



DATA & ANALYTICS

HOMEWORK #5

March 2022



Data 228 - Big Data Technologies and Applications

Department of Applied Data Science

San Jose State University

Faiza Ayoun (015960139)

Harsimran Kaur (016003468)

Pooja Malage (015294760)

Saranya Pandiaraj (015304497)

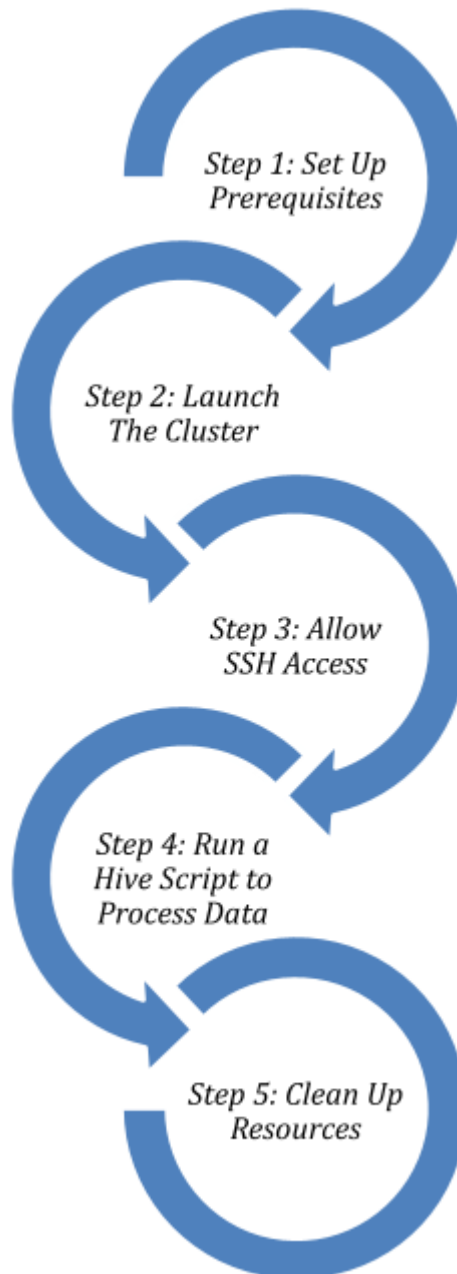
TABLE OF CONTENTS

| | |
|---|-----------|
| Table of Contents | 2 |
| Step 1: Set up Prerequisites | 4 |
| Setting up EMR | 4 |
| Creating an s3 Bucket | 5 |
| Uploading health_violations.py and food_establishment_data.csv to s3 bucket | 5 |
| Step 2: Launch The Cluster | 7 |
| My Account id | 8 |
| Step 3: Allow SSH access | 9 |
| Step 4: Run a hive script to process data | 10 |
| Step 5: Clean up your Amazon EMR resources | 13 |
| Terminating the Cluster | 13 |
| Deleting S3 resources | 13 |
| References: | 15 |

This assignment is designed to walk you through the process of creating a sample Amazon EMR cluster using Quick Create options in the AWS Management Console. After you create the cluster, you submit a Hive script as a step to process sample data stored in Amazon Simple Storage Service (Amazon S3)

Follow the steps and submit screenshots / report of your progress.

<https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-gs.html>



STEP 1: SET UP PREREQUISITES

SETTING UP EMR

Create Key PA

Create key pair [Info](#)

Key pair

A key pair, consisting of a private key and a public key, is a set of security credentials that you use to prove your identity when connecting to an instance.

Name

The name can include up to 255 ASCII characters. It can't include leading or trailing spaces.

Key pair type [Info](#)

- ☒ RSA
☐ ED25519

Private key file format

- ☒ .pem
For use with OpenSSH
☐ .ppk
For use with PuTTY

Tags (Optional)

No tags associated with the resource.

[Add tag](#)

You can add 50 more tags.

[Cancel](#)[Create key pair](#)

✓ Successfully created key pair

Key pairs (1) [Info](#)

| <input type="checkbox"/> | Name | Type | Fingerprint | ID |
|--------------------------|-------------------|------|---|-----------------------|
| <input type="checkbox"/> | DataDivas-KeyPair | rsa | cd:74:6a:cb:0c:d1:34:a9:4e:58:fb:d0:e8... | key-02645fa1f082bd537 |

CREATING AN S3 BUCKET

✔ Successfully created bucket "datadivasbucket"
To upload files and folders, or to configure additional bucket settings choose [View details](#).

