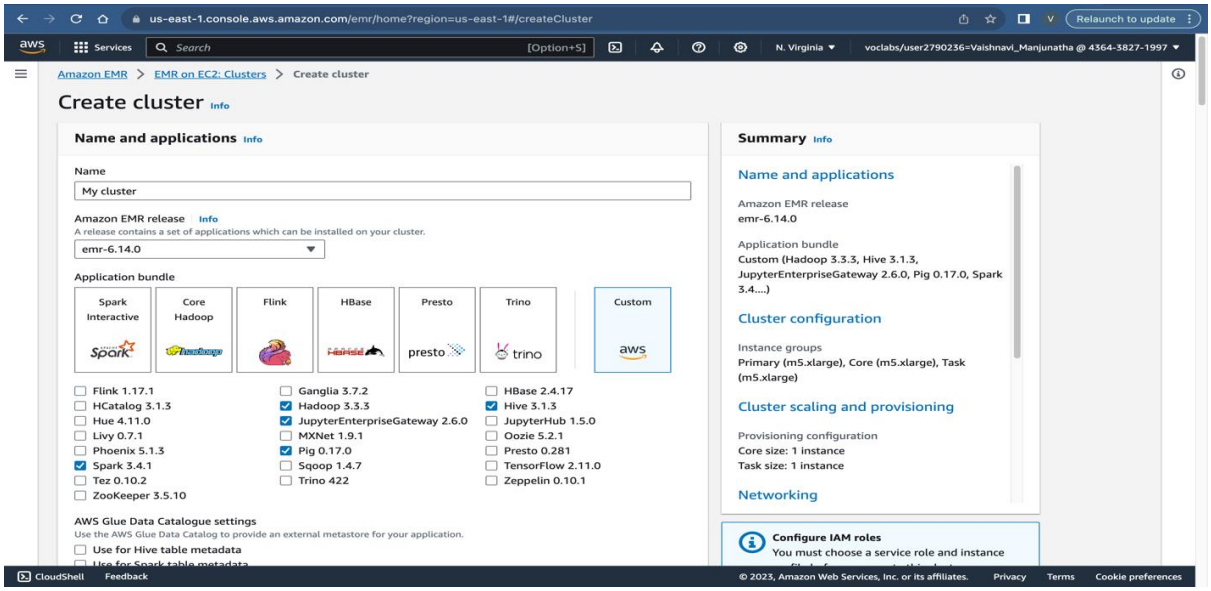
YOUTUBE COMMENTS SPAM DETECTION SYSTEM

Name : Vaishnavi Manjunatha
Student ID : 23260426
Email : Vaishnavi.manjunatha2@mail.dcu.ie
Git Repository Link : [Repo Link]
Dataset : YouTube Spam Collection Data Set (Dataset Link)

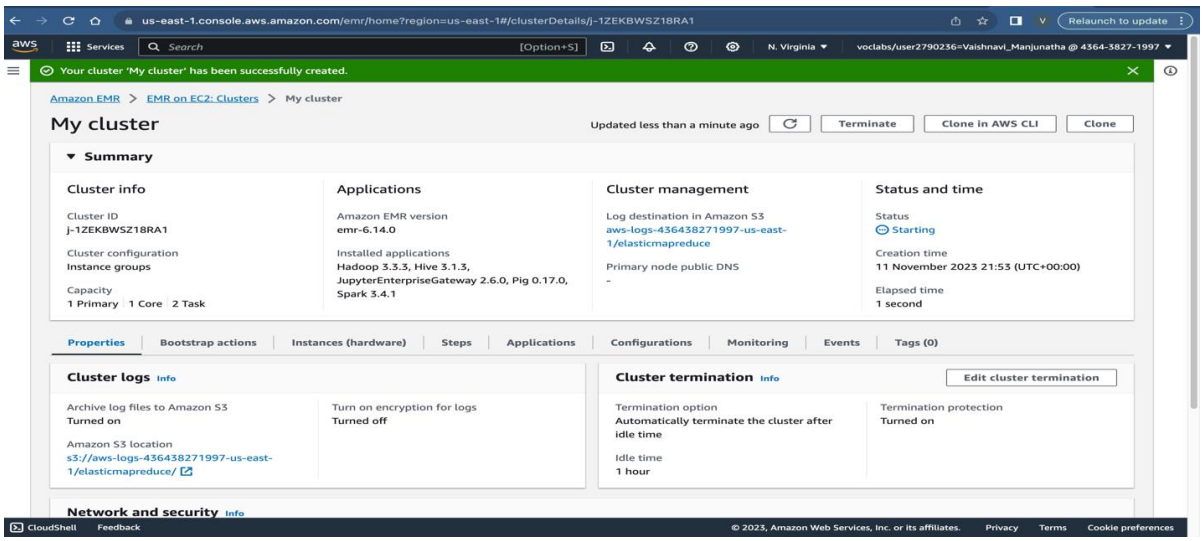
TASK 1 : Cloud Infrastructure Setup (AWS)

