서강대학교 SOGANG UNIVERSITY

Project #6: Hadoop MapReduce

Systems Programming

Department of Computer Science and Engineering
Sogang University

서강대학교 SOGANG UNIVERSITY

Due: June 27 (Thu), 11:59PM (KST)

Goal

1 Goal

The goal of this project is to improve your understanding basic of Hadoop and MapReduce.

1. Write all the code in Python3.
2. Do with Azure HDInsight service.
3. Use MapReduce technique for processing of data.
4. Solve the group by max problem for a large size of data.

2 Group by max

This problem is to find the maximum value in each group. The Input has a one integer and one real number each line, each number is separated by a comma. A integer number part represents a group number. A real number part means a value. You must find the maximum value for each group for given data.

Here is an example of this problem:

Input

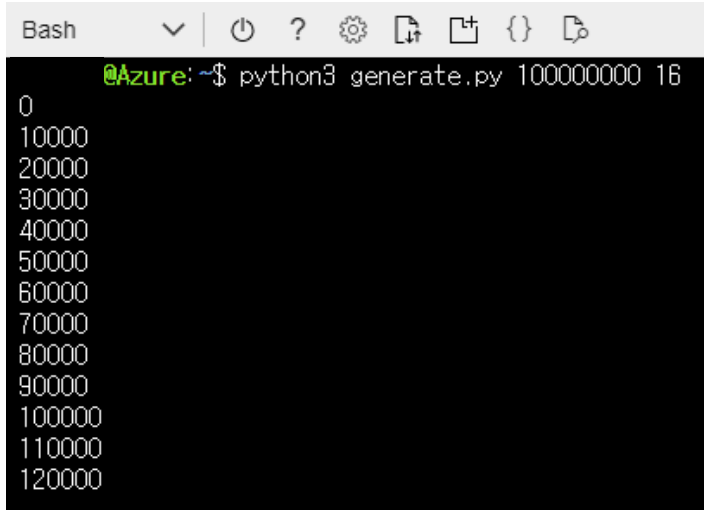
```
2,5.014
3,4.799
2,1.161
1,6.811
3,8.175
1,1.263
3,8.314
```

Output

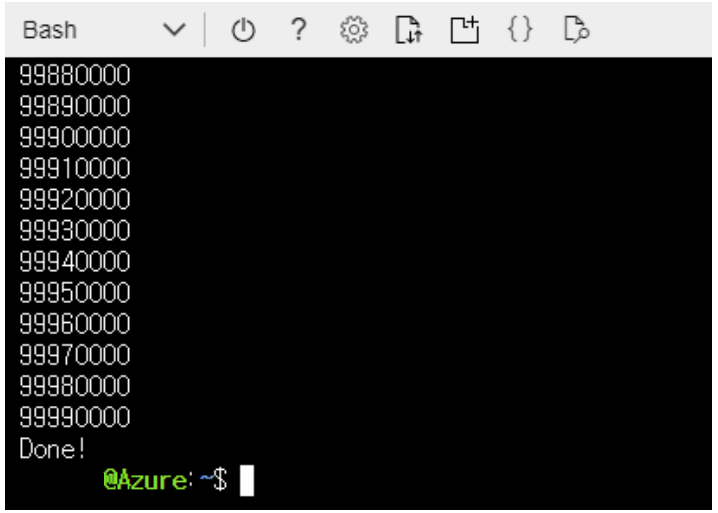
```
1  6.811
2  5.014
3  8.314
```

3 How to generate input.data and upload it to the server

Run the given generate.py file in linux server (cspro, or azure cloud shell) with the command : “python3 generate.py 100000000 16”



