IERG 4300 Fall 2019 Homework #1

Release date: Sep 20, 2019

Due date: Oct 2, 2019 (Wed) 11:59am. (i.e. noon-time)

The solution will be posted soon after the deadline. No late homework will be accepted!

Every Student MUST include the following statement, together with his/her signature in the submitted homework.

I declare that the assignment submitted on Elearning system is original except for source material explicitly acknowledged, and that the same or related material has not been previously submitted for another course. I also acknowledge that I am aware of University policy and regulations on honesty in academic work, and of the disciplinary guidelines and procedures applicable to breaches of such policy and regulations, as contained in the website http://www.cuhk.edu.hk/policy/academichonesty/.

Signed (Student Chim) Date: 30-9-2019

Name Chim Ka Long SID 1155094482

Submission notice:

- Submit your homework via the elearning system

General homework policies:

A student may discuss the problems with others. However, the work a student turns in must be created COMPLETELY by oneself ALONE. A student may not share ANY written work or pictures, nor may one copy answers from any source other than one's own brain.

Each student **MUST LIST** on the homework paper the **name of every person he/she has discussed or worked with**. If the answer includes content from any other source, the student **MUST STATE THE SOURCE**. Failure to do so is cheating and will result in sanctions. Copying answers from someone else is cheating even if one lists their name(s) on the homework.

If there is information you need to solve a problem but the information is not stated in the problem, try to find the data somewhere. If you cannot find it, state what data you need, make a reasonable estimate of its value, and justify any assumptions you make. You will be graded not only on whether your answer is correct, but also on whether you have done an intelligent analysis.

1155094482 Chim ka Long

The overall result:

Part A small dataset:

```
474,606 155
249,610 164
274,414 164
68,414  165
414,610 187
414,599 196
288,414 202
380,414 207
414,448 272
414,474 298
```

Part A large dataset:

```
8811,14463      1055
56707,59269     1061
59269,68259     1070
30723,59269     1157
27468,59269     1218
7795,59269      1356
59269,67385     1429
19635,59269     1577
7795,30723      2152
30687,31327     2330
```

Part B small dataset:

```
582     <123,0.097826> <400,0.08046> <526,0.056604>
182     <64,0.059652> <608,0.059498> <606,0.058324>
282     <354,0.076087> <166,0.071839> <434,0.065217>
82      <354,0.07013> <239,0.053613> <570,0.050992>
382     <220,0.055679> <62,0.044521> <348,0.042553>
482     <363,0.040268> <200,0.0358> <376,0.033195>
```

Part B large dataset:

```
4482    <60423,0.110345> <6339,0.094059> <6420,0.092857>
14482   <37803,0.078845> <38340,0.076394> <70931,0.074074>
24482   <11967,0.078947> <38982,0.065217> <69424,0.065217>
34482   <1822,0.131313> <16834,0.130952> <60127,0.121212>
44482   <8835,0.32906> <54630,0.319249> <22169,0.313653>
54482   <42697,0.066667> <50783,0.066381> <7286,0.056338>
64482   <28285,0.3> <64119,0.294574> <6984,0.286624>
```

Explanation

Part A:

1) Map reduce 0: mapper output everything, but use movie id as key, reducer aggregate the users with same movie in array. Then process the array to emit pair of users if their ratings are same. The output like below

```
270,570 1
270,590 1
132,270 1
270,307 1
270,288 1
270,305 1
217,270 1
214,270 1
```

Map reduce 1: mapper output everything directly. Reducer sum the number and put into array. If array exceed certain number, it will do sorting and only hold the top 10 same rating. Finally, print them out

```
474,606 155
249,610 164
274,414 164
68,414  165
414,610 187
414,599 196
288,414 202
380,414 207
414,448 272
414,474 298
```

Part B:

1) Map reduce 0: Different from part A, reducer only emit pair which include userid same as my last 2 digits. Also, it includes the number of same movies with same rating and the number of same movies regardless to rating. The output like below

```
159,382 0 1
249,382 1 1
246,382 1 1
177,382 1 1
210,382 0 1
```

Map reduce Count: We calculate the sum of movies of every user.

```
175     24
114     31
534     520
412     102
504     87
480     836
```

Map reduce 1: mapper only emit everything, and reducer will sum up the two numbers of same movies and calculate the similarity. Reducer need to read the sum of movies of every user firstly and according to user pair, calculate similarity = {no. of same movies and same rating}/{no. of movies of user A + no. of movies of user B – no. of same movies}. The output like below

```
82,354   0.070130
166,282 0.071839
282,354 0.076087
400,582 0.080460
123,582 0.097826
```