1.1 Hadoop cluster created on Amazon EMR

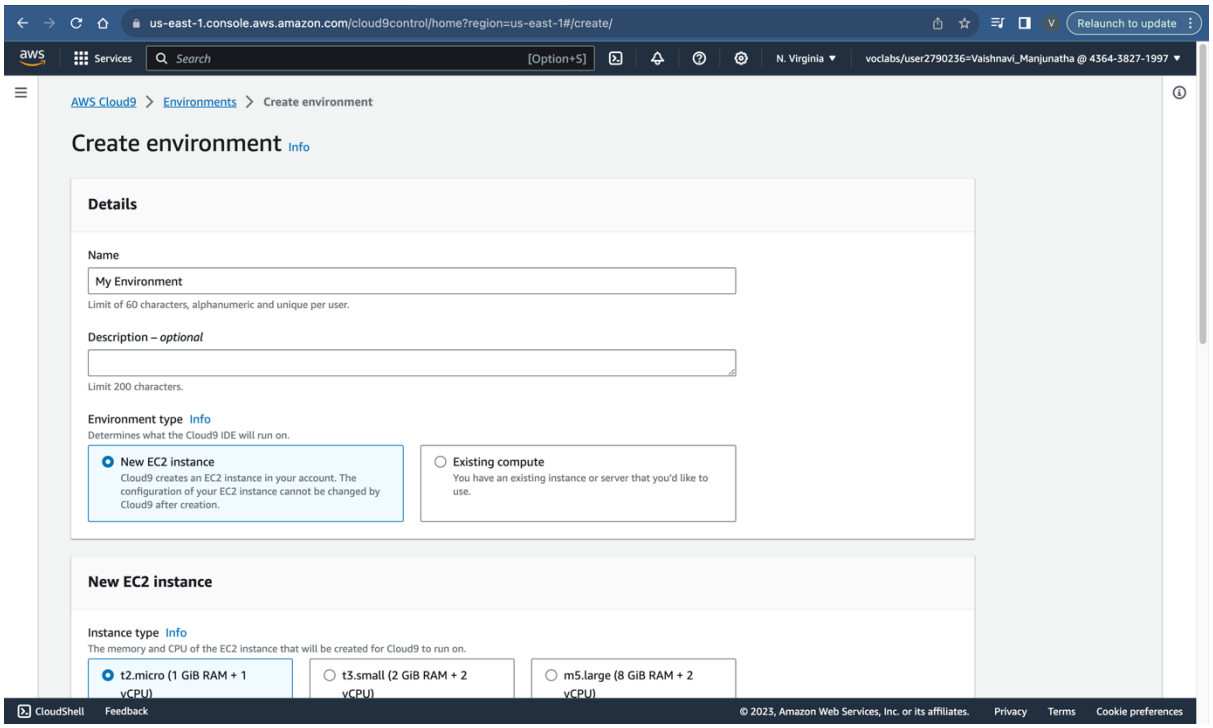
[1] Created a cluster on amazon EMR (name- My cluster) with the installation of spark 3.4.1, Hadoop 3.3.3 JupyterEnterpriseGateway 2.6.0, Pig 0.17.0, Hive 3.1.3



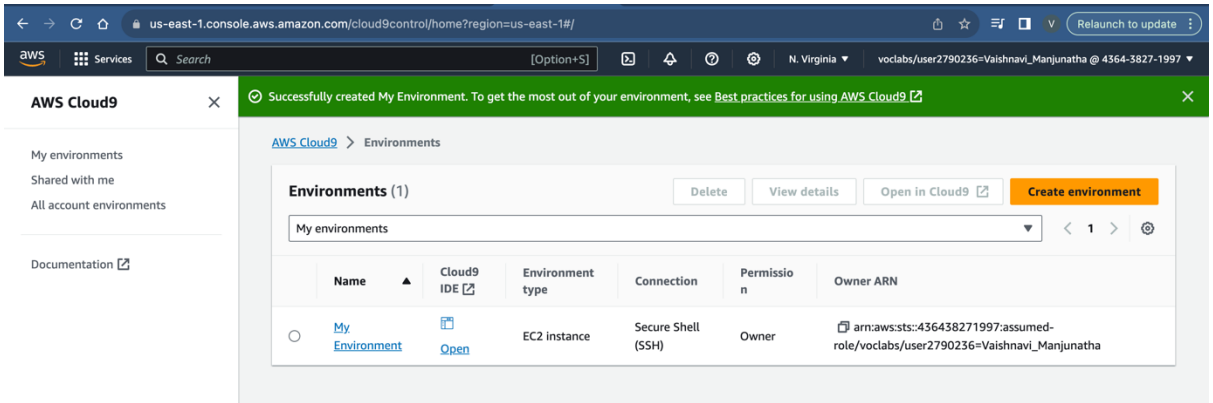
[2] Cluster is created successfully with 1 core node and 2 task nodes with 1 primary, 1 core and 2 task nodes (type m4.large) with the EMR roles properly defined



[3] A cloud 9 environment is created which is of type EC2 Instance with SSH connection

