IFT 598
MANAGING THE CLOUD

TEAM PROJECT

FAKE NEWS DETECTION USING PASSIVE AGGRESSIVE CLASSIFIER

REPORT BY
RASHMI BONGIRWAR
NARASIMHA SIVA SAKETH EMANI

PROJECT MENTOR
THOMAS PEARSON

TABLE OF CONTENTS

ABSTRACT

With the age of digitalization comes its disadvantages too. News these days is more popular in the digital form. Be it social media or official news channel websites, there is a plethora available. But the problem often faced is the widespread amount of fake news available. This low quality and intentional false information can cause major misunderstanding and cause negative impact on society. Some of it is also purposely done because of political motives. What is alarming is the rate with which it spreads. It is a major big data analysis research to get the huge data and detect the difference between fake and real. Various solutions are implemented for this serious issue.

One of the possible solutions is classifying the text which falls under the NLP task. This analysis of text comes under the category of natural language processing. The process we followed is first cleaning the data by removing stop words and then performing TF-IDF vectorization followed by using passive aggressive classifier to predict whether the given labels are real or fake. In the end, we checked the accuracy of the model and showed the confusion matrix.

INTRODUCTION

The first task we did was to find a motivational topic for combining AWS learning opportunity with a good interesting big data problem that can be solved. Previously, we had done big data analysis but never using the cloud. This learning opportunity was perfect to learn how the aws resources could be so useful and fast in analysis. We explored the storage facility, virtual servers and additional presence of benefits for efficiently running big data algorithms . Another inspiration of choosing this topic was with the current coronavirus situation which displayed circulation of some fake news which causes unnecessary anxiety and worry in an already bad situation.

The next task performed was collecting a good dataset followed by applying a classifier and seeing the prediction results. We stored the files on Amazon S3, made an EMR cluster and connected locally to this to perform and get the analysis results.

The used AWS services in this whole project are as follows:

- Amazon S3:It is basically used to store and retrieve any amount of data from anywhere from the web at any time. This can be done by creating buckets and adding files. You can also set any level of privacy permissions and metadata.
- Amazon EMR: It is a cluster platform that makes running big data frameworks simpler to process huge amounts of data. These frameworks include apache spark, hadoop, etc. While making a cluster you can mention your required software and hardware configuration. This service is reliable, flexible, scalable and secured.
- Amazon EC2: This service is responsible for providing scalable computing capacity. Even if you do not have required hardware to process the type of difficult data, EC2 solves it all. It helps to launch as many virtual resources as required.

DATASET

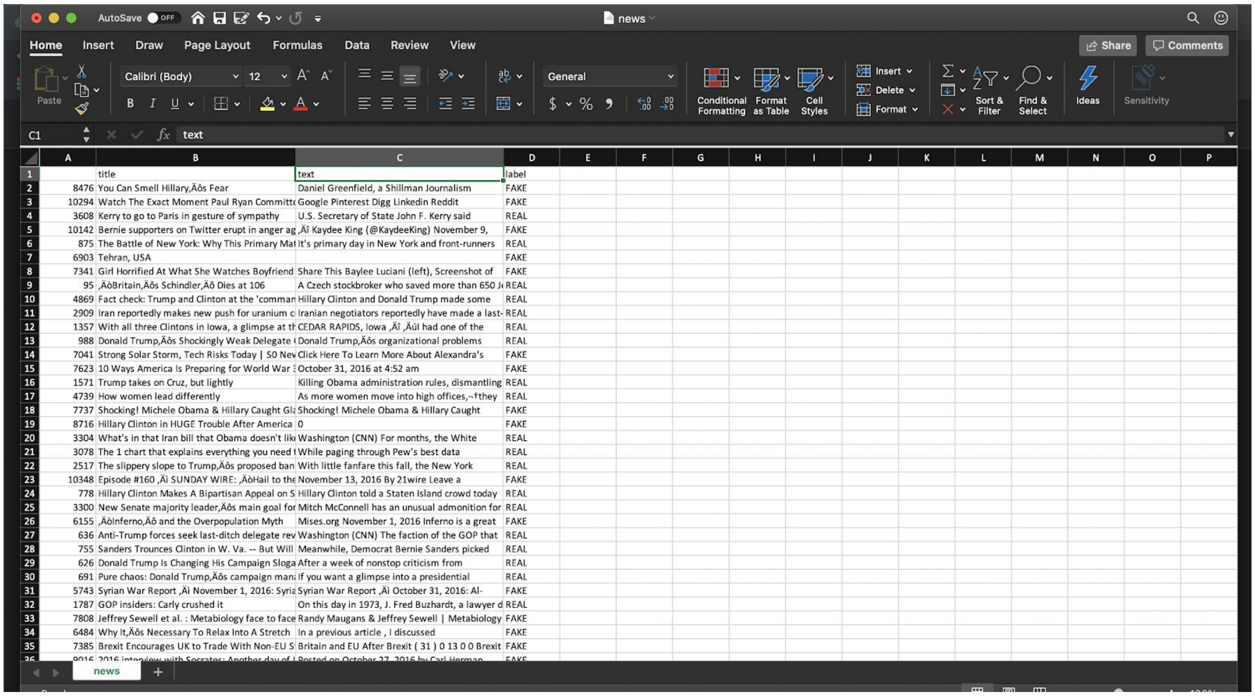
DATA DESCRIPTION

The data set we downloaded is news article texts. These texts include legitimate as well as faux news. The analysis we are planning on doing on this dataset is predicting from these text excerpts whether it is real or fake. We plan on using passive aggressive classifier on this for prediction. We will divide the data into training and testing sets and see the results in the end. We downloaded a huge dataset of 30 mb having shape 7798 x 4. The following columns were present:

- The first column included identification numbers
- The second had title of the news
- The third column consisted of actual news text
- The final column was label whether it is REAL or FAKE.

