

GEORGIA INSTITUTE OF TECHNOLOGY
SCHOOL of ELECTRICAL and COMPUTER ENGINEERING

ECE 4150-A Spring 2021

Lab: Batch Data Analysis using Hadoop, MapReduce, Pig & Hive

References:

- [1] A. Bahga, V. Madiseti, *Cloud Computing Solutions Architect: A Hands-On Approach*, ISBN: 978-0996025591
- [2] <https://pythonhosted.org/mrjob/>
- [3] <http://hadoop.apache.org/>
- [4] <http://storage.googleapis.com/books/ngrams/books/datasetsv2.html>
- [5] <http://pig.apache.org/docs/r0.15.0/basic.html>
- [6] <https://cwiki.apache.org/confluence/display/Hive/LanguageManual>

Due Date:

The lab report will be **due on April 5th, 2021 at 11:59 PM.**

In this lab you will learn how setup a Hadoop cluster and run MapReduce, Pig and Hive job.

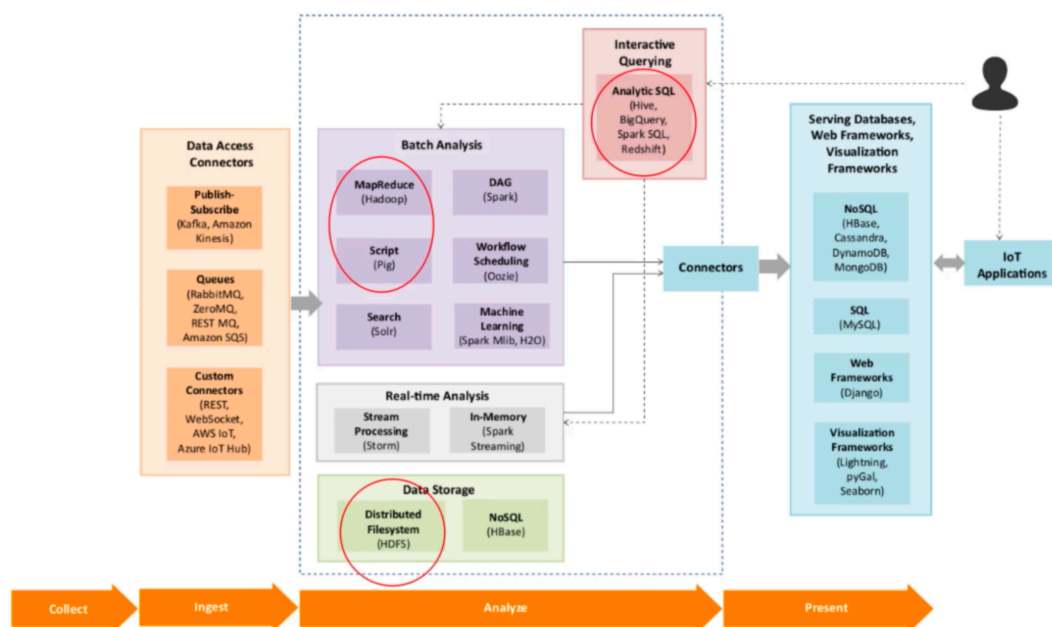


Fig.1 Architecture diagram of data processing in Hadoop

1. Set up a Hadoop Cluster with EMR

Navigate to Amazon EMR console and create a new cluster with the following configurations:

General Configuration

Cluster name

☒ Logging ⓘ

S3 folder ⓘ

Launch mode ☒ Cluster ⓘ ☐ Step execution ⓘ

Software configuration

Release ⓘ

Applications

- ☒ Core Hadoop: Hadoop 2.7.3 with Ganglia 3.7.2, Hive 1.0.0, Hue 3.7.1, Mahout 0.12.2, and Pig 0.14.0
- ☐ HBase: HBase 1.2.2 with Ganglia 3.7.2, Hadoop 2.7.3, Hive 1.0.0, Hue 3.7.1, Phoenix 4.7.0, and ZooKeeper 3.4.9
- ☐ Presto-Sandbox: Presto 0.157.1 with Hadoop 2.7.3 HDFS and Hive 1.0.0 Metastore
- ☐ Spark: Spark 1.6.3 on Hadoop 2.7.3 YARN with Ganglia 3.7.2

Hardware configuration

Instance type ⓘ The selected instance type adds 32 GiB of GP2 EBS storage per instance by default. [Learn more](#) ⓘ

Number of instances (1 master and 0 core nodes)

Security and access

EC2 key pair ⓘ [Learn how to create an EC2 key pair.](#)

Permissions ☒ Default ☐ Custom

Use default IAM roles. If roles are not present, they will be automatically created for you with managed policies for automatic policy updates.

