Discussion Forums

Week 3

Week 3

Revised Instructions for AWS Assignment

Chung Khim Lae Assignment: Graph Analysis in the Cloud · a year ago · Edited

I came across many students having trouble with AWS due to the outdated instructions in GitHub. As such, to post up-to-date instructions, I decided to repeat the assignment. (The last time I did it was 4 years ago!)

The information here should be correct as of 2016/10/01. However, I cannot guarantee the steps will not change, so you should take this post as a guide and adapt accordingly. After all, part of being a data scientist is to experiment through trial and error ☺

For this exercise, you will be using Pig to run MapReduce jobs in AWS to process about 0.5 TB of text. The details and problem statements are provided in assignment4.md with errata given at the end of this post. To minimise the time of reserving AWS machines, I advise that you prepare the Pig scripts for problems 1 to 4 beforehand and call them problem1.pig, problem2.pig, etc. Check the errata when preparing the scripts.

I will be assuming familiarity with Linux and that you already have an AWS account. For Windows users, please install Git which has a Linux terminal emulator known as Git Bash. Contrary to the instructions given in GitHub, if you are using Git Bash, doing this assignment on Windows is as easy as doing it on Linux or Mac. (In fact, I completed the whole assignment on my Android tablet!)

Here are the steps to get started.

1. Complete the opt-in quiz and contact AWS Support to request for credits.
2. Log in to AWS Management Console.
3. To minimise network traffic, at the top of the console (next to your name), change the location to Oregon since all the data are stored there.
4. Create an SSH key pair and download the key (call it

SUBFORUMS

All

Assignment: Graph Analysis in the Cloud

pig.pem) to your project folder.

5. Next, create an EMR cluster. To save on storage, you can disable logging. Choose a current generation instance type like m4.large which has one of the lowest rates available. Use 1 instance (1 master and 0 core node) at the beginning. (Actually, one is enough for problems 0 to 3.) Remember to select the key pair created in the previous step. Leave everything else as default.
6. Wait... The public DNS should appear after 5 min. Take note of the IP address given by the front part of the DNS: ec2-*aa*-*bb*-*cc*-*dd*, where *aa*, *bb*, *cc*, *dd* are 4 numbers for the IP address of the master node.
7. Add an inbound rule to the security group, ElasticMapReduce-master, to open port 22 for SSH login from anywhere.
8. Open Git Bash (for Windows) or your favourite terminal. Run the following commands to copy the Pig scripts to the master node (cd to your project folder first) and to log in (replace *aa.bb.cc.dd* below by the actual IP address of the master node):

```
scp -i pig.pem *.pig hadoop@aa.bb.cc.dd:.
ssh -i pig.pem hadoop@aa.bb.cc.dd
```

At this point, Pig may not be ready yet. (You can try typing *pig* at the command line to see if it works.) The cluster can take up to 20 min for everything to set up properly, whose status is reflected on the website automatically but not instantaneously. To run Pig interactively, simply type *pig* at the command line. The command prompt will be changed to *grunt>* for entering Pig commands. You can then copy and paste the lines from example.pig one by one to execute the script interactively. When you are done, press CTRL-D to exit from Grunt.

If you have prepared the scripts beforehand, you can continue to use the same master node for problems 1 to 3 as each script should take at most 10 min to complete using only one instance. To run a script in the background, type at the command line

```
nohup pig problem1.pig >& problem1.log &
```

The command nohup will allow Pig to continue running in the background when you are disconnected from the master node.

When your script has finished running, open another Git Bash or terminal to copy the log file from the master node to your local machine.

```
scp -i pig.pem hadoop@aa.bb.cc.dd:./problem1.log .
```

[See Appendix below for an extract of my log file for problem 1.]

From the log file, you should be able to answer the first 3 questions of problem 1. Note that this is not the file to be submitted for grading. The uploaded file, as with other parts of the assignment, should contain only *one number* to answer the question asked of each part.

For problem 2, after the script has finished, you can merge and order your results into one file with this command

```
hdfs dfs -getmerge /user/hadoop/problem2-results - | sort -go problem2.txt
```

[See Appendix below for an extract of my results for problem 2.]

