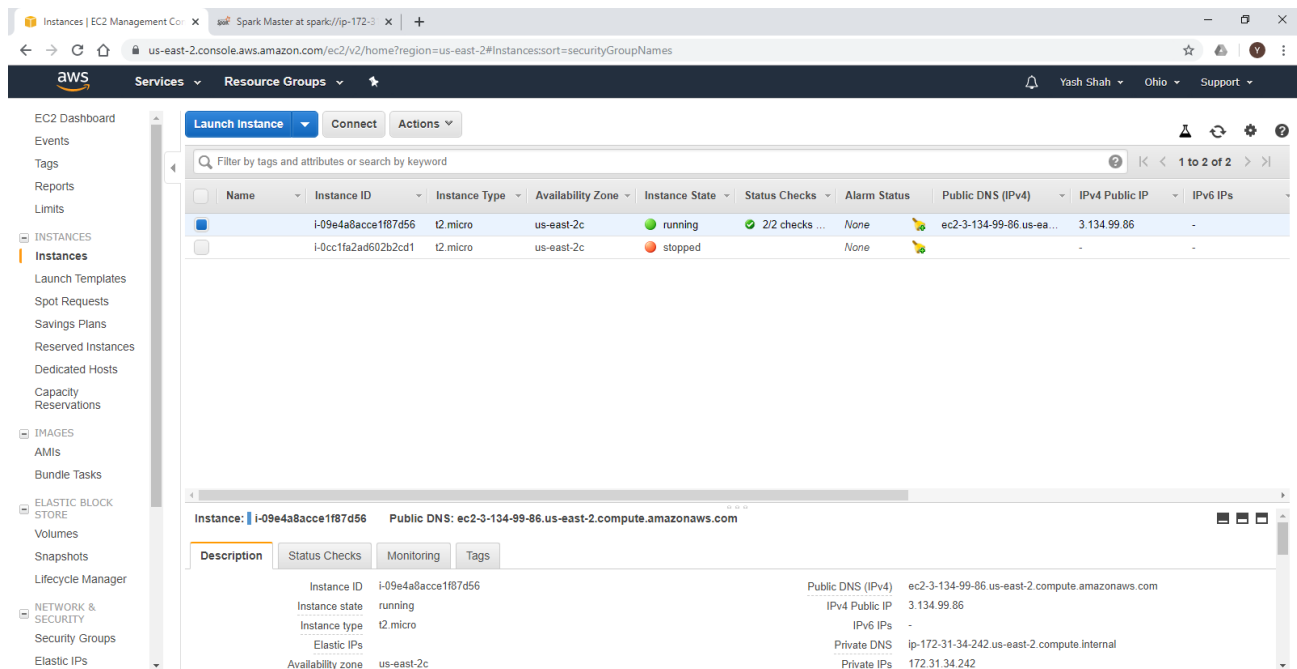


Assignment Report

- Cloud Setup:

- Starting from the basic; the first step in setting up the cloud storage was to create an account on Amazon Website Service (AWS).
- After selecting the most suitable instance type 'Ubuntu Server 18.04 LTS (HVM), SSD Volume Type', I configured it to make the most of this instance.
- For better security, AWS uses asymmetric encryption concept and provides us with a key for secure login. Software named 'PuTTY' is used to create the public and private key of the AWS instance.
- To allow interaction with the instance, I had enabled few ports (8080, 8081, 80, 22) in the inbound rules section under 'Security Groups' option.



1. My Cloud Setup with a running instance

- Post this, in order to run 'Apache Spark' on my instance, I installed Java and Python as the first step. Then, I created a directory 'server' and installed 'Apache Spark 2.4.4' using the following commands – 'wget <http://apache.forsale.plus/spark/spark-2.4.4/spark-2.4.4-bin-hadoop2.7.tgz>' and 'sudo tar zxvf spark-2.4.3-bin-hadoop2.7.tgz'.
- After this, I made some changes in the file name 'profile' in order to use Java and Python with Spark.
- Later, I started a Spark master node and slave node and then I started the Spark shell to run my MapReduce query.