ℹ How to optimize your costs on S3. [Learn more](#)

[Amazon S3](#) > Buckets

▶ **Account snapshot**
Storage lens provides visibility into storage usage and activity trends. [Learn more](#)

[View Storage Lens dashboard](#)

Buckets (1) [Info](#)
Buckets are containers for data stored in S3. [Learn more](#)

[Refresh](#) [Copy ARN](#) [Empty](#) [Delete](#) [Create bucket](#)

< 1 > ⚙

| | Name ▲ | AWS Region ▼ | Access ▼ | Creation date ▼ |
|-----------------------|---------------------------------|--------------------------|---|--------------------------------------|
| <input type="radio"/> | datadivasbucket | US East (Ohio) us-east-2 | Bucket and objects not public | March 25, 2022, 08:57:39 (UTC-07:00) |

UPLOADING HEALTH_VIOLATIONS.PY AND FOOD_ESTABLISHMENT_DATA.CSV TO S3 BUCKET

Upload [Info](#)
Add the files and folders you want to upload to S3. To upload a file larger than 160GB, use the AWS CLI, AWS SDK or Amazon S3 REST API. [Learn more](#)

Drag and drop files and folders you want to upload here, or choose **Add files**, or **Add folders**.

Files and folders (2 Total, 11.3 MB)
All files and folders in this table will be uploaded.

[Remove](#) [Add files](#) [Add folder](#)

< 1 >

| <input type="checkbox"/> | Name ▲ | Folder ▼ | Type ▼ | Size ▼ |
|--------------------------|-----------------------------|----------|----------------------|---------|
| <input type="checkbox"/> | food_establishment_data.csv | - | text/csv | 11.3 MB |
| <input type="checkbox"/> | health_violations.py | - | text/x-python-script | 1.8 KB |

Destination

Destination
[s3://datadivasbucket](#)

▶ **Destination details**
Bucket settings that impact new objects stored in the specified destination.

🟢 Upload succeeded
View details below.

ⓘ The information below will no longer be available after you navigate away from this page.

Summary

Destination
[s3://datadivasbucket](#)

Succeeded
✔ 2 files, 11.3 MB (100.00%)

Failed
❌ 0 files, 0 B (0%)

Files and folders

Configuration

Files and folders (2 Total, 11.3 MB)

🔍 Find by name

< 1 >

| Name | ▲ | Folder | ▼ | Type | ▼ | Size | ▼ | Status | ▼ | Error | ▼ |
|---|---|--------|---|----------------------|---|---------|---|-------------|---|-------|---|
| food_establishment_data.csv | | - | | text/csv | | 11.3 MB | | ✔ Succeeded | | - | |
| health_violations.py | | - | | text/x-python-script | | 1.8 KB | | ✔ Succeeded | | - | |


STEP 2: LAUNCH THE CLUSTER

Create Cluster - Quick Options [Go to advanced options](#)

General Configuration

Cluster name

☒ Logging [i](#)

S3 folder 

Launch mode ☒ Cluster [i](#) ☐ Step execution [i](#)

Software configuration

Release [i](#)

Applications

- ☐ Core Hadoop: Hadoop 2.10.1, Hive 2.3.8, Hue 4.9.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2
- ☐ HBase: HBase 1.4.13, Hadoop 2.10.1, Hive 2.3.8, Hue 4.9.0, Phoenix 4.14.3, and ZooKeeper 3.4.14
- ☐ Presto: Presto 0.261 with Hadoop 2.10.1 HDFS and Hive 2.3.8 Metastore
- ☒ Spark: Spark 2.4.8 on Hadoop 2.10.1 YARN and Zeppelin 0.10.0

☐ Use AWS Glue Data Catalog for table metadata [i](#)

Hardware configuration

Instance type [i](#) The selected instance type adds 64 GiB of GP2 EBS storage per instance by default. [Learn more](#) [Z](#)

Number of instances (1 master and 2 core nodes)

Cluster scaling ☐ scale cluster nodes based on workload

Auto-termination ☒ Enable auto-termination [Learn more](#) [Z](#)

Terminate cluster when it is idle after hours minutes

Security and access

EC2 key pair [i](#) [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](#) [Z](#) ☐ Use EMR_DefaultRole_V2 [i](#)

EC2 instance profile [EMR_EC2_DefaultRole](#) [Z](#) [i](#)

Cancel

Create cluster

✔ Successfully created bucket "datadivasbucket1"
To upload files and folders, or to configure additional bucket settings choose [View details](#).

