Project 3 (Total points: 20)

Due time: April 21th, 2022 11:59pm

The purposes of this project:

1. Understanding the Hadoop distributed file system and mastering the commands used to manipulate the file system
2. Implementing a Hadoop program with the Map Reduce framework
3. Demonstrating the skills to run map-reduce jobs on the department Hadoop cluster (zoinberg.cs.ndsu.nodak.edu)

After you submit your project, **please download it and make sure you submit it successfully and correctly**.

**Biological Data Counting**

A cell of human being contains 23 pairs of chromosomes (from chromosome 1 to chromosome 22, XX for female and XY for male). In Table 1, it contains the number of base pairs of each chromosome. A base pair is a basic building block or unit of a chromosome. For example, the chromosome 3 contains 198,295,559 base pairs. Therefore, on the chromosome 3, the first base pair is indexed as 1 and the last base pair is indexed as 198,295,559. Other chromosomes can be indexed in the same way.

|  |  |
| --- | --- |
| Chromosome | Base pairs |
| 1 | 248,956,422 |
| 2 | 242,193,529 |
| 3 | 198,295,559 |
| 4 | 190,214,555 |
| 5 | 181,538,259 |
| 6 | 170,805,979 |
| 7 | 159,345,973 |
| 8 | 145,138,636 |
| 9 | 138,394,717 |
| 10 | 133,797,422 |
| 11 | 135,086,622 |
| 12 | 133,275,309 |
| 13 | 114,364,328 |
| 14 | 107,043,718 |
| 15 | 101,991,189 |
| 16 | 90,338,345 |
| 17 | 83,257,441 |
| 18 | 80,373,285 |
| 19 | 58,617,616 |
| 20 | 64,444,167 |
| 21 | 46,709,983 |
| 22 | 50,818,468 |
| X | 156,040,895 |
| Y | 57,227,415 |

Table 1: The number of base pairs of each human chromosome

There is a biological experiment, which can detect interactions between chromosome regions. For example, the following is one example of an interaction between two chromosome regions from the input file named “interactions”:

1 566111 571111 5 99380374 99385374

It means one chromosome region on the chromosome 1 from 566,111 to 571,111 interacts with the other chromosome region on the chromosome 5 from 99,380,374 to 99,385,374. The input file contains hundreds of thousands of interactions. In this input file, chromosomes are labeled from 1 to 23. The number 23 is used to represent chromosome X because cells used in the experiment are from a female. (So, there is No chromosome Y)

Now let us divide each chromosome into continuous disjoint bins. Each bin is 100,000 base pairs (except the last one of each chromosome). Therefore, the chromosome 1 has ceiling (248,956,422 /100,000) = 2,490 bins. We index these bins from 1 to 2,490. The chromosome 2 has ceiling (242,193,529/100,000) = 2,422 bins. We index them staring with 2,490+1=2,491 and ends with 2,490+2,422=4,912. In this way, we can continue index each bin of the chromosome 3 until the chromosome X.

After we index these bins from the chromosome 1 to the chromosome X, we want to count in the input file the number of interactions falling into corresponding bin pairs. For example,

1 566111 571111 5 99380374 99385374

falls into the bin 6 and the bin 2,490+2,422+1,983+1,903+994=9,792. So, there is at least one interaction falling into the bin 6 and the bin 9,792. If we have another interaction in the following format

5 99380372 99385372 1 566114 571114

it is also falling into the bin 6 and the bin 9,792. So, for the bin pair (bin 6 and bin 9,792) the number interactions increase by 1.

Please write a Map Reduce program to count the number of interactions falling into the corresponding bin pairs in the input file. Save the bin pairs and their numbers of interactions (frequencies) to the Hadoop distributed file system. You can save bin pairs and frequencies in this format: (6, 9792) 2

If your input file is like:

1          566111            571111            5          99380374        99385374

5          99380372        99385372        1          566114            571114

your output file should be like:

(6, 9792) 2

Note: some rows in your input file are invalid. For example,

1 1132491 1134295 10 213989224 213990924

The above example is not valid because the length of chromosome 10 is 133,797,422. You need remove these rows in your program.

Requirements:

1. Please submit a compressed file which contains your source code and a document. In the document, please include screenshots on how to upload the input file to the Hadoop distributed file system and how to run your Map Reduce program on the CS Hadoop cluster, and results after you execute these commands. The screenshots need contains your username information on the CS Hadoop cluster. In your document, please also include the log information about the number Mappers and the number Reducers.
2. For each bin pair, make sure the first bin is less than the second bin.
3. Please make sure your code is executable and counts correctly. You can use the first 10 lines of the input file to test your code.
4. Please comment your source code briefly.