EMR role [EMR_DefaultRole](#) ⓘ

EC2 Instance profile [EMR_EC2_DefaultRole](#) ⓘ

Wait for the cluster to be created and enter the state of "Waiting", which usually takes 5 minutes to finish.

Amazon EMR

EMR on EC2

Clusters

Notebooks

Git repositories

Security configurations

Block public access

VPC subnets

Events

EMR on EKS

Virtual clusters

Help

What's new

Clone Terminate AWS CLI export

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

Summary Application user interfaces Monitoring Hardware Configurations Events Steps Bootstrap actions

Summary

ID: j-1O1Y9EG2H8J00

Creation date: 2021-03-21 19:26 (UTC-4)

Elapsed time: 31 minutes

After last step completes: Cluster waits

Termination protection: Off [Change](#)

Tags: -- [View All / Edit](#)

Master public DNS: ec2-18-207-109-235.compute-1.amazonaws.com [Connect to the Master Node Using SSH](#)

Configuration details

Release label: emr-4.9.6

Hadoop distribution: Amazon 2.7.3

Applications: Ganglia 3.7.2, Hive 1.0.0, Hue 3.7.1, Mahout 0.12.2, Pig 0.14.0

Log URI: s3://aws-logs-610172127508-us-east-1/elasticmapreduce/ [View](#)

EMRFS consistent view: Disabled

Application user interfaces

On-cluster user interfaces [Not Enabled](#) [Enable an SSH Connection](#)

Network and hardware

Availability zone: us-east-1b

Subnet ID: [subnet-85b38de2](#)

Master: **Running** 1 m4.2xlarge

Core: --

Task: --

Cluster scaling: Not enabled

Security and access

Key name: ece4150

EC2 instance profile: EMR_EC2_DefaultRole

EMR role: EMR_DefaultRole

Visible to all users: All [Change](#)

Security groups for Master: [sg-0e8a1ea10df547022](#) [View](#) (ElasticMapReduce-master)

Security groups for Core & Task: [sg-046614150375a4266](#) [View](#) (ElasticMapReduce-slave)

Navigate to **EC2 instance** and open the one that's running, which holds the cluster that you just created:

EC2 > Instances > i-08d587eff9e14d974

Instance summary for i-08d587eff9e14d974 [Info](#)

Updated less than a minute ago

[Refresh](#) [Connect](#) [Instance state](#)

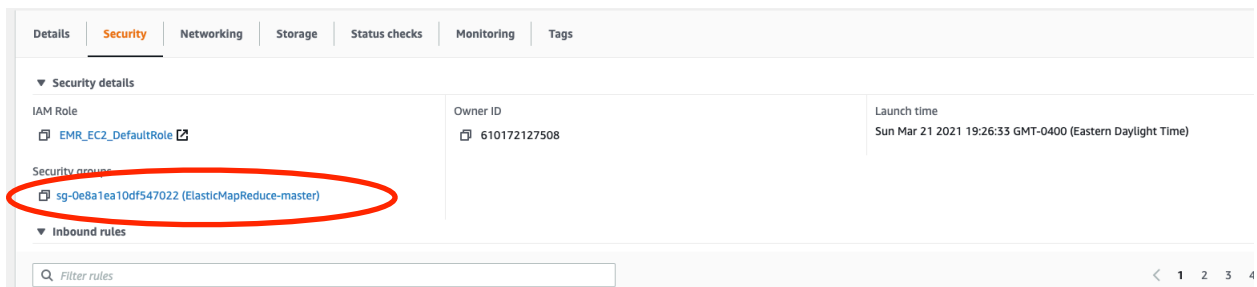
Instance ID i-08d587eff9e14d974	Public IPv4 address 18.207.109.235 open address	Private IPv4 addresses 172.31.5.214
Instance state Running	Public IPv4 DNS ec2-18-207-109-235.compute-1.amazonaws.com open address	Private IPv4 DNS ip-172-31-5-214.ec2.internal
Instance type m4.2xlarge	Elastic IP addresses -	VPC ID vpc-58cd8722
AWS Compute Optimizer finding Opt-in to AWS Compute Optimizer for recommendations. Learn more	IAM Role EMR_EC2_DefaultRole	Subnet ID subnet-85b38de2