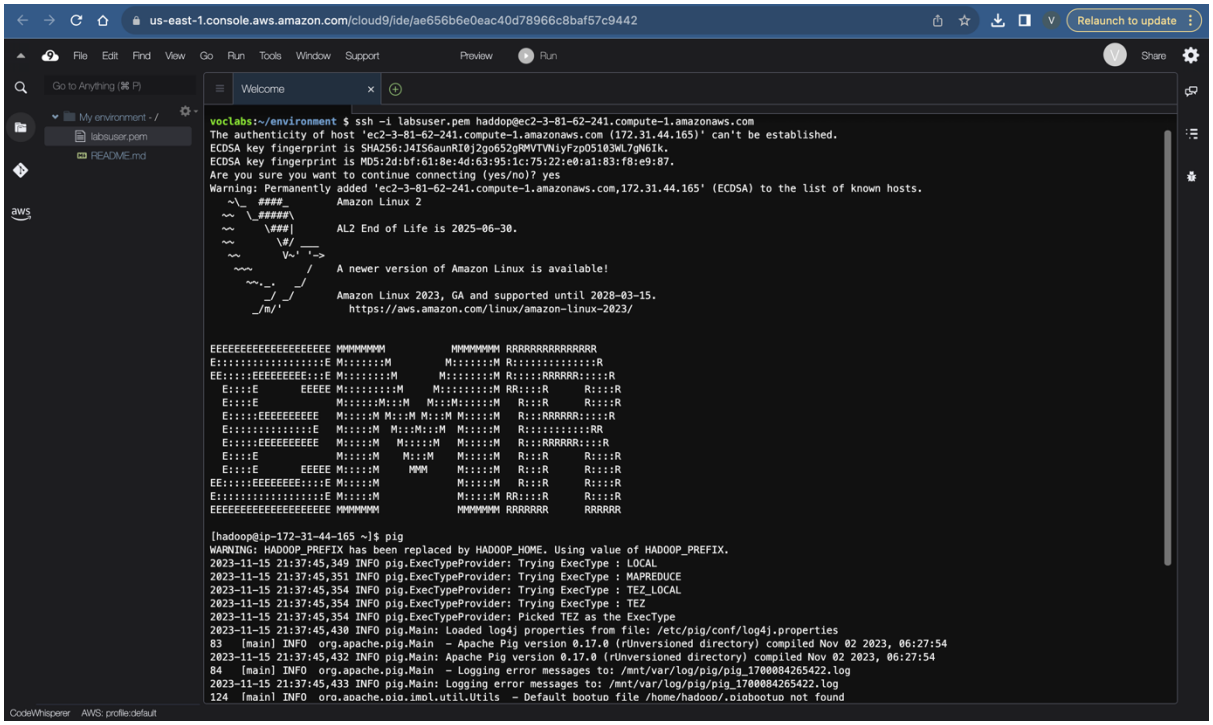


[4] New Environment created successfully. Upload the pem file downloaded from sandbox into the cloud 9 environment. Add the cloud 9 as an inbound rule in the EC2 security groups of primary node of the cluster.



1.2 Hadoop and Pig , hive are installed

- We SSH to connect to the Hadoop primary node-
  ssh -i labsuser.pem hadoop@ec2-54-242-137-63.compute-1.amazonaws.com
- The commands execute below confirm the installation pig, Hive and hadoop



TASK 2 : Dataset

2.1: Choose a relevant dataset

- Dataset chosen - YouTube spam detection dataset taken from Kaggle
- Link - https://www.kaggle.com/datasets/lakshmi25npathi/images
- The dataset is appropriate because of presence of unique ID- author name and youtube comment content, which are the requirements. It consists of 5 columns in total -
  Comment_id(the ID for each comment)
  Author(Unique username of the commentor)
  Date(date of the comment made), content(The text in comment)
  Class(Ham or spam)
- The dataset used is a combination of datasets from five CSV files given. The final CSV file used is available in gitlab.(Dataset file) Ethical concerns are also eliminated as it is a publicly available dataset.

2.2 Dataset was not extracted from any website

2.3 Loaded the dataset into Amazon S3 bucket

[1] New bucket created with name-vaishnavisbucket

s3.console.aws.amazon.com/s3/bucket/create?region=us-east-1

Relaunch to update

aws Services Search [Option+S] Global voclabs/user2790236=Vaishnavi_Manjunatha @ 3246-3554-0431

Amazon S3 > Buckets > Create bucket

Create bucket Info

Buckets are containers for data stored in S3. Learn more

General configuration

Bucket name

vaishnavisbucket

Bucket name must be unique within the global namespace and follow the bucket naming rules. See rules for bucket naming

AWS Region

US East (N. Virginia) us-east-1

Copy settings from existing bucket - *optional*

Only the bucket settings in the following configuration are copied.

Choose bucket

Object Ownership Info

Control ownership of objects written to this bucket from other AWS accounts and the use of access control lists (ACLs). Object ownership determines who can specify access to objects.

- (selected) ACLs disabled (recommended)
  All objects in this bucket are owned by this account. Access to this bucket and its objects is specified using only policies.
- ACLs enabled
  Objects in this bucket can be owned by other AWS accounts. Access to this bucket and its objects can be specified using ACLs.

CloudShell Feedback Privacy Terms Cookie preferences

s3.console.aws.amazon.com/s3/bucket/create?region=us-east-1

Relaunch to update

aws Services Search [Option+S] Global voclabs/user2790236=Vaishnavi_Manjunatha @ 3246-3554-0431

No tags associated with this bucket.

Add tag

Default encryption Info

Server-side encryption is automatically applied to new objects stored in this bucket.

Encryption type Info

- (selected) Server-side encryption with Amazon S3 managed keys (SSE-S3)
- Server-side encryption with AWS Key Management Service keys (SSE-KMS)
- Dual-layer server-side encryption with AWS Key Management Service keys (DSSE-KMS)
  Secure your objects with two separate layers of encryption. For details on pricing, see **DSSE-KMS pricing** on the **Storage** tab of the Amazon S3 pricing page.