The command like below

```
z1550944482@dic14:~/HomeWork1/1b1$ hadoop jar /usr/hdp/2.4.2.0-258/hadoop-mapreduce/hadoop-streaming.jar -D mapred.map.tasks=5 -D mapred.reduce.tasks=5 -files ../small_indi_cnt_out/part-00000#file_sum  -file relation_mapper1.py -mapper relation_mapper1.py -file relation_reducer1.py -reducer relation_reducer1.py -input homework1/small_rela_score_out0/* -output homework1/small_rela_score_out1
```

Map reduce 2: mapper filter the users of pair, if its last 2 digits is my cuid last 2 digits, emit it as Key and the partner userid + similarity as Value. Reducer will only print out the top 3 similar users with their id and similarity. Output like below

```
582     <123,0.097826> <400,0.08046> <526,0.056604>
182     <64,0.059652> <608,0.059498> <606,0.058324>
282     <354,0.076087> <166,0.071839> <434,0.065217>
82      <354,0.07013> <239,0.053613> <570,0.050992>
382     <220,0.055679> <62,0.044521> <348,0.042553>
482     <363,0.040268> <200,0.0358> <376,0.033195>
```

2) Handle large data set is almost same, except we filter with last 4 digits.
   Output:

```
4482    <60423,0.110345> <6339,0.094059> <6420,0.092857>
14482   <37803,0.078845> <38340,0.076394> <70931,0.074074>
24482   <11967,0.078947> <38982,0.065217> <69424,0.065217>
34482   <1822,0.131313> <16834,0.130952> <60127,0.121212>
44482   <8835,0.32906> <54630,0.319249> <22169,0.313653>
54482   <42697,0.066667> <50783,0.066381> <7286,0.056338>
64482   <28285,0.3> <64119,0.294574> <6984,0.286624>
```

Part C:

Our map reduce job has 2 parts

Map reduce 0:

	Max. Mapper time	Min. Mapper time	Average. Mapper time	Max. Reducer time	Min. Reducer time	Average. Reducer time	Total job
5 mappers 5 reducers	8s	7s	7s	1h 29mins	1h 1mins	1h 14mins	1h 41mins
10 mappers 10 reducers	13s	5s	11s	59mins 55s	25mins 9s	43mins 25s	1h 6s
20 mappers 20 reducers	7mins 5s	4s	6s	42mins 4s	11mins 29s	21mins 21s	42mins 18s

Map reduce 1:

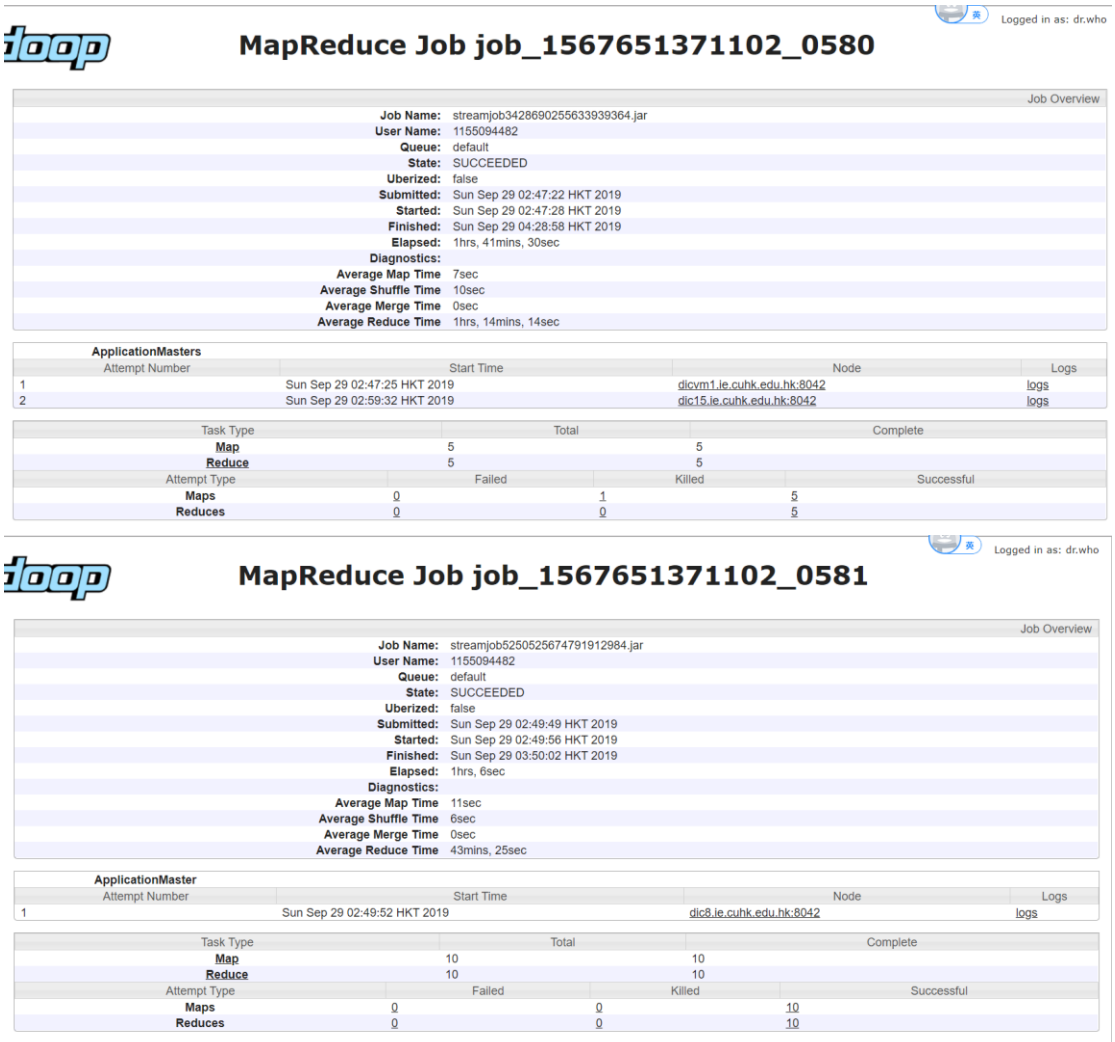
Although I only assign 5 mappers, the input file is too large, it is separated into many map tasks.