Submit your answer for problem 2. Get it right before going to problem 3. Again, your submission for each part of the problem is a text file containing only *one number*.

Important: Before attempting problem 4, make sure you have answered problem 2 correctly.

While the script for problem 3 is running, start up another EMR cluster with 20 instances (1 master, 19 core) for problem 4 following the steps above. 20 is the maximum you can have unless you make a special request. You can use the same SSH key. You will be running over the full data set using the script for problem 2 with minor changes to the input data set, output results location, and perhaps the degree of parallelism. The script should take less than an hour to complete, so check your job every 20 min or so by doing

```
tail problem4.log
```

[See Appendix below for an extract of my results for problem 4.]

Warning: Be sure to go back to the EMR console to TERMINATE your EMR clusters when you are done with Pig. If not, AWS will continue charging you for reserving the instances.

Notes

1. It is not necessary to set up port forwarding or a proxy to monitor your jobs. All required information

to answer the questions can be found in your log files.

2. As of 2017/03/26, the price of launching one m4.large instance in EMR is 15¢ per hour, so a rough estimate of the total cost is 4 USD.
3. Don't forget to copy the log files and results back to your local machine before terminating your EMR clusters. Do not copy the results for problems 1 and 3 as the files are too big and not needed.
4. The command hadoop has changed to hdfs in the new version.

Feel free to leave a comment below. However, if you encounter a problem, it is better to create a new thread for your issue.

Good luck!

Errata

1. For problem 3, the subject should match '.*rdfabout\\.com.*' (2 slashes instead of 1) when running over chunk-000.
2. For problem 4, the name of the data set is 's3n://uw-cse-344-oregon.aws.amazon.com/btc-2010-chunk-*' (no backslash at the end).

⇧ 6 Upvotes Reply Follow this discussion

Earliest Top Most Recent

Timothy Dunbar · a month ago

These updated instructions were incredibly helpful, thanks Chung.

⇧ 0 Upvotes Reply

Daniel Vargas · 2 months ago

I got lost in step 8.

scp -i pig.pem *.pig hadoop@aa.bb.cc.dd:.

ssh -i pig.pem hadoop@aa.bb.cc.dd

git bash gives me an error.

⇧ 0 Upvotes Hide 1 Reply

Chung Khim Lae · 2 months ago

Hi Daniel, aa.bb.cc.dd is the actual IP address of the master node you have. Thanks for pointing this out. I'll add a note to make this clear in my initial post.

⇧ 0 Upvotes

AC

Reply

Reply

Andrés Moreira · 5 months ago

Hi everyone,

Just a comment. I have read a lot here about the problems with Amazon, and the costs.

I have just completed this assignment, using Amazon EMR with m4.large instance (0.030 USD/hour), and in total, including part 4, I used 2 hs. 15 minutes, about 1.8 USD dollars.

I got the answers OK on all the exercises. And I really liked the challenge. Some tips,

- Set up a Vagrant machine with Hadoop & Pig Installed.
- Download the test file from here http://uw-cse-344-oregon.aws.amazon.com.s3.amazonaws.com/cse344-test-file, and then 50,000 rows from http://uw-cse-344-oregon.aws.amazon.com.s3.amazonaws.com/btc-2010-chunk-000 (curl <url> | head -n 50000 > chunk000-50k)
- Build all your scripts in this Vagrant machine, and test them using "pig -x local".
- Once done, and they are working (you can check the results for some of the questions here -the ones that use the file cse344-test-file), then go to Amazon and set up the EMR cluster.
- First, create a single instance m4.large, as Chung suggest, run problem 1 to 3 here. Go to the scores page, and submit the scores early (every time you finish one). After done, SHUT DOWN YOUR INSTANCE.
- Second, create a 20 nodes m4.large cluster. Run the problem4.pig, it took 40 minutes to complete on my case, and I got the right answer.

I hope this helps others, I was afraid at the beginning because of all the comments read here, but now I'm very thankful to Chung and Bill to have put this assignment, it was challenging and easy at the end.