Amazon S3 > Buckets

▶ Account snapshot

Storage lens provides visibility into storage usage and activity trends. [Learn more](#)

View Storage Lens dashboard

Buckets (1) Info

↺

Copy ARN

Empty

Delete

Create bucket

Buckets are containers for data stored in S3. [Learn more](#)

Find buckets by name

< 1 > ⚙

| | Name ▲ | AWS Region ▼ | Access ▼ | Creation date ▼ |
|-----------------------|------------------|--------------------------|-------------------------------|--------------------------------------|
| <input type="radio"/> | datadivasbucket1 | US East (Ohio) us-east-2 | Bucket and objects not public | March 25, 2022, 09:38:16 (UTC-07:00) |

✔ Successfully created key pair

Key pairs (1) Info

↺

Actions ▼

Create key pair

Filter key pairs

< 1 > ⚙

| <input type="checkbox"/> | Name ▼ | Type ▼ | Fingerprint ▼ | ID ▼ |
|--------------------------|------------------|--------|---|-----------------------|
| <input type="checkbox"/> | datadivaskeypair | rsa | b7:2e:99:37:ab:38:72:af:80:08:c3:74:ca... | key-0849837142356e35f |

STEP 3: ALLOW SSH ACCESS

Security Groups (1/2) Info

↻

Actions ▼

Export security groups to CSV ▼

Create security group

Q Filter security groups

< 1 > ⚙

search: sg-024aa88e9b300de25 ✕

Clear filters

| | Name ▼ | Security group ID ▼ | Security group name ▼ | VPC ID ▼ | Description ▼ |
|-------------------------------------|--------|----------------------|-------------------------|---|-----------------------------|
| <input type="checkbox"/> | - | sg-0232d23d46c20e6c6 | ElasticMapReduce-slave | vpc-035410586dbef58c7 🔗 | Slave group for Elastic ... |
| <input checked="" type="checkbox"/> | - | sg-024aa88e9b300de25 | ElasticMapReduce-master | vpc-035410586dbef58c7 🔗 | Master group for Elasti... |

-

SSH ▼

TCP

22

My IP ▼

Q

Delete

71.202.19.197/32 ✕

Add rule

Cancel

Preview changes

Save rules

✓ Inbound security group rules successfully modified on security group (sg-024aa88e9b300de25 | ElasticMapReduce-master)

▶ Details

STEP 4: RUN A HIVE SCRIPT TO PROCESS DATA

Clone

Terminate

AWS CLI export

Cluster: My First EMR Cluster Starting

Summary

Application user interfaces

Monitoring

Hardware

Configurations

Events

Steps

Bootstrap actions

Summary

ID: j-2KSHPAB1GOS0D

Creation date: 2022-03-25 09:41 (UTC-7)

Elapsed time: 30 seconds

After last step completes: Cluster waits

Termination protection: Off [Change](#)

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

Master public DNS: --

Configuration details

Release label: emr-5.34.0

Hadoop distribution: Amazon

Applications: Spark 2.4.8, Zeppelin 0.10.0

Log URI: s3://datadivasbucket1/logs/

EMRFS consistent view: Disabled

Custom AMI ID: --

Application user interfaces

Persistent user interfaces : --

On-cluster user -- interfaces :

Network and hardware

Availability zone: us-east-2b

Subnet ID: [subnet-00568ca32c20e2546](#)

Master: Provisioning 1 m5.xlarge

Core: Provisioning 2 m5.xlarge

Task: --

Cluster scaling: Not enabled

Auto-termination: Terminate if idle for 1 hour

Core Instance Group: Your account is currently being verified. Verification normally takes less than 2 hours. Until your account is verified, you may not be able to launch additional instances or create additional volumes. If you are still receiving this message after more than 2 hours, please let us know by writing to aws-verification@amazon.com. We appreciate your patience..

Master Instance Group: Your account is currently being verified. Verification

Clone

Terminate

AWS CLI export

Cluster: My First EMR Cluster Waiting Cluster ready after last step completed.

Summary

Application user interfaces

Monitoring

Hardware

Configurations

Events

Steps

Bootstrap actions

Summary

ID: j-2KSHPAB1GOS0D

Creation date: 2022-03-25 09:41 (UTC-7)

Elapsed time: 15 minutes

After last step completes: Cluster waits

Termination protection: Off [Change](#)

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

Master public DNS:
ec2-3-21-113-201.us-east-2.compute.amazonaws.com
[Connect to the Master Node Using SSH](#)

Configuration details

Release label: emr-5.34.0

Hadoop distribution: Amazon

Applications: Spark 2.4.8, Zeppelin 0.10.0

Log URI: s3://datadivasbucket1/logs/

EMRFS consistent view: Disabled

Custom AMI ID: --

Application user interfaces

Persistent user interfaces : [Spark history server](#), [YARN timeline server](#)

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

interfaces :

Network and hardware

Availability zone: us-east-2b

Subnet ID: [subnet-00568ca32c20e2546](#)

Master: Running 1 m5.xlarge

Core: Running 2 m5.xlarge

Task: --

Cluster scaling: Not enabled

Auto-termination: Terminate if idle for 1 hour

Security and access

Key name: datadivaskeypair

EC2 instance profile: EMR_EC2_DefaultRole

EMR role: EMR_DefaultRole

Cluster: My First EMR Cluster

Waiting

Cluster ready after last step completed.