Bucket Key

Using an S3 Bucket Key for SSE-KMS reduces encryption costs by lowering calls to AWS KMS. S3 Bucket Keys aren't supported for DSSE-KMS. Learn more

- Disable
- (selected) Enable

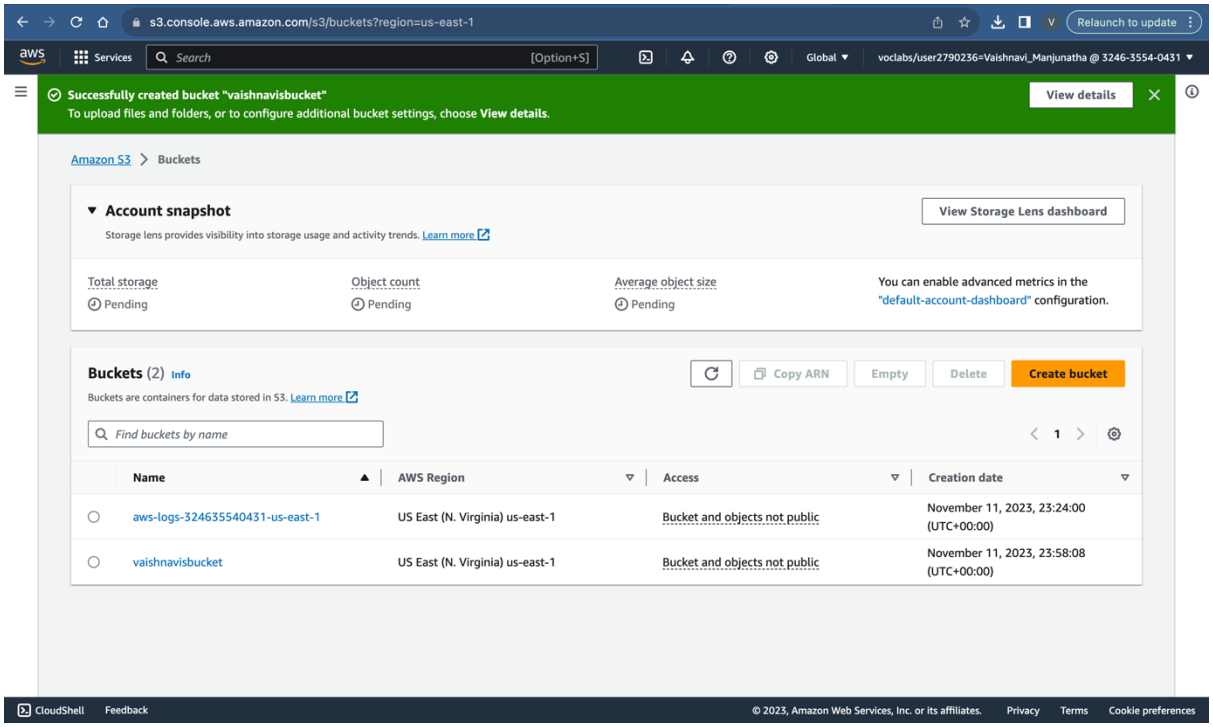
Advanced settings

After creating the bucket, you can upload files and folders to the bucket, and configure additional bucket settings.

