Assignment-1: MapReduce

In this assignment we will be working on the processing of a Movie dataset: https://raw.githubusercontent.com/sidooms/MovieTweetings/master/latest/ratings.dat (https://raw.githubusercontent.com/sidooms/MovieTweetings/master/latest/ratings.dat)

In this dataset, each row contains a movie rating done by a user (e.g., user1 has rated *Titanic* as 10). Here is the format of the dataset: `user_id::movie_id::rating::timestamp`

In this assignment, for each pair of movies A and B, you need to find all the users who rated both movie A and B. For example, given the following dataset (for the sake of illustration we have used U and M to represent users and movies respectively in the example):

```
U1::M1::2::11111111
U2::M2::3::11111111
U2::M3::1::11111111
U3::M1::4::11111111
U4::M2::5::11111111
U5::M2::3::11111111
U5::M1::1::11111111
U5::M3::3::11111111
```

The assumption is that User and Movie names are in String format and Rating is an Integer value. You should ignore the timestamp in the Mapper.

The output of your code should be in the form as below:

```
(M1,M2) [(U5,1,3)]
(M2,M3) [(U5,3,3),(U2,3,1)]
(M1,M3) [(U5,1,3)]
```

where (M,M) shows pairs of movies, [] indicates the list of users and their ratings. For example, (U5,1,3) shows U5 has rated M1 and M2 with 1 and 3 respectively.

(please note that Your output should exactly be formatted as the output example above.)

Tips :

- You may need to implement more than one Mapper/Reducer in this assignment. You need to look at chaining in MapReduce jobs: https://stackoverflow.com/questions/38111700/chaining-of-mapreduce-jobs#answer-38113499 (https://stackoverflow.com/questions/38111700/chaining-of-mapreduce-jobs#answer-38113499)
- You also may need a self-join to find movie pairs, the reduce-side join pattern can help: https://www.edureka.co/blog/mapreduce-example-reduce-side-join/

(https://www.edureka.co/blog/mapreduce-example-reduce-side-join/)

- If the key and the values for a Mapper differ from those of Reduce, you need to set the following configurations:
  job.setMapOutputKeyClass(), job.setOutputKeyClass(), job.setMapOutputValueClass(), job.setapOutputValueClass()
- Do not set **Combiner** in this assignment (~~job.setCombiner()~~)
- If the Value for the Mapper/Reducer is a complex object, you need to implement a Writerable Interface class
- If the Key for the Mapper/Reducer is a complex object, you also need to implement a WritableComparable Interface
- You can use ArrayWritable to store Array values, but you need to implement its toString() function to be able to write the object into a text file.
- Please be aware of iterating over values inside a reducer (Iterable<MyWritable> values). When looping through the Iterable value list, each Object instance is reused internally by the reducer. So if you add them to another list, at the end of the process, all of the elements in the new list will be the same as the last object you added to the list.

Submission Guidelines

Submission Deadline:

Tuesday the 29th of October 2019 17:59

Build your Project

Get and add Hadoop dependency to your project:

1. Create a new Java project in Eclipse

2. Download the dependency to your project: https://github.com/mysilver/COMP9313/raw/master/Hadoop-Core.jar (https://github.com/mysilver/COMP9313/raw/master/Hadoop-Core.jar)

3. Right click on your project and add the Hadoop-Core.jar file to your project:

    Build-Path -> Add External Archives

Submit your Project

Your code must be included (in its entirety) in the file **AssigOne{zid}.java** . Any solution that has compilation errors will receive no more than 5 points for the entire assignment.

You need to test your file before submission, to make sure that it can be compiled/run in the terminal of CSE machines:

```
$ javac -cp ".:Hadoop-Core.jar" AssigOne{zid}.java
$ java -cp ".:Hadoop-Core.jar" AssigOne{zid}  INPUT_PATH OUTPUT_PATH
```

Log in to any CSE server (e.g. williams or wagner) and use the *give command* below to submit your solution:

```
$ give cs9313 assig1 AssigOne{zid}.java
```

where you must replace {zid} above with your own zID.