Andrés

⇧ 2 Upvotes Hide 1 Reply

Chung Khim Lae · 5 months ago · Edited

Thanks for sharing your experience ☺

It's good to know that reading the data locally is faster than reading it from S3, and running time is much shorter on m4.large than on m1.medium instances, which actually lowers the total cost. That is a great tip! I am going to change my instructions to use m4.large instead. Thanks!

⇧ 0 Upvotes

AC

Reply

Reply

SH

Sam Haywood · 5 months ago

Hi, I just wanted to say thanks for posting these instructions. You've made a near impossible assignment into just a difficult one and after struggling on I've managed to finish it!

Perhaps it's time to update the actual course material though? This definitely isn't the most supportive course I've participated in and as a minimum would benefit from the instructions on what is perhaps the most complicated content being up-to-date.

Thanks, SH

⇧ 2 Upvotes Reply

Matt Lewis · 6 months ago

What AWS cluster configuration should be used?

⇧ 0 Upvotes Hide 1 Reply

M

Wesley Engers Mentor · 6 months ago

Hi Matt,

See step 5 for problems 1-3: Next, create an EMR cluster. To save on storage, you can disable logging. Choose the instance type m1.medium which has the lowest rate. Use 1 instance (1 master and 0 core node) at the beginning.

(Actually, one is enough for problems 0 to 3.) Remember to select the key pair created in the previous step. Leave everything else as default.

For problem 4: While the script for problem 3 is running, start up another EMR cluster with 20 instances (1 master, 19 core) for problem 4 following the steps above.

Please read the guide carefully to ensure good results. Good luck!

0 Upvotes

AC

Reply

Reply

MF

Mathew Isabella Francis · 7 months ago

Hi,

Need help with the assignment. I have an issue with SSH connection . seeing connection timed out while trying to connect to ec2 emr instance. Also Added inbound rule to the security group ElasticMapReduce-master, to open port 22 for SSH. what could be the possible reasons for connection timed out. Please help. Thanks

0 Upvotes Hide 4 Replies

M

Wesley Engers Mentor · 7 months ago

Unfortunately, I'm not an expert on AWS. If you can't find help here I'd suggest trying to post your problem to Stack Exchange.

0 Upvotes

Chung Khim Lae · 7 months ago

What is the IP address of your master node? Try to ping it and see if it responds.

0 Upvotes

Abhilash VJ · 7 months ago

can we do this in Azure I am not able to complete the aws registration as i dont have a credit card or net banking.

0 Upvotes

Chung Khim Lae · 7 months ago

You do not need to submit your code, but you will be on your own if you choose to use another cloud provider.

⇧ 0 Upvotes

AC

Reply

Reply

GR

Gregory Ronin · 9 months ago

I got the answer for 3b correct. Then I modified the "LOAD" line in the script to be:

raw = LOAD 's3n://uw-cse-344-oregon.aws.amazon.com/btc-2010-chunk-*' USING TextLoader as (line:chararray);

After running the script, got the result, but the submission shows incorrect.

Has anyone encountered that

Thanks.

⇧ 0 Upvotes Hide 3 Replies

Chung Khim Lae · 9 months ago

Are you referring to problem 2 instead of problem 3b? Note that problem 3b is unrelated to problem 4.

⇧ 1 Upvote

GR

Gregory Ronin · 9 months ago

Yes, you are right, it should be problem 2, not 3. My mistake.

⇧ 0 Upvotes

Chung Khim Lae · 9 months ago

If you compare your results with mine, do you see any difference?

Also, check your log file for any errors.

⇧ 0 Upvotes

AC

Reply

Reply

Yanwen Chen · 10 months ago

Hi Chung,

Thank you for the updated instruction, it's very helpful! I have successfully finished problems 1-3a, but am stuck at 3b: How many records are generated by the join for the btc-2010-chunk-000 dataset?

I used the same script for problem 3a on the test file. Only changed: 1)the source file from test file to chunk-000; 2) subject matches '.*business.*' to subject matches '.*rdfabout\\.com.*'; 3) subject=subject2 to object=subject2.