MongoDB Compass Community - localhost:27017/Assignment-2.Twitter

Connect View Collection Help

My Cluster

DBS COLLECTIONS

filter

Assignment-2

Clean-News

Clean-Twitter

News

Twitter

Data_for_MongoDB

admin

config

local

localhost:27017 STANDALONE

MongoDB 4.0.5 Community

Assignment-2.Twitter

DOCUMENTS 4.3k TOTAL SIZE 1.5MB AVG. SIZE 356B INDEXES 1 TOTAL SIZE 96.0KB AVG. SIZE 96.0KB

Documents Aggregations Explain Plan Indexes

FILTER

INSERT DOCUMENT VIEW LIST TABLE

Displaying documents 1 - 20 of 4279

	_id ObjectId	Time String	Username String	Location Mixed	Tweet Content String
1	5dc39233edf4102f642e6d88	"2019-11-06 20:14:58"	"Rob"	null	"@Avilzco lol england is bad en
2	5dc39233edf4102f642e6d89	"2019-11-06 20:14:58"	"Ace_Bastone"	null	"@Gertrud29409036 @Bradhall386
3	5dc39233edf4102f642e6d8a	"2019-11-06 20:14:57"	"Karin Litzcke"	"Vancouver, BC, Canada"	"@RHDefence There is truly no e
4	5dc39233edf4102f642e6d8b	"2019-11-06 20:14:56"	"Kaiserkriss"	"Calgary, Alberta"	"@VoltaireFreedom @JustinTrudea
5	5dc39233edf4102f642e6d8c	"2019-11-06 20:14:55"	"Екатерина Напачанова"	"Санкт-Петербург, Россия"	"RT @HarvelStudios: Oh snap! Av
6	5dc39233edf4102f642e6d8d	"2019-11-06 20:14:54"	"Don Bacon"	null	"@LevinePMC @NationalPost Canad
7	5dc39233edf4102f642e6d8e	"2019-11-06 20:14:53"	"Terry Martin"	"Vancouver british columbia"	"Yup , and Canada is 1st and cl
8	5dc39233edf4102f642e6d8f	"2019-11-06 20:14:53"	"Belieber ❤️"	null	"RT @Revan2Roxi: Today marks 10
9	5dc39233edf4102f642e6d90	"2019-11-06 20:14:52"	"Gill Stella"	null	"@AamerAnwar 🇵🇰 for kind attn: @
10	5dc39233edf4102f642e6d91	"2019-11-06 20:14:50"	"Shelley Dalton"	"St. John's"	"RT @CBCHL: NEW The cost to h
11	5dc39233edf4102f642e6d92	"2019-11-06 20:14:50"	"girls love beyonce"	"Accident, MD"	"@Lynklix @Slowmortgages @urba
12	5dc39233edf4102f642e6d93	"2019-11-06 20:14:50"	"Екатерина Напачанова"	"Санкт-Петербург, Россия"	"RT @Avengers: Oh snap! @Avenge
13	5dc39233edf4102f642e6d94	"2019-11-06 20:14:50"	"JJ Whelan"	null	"@SteveUBC @ElizabethWay Should
14	5dc39233edf4102f642e6d95	"2019-11-06 20:14:49"	"Musicaum 🇨🇦"	"Calgary, Canada"	"@UncleReel Jeezer. Canada Fede
15	5dc39233edf4102f642e6d96	"2019-11-06 20:14:49"	"Aviva Anders 🇺🇸"	"Nashville & 📧 to you"	"These are always harder than t
16	5dc39233edf4102f642e6d97	"2019-11-06 20:14:49"	"Never mind."	null	"RT @sidney_chest: Lets ban Fre
17	5dc39233edf4102f642e6d98	"2019-11-06 20:14:49"	"Eric Lawton"	"Pontypool ON. See pinned tweet"	"RT @katiepheakoe: So I read t
18	5dc39233edf4102f642e6d99	"2019-11-06 20:14:47"	"Vipul Bhatia"	"New Delhi, India"	"RT @AskAnshul: Punjab needs Se
19	5dc39233edf4102f642e6d9a	"2019-11-06 20:14:46"	"It be like that sometimes"	null	"RT @HarvelStudios: Oh snap! Av