DETAILED STEPS

We searched for the dataset and found a good one on kaggle. The format of the dataset is csv(comma separated values). It helped as it meant more structured data with headers. It also meant simple hassle free processing using any application. We downloaded this dataset in our system and then uploaded it to aws S3. This is the snapshot of the dataset:



	A	B	C	D
1		title	text	label
2	8476	You Can Smell Hillary‚Äôs Fear	Daniel Greenfield, a Shillman Journalism	FAKE
3	10294	Watch The Exact Moment Paul Ryan Committe	Google Pinterest Digg Linkedin Reddit	FAKE
4	3608	Kerry to go to Paris in gesture of sympathy	U.S. Secretary of State John F. Kerry said	REAL
5	10142	Bernie supporters on Twitter erupt in anger ag	‚Äî Kaydee King (@KaydeeKing) November 9,	FAKE
6	875	The Battle of New York: Why This Primary Mat	It's primary day in New York and front-runners	REAL
7	6903	Tehran, USA		FAKE
8	7341	Girl Horrified At What She Watches Boyfriend	Share This Baylee Luciani (left), Screenshot of	FAKE
9	95	‚ÄòBritain‚Äôs Schindler‚Äô Dies at 106	A Czech stockbroker who saved more than 650 J	REAL
10	4869	Fact check: Trump and Clinton at the 'comman	Hillary Clinton and Donald Trump made some	REAL
11	2909	Iran reportedly makes new push for uranium c	Iranian negotiators reportedly have made a last-	REAL
12	1357	With all three Clintons in Iowa, a glimpse at th	CEDAR RAPIDS, Iowa ‚Äî ‚ÄúI had one of the	REAL
13	988	Donald Trump‚Äôs Shockingly Weak Delegate (	Donald Trump‚Äôs organizational problems	REAL
14	7041	Strong Solar Storm, Tech Risks Today | S0 Nev	Click Here To Learn More About Alexandra's	FAKE
15	7623	10 Ways America Is Preparing for World War 3	October 31, 2016 at 4:52 am	FAKE
16	1571	Trump takes on Cruz, but lightly	Killing Obama administration rules, dismantling	REAL
17	4739	How women lead differently	As more women move into high offices,¬†they	REAL
18	7737	Shocking! Michele Obama & Hillary Caught Gla	Shocking! Michele Obama & Hillary Caught	FAKE
19	8716	Hillary Clinton in HUGE Trouble After America	0	FAKE
20	3304	What's in that Iran bill that Obama doesn't like	Washington (CNN) For months, the White	REAL
21	3078	The 1 chart that explains everything you need t	While paging through Pew's best data	REAL
22	2517	The slippery slope to Trump‚Äôs proposed ban	With little fanfare this fall, the New York	REAL
23	10348	Episode #160 ‚Äì SUNDAY WIRE: ‚ÄòHail to the	November 13, 2016 By 21wire Leave a	FAKE
24	778	Hillary Clinton Makes A Bipartisan Appeal on S	Hillary Clinton told a Staten Island crowd today	REAL
25	3300	New Senate majority leader‚Äôs main goal for	Mitch McConnell has an unusual admonition for	REAL
26	6155	‚ÄòInferno‚Äô and the Overpopulation Myth	Mises.org November 1, 2016 Inferno is a great	FAKE
27	636	Anti-Trump forces seek last-ditch delegate rev	Washington (CNN) The faction of the GOP that	REAL
28	755	Sanders Trounces Clinton in W. Va. -- But Will	Meanwhile, Democrat Bernie Sanders picked	REAL
29	626	Donald Trump Is Changing His Campaign Sloga	After a week of nonstop criticism from	REAL
30	691	Pure chaos: Donald Trump‚Äôs campaign mana	If you want a glimpse into a presidential	REAL
31	5743	Syrian War Report ‚Äì November 1, 2016: Syria	Syrian War Report ‚Äì October 31, 2016: Al-	FAKE
32	1787	GOP insiders: Carly crushed it	On this day in 1973, J. Fred Buzhardt, a lawyer d	REAL
33	7808	Jeffrey Sewell et al. : Metabiology face to face	Randy Maugans & Jeffrey Sewell | Metabiology	FAKE
34	6484	Why It‚Äôs Necessary To Relax Into A Stretch	In a previous article , I discussed	FAKE
35	7385	Brexit Encourages UK to Trade With Non-EU St	Britain and EU After Brexit (31) 0 13 0 0 Brexit	FAKE

STORAGE OF DATA ON CLOUD

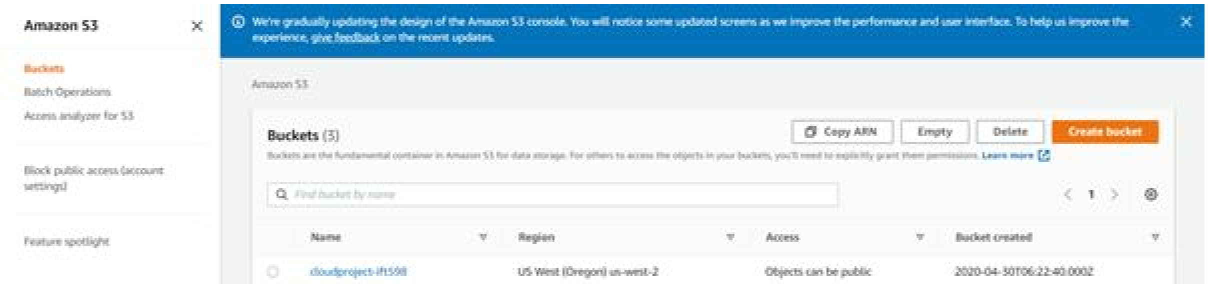
METHOD OF STORAGE

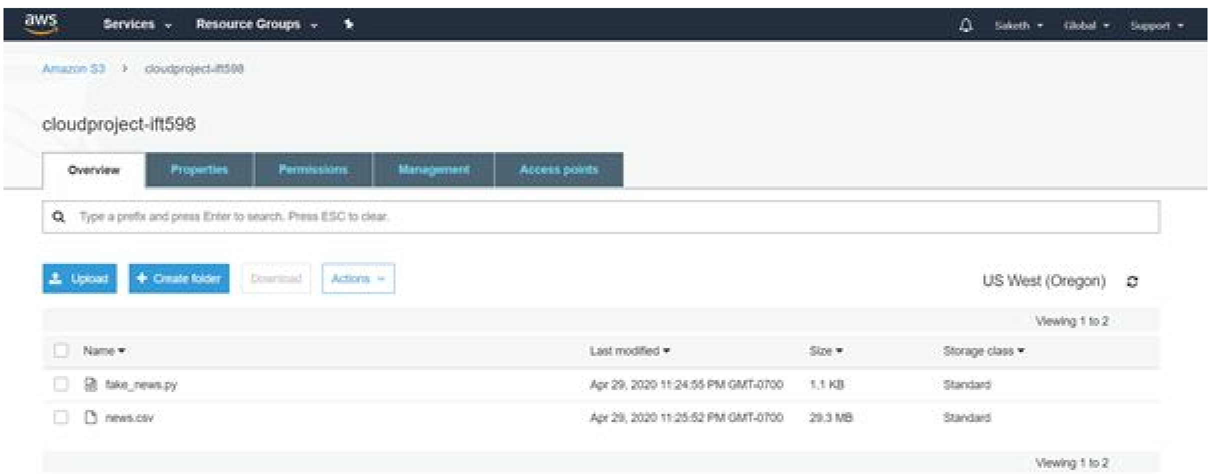
The best method applicable for our use case was to use the simple yet efficient S3. We did not have to make a new database, analysis of available data and focusing on the analysis method was the goal. The cleaning of data was also not required. We chose S3 because it made retrieval and storage of huge size possible. It also allowed us to keep it with our required permission settings. One of the most important points was that it required no cost. It fit nicely in our use case along with using Spark on EMR.

