CSC 4760/6760 Big Data Programming

Spring 2021

All Students

Exam 1

02/24/2021

Name: Tracy Michaels

Panther ID: 002430918

**Note**: You may use additional white papers if the space is not enough.

This table is to be filled by the grader.

|  |  |  |
| --- | --- | --- |
| Problem | Total Score | Score |
| 1 | 15 |  |
| 2 | 30 |  |
| 3 | 20 |  |
| 4 | 5 |  |
| 5 | 25 |  |
| 6 | 5 |  |
| Total | 100 |  |

**Problem 1**. (Hadoop Commands) David is learning Hadoop. David has installed Hadoop in the Ubuntu virtual machine. Please help him with the following questions. Suppose that the full path of hadoop is:

/home/david/hadoop/

1.1The command to start HDFS is

**Answer:**

$ sbin/start-dfs.sh

1.2 David wants to take a look at the folders and files in HDFS. He should run the following command:

**Answer:**

$ hdfs dfs -ls

1.3Suppose there is already a folder“/user/david/” in HDFS. David wants to create a folder “/user/david/data/”. The command is

**Answer:**

$ hdfs dfs -mkdir /user/David/data

1.4David has a file “peterpan.txt” in the local file system. The full directory is

/home/david/data/peterpan.txt

David wants to upload the data from local file system to the folder “/user/david/data/” in HDFS. The command is:

**Answer:**

$ hdfs dfs -put /home/david/data/peterpan.txt /user/david/data

1.5The jar file is in the following folder in the local file system:

/home/david/data/WordCount.jar

David wants to run this .jar file on the file “/user/david/data/peterpan.txt” in HDFS. The output directory is “/user/david/data/out01”. Please provide the command:

**Answer:**

$ hadoop jar /home/david/data/WordCount.jar /user/david/data/peterpan.txt /user/david/data/out01

1.6After hadoop runs successfully, David needs to see the output file. Please provide the command to see that:

**Answer:**

$ hadoop fs -cat /user/david/out01/part-r-00000

1.7David wants to transfer this output file from HDFS to the folder “/home/david/data” in the local file system. Please provide the command to do that.

**Answer:**

$ hadoop fs -get /user/david/out01/\* ~/data

1.8David can also run the .jar file on the file “/home/david/data/peterpan.txt” in local file system. The output folder is “/home/david/data/out02”. The command is:

**Answer:**

$ hadoop jar /home/david/data/WordCount.jar <file:///home/david/data/peterpan.txt> <file:///home/david/data/out02>

**Problem 2**. (WordCount Example) Please design the mapper and reducer for the WordCount problem.

WordCount problem: Given an input text file, count the frequency of each word in the file.

Please design the input and output <key, value> pairs for the Mapper and Reducer.

|  |  |  |
| --- | --- | --- |
|  | Mapper | Reducer |
| Input | Key, value, context | word, value(1), context |
| Output | Key(word),value(1) via context | word, result (added values for each key) via context written to file system |

We get the “WordCount.jar” file by using the above design. It only contains Mapper and Reducer. There is no Combiner and Partitioner in this program.

Please illustrate how the WordCount.jar program works on the following dataset. This text file only contains one line.

the sound sounds sound. it is the right right, right?

Suppose we run the WordCount.jar program on this file. Suppose that the space symbol, “.”, “,”, and “?” are all used for separating the words. The outputs of the Mapper are

|  |
| --- |
| < the , 1 > |
| < sound , 1 > |
| < sounds , 1 > |
| < sound , 1 > |
| < it , 1 > |
| < is , 1 > |
| < the , 1 > |
| < right , 1 > |
| < right , 1 > |
| < right , 1 > |

The outputs of the reducer are

|  |
| --- |
| is 1 |
| it 1 |
| right 3 |
| sound 2 |
| sounds 1 |
| the 2 |

Are the results sorted alphabetically? When was it sorted?

**Answer:**

* During the shuffle between map and reduce

Suppose the chuck size is 64MB. How many mappers and reducers are used? Why?

**Answer:**

* Depends on how big the input file size is. I would imagine that for the one line of text in this file which has 53 characters, each character takes typically 1 - 6 bytes of space depending on how it was stored. So approximately 53 - 318 Bytes, considering a chunk size of 64MB this would require only one mapper this would in turn only have need for one reducer.

Suppose we want to use 2 reducers. How to do that in the “.java” file?