You can also submit your solution using WebCMS, or Give:

https://cgi.cse.unsw.edu.au/~give/Student/give.php (https://cgi.cse.unsw.edu.au/~give/Student/give.php)

If you submit your assignment more than once, we will only consider the last submission. If you face any problem while submitting your code, please e-mail the Course Admin (Maisie Badami, m.badami@student.unsw.edu.au (mailto:m.badami@student.unsw.edu.au))

Assessment

Your source code will be manually inspected and marked based on readability and ease of understanding. We will run your code to verify that it produces correct results. The code documentation (i.e. comments in your source code) is also important. Below, we provide an indicative assessment scheme (maximum mark: 20 points):

Result correctness	17 points
Code structure and source code documentation (comments)	3 points

Late submission penalty

10% reduction of your marks for the 1st day, 30% reduction/day for the following days.

Plagiarism

This is an *individual assignment* . The work you submit must be your own work. Submission of work partially or completely derived from any other person or jointly written with any other person is not permitted. The penalties for such offence may include negative marks, automatic failure of the course and possibly other academic discipline. Assignment submissions will be examined manually.

Do not provide or show your assignment work to any other person - apart from the teaching staff of this course. If you knowingly provide or show your assignment work to another person for any reason, and work derived from it is submitted, you may be penalized, even if the work was submitted without your knowledge or consent. Pay attention that is also your duty to protect your code artifacts. if you are using any online solution to store your code artifacts (e.g., GitHub) then make sure to keep the reposiroty private and do not share access to anyone.

Reminder: Plagiarism is defined as (https://student.unsw.edu.au/plagiarism) using the words or ideas of others and presenting them as your own. UNSW and CSE treat plagiarism as academic misconduct, which means that it carries penalties as severe as being excluded from further study at UNSW. There are several on-line sources to help you understand what plagiarism is and how it is dealt with at UNSW:

- Plagiarism and Academic Integrity (https://student.unsw.edu.au/plagiarism)
- UNSW Plagiarism Procedure (https://www.gs.unsw.edu.au/policy/documents/plagiarismprocedure.pdf)

Make sure that you read and understand these. Ignorance is not accepted as an excuse for plagiarism. In particular, you are also responsible for ensuring that your assignment files are not accessible by anyone but you by setting the correct permissions in your CSE directory and code repository, if using one (e.g., Github and similar). Note also that plagiarism includes paying or asking another person to do a piece of work for you and then submitting it as your own work.

UNSW has an ongoing commitment to fostering a culture of learning informed by academic integrity. All UNSW staff and students have a responsibility to adhere to this principle of academic integrity. Plagiarism undermines academic integrity and is not tolerated at UNSW.

The marking scheme for this assignment is as below:

---

1. The output is generated by Hadoop MapReduce correctly **(7 Marks)**

---

2. The output is formatted correctly as described in the spec **(2 marks)**

- Format must be the same as (M2,M3) [(U5,3,3),(U2,3,1)]
- There should be no empty list like: (M2,M3) []

3.Custom Writables are well defined (e.g., appropriate data-types have been used without wasting memory) **(4 marks)**

4. The code could be executed in CSE machines as described in the specification of the assignment **(2 Marks)**

5. The maximum number of Mappers and Reducers is 2 (2 Mappers and 2 Reducers) **(2 Marks)**

6. Documentation and code structure. here you can clearly explain you solution in a short paragraph in the beginning of the program (no more than 300 words) and provide comments describing what each class is doing. **(3 Marks)**

Resource created about a month ago (Thursday 10 October 2019, 12:01:33 PM), last modified 12 days ago (Monday 28 October 2019, 09:04:26 PM).