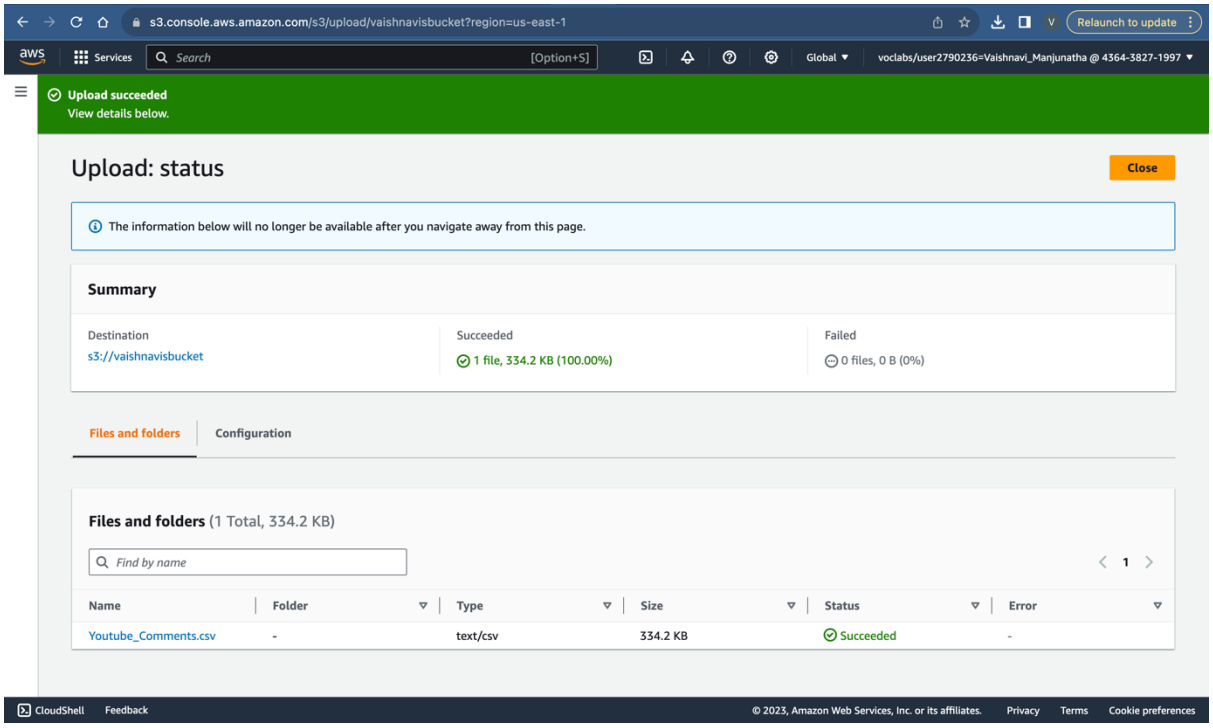
Cancel Create bucket

CloudShell Feedback © 2023, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

[2] Bucket creation is successful



[3] The Youtube comments CSV file is uploaded to the S3 bucket



Task 3: Clean and process the data using Hive

The process of cleaning involves loading the data from the CSV file present in the s3 bucket to a Hive table. Various cleaning techniques are applied depending on the requirement.

[1] Created a hive table in the location of the new S3 bucket

[2] Loading data into the table from the dataset in S3 bucket

```
[hadoop@ip-172-31-7-30 ~]$ hive
Hive Session ID = 6e933fd0-5945-40a4-8295-386c4d9a2223

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: true
hive> CREATE EXTERNAL TABLE YTcomments_data (
    >     comment_id STRING,
    >     author STRING,
    >     commentDate DATE,
    >     content STRING,
    >     spamLabel INT
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > LOCATION 's3://vaishnavisbucket'
    > TBLPROPERTIES('skip.header.line.count' = '1');
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. AlreadyExistsException(message:Table hive.default.YTcomments_data alread
y exists)
hive> LOAD DATA INPATH 's3://vaishnavisbucket/Youtube_Comments.csv' INTO TABLE YTcomments_data;
Loading data to table default.ytcomments_data
OK
Time taken: 3.866 seconds
hive> select count(*) from YTcomments_data;
Query ID = hadoop_20231116014127_efc8c1f6-dc5b-4916-9d14-d8d6099170c3
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1700096042925_0002)

--------------------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------------------------------
Map 1 ..........  container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......  container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 19.84 s
--------------------------------------------------------------------------------------------------------
OK
1961
Time taken: 25.717 seconds, Fetched: 1 row(s)
hive>
```

[3] Removed null values in the columns- author, content and spam label and created a new table to store the data. Notice that the new table (YTcomments_Null_removed) now has only the 3 mentioned columns with null-free data.

```
hadoop@ip-172-31-7-30:~ ×
hive> CREATE TABLE YTcomments_Null_removed AS
    > SELECT
    >     author,
    >     LOWER(content) AS content, spamLabel
    > FROM YTcomments_data where content is not null and author is not null;
Query ID = hadoop_20231116021501_4baef278-4bad-4004-b4cd-307122661602
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1700096042925_0005)

--------------------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------------------------------
Map 1 ..........  container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 15.89 s
--------------------------------------------------------------------------------------------------------
Moving data to directory hdfs://ip-172-31-7-30.ec2.internal:8020/user/hive/warehouse/ytcomments_null_removed
OK
Time taken: 24.782 seconds
hive> select * from YTcomments_Null_removed LIMIT 10;
OK
Julius NM       huh anyway check out this you[tube] channel: kobyoshi02 1
adam riyati     hey guys check out my new channel and our first vid this is us the  monkeys!!! i'm the monkey in the white shirtplease leave a like comment
 and please subscribe!!!!        1
Evgeny Murashkin        just for test i have to say murdev.com  1
ElNino Melendez me shaking my sexy ass on my channel enjoy ^_^  1
GsMega  watch?v=vtarggvgtwq   check this out .  1
Jason Haddad    hey check out my new website!! this site is about kids stuff. kidsmediausa  . com       1
ferleck ferles  subscribe to my channel         1
Bob Kanowski    i turned it on mute as soon is i came on i just wanted to check the  views...   0
Cony    you should check my channel for funny videos!!  1
BeBe Burkey     and u should.d check my channel and tell me what i should do next!      1
Time taken: 0.103 seconds, Fetched: 10 row(s)
hive>
```

[4] Removed URL values from the previous table, created a new table (YTcomments_htmlTags_removed) to store the data

```
hadoop@ip-172-31-7-30:~ ×
hive> CREATE TABLE YTcomments_url_removed AS
    > SELECT
    >     author,
    >     regexp_replace(content, 'http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\\(\\),]|(?:%[0-9a-fA-F][0-9a-fA-F]))+', '') AS content, spamLabel
    > FROM YTcomments_Null_removed;
Query ID = hadoop_20231116021729_6bdbdf81-7583-4e89-8bba-d681676e9a28
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1700096042925_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 7.20 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://ip-172-31-7-30.ec2.internal:8020/user/hive/warehouse/ytcomments_url_removed
OK
Time taken: 7.715 seconds
hive> select * from YTcomments_url_removed LIMIT 10;
OK
Julius NM       huh anyway check out this you[tube] channel: kobyoshi02 1
adam riyati     hey guys check out my new channel and our first vid this is us the  monkeys!!! i'm the monkey in the white shirtplease leave a like comment
 and please subscribe!!!!       1
Evgeny Murashkin        just for test i have to say murdev.com  1
ElNino Melendez me shaking my sexy ass on my channel enjoy ^_^  1
GsMega  watch?v=vtarggvgtwq   check this out .  1
Jason Haddad    hey check out my new website!! this site is about kids stuff. kidsmediausa  . com       1
ferleck ferles  subscribe to my channel         1
Bob Kanowski    i turned it on mute as soon is i came on i just wanted to check the  views...   0
Cony    you should check my channel for funny videos!!  1
BeBe Burkey     and u should.d check my channel and tell me what i should do next!      1
Time taken: 0.182 seconds, Fetched: 10 row(s)
hive>
```