DETAILED STEPS

1. First, we signed into our aws account.
2. In services, we chose S3
3. We clicked on the **Create Bucket** option.

4. Next, named the bucket as **cloudproject-ift598** and unchecked the option to block it for public access. This is unchecked so that there is no problem while accessing the files uploaded on the bucket.
5. Next, uploaded the dataset and our python script having code for analysis.





DISTRIBUTED DATA PROCESSING SYSTEM

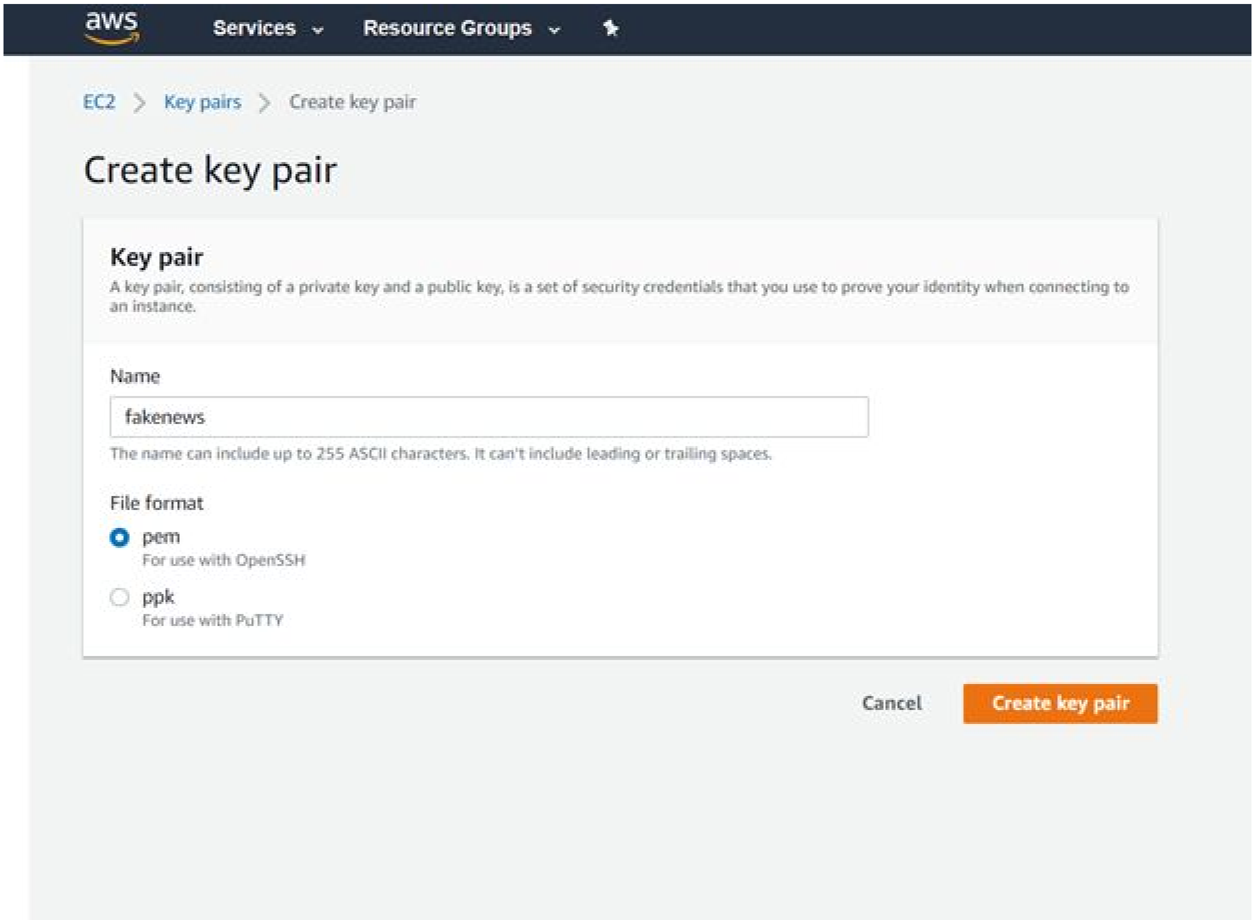
METHOD USED

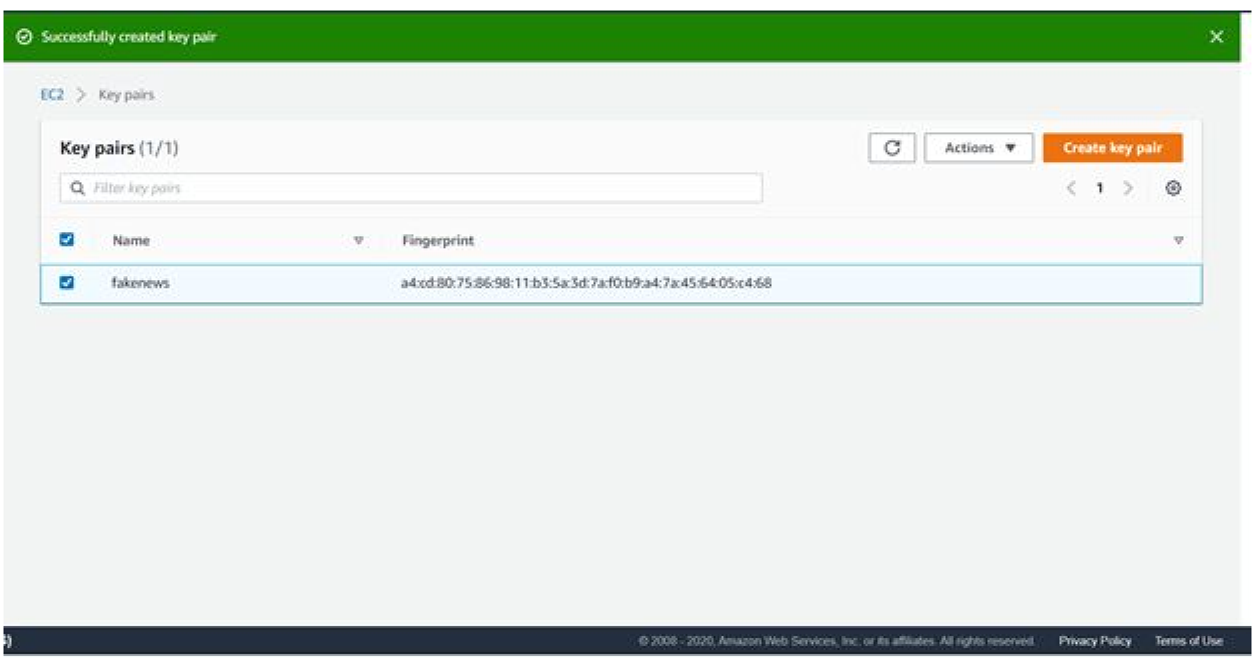
The data processing system is an important part as it is the deciding factor of which system will fit for use cases and optimize the whole process of analysis. We had the option of using Apache Hadoop, Storm, Spark. We chose to use Apache Spark as we wanted to exhaust the high processing speed advantage feature of it. The efficiency of Spark providing more types of computation calculations power was useful for our project. We had to perform tf idf vectorization, stop words removal, etc for which Spark would give better overall advantage.

