

Hadoop Hive Lab

The goal of this lab is to gain familiarity with Hadoop and Hive. Additional documentation that is useful for this homework is available at: <https://hive.apache.org/>.

Access to Hadoop

- 1) Option #1: Amazon Elastic MapReduce (Amazon EMR) is a web service that makes it easy to quickly and cost-effectively process vast amounts of data. Amazon EMR uses Hadoop, an open source framework, to distribute your data and processing across a resizable cluster of Amazon EC2 instances. Please register an Amazon AWS account using your IUPUI email and request student credit. You should be able to receive \$40 credit for your account. **Please remember to terminate your cluster when you finish using it to avoid unnecessary charges.**
- 2) Option #2: Download and install on your desktop or laptop Hortonworks HDP Sandbox, a self-contained VM image pre-loaded with all the needed Hadoop software.
- 3) Option #3: Download and install on your desktop or laptop Cloudera QuickStart Virtual Machines.

If you choose Option #2 or Option #3, the machine you install Hadoop virtual machine is usually required to have at least 8G-10G of RAM.

Exercise : Calls and Messages Data Analysis

In this homework, a calls.csv and a messages.csv are given. The calls.csv file has the following fields:

1	callernumber	phone number of the caller
2	receivernumber	phone number of the receiver
3	duration	the duration of the call
4	year	the year when the call was made
5	month	the month when the call was made
6	day	the day when the call was made
7	hour	the hour when the call was made
8	minute	the minute when the call started

The messages.csv file has the following fields:

1	callernumber	phone # of the person who sends the message
2	receivernumber	phone # of the person who receives the msg
3	length	the length of the message
4	year	the year when the message was sent
5	month	the month when the message was sent
6	day	the day when the message was sent
7	hour	the hour when the message was sent
8	minute	the minute when the message was sent

Based on these files, write HIVE scripts for the following queries:

1. Find the number of distinct callernumbers in calls.csv.
2. Find callernumbers which have made more than 10 calls.
3. Find callernumbers which have called for the longest period of time (the sum of the duration of the calls.)
4. Find the receivernumber which has received the most calls.
5. Find the top 10 longest messages.
6. Find the peak messaging hour, i.e. the hour when most messages were sent.
7. Find the average duration of calls of each hour.
8. Identify which numbers made calls as well as sent messages, where number of calls made should be more than 3 and the number of messages should be more than 2.

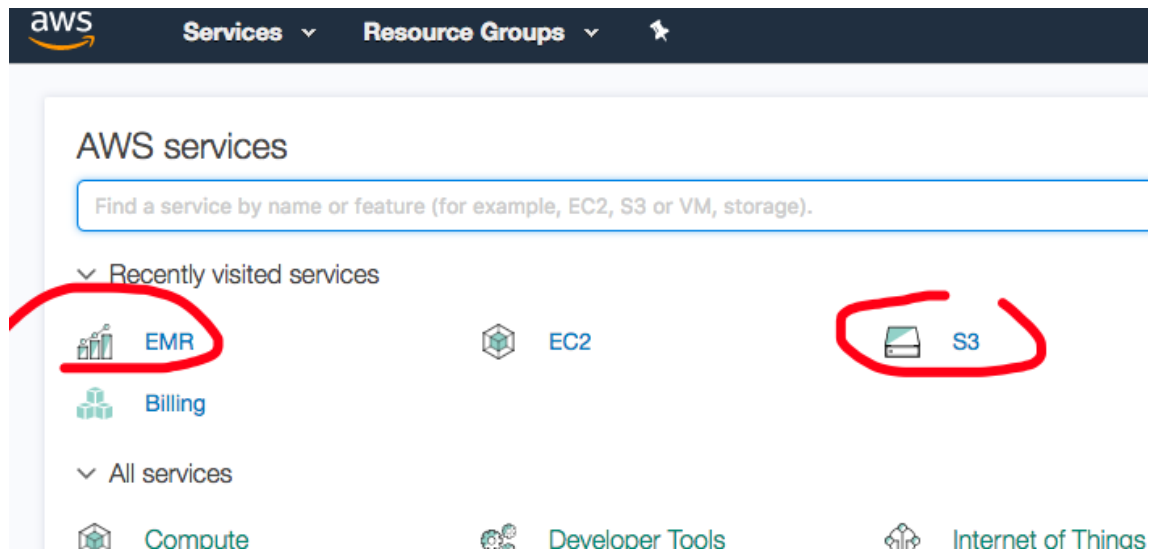
This assignment should be completed independently. No collaboration is allowed.

Please submit your HIVE scripts for all queries. Your scripts should be clearly written and ready for the TA to run.