I got right answer for test but wrong for chunk-000. I noticed in the suggested steps from the instructor, the last step is: Remove duplicate tuples from the result of the join. I tried to do that, but there's no duplicates in my result. Does this sound right to you? Other than that I really don't know what went wrong.

Any help would be appreciated!

⇧ 0 Upvotes Hide 5 Replies

Chung Khim Lae · 10 months ago

For 3B, there are duplicates after the join which need to be removed by using DISTINCT. Hope it helps!

⇧ 0 Upvotes

Yanwen Chen · 10 months ago

Thank you for the quick response! I did try using DISTINCT, but the result has the same number of records as the joined data. That's way I thought there was no duplicate in my result.

I'm also trying to think why would there be duplicates? I'm assuming the original file does not have duplicate since it's basically describing a graph of the web. If there's no duplicate in the original file, why would there be duplicates in the joined data?

Thanks again for your help.

⇧ 0 Upvotes

Chung Khim Lae · 10 months ago · Edited

Hmm, I did get duplicates after the join, but it could be the way I set up the script is different from yours. I only use DISTINCT at the very end.

By the way, are you sure all the jobs were run successfully? Can you check your log file for failed jobs?

⇧ 0 Upvotes

BT

Bala Subrahmanyam Tubati · 10 months ago

Hi Chen/Lea,

Even I am stuck with 3B. I am able to provide correct answers to rest of the questions. Please provide more clear instructions for Problem 3B.

Thanks and regards,

Bala

⇧ 0 Upvotes

Chung Khim Lae · 10 months ago

Hi Bala,

I'm not sure the cause of your problem, but if you check your log file, do you see any errors?

You may want to check the Pig documentation for hints on using join or distinct in a Pig script. Make sure your subject matches '.*rdfabout\\.com.*' for part 3B.

Hope it helps!

⇧ 0 Upvotes

AC

Reply

Reply

AS

Angelo Sebastianelli · 10 months ago

a dummy question

with " How many records are there in count_by_object?"

do you mean the number of "rows" or the sum of the count generated in the pig script?

⇧ 0 Upvotes Hide 10 Replies

Chung Khim Lae · 10 months ago

The number of records is the number of rows or lines in the output file after merging all the results together. This number is also indicated in the summary table near the end of the log file. Have a look at the Appendix to see an example of my log file. Hope it helps!

⇧ 0 Upvotes

AS

Angelo Sebastianelli · 10 months ago

ok. i am at the problem 2

am i wrong or this requires just another group? from the previous script?

⇧ 0 Upvotes

Chung Khim Lae · 10 months ago

Sorry, I don't understand your question. What do you mean by "group"?

You will need to modify example.pig to solve Problem 2, but the steps to run the script remain the same.

⇧ 0 Upvotes

AS

Angelo Sebastianelli · 10 months ago

i did add a second foreach ...group and i got the same result as the "debug".

⇧ 0 Upvotes

AS

Angelo Sebastianelli · 10 months ago

what is not clear to me is what we need to turn in..

the answer to the question or the result of the script?

I mean for problem 2 you submited the first 5 points.. but the question is how many.. i am a bit confused

⇧ 0 Upvotes

Chung Khim Lae · 10 months ago · Edited

Sorry to confuse you with my sample output. What I have shown is the first 5 lines of my final result (or the first 5 bins of the histogram). Of course, I can't show the whole output, else you know how many lines there are and hence the answer to problem 2 ☺

My intention is for you to check your result with mine, so if you get the same numbers for the first 5 lines, you can be sure that your script is working correctly and can go ahead with problem 4.

Let me know if you still have doubts. Thanks!

⇧ 0 Upvotes

AS

Angelo Sebastianelli · 10 months ago

thanks. i completed it, but it would be nice to be a bit more clear. the lectures are quite interesting..