Summary

Application user interfaces

Monitoring

Hardware

Configurations

Events

Steps

Bootstrap actions

Concurrency: 1

Change

After last step completes:

Cluster waits

Add step

Clone step

Cancel step

View Jobs in the Application History Tab

Filter: All steps

Filter steps ...

2 steps (all loaded)

| | ID | Name | Status | Start time (UTC-7) | Elapsed time | Log files |
|--|----------------|-------------------|-----------|--------------------------|--------------|---------------------------|
| | s-HJ2QHVHM2QMS | Spark application | Completed | 2022-03-25 10:00 (UTC-7) | 32 seconds | View logs |

myOutputFolder/

Copy S3 URI

Objects

Properties

Objects (2)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

Copy S3 URI

Copy URL

Download

Open

Delete

Actions

Create folder

Upload

Find objects by prefix

< 1 >

| | Name | Type | Last modified | Size | Storage class |
|--|--|------|--------------------------------------|---------|---------------|
| | _SUCCESS | - | March 25, 2022, 10:00:46 (UTC-07:00) | 0 B | Standard |
| | part-00000-044d1f78-53ef-4876-a399-b3c8cb2b44f9-c000.csv | csv | March 25, 2022, 10:00:46 (UTC-07:00) | 219.0 B | Standard |

part-00000-044d1f78-53ef-4876-a399-b3c8cb2b44f9-c000

125%

Add Category

Insert

Table

Chart

Text

Shape

Media

Comment

Sheet 1

A

B

part-00000-044d1f78-53ef-4876-a399-

| | name | total_red_violations |
|----|-----------------------|----------------------|
| 1 | SUBWAY | 322 |
| 2 | T-MOBILE PARK | 315 |
| 3 | WHOLE FOODS MARKET | 299 |
| 4 | PCC COMMUNITY MARKETS | 251 |
| 5 | TACO TIME | 240 |
| 6 | MCDONALD'S | 177 |
| 7 | THAI GINGER | 153 |
| 8 | SAFEWAY INC #1508 | 143 |
| 9 | TAQUERIA EL RINCONITO | 134 |
| 10 | HIMITSU TERIYAKI | 128 |

Availability zone: us-east-2b

Subnet ID: [subnet-00568ca32c20e2546](#)

Master: **Running** 1 m5.xlarge

Core: **Running** 2 m5.xlarge

Task: --

Cluster scaling: Not enabled

Auto-termination: Terminate if idle for 1 hour

STEP 5: CLEAN UP YOUR AMAZON EMR RESOURCES

TERMINATING THE CLUSTER

Terminate clusters

×

Are you sure you want to terminate this cluster?

- j-2KSHPAB1GOS0D (My First EMR Cluster)

Any pending work or data residing on the cluster will be lost, such as data stored in HDFS. This action is irreversible.

[Cancel](#) [Terminate](#)

DELETING S3 RESOURCES

Empty bucket

Info

⚠

- Emptying the bucket deletes all objects in the bucket and cannot be undone.
- Objects added to the bucket while the empty bucket action is in progress might be deleted.
- To prevent new objects from being added to this bucket while the empty bucket action is in progress, you might need to update your bucket policy to stop objects from being added to the bucket.

[Learn more](#)

ℹ

If your bucket contains a large number of objects, creating a lifecycle rule to delete all objects in the bucket might be a more efficient way of emptying your bucket. [Learn more](#)

[Go to lifecycle rule configuration](#)

Permanently delete all objects in bucket "datadivasbucket1"?

To confirm deletion, type *permanently delete* in the text input field.

[Cancel](#) [Empty](#)



Successfully emptied bucket "datadivasbucket1"

View details below. If you want to delete this bucket, use the [delete bucket configuration](#).

Delete bucket Info



- Deleting a bucket cannot be undone.
- Bucket names are unique. If you delete a bucket, another AWS user can use the name.

[Learn more](#) 

Delete bucket "datadivasbucket1"?

To confirm deletion, enter the name of the bucket in the text input field.

Cancel

Delete bucket

REFERENCES:

- (2022). Retrieved 2 April 2022, from <https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-gs.html>.