[5] Removed HTML tags from the previous table, created a new table (YTcomments_htmlTags_removed) to store the data

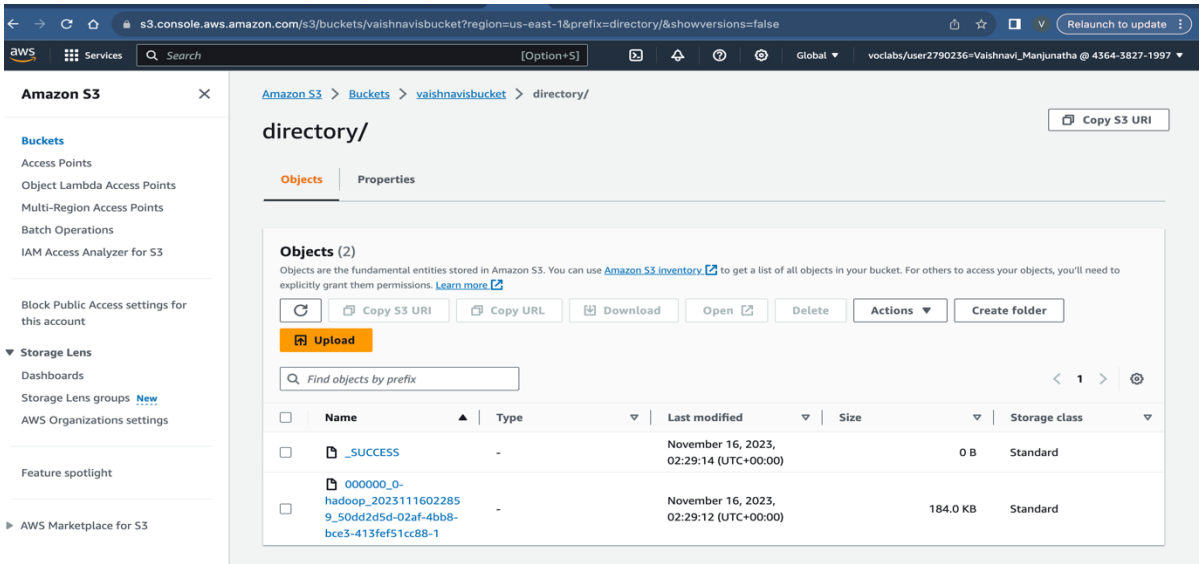
```
hadoop@ip-172-31-7-30:~ ×
hive> CREATE TABLE YTcomments_htmlTags_removed AS
    > SELECT
    >     author,
    >     regexp_replace(content, '<[^>]+>', '') AS content,spamLabel
    > FROM
    >     YTcomments_url_removed;
Query ID = hadoop_20231116021940_f8bdb37e-7842-4cdc-8ff3-bdfb3f999439
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1700096042925_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 7.11 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://ip-172-31-7-30.ec2.internal:8020/user/hive/warehouse/ytcomments_htmltags_removed
OK
Time taken: 7.561 seconds
hive> select * from YTcomments_htmlTags_removed LIMIT 10;
OK
Julius NM       huh anyway check out this you[tube] channel: kobyoshi02 1
adam riyati     hey guys check out my new channel and our first vid this is us the  monkeys!!! i'm the monkey in the white shirtplease leave a like comment
 and please subscribe!!!!       1
Evgeny Murashkin        just for test i have to say murdev.com  1
ElNino Melendez me shaking my sexy ass on my channel enjoy ^_^  1
GsMega  watch?v=vtarggvgtwq   check this out .  1
Jason Haddad    hey check out my new website!! this site is about kids stuff. kidsmediausa  . com       1
ferleck ferles  subscribe to my channel         1
Bob Kanowski    i turned it on mute as soon is i came on i just wanted to check the  views...   0
Cony    you should check my channel for funny videos!!  1
BeBe Burkey     and u should.d check my channel and tell me what i should do next!      1
Time taken: 0.084 seconds, Fetched: 10 row(s)
hive>
```