DETAILED STEPS TO BUILD THE SYSTEM

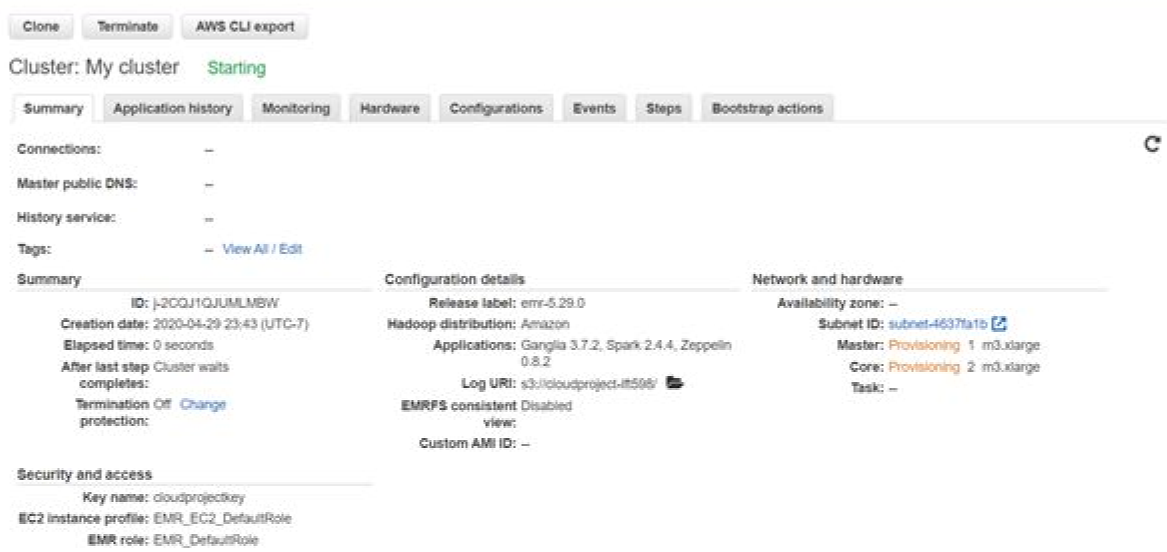
The series of steps we followed were:

1. Opened the services tab and clicked on **EMR**.
2. Clicked on the **create cluster** button.
3. We kept the name as default 'My cluster'.
4. Checked the option of logging.
5. For the storage, specified the path of the S3 bucket we created above where files were stored.
6. This was all General configuration. Under Software configuration, selected **Spark**.
7. Under hardware configuration, selected **ms.xlarge**. This assigned EC2 instances for us which included one master and two core nodes.
8. Then we created a new key pair by following the steps provided to do so in a new popup window.

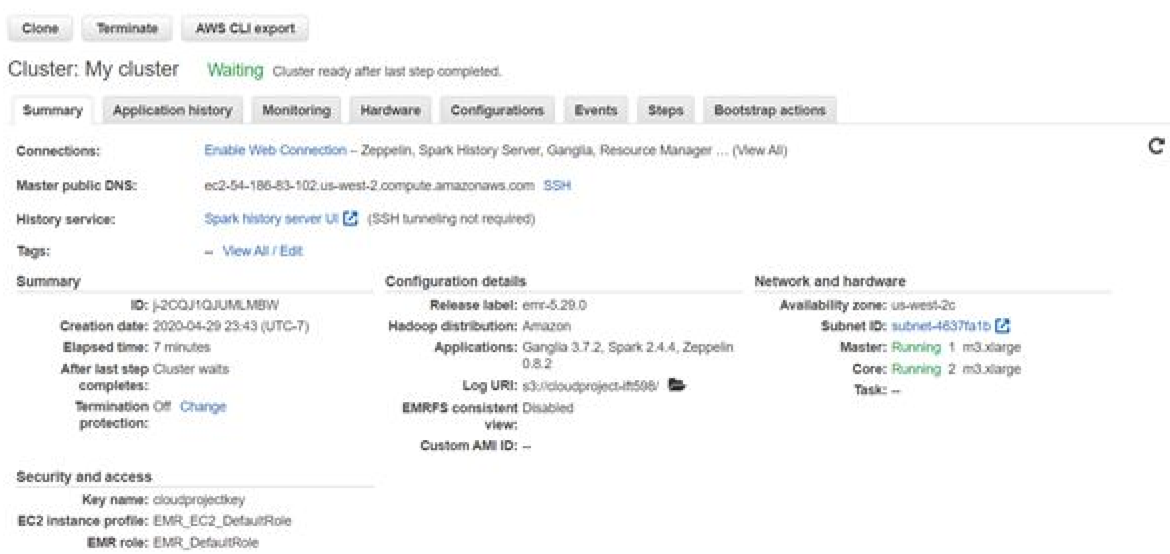

aws
Services
Resource Groups
EC2 > Key pairs > Create key pair
Create key pair
Key pair
A key pair, consisting of a private key and a public key, is a set of security credentials that you use to prove your identity when connecting to an instance.
Name
fakenews
The name can include up to 255 ASCII characters. It can't include leading or trailing spaces.
File format
pem
For use with OpenSSH
ppk
For use with PuTTY
Cancel
Create key pair