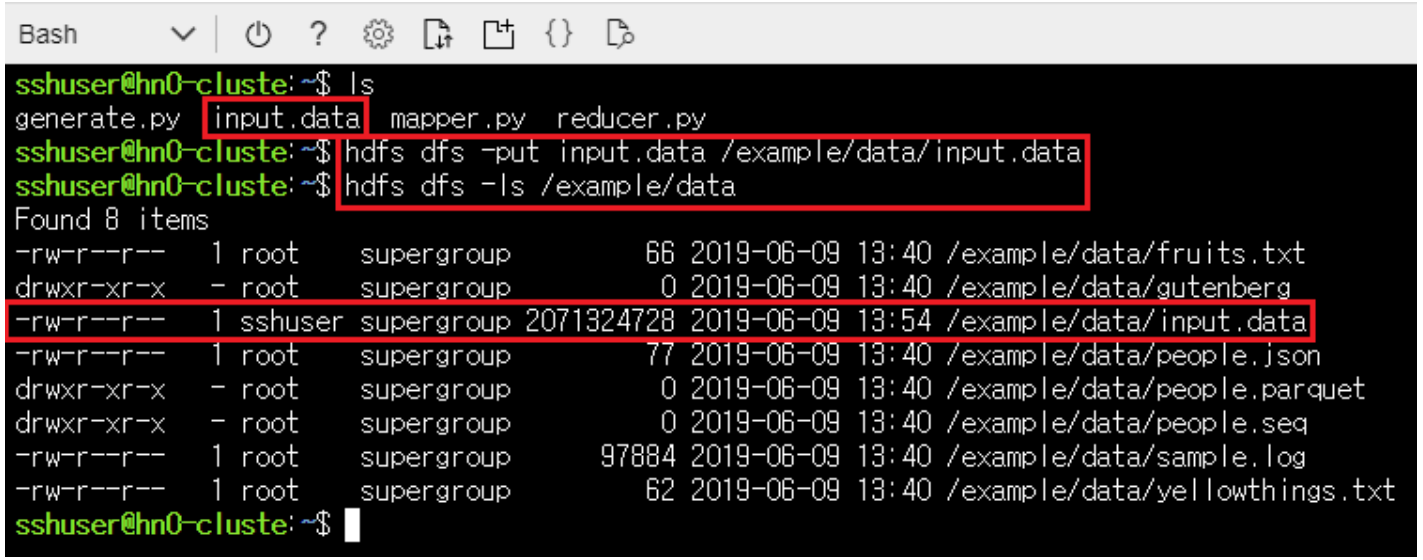
… …



The dataset ‘input.data’ to be given consists of 100 million real numbers from 0 to 10000. And there are 16 group keys.

3 How to generate input.data and upload it to the server

Upload the input.data file in your local directory to the hdfs environment with the command : “hdfs dfs - put input.data /example/data/input.data”



Check that the file is uploaded well into the hdfs environment with the command : hdfs dfs - ls /example/data

4 Hadoop File System command

```
hdfs dfs -put <local_src> . . . <dst>
```

```
hdfs dfs -put input.data /example/data/input.data
```

Copy single src, or multiple srcs from local file system to the destination file system.

```
hdfs dfs -get <src> <local_dst>
```

```
hdfs dfs -put /example/data/output.data output.data
```

Copy files to the local file system.

```
hdfs dfs -ls <args>
```

```
hdfs dfs -ls /example/wordcountout/
```

Files within a directory are order by filename by default.

```
hdfs dfs -text <src>
```

```
hdfs dfs -text /example/wordcountout/part-00000
```

Takes a source file and outputs the file in text format.

```
hdfs dfs -rm [-R] URI [URI . . .]
```

```
hdfs dfs -rm -R /example/wordcountout*
```

Delete files specified as args. The -R option deletes the directory and any content under it recursively.

5 Requirements

You need to resolve the group by max for the input.data you created yourself.

Generate 10000000 numbers for 16 groups using the given code.

You must include your student ID in output file name when use Hadoop MapReduce.

```
yarn jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar ₩
-files mapper.py,reducer.py-mapper mapper.py -reducer reducer.py ₩
-input /example/data/input.data ₩
-output /example/result[your Student ID]
```

If your student ID is 20161234, then output filed is “result2016124”

Submission

1 Things

(1) Python codes

- Each python code file named “mapper.py”, “reducer.py”

(2) A document file

- This document file should describe how you implemented your programs.
- You should insert a data flow diagram that shows how MapReduce worked.
- A sample document will be posted on cyber campus.