⇧ 0 Upvotes

AS

Angelo Sebastianelli · 10 months ago

probably a quiz would be easier to answer

⇧ 0 Upvotes

Chung Khim Lae · 10 months ago

Yes, I agree. Submitting a text file with a number in it is an overkill. Would you like to use the flag icon at the bottom of the assignment page to feedback to the teaching staff?

I will update my instructions to be clear on the submission. Thanks!

⇧ 0 Upvotes

Chung Khim Lae · 10 months ago

Hi Angelo, just curious, how long does it take for you to complete problem 4? What instance type did you use?

⇧ 0 Upvotes

AC

Reply

Reply

RR

Ricardo Ríos · 10 months ago

Thanks for your indications, I have three questions when I run the following:

scp -i pig.pem *.pig hadoop@aa.bb.cc.dd:.

I can't get the files, is there other way to access those files?

If my cluster is running, I am being billed?

If my cluster is terminated, I am not being billed?

I really appreciate any help or advice. Thanks in advance.

⇧ 0 Upvotes Hide 2 Replies

Chung Khim Lae · 10 months ago

scp is a Linux command to copy files from your local machine to a remote machine or vice versa. Which files are you referring to?

You will be billed as long as your cluster is **not** terminated.

⇧ 0 Upvotes

RR

Ricardo Ríos · 10 months ago

Thanks for your response Chung , I have understood this step.

⇧ 0 Upvotes

AC

Reply

Reply

Chung Khim Lae · 10 months ago · Edited

Appendix

Here is the end of my log file for problem 1 (edited to hide the answer):

```
VertexId Parallelism TotalTasks   InputRecords
ReduceInputRecords  OutputRecords  FileBytesRead
FileBytesWritten  HdfsBytesRead HdfsBytesWritten
Alias    Feature Outputs
scope-19          33         33       10000000
                 0       10000000          79728
        59815300            0               0
count_by_object,ntriples,objects,raw
scope-20          50         50              0
           3256733        1627294       43329535
        68791529            0               0
count_by_object,count_by_object_ordered
GROUP_BY,SAMPLER
scope-29           1          1              0
              5000              1          10994
           10064            0               0

scope-39          50         50        1622313
                 0        X       28863196
  42134163            0               0
count_by_object_ordered
scope-41          50         50              0
           X        X       53577335
37315412            0         89971068
ORDER_BY    /user/hadoop/problem1-results,

Input(s):
Successfully read 10000000 records from:
"s3n://uw-cse-344-oregon.aws.amazon.com/btc-2010-
chunk-000"

Output(s):
Successfully stored X records (89971068 bytes)
in: "/user/hadoop/problem1-results"

187796 [main] INFO  org.apache.pig.Main  - Pig
script completed in 3 minutes, 8 seconds and 330
milliseconds (188330 ms)
16/09/30 01:29:06 INFO pig.Main: Pig script
completed in 3 minutes, 8 seconds and 330
milliseconds (188330 ms)
187796 [main] INFO
org.apache.pig.backend.hadoop.executionengine.tez
.TezLauncher  - Shutting down thread pool
16/09/30 01:29:06 INFO tez.TezLauncher: Shutting
down thread pool
187852 [Thread-19] INFO
org.apache.pig.backend.hadoop.executionengine.tez
.TezSessionManager  - Shutting down Tez session
org.apache.tez.client.TezClient@3e2351b2
16/09/30 01:29:06 INFO tez.TezSessionManager:
Shutting down Tez session
org.apache.tez.client.TezClient@3e2351b2
16/09/30 01:29:06 INFO client.TezClient: Shutting
down Tez Session,
sessionName=PigLatin:problem1.pig,
applicationId=application_1475197660777_0001
```

Here are the first 5 lines of my results for problem 2 over chunk-000:

```
1 20430
2 21865
3 77726
4 32635
5 82351
```

Here are the first 5 lines of my results for problem 4:

```
  1   1   8950222
  2   2   10290572
  3   3   49171908
  4   4   11692376
  5   5   4945471
```

Note: These are not the files expected by the grader. They are shown here so you can check your results with mine.

⇧ 3 Upvotes Reply

‹ 1 ›

AC

Reply

Reply