[6] Removed special characters from the previous table, created a new table (YTcomments_cleaned) to store the data

```
hadoop@ip-172-31-7-30:~ ×

hive> CREATE TABLE YTcomments_cleaned AS
    > SELECT
    >     author,
    >     REGEXP_REPLACE(content, '[^a-zA-Z0-9\\s]', '') AS content, spamLabel
    > FROM YTcomments_htmlTags_removed;
Query ID = hadoop_20231116022359_522d5795-4743-47be-9aa7-09759bc30dea
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1700096042925_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 6.69 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://ip-172-31-7-30.ec2.internal:8020/user/hive/warehouse/ytcomments_cleaned
OK
Time taken: 7.228 seconds
hive> select * from YTcomments_cleaned LIMIT 10;
OK
Julius NM       huh anyway check out this youtube channel kobyoshi02    1
adam riyati     hey guys check out my new channel and our first vid this is us the  monkeys im the monkey in the white shirtplease leave a like comment  and
 please subscribe       1
Evgeny Murashkin        just for test i have to say murdevcom   1
ElNino Melendez me shaking my sexy ass on my channel enjoy      1
GsMega  watchvvtarggvgtwq   check this out      1
Jason Haddad    hey check out my new website this site is about kids stuff kidsmediausa   com   1
ferleck ferles  subscribe to my channel         1
Bob Kanowski    i turned it on mute as soon is i came on i just wanted to check the  views      0
Cony    you should check my channel for funny videos    1
BeBe Burkey     and u shouldd check my channel and tell me what i should do next        1
Time taken: 0.122 seconds, Fetched: 10 row(s)
hive>
```

[7] The cleaned file is now saved to a text file

```
hadoop@ip-172-31-21-12 ×   Immediate (Javascript (brov ×

hive> INSERT OVERWRITE LOCAL DIRECTORY 's3://vaishnavisbucket/directory/'
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
    > SELECT * FROM YTcomments_cleaned;
Query ID = hadoop_20231116153852_08dac9f1-64ad-4832-a420-5d47640ca11b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1700148220065_0001)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 11.81 s
--------------------------------------------------------------------------------
OK
Time taken: 13.04 seconds
hive>
```



Task 4: Ham and Spam using Hive

- The cleaned dataset is divided into 2 parts ham dataset and spam dataset. Written hive queries to get the top 10 ham and spam accounts.
- The following bag of words are considered for the prediction of ham or spam- subscribe, youtube, free, amazing, win, my channel, click here, check, limited time offer, promo, url, warning.

```
hadoop@ip-172-31-35-24: ×   Immediate
hive> CREATE TABLE spam_accounts AS
    > SELECT
    >     author,content,
    >     CASE
    >         WHEN content LIKE '%subscribe%' OR
    >             content LIKE '%youtube%' OR
    >             content LIKE '%free%' OR
    >             content LIKE '%amazing%' OR
    >             content LIKE '%win%' OR
    >             content LIKE '%my channel%' OR
    >             content LIKE '%click here%' OR
    >             content LIKE '%check%' OR
    >             content LIKE '%limited time offer%' OR
    >             content LIKE '%promo%' OR
    >             content LIKE '%url%' OR
    >             content LIKE '%warning%'
    >         THEN 1
    >         ELSE 0
    >     END AS is_spam
    > FROM YTcomments_cleaned;
Query ID = hadoop_20231116220559_39410154-c8b0-457d-a7dd-786b48f1f72f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1700171172617_0002)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 7.10 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://ip-172-31-35-24.ec2.internal:8020/user/hive/warehouse/spam_accounts
OK
Time taken: 7.694 seconds
hive>
```

```
hadoop@ip-172-31-35-24: ×   Immediate
hive> select * from spam_accounts LIMIT 10;
OK
Julius NM       huh anyway check out this youtube channel kobyoshi02    1
adam riyati     hey guys check out my new channel and our first vid this is us the  monkeys im the monkey in the white shirtplease leave a like comment  and
 please subscribe        1
Evgeny Murashkin        just for test i have to say murdevcom   0
ElNino Melendez me shaking my sexy ass on my channel enjoy      1
GsMega  watchvvtarggvgtwq   check this out      1
Jason Haddad    hey check out my new website this site is about kids stuff kidsmediausa   com   1
ferleck ferles  subscribe to my channel         1
Bob Kanowski    i turned it on mute as soon is i came on i just wanted to check the  views      1
Cony    you should check my channel for funny videos    1
BeBe Burkey     and u shouldd check my channel and tell me what i should do next        1
Time taken: 0.186 seconds, Fetched: 10 row(s)
hive>
```

[1] Picked and stored the ham data from the main dataset

```
hive> CREATE TABLE ham_dataset AS
    > SELECT *
    > FROM spam_accounts
    > WHERE is_spam =0;
Query ID = hadoop_20231116221350_9b881909-39f9-47f9-b181-9339f464a58f
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1700171172617_0003)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ......... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 7.79 s
--------------------------------------------------------------------------------
Moving data to directory hdfs://ip-172-31-35-24.ec2.internal:8020/user/hive/warehouse/ham_dataset
OK
Time taken: 17.128 seconds
hive> select * from ham_dataset LIMIT 10;
OK
Evgeny Murashkin        just for test i have to say murdevcom   0
Archie Lewis            0
Francisco Nora  please like d   0
OutrightIgnite          0
Tony K Frazier          0
OutrightIgnite  show your auburn pride here     0
Living4Techno   marketglory  comstrategygameandrijamatf earn real money from game       0
Tasha Lucius    2 billioncoming soon    0
Eugene Kalinin  the projects after effects music foto web sites and another you can find  and buy here          0
cake cat        if you like roblox minecraft world of warcraft gta5 mario suscribe to my  channel       0
Time taken: 0.113 seconds, Fetched: 10 row(s)
hive>
```