3.Unformatted Twitter Dataset

MongoDB Compass Community - localhost:27017/Assignment-2.Clean-Twitter

Connect View Collection Help

My Cluster

DBS COLLECTIONS

filter

Assignment-2

Clean-News

Clean-Twitter

News

Twitter

Data_for_MongoDB

admin

config

local

localhost:27017 STANDALONE

MongoDB 4.0.5 Community

Assignment-2.Clean-Twitter

DOCUMENTS 4.3k TOTAL SIZE 1.4MB AVG. SIZE 346B INDEXES 1 TOTAL SIZE 96.0KB AVG. SIZE 96.0KB

Documents Aggregations Explain Plan Indexes

FILTER

INSERT DOCUMENT VIEW LIST TABLE

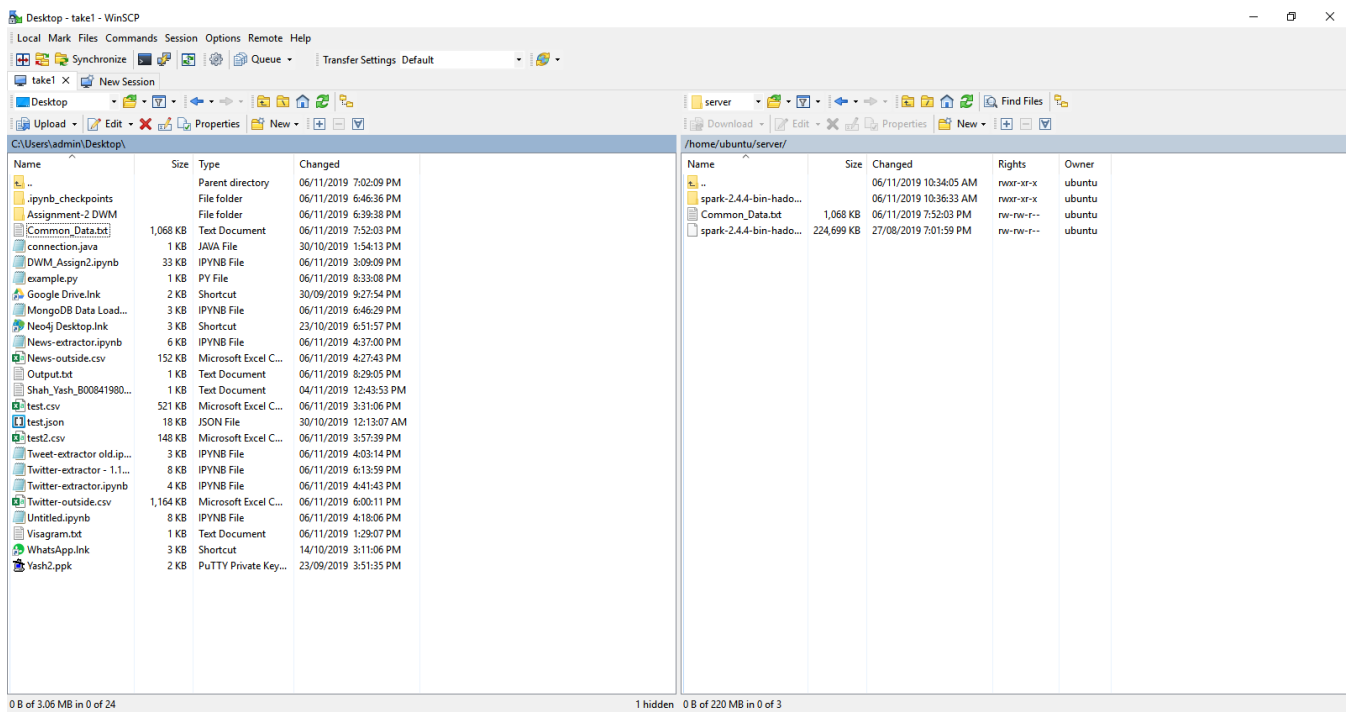
Displaying documents 1 - 20 of 4279

	_id ObjectId	Time String	Username String	Location Mixed	Tweet Content String
1	5dc39234edf4102f642e8035	"2019-11-06 20:14:58"	"Rob"	null	"@Avilzco lol england is bad en
2	5dc39234edf4102f642e8036	"2019-11-06 20:14:58"	"Ace Bastone"	null	"@Gertrud29409036 @Bradhall386
3	5dc39234edf4102f642e8037	"2019-11-06 20:14:57"	"Karin Litzcke"	"Vancouver BC Canada"	"@RHDefence There is truly no e
4	5dc39234edf4102f642e8038	"2019-11-06 20:14:56"	"Kaiserkriss"	"Calgary Alberta"	"@VoltaireFreedom @JustinTrudea
5	5dc39234edf4102f642e8039	"2019-11-06 20:14:55"	" "	" "	"RT @HarvelStudios: Oh snap Ave
6	5dc39234edf4102f642e803a	"2019-11-06 20:14:54"	"Don Bacon"	null	"@LevinePMC @NationalPost Canad
7	5dc39234edf4102f642e803b	"2019-11-06 20:14:53"	"Terry Martin"	"Vancouver british columbia"	"Yup and Canada is 1st and clos
8	5dc39234edf4102f642e803c	"2019-11-06 20:14:53"	"Belieber "	null	"RT @Revan2Roxi: Today marks 10
9	5dc39234edf4102f642e803d	"2019-11-06 20:14:52"	"Gill Stella"	null	"@AamerAnwar For kind attn: @Du
10	5dc39234edf4102f642e803e	"2019-11-06 20:14:50"	"Shelley Dalton"	"St John s"	"RT @CBCHL: NEW The cost to hou
11	5dc39234edf4102f642e803f	"2019-11-06 20:14:50"	"girls love beyonce"	"Accident MD"	"@Lynklix @Slowmortgages @urba
12	5dc39234edf4102f642e8040	"2019-11-06 20:14:50"	" "	" "	"RT @Avengers: Oh snap Avengers
13	5dc39234edf4102f642e8041	"2019-11-06 20:14:50"	"JJ Whelan"	null	"@SteveUBC @ElizabethWay Should
