bin.tar.gz  hadoop-3.3.2.tar.gz  pig-0.17.0.tar.gz  test.sh
derby                        hive            ProcessUnits.java  tmp
hadoopminhchau@minhchau-master:~$

```

IV. Thông dịch và thực thi chương trình Process Units

1. Tạo thư mục units chứa các file sau khi thông dịch ProcessUnits.java

```
$ mkdir units
```

```

hadoopminhchau@minhchau-master:~$ mkdir units
hadoopminhchau@minhchau-master:~$ ls
apache-hive-3.1.3-bin.tar.gz  hadoop          pig              sample.txt  units
db-derby-10.15.2.0-bin.tar.gz  hadoop-3.3.2.tar.gz  pig-0.17.0.tar.gz  test.sh
derby                        hive            ProcessUnits.java  tmp
hadoopminhchau@minhchau-master:~$

```

2. Download hadoop-core-1.2.1.jar

```
$ wget http://www.java2s.com/Code/JarDownload/hadoop-core/hadoop-core-1.2.1.jar.zip
```

```
$ unzip hadoop-core-1.2.1.jar.zip
```

```

hadoopminhchau@minhchau-master:~$ wget http://www.java2s.com/Code/JarDownload/hadoop-core/hadoop-core-1.2.1.jar.zip
--2022-04-18 16:36:11-- http://www.java2s.com/Code/JarDownload/hadoop-core/hadoop-core-1.2.1.jar.zip
Resolving www.java2s.com (www.java2s.com)... 52.217.136.149
Connecting to www.java2s.com (www.java2s.com)|52.217.136.149|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 3899039 (3.7M) [application/zip]
Saving to: 'hadoop-core-1.2.1.jar.zip'

hadoop-core-1.2.1.jar.zip 100%[=====>] 3.72M 1.42MB/s in 2.6s

2022-04-18 16:36:14 (1.42 MB/s) - 'hadoop-core-1.2.1.jar.zip' saved [3899039/3899039]

hadoopminhchau@minhchau-master:~$ ls
apache-hive-3.1.3-bin.tar.gz  hadoop-3.3.2.tar.gz      pig-0.17.0.tar.gz  tmp
db-derby-10.15.2.0-bin.tar.gz  hadoop-core-1.2.1.jar.zip  ProcessUnits.java  units
derby                          hive                      sample.txt
hadoop                         pig                       test.sh
hadoopminhchau@minhchau-master:~$ unzip hadoop-core-1.2.1.jar.zip
Archive:  hadoop-core-1.2.1.jar.zip
  inflating: hadoop-core-1.2.1.jar
hadoopminhchau@minhchau-master:~$ ls
apache-hive-3.1.3-bin.tar.gz  hadoop-3.3.2.tar.gz      pig                test.sh
db-derby-10.15.2.0-bin.tar.gz  hadoop-core-1.2.1.jar    pig-0.17.0.tar.gz  tmp
derby                          hadoop-core-1.2.1.jar.zip  ProcessUnits.java  units
hadoop                         hive                      sample.txt
hadoopminhchau@minhchau-master:~$

```

3. Thông dịch file nguồn và tạo file jar

```
$ javac -classpath hadoop-core-1.2.1.jar -d units
```

ProcessUnits.java

hadoop-core-1.2.1.jar chứa các class file cần thiết để thông dịch chương trình, -d xác định thư mục chứa các class file được tạo ra

```
$ jar -cvf units.jar -C units/ .
```

Tạo file units.jar chứa tất cả (dấu .) file trong thư mục units (-C là lấy trong đường dẫn hiện tại)

```

hadoopminhchau@minhchau-master:~$ javac -classpath hadoop-core-1.2.1.jar -d units ProcessUnits.java
hadoopminhchau@minhchau-master:~$ jar -cvf units.jar -C units/ .
added manifest
adding: hadoop/(in = 0) (out= 0)(stored 0%)
adding: hadoop/ProcessUnits$E_EMapper.class(in = 1984) (out= 817)(deflated 58%)
adding: hadoop/ProcessUnits.class(in = 1591) (out= 779)(deflated 51%)
adding: hadoop/ProcessUnits$E_EReducer.class(in = 1680) (out= 692)(deflated 58%)
hadoopminhchau@minhchau-master:~$

```

4. Tạo thư mục input trong HDFS

```
$ hdfs dfs -mkdir input_dir
```

```

hadoopminhchau@minhchau-master:~$ hdfs dfs -mkdir input_dir
hadoopminhchau@minhchau-master:~$ _

```

5. Đưa dữ liệu sample.txt vào thư mục input_dir trong HDFS

```
$ hdfs dfs -put sample.txt input_dir
```

```
hadoopminhchau@minhchau-master:~$ hdfs dfs -put sample.txt input_dir
hadoopminhchau@minhchau-master:~$ hdfs dfs -ls input_dir
Found 1 items
-rw-r--r--  2 hadoopminhchau supergroup      222 2022-04-19 08:36 input_dir/sample.txt
hadoopminhchau@minhchau-master:~$
```

6. Thực thi chương trình Eleunit_max trong Hadoop

\$ `hadoop jar units.jar hadoop.ProcessUnits input_dir output_dir`

Đợi cho chương trình thực thi xong sẽ cho ra kết quả số lượng input được chia tách, số lượng tác vụ Map và Reduce...