Comments

(/COMP9313/19T3/forums/search?forum_choice=resource/36376)

(/COMP9313/19T3/forums/resource/36376)

Add a comment

Binru Wang (/users/z5179653) 10 days ago (Wed Oct 30 2019 13:01:40 GMT+1100 (澳大利亚东部夏令时间))

```
Session: 19T3
Assignment: assign1  Submission ID: 5179653


Current day and time:   Wed Oct 30 12:56:05 2019
Assignment deadline:    Tue Oct 29 17:59:59 2019
A submission now would be 19 hours late
Event           Day and time              Details
================================================================================
Submission      Sun Oct 27 20:30:13 2019  5179653 all assign1 [21 hours early]
Submission      Sun Oct 27 21:31:27 2019  5179653 all assign1 [20 hours early]
Submission      Mon Oct 28 19:12:44 2019  5179653 all assign1 [2 hours late]

Most recent submission:
-rw-r----- z5179653/z5179653 8987 2019-10-28 19:10 AssigOnez5179653.java
```

Hi,

Just make sure that my last submission is actually earlier than the deadline. I think the system reacts according to the old deadline.

Best,

Reply

Maisie Badami (/users/z3500952) 10 days ago (Wed Oct 30 2019 14:23:54 GMT+1100 (澳大利亚东部夏令时间))

That's fine, we won't consider it as a late submission, although I would recommend not to leave the submission to the last day or hours of deadline

Thanks

Maisie

Reply

Run Wang (/users/z5178164) 10 days ago (Wed Oct 30 2019 10:59:28 GMT+1100 (澳大利亚东部夏令时间))

This is my latest submission, although the date is before the due, but the system shows I submitted late, does that matters?

```
Most recent submission: -rw------- z5178164/z5178164 12140 2019-10-29 00:
```

Reply

Maisie Badami (/users/z3500952) 10 days ago (Wed Oct 30 2019 14:21:36 GMT+1100 (澳大利亚东部夏令时间))

That's fine, we won't consider it as a late submission.

Reply

Xiuye Yuan (/users/z3485805) 11 days ago (Tue Oct 29 2019 13:24:45 GMT+1100 (澳大利亚东部夏令时间))

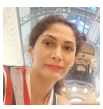
Hi,

As the spec states: "The assumption is that User and Movie names are in String format and Rating is an Integer value."

May I also make an assumption that the Rating is in string format?

Best Regards.

Reply



Maisie Badami (/users/z3500952) 11 days ago (Tue Oct 29 2019 15:32:30 GMT+1100 (澳大利亚东部夏令时间))

no you can't, it is what is clearly stated in the spec.

Reply

Xiuye Yuan (/users/z3485805) 11 days ago (Tue Oct 29 2019 15:54:14 GMT+1100 (澳大利亚东部夏令时间))

Got it, Thanks for your reply!

Reply

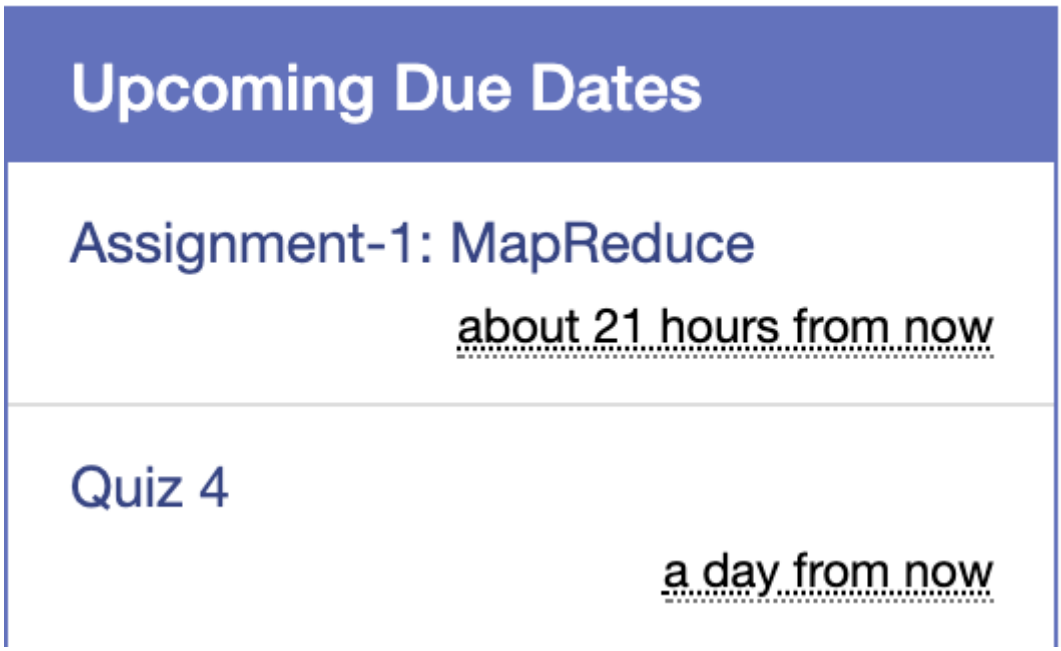
Jacob Sturges (/users/z5059632) 12 days ago (Mon Oct 28 2019 21:29:43 GMT+1100 (澳大利亚东部夏令时间))

