Institute of Business Administration

Project report

Instructor: Dr. Tariq Mehmood

Submitted by: Mehak Jamil Khatri (23653)
Spring'22

Mehak Jamil Khatri
23653

Download the data from:

https://www.kaggle.com/datasets/jannalipenkova/covid19-public-media-dataset

SPARKSQL

Download the project from

https://github.com/big-data-europe/docker-hadoop-spark-workbench

Go to the yml directory and type this on cmd

```
docker-compose up –d
```

```
Digest: sha256:95b2b8b2e622b474851cb362934592fc22e520f608e891f9aa4be45032b23cc
Status: Downloaded newer image for bde2020/hdfs-filebrowser:3.11
Creating spark-notebook                                  ... done
Creating spark-master                                    ... done
Creating docker-hadoop-spark-workbench-master_hue_1      ... done
Creating namenode                                        ... done
Creating docker-hadoop-spark-workbench-master_datanode_1     ... done
Creating docker-hadoop-spark-workbench-master_spark-worker_1 ... done
```

```
root@4cc49ee2e0a2:/# hadoop fs -mkdir /user
root@4cc49ee2e0a2:/# hadoop fs -put -f /home/data /user/data
root@4cc49ee2e0a2:/# hadoop fs -ls /user/data
Found 1 items
-rw-r--r--   3 root supergroup 2802728054 2022-06-01 13:33 /user/data/covid19_articles.csv
root@4cc49ee2e0a2:/#
```

```
root@4cc49ee2e0a2:/home/data# /spark/bin/spark-shell --master=local spark://spark-master:7077
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/06/01 13:54:18 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
22/06/01 13:54:25 WARN metastore.ObjectStore: Version information not found in metastore. hive.metastore.schema.verification is not enabled so recordin
g the schema version 1.2.0
22/06/01 13:54:25 WARN metastore.ObjectStore: Failed to get database default, returning NoSuchObjectException
22/06/01 13:54:27 WARN metastore.ObjectStore: Failed to get database global_temp, returning NoSuchObjectException
Spark context Web UI available at http://172.21.0.4:4040
Spark context available as 'sc' (master = local, app id = local-1654091658990).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.1.2-SNAPSHOT
      /_/

Using Scala version 2.11.8 (OpenJDK 64-Bit Server VM, Java 1.8.0_121)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

Write following commands on scala to load data from hdfs

```
scala> val dataFrame1 = spark.read.format("csv").option("header","true").load("hdfs://namenode:8020/user/data/covid19_articles.csv");
dataFrame1: org.apache.spark.sql.DataFrame = [_c0: string, author: string ... 8 more fields]
```

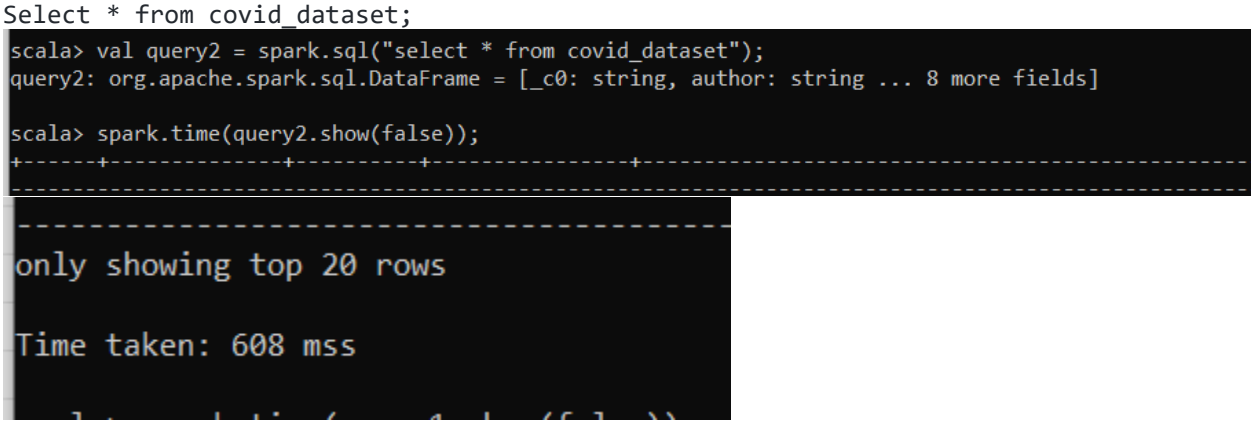
Set the name of the table