**Answer:**

* job.setNumReduceTasks(2);

Suppose the key-value pairs with the keys “the”, “sound”, “is” are processed by reducer 1. Suppose the key-value pairs with the keys “it”, “right”, “sounds” are processed by reducer 2. Please determine the outputs of the two reducers.

|  |  |
| --- | --- |
| **Reducer 1** | **Reducer 2** |
| is 1 | it 1 |
| sound 2 | right 3 |
| the 2 | sounds 1 |

From the above results, we can see that

1. Are the records in the output file of one reducer sorted?

**Answer:**

* yes

2. Are the records in the output file of Reducer 1 greater alphabetically than those in the output file of Reducer 2?

**Answer:**

* no, they are about the same

If we want that the answer to the second question is “Yes”, how can we design MapReduce program?

**Answer:**

* you can use a partitioner before the shuffle so that each reducer would be given words based on alphabetical order

Please design your partitioner with the following boundary:

If the first letter of a word is less than “s”, this word belongs to partition 1; otherwise, this word belongs to partition 2. What are the partitions in the above example?

|  |  |
| --- | --- |
| Partition 1 | If key < s |
| Partition 2 | If key > s |

What are the outputs of reducer 1 and 2?

|  |  |
| --- | --- |
| **Reducer 1** | **Reducer 2** |
| is 1 | right 3 |
| it 1 | sound 2 |
| the 2 | sounds 1 |

Suppose we add the combiner to the program. Can we use the code for the reducer as the combiner?

**Answer:**

* yes, it does the same function in an intermediary step

Will the combiner helps improve the computing efficiency and why?

**Answer:**

* yes, because it will reduce the amount of traffic between nodes

How many combiners do we have?

**Answer:**

* 1 since there is only one mapper

Please Indicate the output of the combiner.

|  |
| --- |
| < the , 2 > |
| < sound, 2 > |
| < sounds, 1> |
| < it, 1> |
| <is, 1> |
| <right, 3> |

**Problem 3**. (Number of Mappers and Reducers) Suppose we have a cluster with 240 rack servers, and each rack server has 4 cores. We want to run “WordCount.jar” on a text file. Suppose the size of the input file is 36,853 MB. The chunk size in the HDFS is 64MB. Please answer the following questions.

1. How many mappers will be created when you run a Hadoop program on this cluster?

**Answer:**

- 36853/64 = 575.83 therefore 576 maps

- max number of possible mappers 240 \* 4 = 960

- since only 576 maps are required then you can set the number of mapper with job.setNumMapTasks(576)

2. Can we set the number of reducers to 0.9 \* 240 \* 4 = 864? If so, how many reducers will run simultaneously?

**Answer:**

- all of them (864) since we have 240 nodes with 4 cores each = 960

3. Can we set the number of reducers to 1.7 \* 240 \* 4 = 1632? If so, how many reducers will run simultaneously?

**Answer:**

* At most 960

4. Describe the differences of the two strategies for setting the number of reducers? One is 864, the other is 1,632. Which one is better and why?

**Answer:**

* With 864 reducers all reducers will launch immediately
  + This will have a lower overhead
* With 1632 reducers, the faster nodes will finish their first round then start on another
  + This increases overhead, but also increases load balancing and lowers cost of failures

5. When we set the number of reducers to 864, how many partitions are there in the program? How about 1,632 reducers?

**Answer:**

* The number of partitioners is the same as the number of reducers
  + 864 reducers = 864 partitioners
  + 1632 reducers = 1632 partitioners

6. Suppose we use “RandomPartitioner” and we set the number of reducers to 864. Then how many partitions do we have?

**Answer:**

* 864

**Problem 4**. (Passing Parameters)

In the top-k query problem, we need to pass the parameter k from the command line to the mapper or reducer. Please describe how to pass the parameters in Hadoop program.

**Answer:**

In the main() function:

Configuration conf = ­­­­­­­­­­­­­­­­­\_\_\_\_\_new Configuration();\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

conf. ­­­­­­­­­­­­­­­­­\_\_\_set\_\_\_ ("k value for top-k query", otherArgs[0]);

In the map() or reduce() functions:

Configuration conf = ­­­­­­­­­­­­­­­­­\_\_\_\_\_context.getConfiguration();\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Long nKValue = Long.parseLong(conf. ­­­­­­­­­­­­­­­­­\_\_\_get\_\_\_\_\_ ("k value for top-k query"));

We use \_\_\_\_\_\_\_\_arguments\_\_\_\_\_\_\_ in Hadoop to pass the parameter.