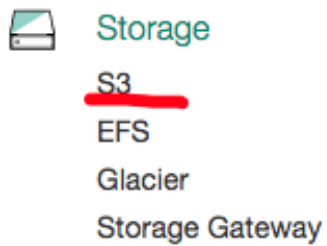
More Information:

If you choose to use Amazon AWS, the very first step you should apply for Amazon EMR account at: <https://aws.amazon.com/emr/>

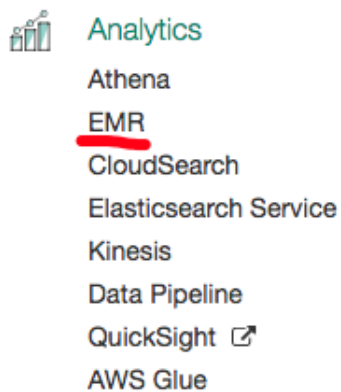
Once you have your account and enter the AWS main page, the service you need to use for this homework is S3 and EMR.



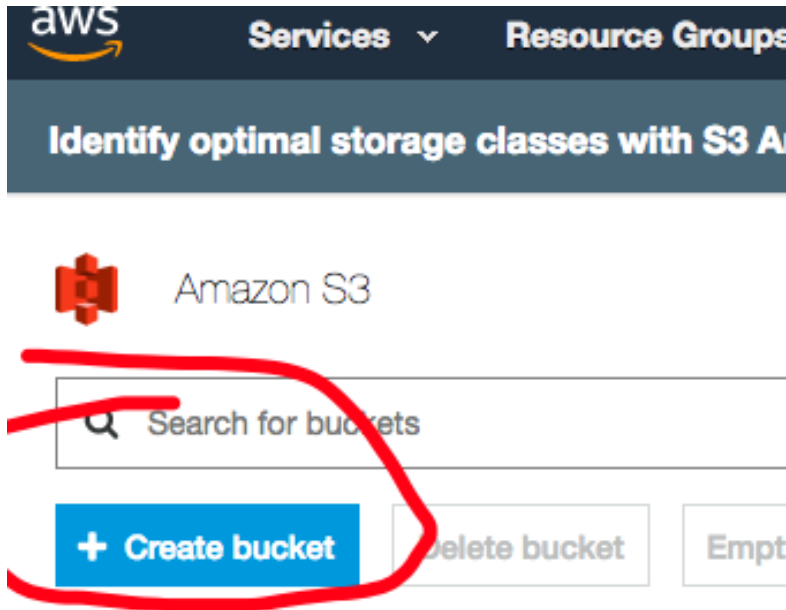
S3 is the storage system in AWS, it is under storage service:



Amazon EMR is a managed cluster platform that simplifies running big data frameworks, such as Apache Hadoop. It is under “Analytics”:



If you go to S3, just click the “create bucket” button:



Type a bucket name you want and then keep clicking “next” until finished.

Create bucket

1

Name and region

2

Set properties

3

Set permissions

4

Review

Name and region

Bucket name ⓘ

derenli700v2

Region

US East (Ohio)

Copy settings from an existing bucket

Select bucket (optional)

1 Buckets

Create

Cancel

Next

Once you create the bucket, click on the bucket and you can create a folder:

Amazon S3 > derenli700

Overview Properties Permissions Management

Q Type a prefix and press Enter to search. Press ESC to clear.

Upload + Create folder More ▾

<input type="checkbox"/>	Name ↑ ▾	Last modified ↑ ▾
<input type="checkbox"/>	Assignment1	--
<input type="checkbox"/>	elasticmapreduce	--

You can then upload all your data and script files here:

Amazon S3 > derenli700 / Assignment1

Overview

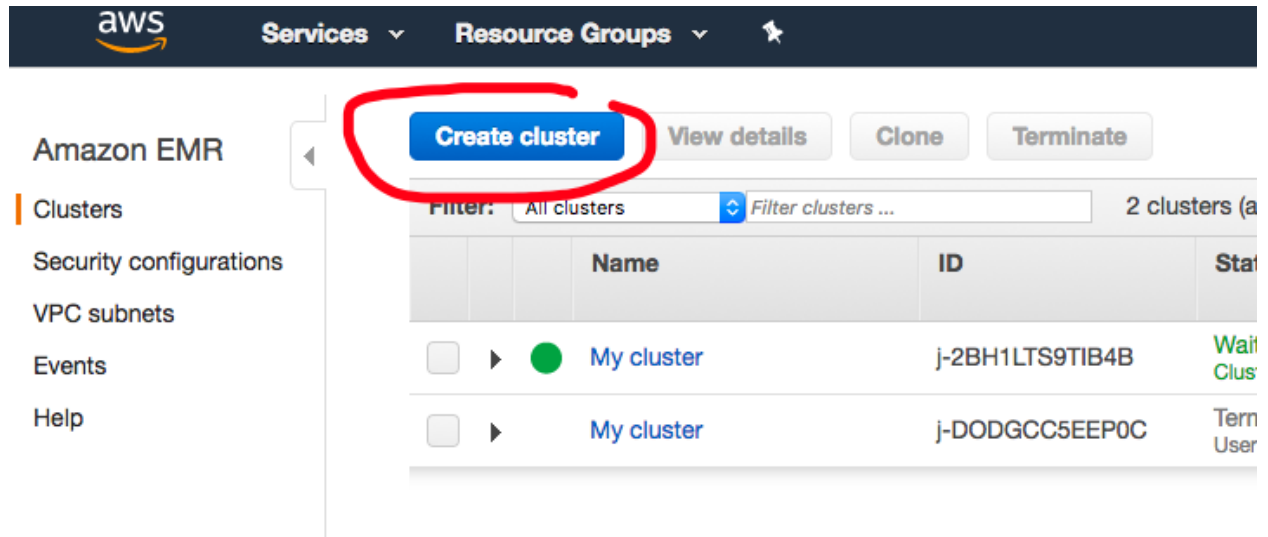
Q Type a prefix and press Enter to search. Press ESC to clear.

Upload + Create folder More ▾

<input type="checkbox"/>	Name ↑ ▾
<input type="checkbox"/>	airports.csv
<input type="checkbox"/>	carriers.csv
<input type="checkbox"/>	exercise1.txt
<input type="checkbox"/>	exercise2.csv

After uploading the data files and scripts, you can go to EMR to run the job.

In EMR, click on “create a cluster”:



Once you click “create cluster”, you will see:

General Configuration

Cluster name

☒ **Logging** ⓘ

S3 folder ⓘ

Launch mode ☒ **Cluster** ⓘ ☐ **Step execution** ⓘ

After you create the cluster, you may still need to wait for a while to get the status to waiting, which means your cluster is ready for use:

Clone
Terminate
AWS CLI export

Cluster: My cluster **Waiting** Cluster ready after last step completed.

Summary
Application history
Monitoring
Hardware
Events
Steps
Configurations
Bootstrap actions

Connections: Enable Web Connection – Hue, Ganglia, Resource Manager ... (View All)
Master public DNS: ec2-18-221-180-149.us-east-2.compute.amazonaws.com SSH
Tags: -- View All / Edit

Summary	Configuration details
ID: j-2BH1LTS9TIB4B Creation date: 2017-09-28 16:48 (UTC-4) Elapsed time: 38 minutes Auto-terminate: No Termination protection: Off <a>Change	Release label: emr-5.8.0 Hadoop distribution: Amazon 2.7.3 Applications: Ganglia 3.7.2, Hive 2.3.0, Hue 3.12.0, Mahout 0.13.0, Pig 0.16.0, Tez 0.8.4 Log URI: s3://derenli700/elasticmapreduce/ EMRFS consistent view: Disabled Custom AMI ID: --
Network and hardware	Security and access
Availability zone: us-east-2a Subnet ID: subnet-e66a55e2	Key name: -- EC2 instance profile: EMP-EC2-DefaultRole

Then click on step panel and click “add step”:

Clone
Terminate
AWS CLI export

Cluster: My cluster **Waiting** Cluster ready after last step completed.

Summary
Application history
Monitoring
Hardware
Events
Steps
Configurations
Bootstrap actions

Add step
Clone step
Cancel step

View all interactive jobs

Filter:	All steps	Filter steps ...	1 step (all loaded)		
ID	Name	Status	Start time (UTC-4)	Elapsed time	Log files
s-1AK7EIFCVSH59	Setup hadoop debugging	Completed	2017-09-28 16:55 (UTC-4)	2 seconds	<a>View logs

Then choose “Hive program” and choose your script file, input location and output location, then click add. You job is now running. (You will need upload your script and datasets on S3 server first.)

Add step

Step type

Hive program

Name

Hive program

Script S3 location*

s3://

S3 location of your Hive script.

Input S3 location

s3://

S3 location of your Hive input files.

Output S3 location

s3://

S3 location of your Hive output files.

Arguments

Specify optional arguments for your script.

Action on failure

Continue

What to do if the step fails.

Cancel

Add

You can always track your status under event panel. Once you terminate your cluster, you may not be able to “restart” it again. However, as all your files are on S3 server, so next time when you need, simply start a new cluster again.

You can find lots of useful information here:

<http://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-manage.html>