14	5dc39234edf4102f642e8042	"2019-11-06 20:14:49"	"Musicaum "	"Calgary Canada"	"@UncleReel Jeezer. Canada Feder
15	5dc39234edf4102f642e8043	"2019-11-06 20:14:49"	"Aviva Anders "	"Nashville to you"	"These are always harder than t
16	5dc39234edf4102f642e8044	"2019-11-06 20:14:49"	"Never mind "	null	"RT @sidney chest: Lets ban Fre
17	5dc39234edf4102f642e8045	"2019-11-06 20:14:49"	"Eric Lawton"	"Pontypool ON See pinned tweet"	"RT @katiepheakoe: So I read t
18	5dc39234edf4102f642e8046	"2019-11-06 20:14:47"	"Vipul Bhatia"	"New Delhi India"	"RT @AskAnshul: Punjab needs Se
19	5dc39234edf4102f642e8047	"2019-11-06 20:14:46"	"It be like that sometimes"	null	"RT @HarvelStudios: Oh snap Ave

4.Cleaned Twitter Dataset

Query Execution:

- Only the tweets/retweets and news article content were transferred into a text file named 'Common_Data.txt'.
- This text file was uploaded to my cloud instance using a software 'WinSCP'.
- The MapReduce query was executed on this file and the output was stored in another text file named 'Output.txt'.



5. Operating window of WinSCP

- Note:

- File used for extraction of tweets is named 'Twitter-extractor.ipynb' and that used for news article is named 'News-extractor.ipynb'.
- File used for cleaning the csv files is named 'Content-cleaner.ipynb'.
- File used for loading data to MongoDB is named 'MongoDB Data Loader.ipynb'.
- The unformatted csv files are named 'Twitter.csv' and 'News.csv'. The cleaned csv files are named 'Clean-Twitter.csv' and 'Clean-News.csv'.

- References:

- "API Reference for Tweepy", Joshua Roesslein, 2019. [Online]. Available: <http://docs.tweepy.org/en/v3.8.0/api.html#api-reference>. [Accessed: 04 November, 2019].
- "Extracting Twitter Data, Preprocessing and Sentiment Analysis using Python 3.0", Dilan Jayasekara, 2019. [Online]. Available: <https://towardsdatascience.com/extracting-twitter-data-pre-processing-and-sentiment-analysis-using-python-3-0-7192bd8b47cf>. [Accessed: 04 November, 2019].