```
scala> dataFrame1.createOrReplaceTempView("covid_dataset");
```

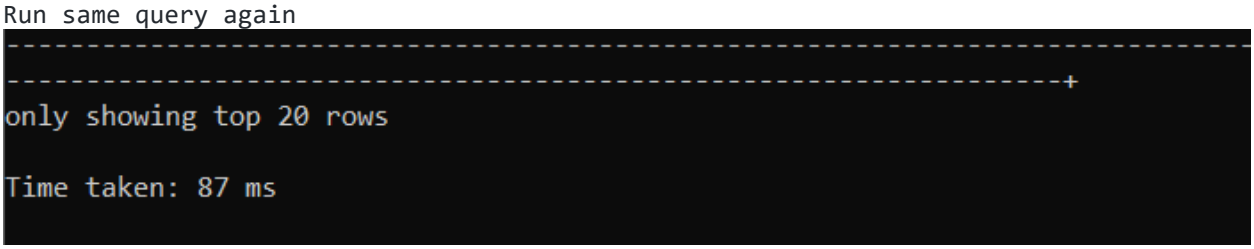
Execute the query

```
Val query1 = spark.sql(“select * from covid_dataset limit 50”);
Spark.time(query1.show(false));
```

Select * from covid_dataset;



Run same query again



Select * from covid_dataset where domain=’digitaljournal’

Mehak Jamil Khatri
23653

```
scala> val scType1 = spark.sql("select * from covid_dataset where domain='digitaljournal'");
scType1: org.apache.spark.sql.DataFrame = [_c0: string, author: string ... 8 more fields]

scala> spark.time(scType1.show(false));
+------+------+----------+-------------+------------------------------------------------------------
|_c0   |author|date      |domain       |title                                                         |content
                                                                                          |datatype
                                                      |companies
                                                  |locations
                                                                  |sdgs
          |
+------+------+----------+-------------+------------------------------------------------------------
|389108|null  |2020-01-02|digitaljournal|Looking into the truth about modern workplace environments                    |"['Hi, what are you looking for?', 'By', 'Published', 'Workplaces are being transf
ormed, according to the Olivet Nazarene University study ( “ The Truth about Modern Offices “). The days of men and women fitted in formal clothing and working under the harsh glare fluorescent lighting are
slipping away. The trends of today are reflective of casual dress codes and soft furnishings.', 'There are signs too that the ‘ animal penning’ ( read cubicles) are starting to be withdrawn, and replaced wit
h open floor plans ( indications are that over half of workplaces – 51 percent – have gravitated towards the ‘ open office’ or open floor plan concept).', 'The survey also showed that 77 percent of U.S. loca
ted workers are reportedly happy with the way their office is set up. However, as an indication of thoughts about open-plan areas, those with private offices stated they were happiest', 'Through changing tec
hnology, more employees are now having more daily conversations via messaging apps rather than face-to-face interactions ( for every 17 conversations, nine are via messaging systems against eight that are fa
ce-to-face). An example of such a system is Skype for Business.', 'The poll also found that 58 percent of U.S. citizens self-report they are less productive when working from home. The main reasons for this
loss of productivity include distractions, trouble signing off at the end of the day and difficulty communicating and collaborating with coworkers. Furthermore, 80 percent of those ‘ homeworking’ freely admi
tted to multitasking. By this, people are choosing to focus on things like laundry, family, or television. So while remote working takes off, there are some apparent downsides to this trend.', 'The data was
compiled between November 15 and November 17, 2019. Encompassed in the survey were 2,009 people, who were polled about their current workplace environments. The mean age of the participants was 37 years. By
gender, 55 percent of the respondents were female and 45 percent were male.', ""Dr. Tim Sandle is Digital Journal's Editor-at-Large for science news. Tim specializes in science
```