Details | Security | Networking | Storage | Status checks | Monitoring | Tags

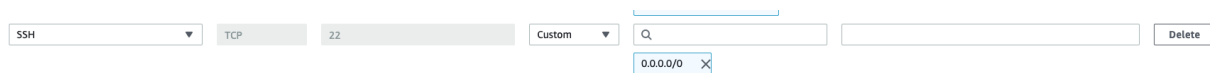
Instance details [Info](#)

Platform Linux/UNIX (Inferred)	AMI ID ami-0a275fa45e26f86ef	Monitoring disabled
Platform details Linux/UNIX	AMI name Amazon Elastic MapReduce 2019-11-23-00-32-15 hvm/ebs - 4.9.6	Termination protection Disabled

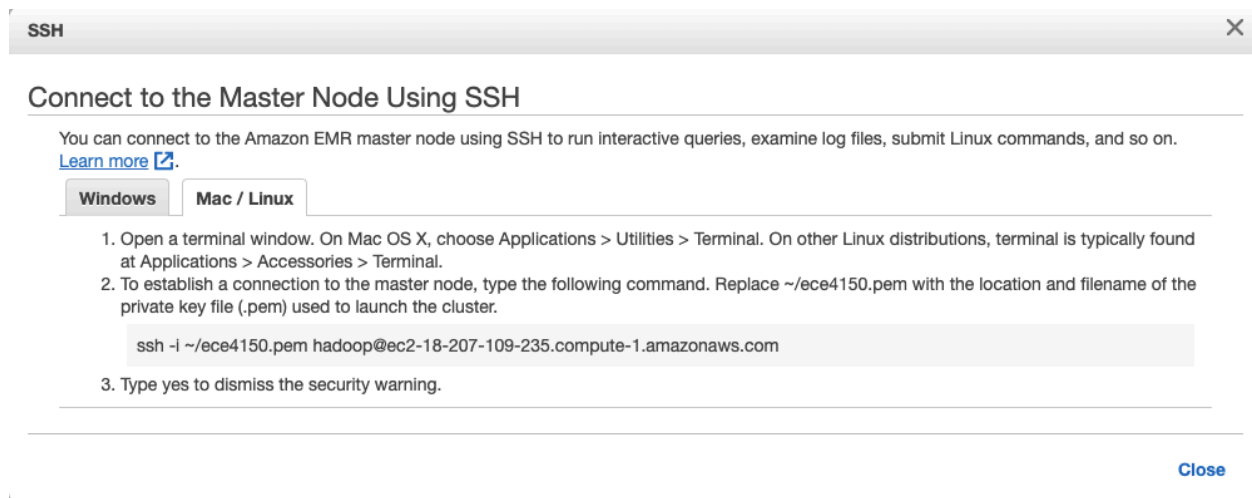
Go to **Security-Security groups** and open the security group for the master cluster, in this case **ElasticMapReduce-master**:



Edit the inbound rules and add a rule for SSH and save it:



Now go back to your **EMR cluster**, click on “**connect to the master node using SSH**” and follow the instruction to connect to the master node:



For example, a successful connection would appear like:

```
(base) [19:59] (/>w<)/ ~/desktop $ chmod 400 ece4150.pem
(base) [19:59] (/>w<)/ ~/desktop $ ssh -i ece4150.pem hadoop@ec2-18-207-109-235.compute-1.amazonaws.com
Last login: Sun Mar 21 23:54:02 2021

  _-|  _-|_ )
  _| (  _/
 _-|\___|___|
                Amazon Linux AMI

https://aws.amazon.com/amazon-linux-ami/2017.03-release-notes/
Amazon Linux version 2018.03 is available.

EEEEEEEEEEEEEEEEEEEE MMMMMMM      MMMMMMM RRRRRRRRRRRRRRR
E:::EE:::EE:::EE::: M:::EE::: M:::EE::: R:::EE:::EE:::EE:::R
EE:::EE:::EE:::EE::: M:::EE::: M:::EE::: R:::EE:::RRRRRR:::R
E:::E      EEEEE M:::EE::: M:::EE::: RR:::R      R:::R
E:::E      M:::EE::: M:::EE::: M:::EE::: R:::R      R:::R
E:::EE:::EE:::EE::: M:::EE::: M:::EE::: M:::EE::: R:::RRRRRR:::R
E:::EE:::EE:::EE::: M:::EE::: M:::EE::: M:::EE::: R:::EE:::EE:::RR
E:::EE:::EE:::EE::: M:::EE::: M:::EE::: M:::EE::: R:::RRRRRR:::R
E:::E      M:::EE::: M:::EE::: M:::EE::: R:::R      R:::R
E:::E      EEEEE M:::EE::: M:::EE::: M:::EE::: R:::R      R:::R
EE:::EE:::EE:::EE::: M:::EE::: M:::EE::: M:::EE::: R:::R      R:::R
E:::EE:::EE:::EE::: M:::EE::: M:::EE::: M:::EE::: RR:::R      R:::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM      MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-5-214 ~]$
```

2. Upload Datasets to HDFS

First, we need to enable SSH tunnel in the browser. Navigate to EMR cluster and click on **“Enable an SSH Connection”** in Application user interface:

Clone

Terminate

AWS CLI export

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

Summary

Application user interfaces

Monitoring

Hardware

Configurations

Events

Steps

Bootstrap actions

Summary

ID: j-1O1Y9EG2H8J00

Creation date: 2021-03-21 19:26 (UTC-4)

Elapsed time: 4 hours, 9 minutes

After last step completes: Cluster waits

Termination protection: Off [Change](#)

Tags: -- [View All / Edit](#)

Master public DNS: ec2-18-207-109-235.compute-1.amazonaws.com [Connect to the Master Node Using SSH](#)

Configuration details

Release label: emr-4.9.6

Hadoop distribution: Amazon 2.7.3

Applications: Ganglia 3.7.2, Hive 1.0.0, Hue 3.7.1, Mahout 0.12.2, Pig 0.14.0

Log URI: s3://aws-logs-610172127508-us-east-1/elasticmapreduce/

EMRFS consistent view: Disabled

Application user interfaces

On-cluster user Not Enabled

Interfaces

[Enable an SSH Connection](#)

Network and hardware

Availability zone: us-east-1b

Subnet ID: [subnet-85b38de2](#)

Master: Running 1 m4.2xlarge

Core: --

Task: --

Cluster scaling: Not enabled

Security and access

Key name: ece4150

EC2 Instance profile: EMR_EC2_DefaultRole

EMR role: EMR_DefaultRole

Visible to all users: All [Change](#)

Security groups for Master: [sg-0e8a1ea10df547022](#) (ElasticMapReduce-master)

Security groups for Core & Task: [sg-046614150375a4266](#) (ElasticMapReduce-slave)

First, follow the instructions to enable an SSH tunnel to the EMR Master Node:

Enable an SSH Connection

Enable an SSH Connection

EMR applications publish user interfaces as web sites hosted on the master node. For security reasons, these web sites are only available on the master node's local web server.

To reach the web interfaces, you must establish an SSH tunnel with the master node using either dynamic or local port forwarding. If you use dynamic port forwarding, you must also configure a proxy server to view the web interfaces.