	Max. Mapper time	Min. Mapper time	Average. Mapper time	Max. Reducer time	Min. Reducer time	Average. Reducer time	Total job
5 mappers 5 reducers	19mins 39s	13s	56s	2h 34mins	1h 46mins	1h 2mins	2h 44mins
10 mappers 10 reducers	9mins 15sec	14s	48s	1h 22mins	57mins 9s	25mins 24s	1h 35mins
20 mappers 20 reducers	2mins 2s	6s	46s	57mins 16s	24mins 36s	12mins 31s	1h 6mins

Obviously, more mappers and reducers can make entire job faster. But it is not

linearly. For example, if the mappers and reducers increase become 2 times, the total job elapsed time will not become half. There are mappers or reducers which completing job slowly, because their hardware is worse than others, or their jobs is much difficult. For example, in map reducer1, if the movie was watched by many users, the number of pairs emitted will exponentially increase. The job will become more difficult.

Reference for part C:



Logged in as: dr.who

MapReduce Job job_1567651371102_0580

Job Overview	
Job Name:	streamjob3428690255633939364.jar
User Name:	1155094482
Queue:	default
State:	SUCCEEDED
Uberized:	false
Submitted:	Sun Sep 29 02:47:22 HKT 2019
Started:	Sun Sep 29 02:47:28 HKT 2019
Finished:	Sun Sep 29 04:28:58 HKT 2019
Elapsed:	1hrs, 41mins, 30sec
Diagnostics:	
Average Map Time	7sec
Average Shuffle Time	10sec
Average Merge Time	0sec
Average Reduce Time	1hrs, 14mins, 14sec

ApplicationMasters

Attempt Number	Start Time	Node	Logs
1	Sun Sep 29 02:47:25 HKT 2019	dicvm1.ie.cuhk.edu.hk:8042	logs
2	Sun Sep 29 02:59:32 HKT 2019	dic15.ie.cuhk.edu.hk:8042	logs

Task Type	Total	Complete
Map	5	5
Reduce	5	5

Attempt Type	Failed	Killed	Successful
Maps	0	1	5
Reduces	0	0	5

Logged in as: dr.who

MapReduce Job job_1567651371102_0581

Job Overview	
Job Name:	streamjob5250525674791912984.jar
User Name:	1155094482
Queue:	default
State:	SUCCEEDED
Uberized:	false
Submitted:	Sun Sep 29 02:49:49 HKT 2019
Started:	Sun Sep 29 02:49:56 HKT 2019
Finished:	Sun Sep 29 03:50:02 HKT 2019
Elapsed:	1hrs, 6sec
Diagnostics:	
Average Map Time	11sec
Average Shuffle Time	6sec
Average Merge Time	0sec
Average Reduce Time	43mins, 25sec