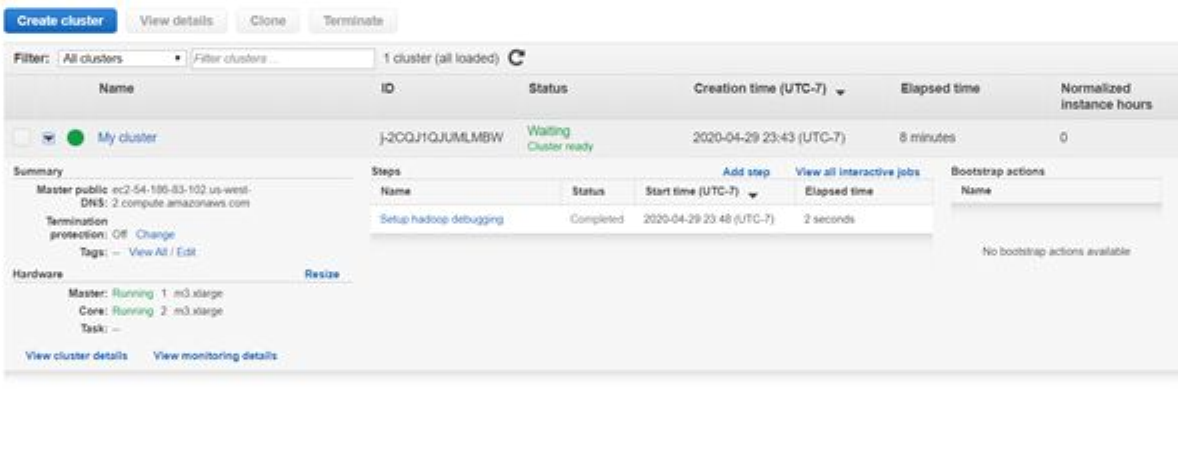

Successfully created key pair
EC2 > Key pairs
Key pairs (1/1)
Actions
Create key pair
Filter key pairs
Name
Fingerprint
fakenews
a4:cd:80:75:86:98:11:b3:5a:3d:7a:f0:b9:a4:7a:45:64:05:c4:68
Privacy Policy
Terms of Use

9. After successfully creating the key pair and keeping it in a secure place, we selected the name of the key pair we created to be associated with our cluster.
10. Then we clicked on the **create cluster** button.
11. Then we waited for our resources to be allocated.