```
hadoopminhchau@minhchau-master:~$ hadoop jar units.jar hadoop.ProcessUnits input_dir output_dir
2022-04-19 09:02:15,454 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceMan
r at minhchau-master/192.168.86.1:9003
2022-04-19 09:02:15,600 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceMan
r at minhchau-master/192.168.86.1:9003
2022-04-19 09:02:15,772 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing no
performed. Implement the Tool interface and execute your application with ToolRunner to remedy thi
2022-04-19 09:02:15,786 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /t
hadoop-yarn/staging/hadoopminhchau/.staging/job_1650357234807_0002
2022-04-19 09:02:15,968 INFO mapred.FileInputFormat: Total input files to process : 1
2022-04-19 09:02:16,048 INFO mapreduce.JobSubmitter: number of splits:2
2022-04-19 09:02:16,240 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1650357234807
02
2022-04-19 09:02:16,240 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-04-19 09:02:16,381 INFO conf.Configuration: resource-types.xml not found
2022-04-19 09:02:16,381 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-04-19 09:02:16,440 INFO impl.YarnClientImpl: Submitted application application_1650357234807
02
2022-04-19 09:02:16,468 INFO mapreduce.Job: The url to track the job: http://minhchau-master:9004
oxy/application_1650357234807_0002/
2022-04-19 09:02:16,469 INFO mapreduce.Job: Running job: job_1650357234807_0002
2022-04-19 09:02:21,542 INFO mapreduce.Job: Job job_1650357234807_0002 running in uber mode : fa
2022-04-19 09:02:21,542 INFO mapreduce.Job:  map 0% reduce 0%
2022-04-19 09:02:27,614 INFO mapreduce.Job:  map 100% reduce 0%
```

```

Map-Reduce Framework
  Map input records=5
  Map output records=5
  Map output bytes=45
  Map output materialized bytes=45
  Input split bytes=240
  Combine input records=5
  Combine output records=3
  Reduce input groups=3
  Reduce shuffle bytes=45
  Reduce input records=3
  Reduce output records=3
  Spilled Records=6
  Shuffled Maps =2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=69
  CPU time spent (ms)=1400
  Physical memory (bytes) snapshot=720945152
  Virtual memory (bytes) snapshot=8159666176
  Total committed heap usage (bytes)=675282944
  Peak Map Physical memory (bytes)=271699968
  Peak Map Virtual memory (bytes)=2717286400
  Peak Reduce Physical memory (bytes)=201023488
  Peak Reduce Virtual memory (bytes)=2726711296
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=333
File Output Format Counters
  Bytes Written=24
hadoopminhchau@minhchau-master:~$ _

```

7. Xuất kết quả

```
$ hdfs dfs -cat output_dir/part-00000
```

```

hadoopminhchau@minhchau-master:~$ hdfs dfs -ls output_dir
Found 2 items
-rw-r--r--  2 hadoopminhchau supergroup      0 2022-04-19 09:02 output_dir/_SUCCESS
-rw-r--r--  2 hadoopminhchau supergroup    24 2022-04-19 09:02 output_dir/part-00000
hadoopminhchau@minhchau-master:~$ hdfs dfs -cat output_dir/part-00000
1981      34
1984      40
1985      45
hadoopminhchau@minhchau-master:~$

```

8. Copy kết quả từ Hadoop ra thư mục bên ngoài

```
$ hdfs dfs -get output_dir /home/hadoopminhchau
```

```
hadoopminhchau@minhchau-master:~$ hdfs dfs -get output_dir /home/hadoopminhchau/
hadoopminhchau@minhchau-master:~$ ls
apache-hive-3.1.3-bin.tar.gz  hadoop-3.3.2.tar.gz      output_dir      sample.txt  units.jar
db-derby-10.15.2.0-bin.tar.gz  hadoop-core-1.2.1.jar    pig             test.sh
derby                        hadoop-core-1.2.1.jar.zip  pig-0.17.0.tar.gz  tmp
hadoop                      hive                     ProcessUnits.java  units
hadoopminhchau@minhchau-master:~$ ls output_dir/
part-000000 _SUCCESS
hadoopminhchau@minhchau-master:~$
```

```
hadoopminhchau@minhchau-master:~$ ls
apache-hive-3.1.3-bin.tar.gz  hadoop-3.3.2.tar.gz      output_dir      sample.txt  units.jar
db-derby-10.15.2.0-bin.tar.gz  hadoop-core-1.2.1.jar    pig             test.sh
derby                        hadoop-core-1.2.1.jar.zip  pig-0.17.0.tar.gz  tmp
hadoop                      hive                     ProcessUnits.java  units
hadoopminhchau@minhchau-master:~$ cat output_dir/part-000000
1981      34
1984      40
1985      45
hadoopminhchau@minhchau-master:~$
```