Hey I just submitted my assignment. According to webCMS I still have several hours to submit the assignment. Though When I look at my submission give it says my submission is 4 hours late. I was just wondering which of these is correct

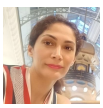
```
Session: 19T3
Assignment: assign1  Submission ID: 5059632


Current day and time:   Mon Oct 28 21:27:45 2019
Assignment deadline:    Mon Oct 28 17:59:59 2019
A submission now would be 4 hours late
Event           Day and time              Details
========================================================================
Submission      Mon Oct 28 21:26:38 2019  5059632 all assign1 [4 hours late]

Most recent submission:
-rw-r--r-- z5059632/z5059632 11701 2019-10-28 21:26 AssigOnez5059632.java
```



Reply



Maisie Badami (/users/z3500952) 11 days ago (Tue Oct 29 2019 09:41:51 GMT+1100 (澳大利亚东部夏令时间))

it was a system issue that is fixed now, would you please check again and let me know if it is still showing the deley

thanks

Maisie

Reply

Jacob Sturges (/users/z5059632) 11 days ago (Tue Oct 29 2019 09:50:40 GMT+1100 (澳大利亚东部夏令时间))

Hey it looks like the still says my submission was late. Though it says a submission now would be early.

```
Session: 19T3
Assignment: assign1  Submission ID: 5059632


Current day and time:   Tue Oct 29 09:49:14 2019
Assignment deadline:    Tue Oct 29 17:59:59 2019
A submission now would be 8 hours early
Event           Day and time              Details
================================================================================
Submission      Mon Oct 28 21:26:38 2019  5059632 all assign1 [4 hours late]

Most recent submission:
-rw-r--r-- z5059632/z5059632 11701 2019-10-28 21:26 AssigOnez5059632.java
```

Reply

Maisie Badami (/users/z3500952) 11 days ago (Tue Oct 29 2019 10:00:06 GMT+1100 (澳大利亚东部夏令时间))

maybe submit it again and see what it says, it should say it is early.

Reply

Yijia Huang (/users/z5198992) 12 days ago (Mon Oct 28 2019 20:14:17 GMT+1100 (澳大利亚东部夏令时间))

Can I use an Text format replace the Sequence format as an output format for the first map. Does it waste more or not?
Reply

Mohammadali Yaghoubzadehfard (/users/z5138589) 11 days ago (Tue Oct 29 2019 11:00:11 GMT+1100 (澳大利亚东部夏令时间))

Since I do not know what you have done in your mapper/reducer, I cannot say much. But the safe side is to use custom writables for any pair/triple/list object.

Reply

Yuehui Chu (/users/z5180907) 12 days ago (Mon Oct 28 2019 15:22:06 GMT+1100 (澳大利亚东部夏令时间))

Hi, I have an trouble with my output . I do know why there are two (m2,m3). can anyone tell me why don't this two same key merged after map?

Reply

Mohammadali Yaghoubzadehfard (/users/z5138589) 11 days ago (Tue Oct 29 2019 10:57:43 GMT+1100 (澳大利亚东部夏令时间))

Most probably, your comareto function is not working as you expect. It might be because of white spaces in strings, or using == instead of equals to compare string values.