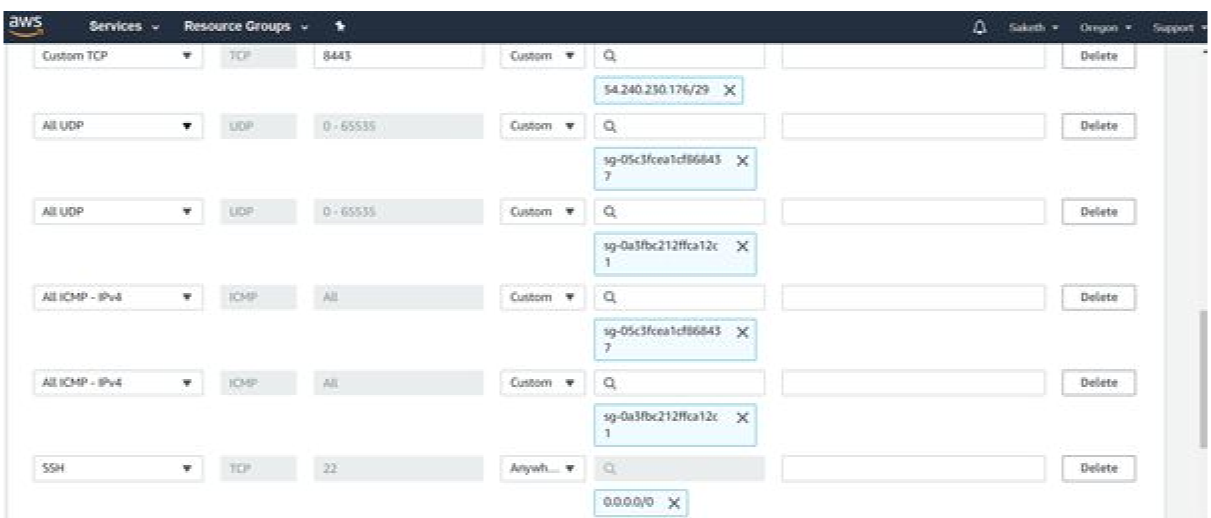


12. Successful creation of the cluster showed status change from **Starting** to **Waiting**, **Cluster ready**.

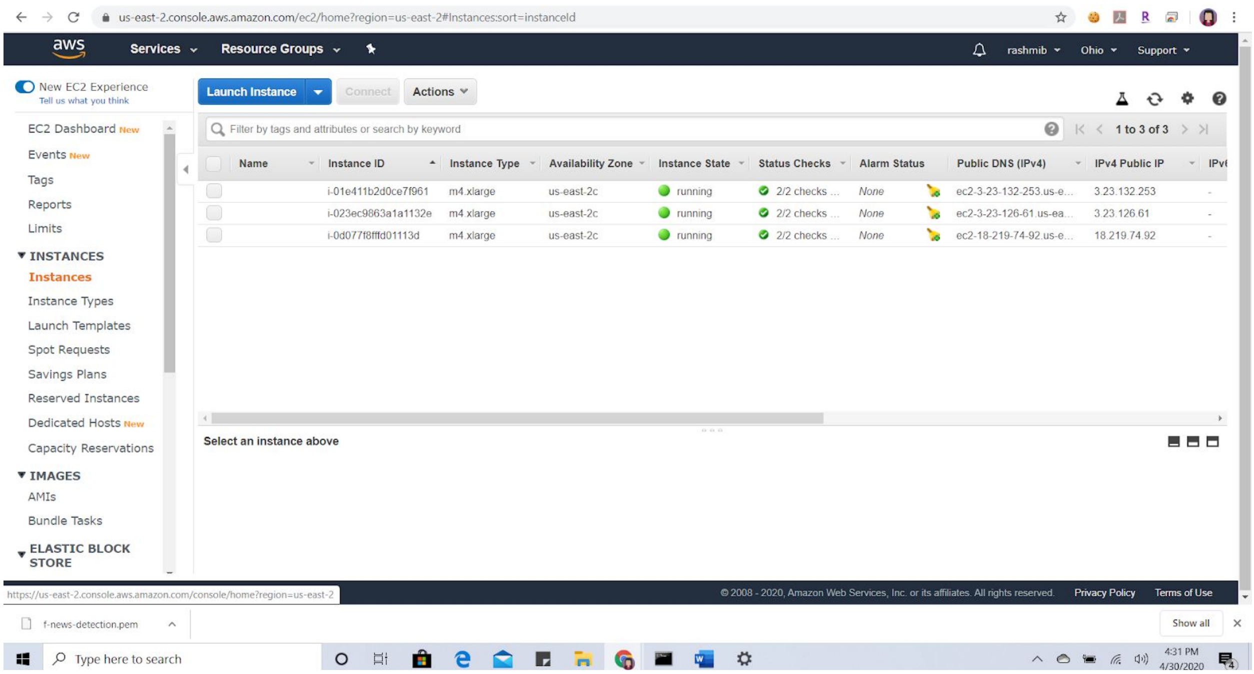




13. Then, in the summary, we clicked on the **security group of master** under **Security and Access**. This popped open a new window showing the security groups.
14. We opened this to edit the inbound rules. We needed to add an **SSH rule** so that we could maintain communication between the instance and our EMR cluster. For that reason, we kept the IP source as **anywhere**.



This shows the view of 1 master and two core nodes created after successful creation of the EMR cluster:



We made sure that these were running and then proceeded forward to perform the analysis of data.

DATA ANALYSIS

DATA ANALYSIS MOTIVE

The analysis that we performed on the data is to find whether a given news is fake or not. Inorder to do this, we ran our dataset against a python Machine Learning script. We have the dataset divided into a test set which already has the ground truth values that is whether the news is fake or not. We analyzed the news against our trained model and saw what the model predicted. The accuracy of the model was measured in terms of the number of correct predictions on the total predictions.

Inorder to do the data analysis, we first connected to our remote EC2 machine. The Elastic Map Reduce of the AWS is a distributed algorithm based process using Hadoop. This EMR is very useful in processing large datasets. The same dataset took comparatively longer processing time in the local machine than the EC2 remote machine. EMR that we used comprises the S3 storage system and a cluster of EC2 instances - a master node and two slave nodes. This parallel processing across multiple nodes increases processing speed.

Inorder to connect to the remote machine we ssh into the machine using the secure shell. The process to connect to the remote machine is described below:

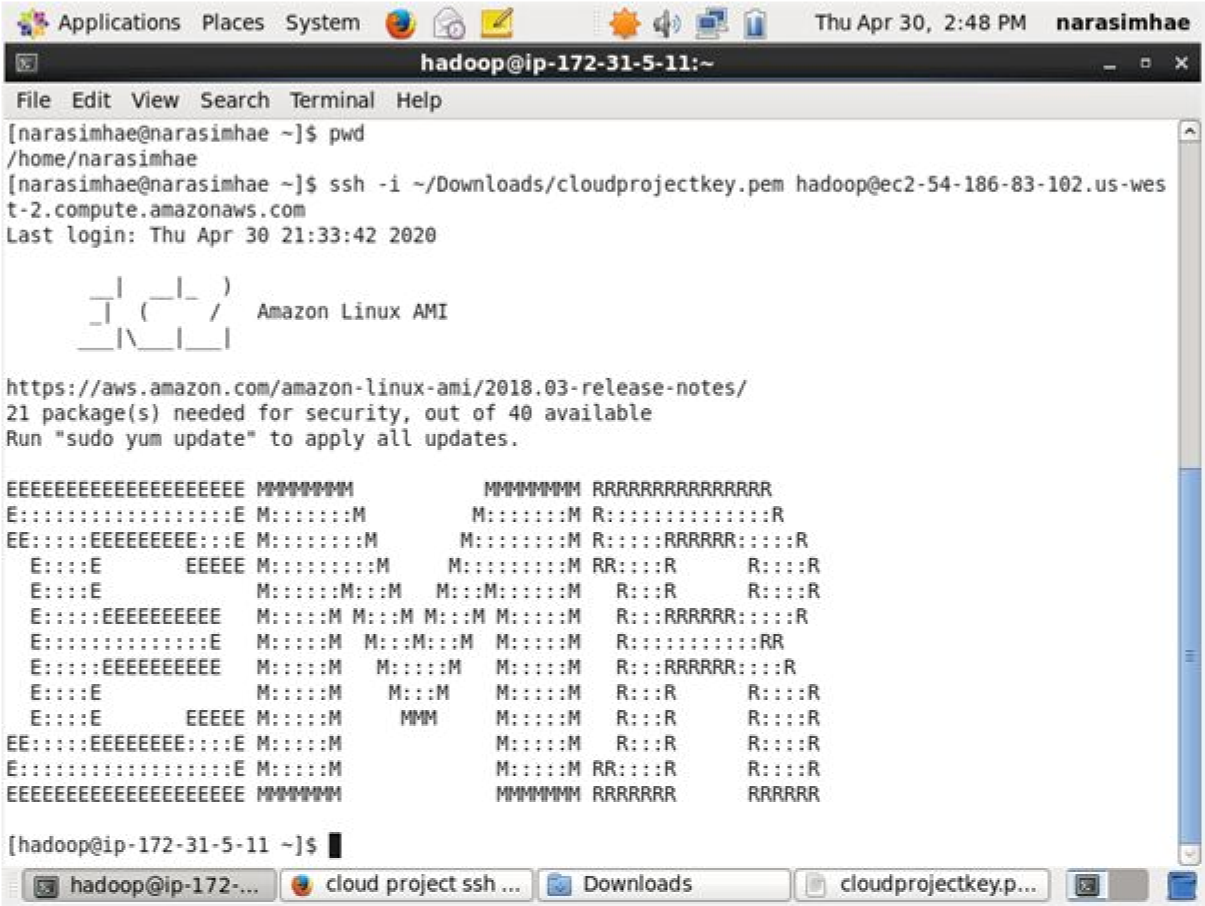
STEPWISE PROCESS TO SSH INTO EC2

Step 1: Open the linux terminal.
Step 2: Give the following command:

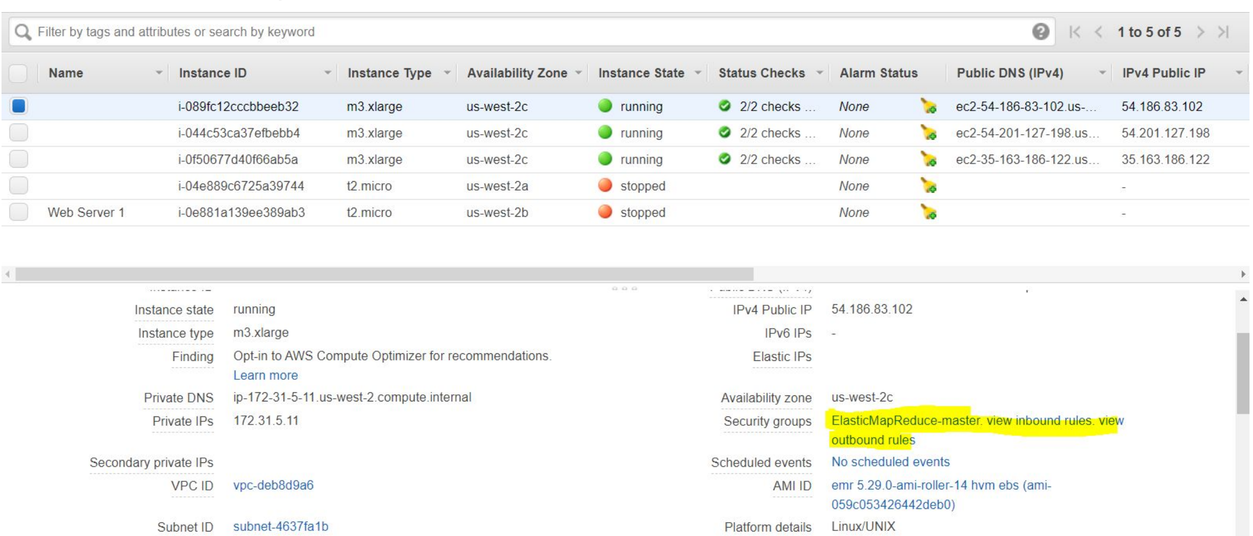
Ssh -i ~/Downloads/cloudprojectkey.pem
hadoop@ec2-54-186-83-102.us-west-2.compute.amazonaws.com

Step 3: Type yes to the warning that is thrown. That does not cause any issue.

After successful login to the EC2 machine in the EMR, we get the following confirmation.



The three running instances in EMR are shown below



RESULTS

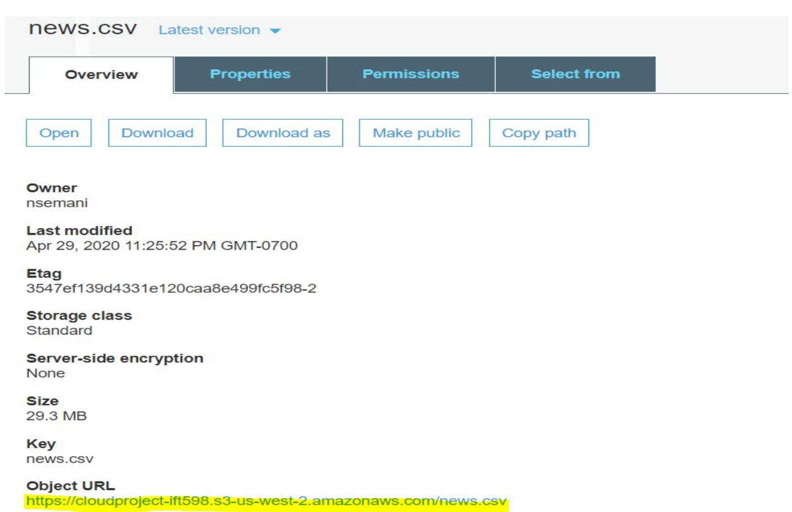
Once the data analysis system is ready, our job is now to see the results. Inorder to see the results, we run our python script with the associated data in the EC2 remote machine. Inorder to do that, first we need to copy our files from S3 to the EC2 machine.

COPYING FILES FROM S3 TO EC2 USING WGET

That we can do using wget command. The wget command copies the files form the S3 storage link to the EC2 instance. The following command does the needful.

wget https://<url of the file>

Both the data file and the python script using the above command.



Both files copied to the remote machine on cloud

```
[hadoop@ip-172-31-5-11 ~]$ pwd
/home/hadoop
[hadoop@ip-172-31-5-11 ~]$ ls
fake_news.py  news.csv
[hadoop@ip-172-31-5-11 ~]$
```

INSTALLATIONS OF NECESSARY PACKAGES IN THE EC2 INSTANCE

Python

```
  Installing : python35-3.5.7-1.25.amzn1.x86_64
  Installing : python35-setuptools-36.2.7-1.33.amzn1.noarch
  Installing : python35-pip-9.0.3-1.27.amzn1.noarch
  Verifying  : python35-pip-9.0.3-1.27.amzn1.noarch
  Verifying  : python35-3.5.7-1.25.amzn1.x86_64
  Verifying  : python35-setuptools-36.2.7-1.33.amzn1.noarch
  Verifying  : python35-libs-3.5.7-1.25.amzn1.x86_64

Installed:
  python35-pip.noarch 0:9.0.3-1.27.amzn1

Dependency Installed:
  python35.x86_64 0:3.5.7-1.25.amzn1            python35-libs.x86_64 0:3
  python35-setuptools.noarch 0:36.2.7-1.33.amzn1

Complete!
[root@ip-172-31-5-11 hadoop]# python
Python 2.7.16 (default, Oct 14 2019, 21:26:56)
[GCC 4.8.5 20150623 (Red Hat 4.8.5-28)] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>> exit()
[root@ip-172-31-5-11 hadoop]# sudo alternatives --set python /usr/bin/python3
[root@ip-172-31-5-11 hadoop]# python --version
Python 3.5.7
[root@ip-172-31-5-11 hadoop]#
```