Step 1: Open an SSH Tunnel to the Amazon EMR Master Node - [Learn more](#)

Windows

Mac / Linux


1. Open a terminal window. On Mac OS X, choose Applications > Utilities > Terminal. On other Linux distributions, terminal is typically found at Applications > Accessories > Terminal.
2. To establish an SSH tunnel with the master node using dynamic port forwarding, type the following command. Replace ~/ece4150.pem with the location and filename of the private key file (.pem) used to launch the cluster.

```
ssh -i ~/ece4150.pem -ND 8157 hadoop@ec2-18-207-109-235.compute-1.amazonaws.com
```

Note: Port 8157 used in the command is a randomly selected, unused local port.

3. Type yes to dismiss the security warning.

Then, install FoxyProxy in Chrome. It's worth noticing that the url provided in the instruction doesn't work, so please manually install it as an extension on you browser. Take Chrome as an example: go to **chrome store**, search for "**foxyproxy basic**", install and add to chrome, **restart** chrome after installing. Create a file named foxyproxy-settings.xml as suggested and import it to your FoxyProxy. At the top of your FoxyProxy page, choose "use proxy emr-socks-proxy for all URLs" (not the same as AWS instruction!!)



Proxies

- Global Settings
- Import/Export
- About

Proxy mode: Use proxy emr-socks-proxy for all URLs

Proxies

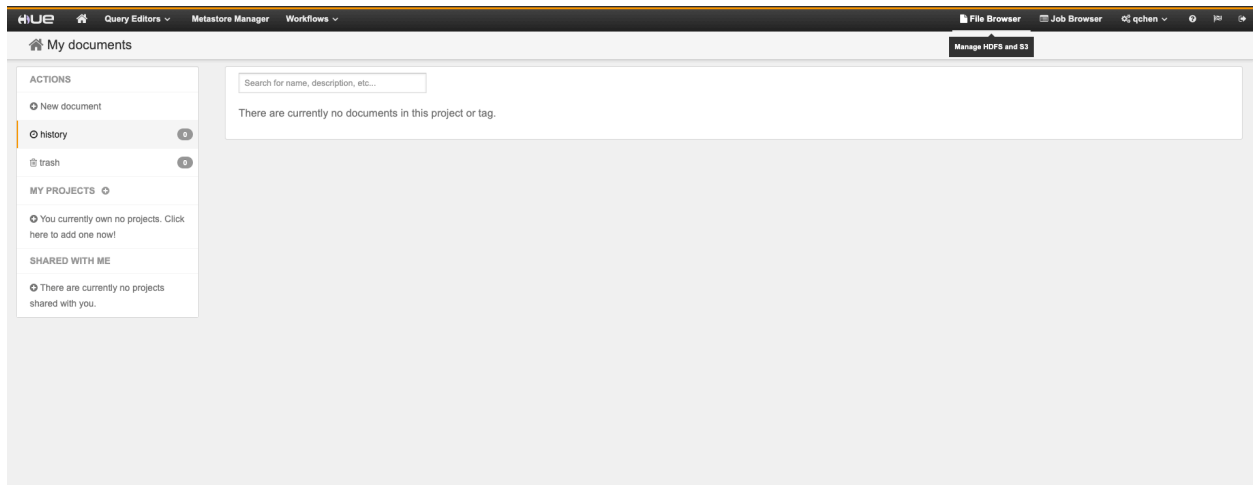
Enabled	Color	Proxy Name	Proxy Notes	Host or IP Address	Port	SOCKS proxy?	SOCKS Version	Auto PAC URL	
✓	Blue	emr-socks-proxy		localhost	8157	✓	5		Move Up Move Down
✓	Blue	Default	These are the settings that are used when no patterns match an URL				5		Add New Proxy Edit Selection Copy Selection Delete Selection

[Import your proxies from FoxyProxy on Mozilla Firefox or from another computer.](#)