(3) Capture images

- Two capture images for execution and output about MapReduce.
- “20161234_1.png” is captured image about with execution command.
- “20161234_2.png” is captured image about MapReduce result.
- The numeric part should be **your student ID**.

2 Example of a captured image “20161234_1.png”

서강대학교 SOGANG UNIVERSITY

Bash

```
sshuser@hn0-cluste:~$ yarn jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar -files mapper.py,reducer
.py -mapper mapper.py -reducer reducer.py -input /example/data/input.data -output /example/data/result20161234
packageJobJar: [] [/usr/hdp/2.6.5.3008-11/hadoop-mapreduce/hadoop-streaming-2.7.3.2.6.5.3008-11.jar] /tmp/streamjob47
43914934645318502.jar tmpDir=null
19/06/09 14:04:38 INFO client.AHSProxy: Connecting to Application History server at headnodehost/10.0.0.21:10200
19/06/09 14:04:38 INFO client.AHSProxy: Connecting to Application History server at headnodehost/10.0.0.21:10200
19/06/09 14:04:39 INFO client.RequestHedgingRMFailoverProxyProvider: Looking for the active RM in [rm1, rm2]...
19/06/09 14:04:39 INFO client.RequestHedgingRMFailoverProxyProvider: Found active RM [rm2]
19/06/09 14:04:40 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library
19/06/09 14:04:40 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev b5efb3e531b
c1558201462b8ab15bb412ffa6b89]
19/06/09 14:04:40 INFO mapred.FileInputFormat: Total input paths to process : 1
19/06/09 14:04:40 INFO mapreduce.JobSubmitter: number of splits:4
19/06/09 14:04:40 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1560086904039_0002
19/06/09 14:04:41 INFO impl.YarnClientImpl: Submitted application application_1560086904039_0002
19/06/09 14:04:41 INFO mapreduce.Job: The url to track the job: http://hn1-cluste.3oquisiz15gu3pixh02cs3blzc.psx.inte
rnal.cloudapp.net:8088/proxy/application_1560086904039_0002/
19/06/09 14:04:41 INFO mapreduce.Job: Running job: job_1560086904039_0002
19/06/09 14:05:00 INFO mapreduce.Job: Job job_1560086904039_0002 running in uber mode : false
19/06/09 14:05:00 INFO mapreduce.Job:  map 0% reduce 0%
19/06/09 14:05:27 INFO mapreduce.Job:  map 5% reduce 0%
19/06/09 14:05:30 INFO mapreduce.Job:  map 9% reduce 0%
```

Please name the result file 'result + student ID'

2 Example of a captured image “20161234_2.png”

서강대학교 SOGANG UNIVERSITY

Bash

```
sshuser@hn0-cluste:~$ hdfs dfs -text /example/data/result20161234/part-00000
19/06/09 14:16:31 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library
19/06/09 14:16:31 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev b5efb3e531b
c1558201462b8ab15bb412ffa6b89]
11      9997.47050915
10      9996.36036315
13      9993.13887596
12      9989.68431495
15      9999.11036256
14      9994.30161411
1       9997.27526803
0       9994.8446836
3       9992.5913585
2       9991.17793128
5       9992.72677359
4       9996.35257724
7       9997.67399153
6       9997.73439591
9       9998.20677729
8       9994.16280426
sshuser@hn0-cluste:~$
```

3 Instructions

- Make a directory named “sp20161234_proj6”. The numeric part should be **your student ID**.
- Put all the files in the directory, and compress the directory itself using tar or zip.
- When you make a tar file, do NOT use the z option (which makes a gz compressed file.)

Example:

sp20161234_proj6/

document.docx

20161234_1.png

20161234_2.png

mapper.py

reducer.py

3 Instructions

The file for submission

sp20161234_proj6.tar or sp20161234_proj5.zip

Upload this file on the cyber campus.

Late Submission

No late submissions accepted for this project