Numpy and Sklearn

```
[root@ip-172-31-5-11 hadoop]# pip install sklearn
Collecting sklearn
  Downloading https://files.pythonhosted.org/packages/1e/7a/dbb3be0ce9bd5c8b7
b2b1bfa4774cb1147bfcd3f/sklearn-0.0.tar.gz
Collecting scikit-learn (from sklearn)
  Downloading https://files.pythonhosted.org/packages/42/ec/32310181e803f5d22
d08cf5b6e116a93a6a5d1c6/scikit_learn-0.22.2.post1-cp35-cp35m-manylinux1_x86_6
    100% |████████████████████████████████| 7.0MB 189kB/s
Collecting scipy>=0.17.0 (from scikit-learn->sklearn)
  Downloading https://files.pythonhosted.org/packages/c1/60/8cbf00c0deb50a971
2867df979870a454481817c/scipy-1.4.1-cp35-cp35m-manylinux1_x86_64.whl (26.0MB)
    100% |████████████████████████████████| 26.0MB 46kB/s
Collecting joblib>=0.11 (from scikit-learn->sklearn)
  Downloading https://files.pythonhosted.org/packages/28/5c/cf6a2b65a321c4a20
62661f8f6f4bb28547cf1bf/joblib-0.14.1-py2.py3-none-any.whl (294kB)
    100% |████████████████████████████████| 296kB 4.4MB/s
Requirement already satisfied: numpy>=1.11.0 in /usr/lib64/python3.5/dist-pac
earn->sklearn)
Installing collected packages: scipy, joblib, scikit-learn, sklearn
  Running setup.py install for sklearn ... done
Successfully installed joblib-0.14.1 scikit-learn-0.22.2.post1 scipy-1.4.1 sk
You are using pip version 9.0.3, however version 20.1 is available.
You should consider upgrading via the 'pip install --upgrade pip' command.
```

Pandas

```
[root@ip-172-31-5-11 hadoop]# pip install pandas
Collecting pandas
  Downloading https://files.pythonhosted.org/packages/a9/55/e3f34ad611f703454
92994cebc4d8e0ec0af38c4/pandas-0.25.3-cp35-cp35m-manylinux1_x86_64.whl (10.3M
    100% |████████████████████████████████| 10.3MB 123kB/s
Requirement already satisfied: numpy>=1.13.3 in /usr/lib64/python3.5/dist-pac
Collecting pytz>=2017.2 (from pandas)
  Using cached https://files.pythonhosted.org/packages/4f/a4/879454d49688e2fa
3c745fd2ec2a3adf87b0808d/pytz-2020.1-py2.py3-none-any.whl
Collecting python-dateutil>=2.6.1 (from pandas)
  Using cached https://files.pythonhosted.org/packages/d4/70/d60450c3dd48ef87
0b306af2bce5d134d78615cb/python_dateutil-2.8.1-py2.py3-none-any.whl
Collecting six>=1.5 (from python-dateutil>=2.6.1->pandas)
  Downloading https://files.pythonhosted.org/packages/65/eb/1f97cb97bfc2390a2
f5058082d4cb10c6c5c1dba/six-1.14.0-py2.py3-none-any.whl
Installing collected packages: pytz, six, python-dateutil, pandas
Successfully installed pandas-0.25.3 python-dateutil-2.8.1 pytz-2020.1 six-1.
You are using pip version 9.0.3, however version 20.1 is available.
You should consider upgrading via the 'pip install --upgrade pip' command.
```

PYTHON CODE

```
# Importing necessary Packages
import numpy as np
import pandas as pd
import itertools
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import PassiveAggressiveClassifier
from sklearn.metrics import accuracy_score, confusion_matrix

# Converting the .csv file to pandas dataframe
df=pd.read_csv('C:/Users/saket/Cloud_Project/news.csv')
df.shape
df.head()

labels=df.label
labels.head()

# dividing the dataset into train and test dataset
x_train,x_test,y_train,y_test=train_test_split(df['text'], labels, test_size=0.2, random_state=7)

# Creating the tfidf vectorizer to convert the news
# articles to vectors for mathematical calculations and deriving relation
tfidf_vectorizer=TfidfVectorizer(stop_words='english', max_df=0.7)

tfidf_train=tfidf_vectorizer.fit_transform(x_train)
tfidf_test=tfidf_vectorizer.transform(x_test)

#Passive aggressive classifier model created and trained based on the training vectors created above
pac=PassiveAggressiveClassifier(max_iter=50)
pac.fit(tfidf_train,y_train)
```

```
# Predicting the output of the test data and calculating the accuracy
y_pred=pac.predict(tfidf_test)
score=accuracy_score(y_test,y_pred)
print(y_pred)
print(f'Accuracy: {round(score*100,2)}%')

confusion_matrix(y_test,y_pred, labels=['FAKE','REAL'])
print(str(confusion_matrix))
```

Note: In the above code, instead of the local path of the data file, the remote system's path is replaced when run on cloud. It is done using vim editor.

Code in vim editor in EC2 machine

hadoop@ip-172-31-5-11:~

File Edit View Search Terminal Help

```python
import numpy as np
import pandas as pd
import itertools
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import PassiveAggressiveClassifier
from sklearn.metrics import accuracy_score, confusion_matrix



df=pd.read_csv('/home/hadoop/news.csv')
df.shape
df.head()


labels=df.label
labels.head()


x_train,x_test,y_train,y_test=train_test_split(df['text'], labels, test_size=0.2, random_state=7)


tfidf_vectorizer=TfidfVectorizer(stop_words='english', max_df=0.7)
tfidf_train=tfidf_vectorizer.fit_transform(x_train)
tfidf_test=tfidf_vectorizer.transform(x_test)


pac=PassiveAggressiveClassifier(max_iter=50)
pac.fit(tfidf_train,y_train)
y_pred=pac.predict(tfidf_test)
score=accuracy_score(y_test,y_pred)
print("PREDICTIONS")
print(y_pred)
print("\nAccuracy:",round(score*100,2),"%")

print("\nCONFUSION MATRIX")
confusion_matrix = confusion_matrix(y_test,y_pred, labels=['FAKE','REAL'])
print(confusion_matrix)
```

OUTPUT

The Prediction array, the accuracy and the confusion matrix are analyzed and printed in the output

```
[hadoop@ip-172-31-5-11 ~]$ python fake_news.py
PREDICTIONS
['REAL' 'FAKE' 'REAL' ... 'REAL' 'FAKE' 'REAL']

Accuracy: 92.82 %

CONFUSION MATRIX
[[591  47]
 [ 44 585]]
```

CONCLUSION

In summary, this project helped us understand how the EMR helps do parallel processing and can thus improve the efficiency and save run time while dealing with large datasets. Though the difference is not significant in this project we could still identify the efficiency and could extrapolate how this works with large datasets.The course in total gave us a very clear understanding of the basics of Cloud, AWS in specific. We understood how the cloud functions, what a remote machine is and how to create such instances in the cloud. The best part of the course is that we got hands on experience with multiple labs that helped while working on this project.

REFERENCES

Aws documentation:
https://docs.aws.amazon.com/

Sklearn documentation:
https://scikit-learn.org/stable/user_guide.html

Stackoverflow:
https://stackoverflow.com/questions

Kaggle:
https://www.kaggle.com/