**Problem 5**. (PageRank)

The math equation of one page for PageRank is:

where represents the PageRank value of , represents the set of pages pointing into , represents the number of outlinks from , represents the decay factor, and represents the number of web pages in the web graph. The decay factor is usually set to 0.85, i.e., .

If we write out the equations for all the nodes, we have a system of linear equations. \_\_\_\_\_\_\_\_\_\_\_Power iteration\_\_\_\_\_\_\_ method is usually used to solve the system of linear equations. The time complexity is , where represents the number of iterations, and represents the number of edges in the graph.

To design the MapReduce algorithm for implementing the power iteration method for computing PageRank, we need to first figure out the Mapper and Reducer.

Let us use to represent the set of \_\_\_\_\_\_outlinks\_\_\_\_\_\_\_\_\_ from and to represent the set of \_\_\_\_\_\_\_\_\_pages\_\_\_\_\_\_\_\_ pointing into .

Each mapper will read one line of the adjacency list file. That means each mapper knows all the outlinks of a node.

Mapper:

|  |  |
| --- | --- |
| Input | PageRank value of node  Outlinks of node (one row of the adjacency matrix) |
| Output | For each node , output < \_\_node j\_\_ , \_\_\_r­it-1 ∙ pi,j\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ > |

Each mapper processes a single node . Each mapper knows the outlinks of node . For each out neighbor node , the mapper will output a key-value pair, where the key is the index of node , and the value is the value .

Reducer:

|  |  |
| --- | --- |
| Input | <\_\_\_\_node j\_\_\_\_ , \_a list of values \_ r­it-1 ∙ pi,j \_\_\_\_> , where node |
| Output | < \_\_\_\_\_\_node j\_\_\_\_\_\_\_\_\_\_\_, \_\_\_new PageRank Value r­jt\_\_\_\_\_ >,  where |

Each reducer processes a single node . During the shuffle processes, all the key-value pairs with the same \_\_\_key\_\_\_\_\_ will be sent to the same \_\_\_\_\_reducer\_\_\_\_. Therefore, each reducer knows the \_\_\_\_\_\_\_output\_\_\_\_\_ of node . For each in-neighbor node , the reducer knows the value . The reducer will sum these values together. The reducer will then calculates the value of the PageRank equation . The reducer will output a key-value pair, where the key is the index of node, and the value is the value .

Since the mapper needs to read two files, therefore we use the \_\_\_\_\_\_power iteration\_\_\_\_\_\_\_ technique to read the PageRank value file, and use the InputFileFormat to read one line in the adjacency list file.

The MapReduce program will repeat this process many times in order to calculate the final PageRank values. After each iteration, the intermediate PageRank values will written into the \_\_\_\_\_\_\_\_\_\_\_\_\_disk\_\_\_\_\_\_\_\_\_\_. At the beginning of each iteration, the program will read the intermediate PageRank values from the \_\_\_\_\_\_\_\_disk\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

Can we use combiner in the program? \_\_\_\_if there are multiple key-value pairs with the same key than part of the reducer function could be used in a combiner step\_\_\_\_\_\_\_\_\_.

If so, can we use the code of reducer as the combiner? \_\_\_\_\_since the combiner performs the same functions as the reducer we could use the code of the reducer in the combiner \_\_\_\_\_\_\_\_.

If not, please describe the idea of combiner.

Combiner:

|  |  |
| --- | --- |
| Input | <\_\_\_\_node j\_\_\_\_ , \_a list of values \_ r­it-1 ∙ pi,j \_\_\_\_> , where node |
| Output | < \_\_\_\_\_\_node j\_\_\_\_\_\_\_\_\_\_\_, \_\_\_new PageRank Value r­jt\_\_\_\_\_ >,  where |

Do we need to change the code of reducer if we use this combiner? \_\_\_\_\_\_\_it should be the same or similar\_\_\_\_\_\_\_\_\_\_\_.

**Problem 6**. (Skewed Distribution)

In the WordCount example, we have the skewed distribution problem. The frequency of some commonly used words are much higher than that of barely used words. The tasks for processing some keys take significantly longer time than other tasks.

Please briefly describe your idea for solving the skewed distribution problem such that the work loads are balanced across all the reducers.

**Answer:**

Use a combiner after the map task then use a partitioner and check if the value for each key is is above a specific threshold and if it is then send smaller amounts of those keys to the reducer while if it is under that threshold send larger amounts of those keys to the reducer