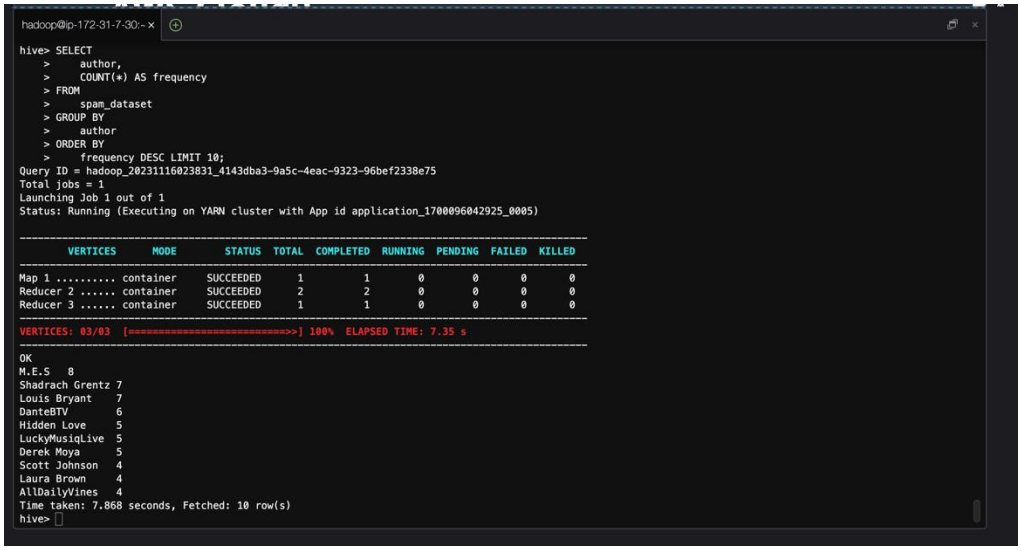
[2] Picked and stored the spam data from the main dataset

```
hadoop@ip-172-31-35-24: ×   Immediate
hive> CREATE TABLE spam_dataset AS
    > SELECT *
    > FROM spam_accounts
    > WHERE is_spam =1;
Query ID = hadoop_20231116221557_1666fc41-24e5-40e2-941d-69a38126ddcf
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1700171172617_0003)

--------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------------------
Map 1 ........... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 6.55 s
--------------------------------------------------------------------------------------------
Moving data to directory hdfs://ip-172-31-35-24.ec2.internal:8020/user/hive/warehouse/spam_dataset
OK
Time taken: 7.166 seconds
hive> select * from spam_dataset LIMIT 10;
OK
Julius NM       huh anyway check out this youtube channel kobyoshi02    1
adam riyati     hey guys check out my new channel and our first vid this is us the  monkeys im the monkey in the white shirtplease leave a like comment  and
 please subscribe       1
ElNino Melendez me shaking my sexy ass on my channel enjoy      1
GsMega  watchvvtarggvgtwq   check this out      1
Jason Haddad    hey check out my new website this site is about kids stuff kidsmediausa   com   1
ferleck ferles  subscribe to my channel         1
Bob Kanowski    i turned it on mute as soon is i came on i just wanted to check the  views      1
Cony    you should check my channel for funny videos    1
BeBe Burkey     and u shouldd check my channel and tell me what i should do next        1
Huckyduck       hey subscribe to me     1
Time taken: 0.149 seconds, Fetched: 10 row(s)
hive>
```

4.2 Selected the top 10 ham accounts

```
hadoop@ip-172-31-7-30:~ ×
hive> SELECT
    >     author,
    >     COUNT(*) AS frequency
    > FROM
    >     ham_dataset
    > GROUP BY
    >     author
    > ORDER BY
    >     frequency DESC LIMIT 10;
Query ID = hadoop_20231116023642_2cd5c266-3cd1-4e40-bd71-6ce7451f2f0b
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1700096042925_0005)

--------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------------------
Map 1 ........... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 7.02 s
--------------------------------------------------------------------------------------------
OK
5000palo        7
Marshmallow Kingdom     3
Seth Ryan       3
janet rangel    2
Alain Bruno     2
Brian Brai      2
Paul Crowder    2
Pepe The Meme King      2
Sonny Carter    2
Eric Gonzalez   2
Time taken: 7.571 seconds, Fetched: 10 row(s)
hive>
```

4.2 Selected the top 10 spam accounts



```
hadoop@ip-172-31-7-30:~ ×
hive> SELECT
    >     author,
    >     COUNT(*) AS frequency
    > FROM
    >     spam_dataset
    > GROUP BY
    >     author
    > ORDER BY
    >     frequency DESC LIMIT 10;
Query ID = hadoop_20231116023831_4143dba3-9a5c-4eac-9323-96bef2338e75
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1700096042925_0005)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 7.35 s
--------------------------------------------------------------------------------
OK
M.E.S   8
Shadrach Grentz 7
Louis Bryant    7
DanteBTV        6
Hidden Love     5
LuckyMusiqLive  5
Derek Moya      5
Scott Johnson   4
Laura Brown     4
AllDailyVines   4
Time taken: 7.868 seconds, Fetched: 10 row(s)
hive>
```

REFERENCES

[1] Hive queries- https://docs.cloudera.com/cdw-runtime/cloud/using-hiveql/topics/hive_query_information_schema.html
[2] AWS EMR- https://docs.aws.amazon.com/emr/
[3] Errors related to Hive queries- https://stackoverflow.com/
[4] Dataset- https://www.kaggle.com/datasets