Reply

Yuexuan Liu (/users/z5093599) 13 days ago (Sun Oct 27 2019 22:33:06 GMT+1100 (澳大利亚东部夏令时间))

Do we have to sort the movie pairs inside the parentheses? Like in the example, currently what I get is

```
(M2,M1) [(U5,3,1)]
```

Do I have sort it into

```
(M1,M2) [(U5,1,3)]
```

Or we just deal with the dataset, assume that movieid is integer?

Reply

Maisie Badami (/users/z3500952) 12 days ago (Mon Oct 28 2019 10:06:19 GMT+1100 (澳大利亚东部夏令时间)), last modified 11 days ago (Tue Oct 29 2019 10:20:46 GMT+1100 (澳大利亚东部夏令时间))

Movie IDs should be considered as string, however sorting the output is not nesesory

Reply

Sankalp Shukla (/users/z3462149) 13 days ago (Sun Oct 27 2019 20:51:31 GMT+1100 (澳大利亚东部夏令时间))

Hi,

I have implemented an ArrayWritable variable to display the list of triples (UID, Rating, Rating) in the final output. However my output is in the form:

(M2,M3)SPACE[(U2,3,1),SPACE(U5,3,3)] instead of (M2,M3)SPACE[(U2,3,1),(U5,3,3)].

I can output it in the second form if I use something like:

new Text("[" + output + "]")

According to the spec, this isn't allowed as it is a complex output. So I'm wondering whether it is acceptable to output it the first way (with spaces between array elements) or whether I can use the Text function to obtain the required output.

Cheers,

Reply

Con Tieu-Vinh (/users/z5245136) 12 days ago (Mon Oct 28 2019 09:09:13 GMT+1100 (澳大利亚东部夏令时间)), last modified 12 days ago (Mon Oct 28 2019 09:09:52 GMT+1100 (澳大利亚东部夏令时间))

Hi Sankalp,

Rather than outputting Text from your final reducer, you could still output an ArrayWritable but modify the toString() method so that it does the same thing as "[" + output + "]" without using the Text type.

Reply

Sankalp Shukla (/users/z3462149) 12 days ago (Mon Oct 28 2019 14:22:36 GMT+1100 (澳大利亚东部夏令时间))

That works! Thanks!

Reply

Roy Aranda (/users/z5075620) 13 days ago (Sun Oct 27 2019 12:41:17 GMT+1100 (澳大利亚东部夏令时间))

Has anyone run into the problem where it gets stuck on map 100% reduce 0%? Trying to see if anyone has run into this and has a solution.

Reply

Xiao Xiao (/users/z5219036) 13 days ago (Sun Oct 27 2019 21:38:42 GMT+1100 (澳大利亚东部夏令时间))

having the same issue have u come to any solutions?

Reply

Ian Commerford (/users/z3207996) 12 days ago (Mon Oct 28 2019 07:26:22 GMT+1100 (澳大利亚东部夏令时间))

It could be a limited disk space issue, try with a smaller dataset

Reply

Wanting Liu (/users/z5108679) 14 days ago (Sun Oct 27 2019 01:09:08 GMT+1100 (澳大利亚东部夏令时间)), last modified 14 days ago (Sun Oct 27 2019 01:15:44 GMT+1100 (澳大利亚东部夏令时间))

Hi,

I use following on the terminal to test my code:

```
javac -cp ".:Hadoop-Core.jar" Assig0nez5******.java
java -cp ".:Hadoop-Core.jar" Assig0nez5****** input.txt output
```

But it shows the error:

```
Error: Could not find or load main class Assig0nez5******
```

How could I fix it?

Thanks.

Reply

Chihao Chen (/users/z5185341) 14 days ago (Sun Oct 27 2019 01:18:49 GMT+1100 (澳大利亚东部夏令时间))

your class name needs to be exactly the same as you java file name, and don't forget to remove package declaration in you java file, we don't need it in this case.

Reply

Wanting Liu (/users/z5108679) 14 days ago (Sun Oct 27 2019 01:36:51 GMT+1100 (澳大利亚东部夏令时间))

It works, thank you!

Reply

Load More Comments