ApplicationMaster

Attempt Number	Start Time	Node	Logs
1	Sun Sep 29 02:49:52 HKT 2019	dic8.ie.cuhk.edu.hk:8042	logs

Task Type	Total	Complete
Map	10	10
Reduce	10	10

Attempt Type	Failed	Killed	Successful
Maps	0	0	10
Reduces	0	0	10

Logged in as: dr.who

MapReduce Job job_1567651371102_0582

Job Overview	
Job Name:	streamjob8506691652159702830.jar
User Name:	1155094482
Queue:	default
State:	SUCCEEDED
Uberized:	false
Submitted:	Sun Sep 29 02:52:12 HKT 2019
Started:	Sun Sep 29 02:52:25 HKT 2019
Finished:	Sun Sep 29 03:34:43 HKT 2019
Elapsed:	42mins, 18sec
Diagnostics:	
Average Map Time	6sec
Average Shuffle Time	7sec
Average Merge Time	0sec
Average Reduce Time	21mins, 21sec

ApplicationMaster

Attempt Number	Start Time	Node	Logs
	Sun Sep 29 02:52:17 HKT 2019	dic18.ie.cuhk.edu.hk:8042	logs

Task Type	Total	Complete
Map	20	20
Reduce	20	20

Attempt Type	Failed	Killed	Successful
Maps	0	1	20
Reduces	0	1	20

Logged in as: dr.who

MapReduce Job job_1567651371102_0745

Job Overview	
Job Name:	streamjob8980375982397334413.jar
User Name:	1155094482
Queue:	default
State:	SUCCEEDED
Uberized:	false
Submitted:	Sun Sep 29 14:05:04 HKT 2019
Started:	Sun Sep 29 14:05:25 HKT 2019
Finished:	Sun Sep 29 16:49:50 HKT 2019
Elapsed:	2hrs, 44mins, 24sec
Diagnostics:	
Average Map Time	56sec
Average Shuffle Time	58mins, 39sec
Average Merge Time	10sec
Average Reduce Time	1hrs, 2mins, 59sec

ApplicationMaster

Attempt Number	Start Time	Node	Logs
	Sun Sep 29 14:05:14 HKT 2019	dic19.ie.cuhk.edu.hk:8042	logs

Task Type	Total	Complete
Map	915	915
Reduce	5	5

Attempt Type	Failed	Killed	Successful
Maps	0	25	915
Reduces	0	0	5

Logged in as: dr.wh

MapReduce Job job_1567651371102_0748

Job Overview	
Job Name:	streamjob5609767518110085365.jar
User Name:	1155094482
Queue:	default
State:	SUCCEEDED
Uberized:	false
Submitted:	Sun Sep 29 14:06:59 HKT 2019
Started:	Sun Sep 29 14:15:57 HKT 2019
Finished:	Sun Sep 29 15:50:59 HKT 2019
Elapsed:	1hrs, 35mins, 1sec
Diagnostics:	
Average Map Time	48sec
Average Shuffle Time	41mins, 20sec
Average Merge Time	4sec
Average Reduce Time	25mins, 24sec

ApplicationMaster

Attempt Number	Start Time	Node	Logs
	Sun Sep 29 14:15:46 HKT 2019	dic6.ie.cuhk.edu.hk:8042	logs

Task Type	Total	Complete
Map	918	918
Reduce	10	10

Attempt Type	Failed	Killed	Successful
Maps	0	0	918
Reduces	0	0	10

Logged in as: dr.w

MapReduce Job job_1567651371102_0750

Job Overvie

Job Name:	streamjob3569073320037543332.jar
User Name:	1155094482
Queue:	default
State:	SUCCEEDED
Uberized:	false
Submitted:	Sun Sep 29 14:08:50 HKT 2019
Started:	Sun Sep 29 14:28:35 HKT 2019
Finished:	Sun Sep 29 15:35:11 HKT 2019
Elapsed:	1hrs, 6mins, 36sec
Diagnostics:	
Average Map Time	46sec
Average Shuffle Time	18mins, 14sec
Average Merge Time	3sec
Average Reduce Time	12mins, 31sec

ApplicationMaster

Attempt Number	Start Time	Node	Logs
	Sun Sep 29 14:28:24 HKT 2019	dic10.ie.cuhk.edu.hk:8042	logs

Task Type	Total	Complete
Map	923	923
Reduce	20	20

Attempt Type	Failed	Killed	Successful
Maps	0	0	923
Reduces	0	0	20