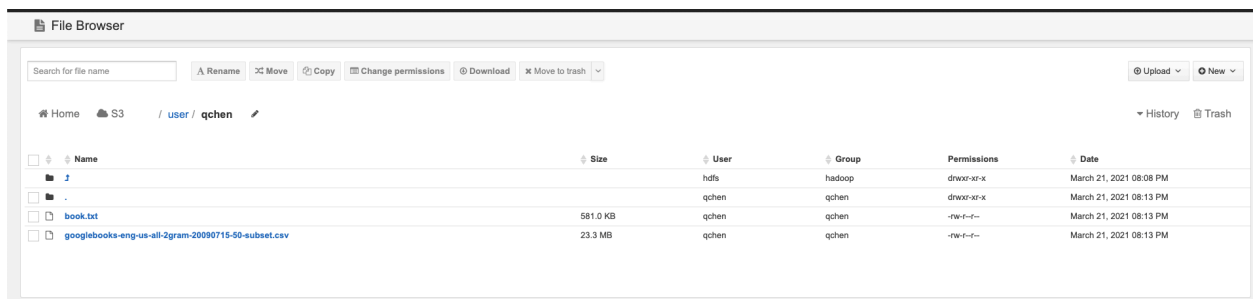
[Please Donate](#) [Buy Proxy Service](#)

Go back to the EMR cluster and navigate to **Application user interfaces**, copy the url of Hue and open it in the browser in which you just installed the FoxyProxy, create a new account in Hue and **save the username**.

Follow the default setting in quick start and choose **Hue Home** in step 4 to enter the home page of Hue:



Upload the provided dataset to File Browser in the following structure (don't upload the zip file or folder, use only separate files)



Next, create the mrjob configuration file on the Hadoop master node with vim using the provided mrjob.conf, save and exit:

```

runners:
  hadoop:
    hadoop_home: /usr/lib/hadoop
    hadoop_streaming_jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.jar

```

"mrjob.conf" 5L, 136C 5,0-1 All

Create wordcount-mr.py on Hadoop master node with vim using the provided wordcount-mr.py, save and exit:

```

"""
python wordcount-mr.py -r hadoop hdfs:///user/clouduser/book.txt --output-dir=hdfs:///user/clouduser/wordcountoutput --conf-path=mrjob.conf
"""

from mrjob.job import MRJob

class MRmyjob(MRJob):
    def mapper(self, _, line):
        wordlist = line.split()
        for word in wordlist:
            yield word,1

    def reducer(self, key, list_of_values):
        yield key,sum(list_of_values)

if __name__ == '__main__':
    MRmyjob.run()

```

"wordcount-mr.py" 17L, 414C 17,5 All

3. Setup mrJob on Master Node

In the Hadoop instance, run the following commands to set up mrJob:

```

sudo yum install python-pip
sudo pip install mrjob

```

4. Run MapReduce Program

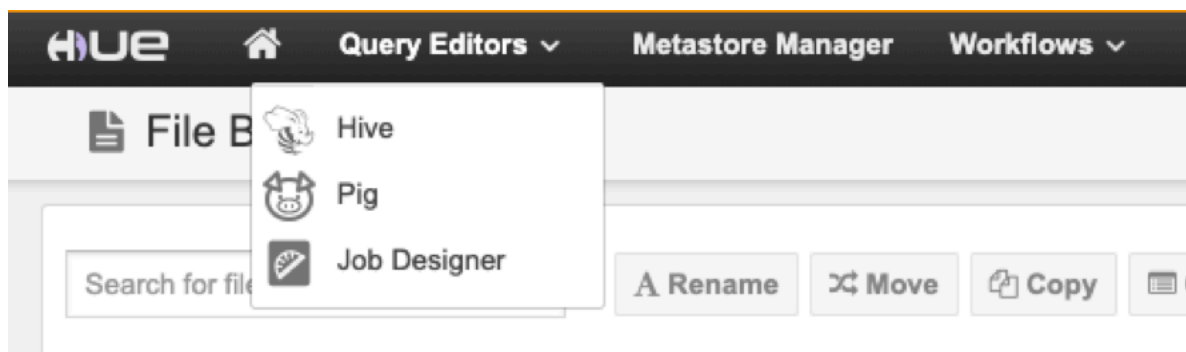
In the Hadoop instance, paste the command line from wordcount-mr.py and change the clouduser to your username. After the program finishes execution, you can view the results in the file browser:

Home S3 / user / qchen

Name	Size	
.		hd
.		qc
book.txt	581.0 KB	qc
googlebooks-eng-us-all-2gram-20090715-50-subset.csv	23.3 MB	qc
wordcountoutput		ha

5. Run Pig Program

Find the provided code in wordcount-pig.txt, navigate to pig editor in Hue interface:



Paste the code in script, change "clouduser" to your own username and run:



After the job is finished, you can view the output in file browser:

Unsaved script

Progress: 100% Status: OK

Job Stats (time in seconds):

JobId	Maps	Reduces	MaxMapTime	MinMapTime	AvgMapTime	MedianMapTime	MaxReduceTime	MinReduceTime	AvgReduceTime	MedianReduceTime	Alias	Feature	Outputs
job_1616369296561_0005	1	1	4	4	4	4	4	4	a,b,c,d	GROUP_BY, COMBINER	/user/qchen/pig_wordcount,		

Input(s):
Successfully read 13052 records (595320 bytes) from: "/user/qchen/book.txt"

Output(s):
Successfully stored 12272 records (126320 bytes) in: "/user/qchen/pig_wordcount"

Counters:
Total records written : 12272
Total bytes written : 126320
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_1616369296561_0005

2021-03-22 03:36:12,081 [uber-SubtaskRunner] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at ip-172-31-5-214.ec2.internal/172.31.5.214:8032
2021-03-22 03:36:12,084 [uber-SubtaskRunner] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2021-03-22 03:36:12,112 [uber-SubtaskRunner] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at ip-172-31-5-214.ec2.internal/172.31.5.214:8032
2021-03-22 03:36:12,116 [uber-SubtaskRunner] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2021-03-22 03:36:12,144 [uber-SubtaskRunner] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at ip-172-31-5-214.ec2.internal/172.31.5.214:8032
2021-03-22 03:36:12,147 [uber-SubtaskRunner] INFO org.apache.hadoop.mapred.ClientServiceDelegate - Application state is completed. FinalApplicationStatus=SUCCEEDED. Redirecting to job history server
2021-03-22 03:36:12,178 [uber-SubtaskRunner] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2021-03-22 03:36:12,198 [uber-SubtaskRunner] INFO org.apache.pig.Main - Pig script completed in 28 seconds and 42 milliseconds (28042 ms)
Hadoop Job IDs executed by Pig: job_1616369296561_0005

Search for file name

Rename Move Copy Change permissions Download Move to trash

Home S3 / user / qchen

Name	Size	User
+		hdfs
.		qchen
.Trash		qchen
book.txt	581.0 KB	qchen
googlebooks-eng-us-all-2gram-20090715-50-subset.csv	23.3 MB	qchen
oozie-oozi		qchen
pig_wordcount		qchen
wordcountoutput		hadoop

6. Challenges (75%)

1. Implement a MapReduce program that emits the bigrams which were coined after year 1992 (or which started appearing after the year 1992).

Output of the program should include: (bigram, year)

Example output: (mobile phone, 1996) means that the bigram 'mobile phone' first appeared in the dataset in the year 1996.

2. Implement a MapReduce program that emits the average number of times each bigram appears in a book (over all the years). [Average for a particular n-gram is the total count of n-gram (over all the years) divided by the total number of books in which the n-gram appeared (over all the years)]
Output of the program should include: (bigram, average)
Example output: (how are, 6) means that the bigram 'how are' appears on average 6 times in a book (over all the years).
3. Implement a Pig program that computes the most common bigram in the year 2003 in the dataset (as determined by the count field).
Output of the program should include: (bigram, count)
Example output: (how many, 5001) means that the bigram 'how many' was the most popular bigram in the year 2002 and it appeared a total of 5001 times in all the books in that year.
4. Implement a Pig program that computes the most common bigram in each year in the dataset (as determined by the count field).
Output of the program should include: (year, bigram, count)
Example output: (2003, mobile phone, 3012) means that in the year 2003 the most popular bigram was 'mobile phone' and it appeared 3012 times in all the books in that year. Emit such tuples for each year in the dataset.
5. Create a Hive meta-store table from the N-Gram dataset (CSV) file from the Hue web interface.
Implement a Hive query (in the SQL-like Hive Query Language) to find the most popular bigram (over all the years).

Deliverables

1. The complete code with the modifications needed to complete each exercise, including the new lambda function.
2. Output files (.txt or .csv) that contain results for each exercise program.