```
----------------------------------------------------------------------------------------------------+
only showing top 20 rows

Time taken: 115 ms

scala>
```

Select * from covid_dataset where author is null limit 5;

```
scala> spark.time(scType4.show(false));
+------+
|author|
+------+
|null  |
|null  |
|null  |
|null  |
|null  |
+------+

Time taken: 122 ms
```

```
Select * from covid_dataset where author is not null limit 5;
```

```
scala> val scType4 = spark.sql("select author From covid_dataset where author is not null limit 5");
scType4: org.apache.spark.sql.DataFrame = [author: string]

scala> spark.time(scType4.show(false));
+--------------+
|author        |
+--------------+
|Thomas Hughes |
|Helen Ziatyk  |
|Roberto Torres|
|Thomas Hughes |
|Jenna Fletcher|
+--------------+

Time taken: 110 ms
```

```
val scType5 = spark.sql("select title From covid_dataset where date between '2020-01-02' and '2020-02-02'");
```

```
scala> val scType5 = spark.sql("select title From covid_dataset where date between '2020-01-02' and '2020-02-02'");
scType5: org.apache.spark.sql.DataFrame = [title: string]

scala> spark.time(scType5.show(false));
+-----------------------------------------------------------------------+
|title                                                                  |
+-----------------------------------------------------------------------+
|Looking into the truth about modern workplace environments             |
|Hexo refiles financial statements                                      |
|Japan raid, Turkey arrests in widening Ghosn probe                     |
|Pope's bodyguards criticised over slapping incident                    |
|Lebanon denies president welcomed fugitive Ghosn                       |
|Lebanese lawyers want Ghosn prosecuted over Israel trip                |
|' I did it alone ', Ghosn says of Japan escape                         |
|Ghosn escape sparks calls to toughen Japan's bail system               |
|Mexico's Santiago River has become a toxic 'environmental hell '       |
|Fired British vegan files landmark discrimination case                 |
|US places ban on Cuban defense chief                                   |
|Medicinal cannabis substitute can treat Parkinson's disease            |
|2020 trends to watch in US beverage                                    |
|Madonna rings in the New Year, ready for another decade of dreams      |
|Three Industrial Giants You Should Own In 2020                         |
|My experience of surviving cancer twice                                |
|Alcohol-free craft beer from Partake looks to fill whitespace of US market|
|Buonafide 0.0 imports alcohol-free wine from Italy                     |
|Oil prices soar as US kills top Iranian general, fans war fears        |
|On the road to AI adoption, execs grapple with expertise and data      |
+-----------------------------------------------------------------------+
only showing top 20 rows

Time taken: 165 ms
```

```
val scType6 = spark.sql("select title From covid_dataset where date like '2020-01-%'");
```

```
scala> val scType6 = spark.sql("select title From covid_dataset where date like '2020-01-%'");
scType6: org.apache.spark.sql.DataFrame = [title: string]

scala> spark.time(scType6.show(false));
+-----------------------------------------------------------------------+
|title                                                                  |
+-----------------------------------------------------------------------+
|Looking into the truth about modern workplace environments             |
|Hexo refiles financial statements                                      |
|Japan raid, Turkey arrests in widening Ghosn probe                     |
|Pope's bodyguards criticised over slapping incident                    |
|Lebanon denies president welcomed fugitive Ghosn                       |
|Lebanese lawyers want Ghosn prosecuted over Israel trip                |
|' I did it alone ', Ghosn says of Japan escape                         |
|Ghosn escape sparks calls to toughen Japan's bail system               |
|Mexico's Santiago River has become a toxic 'environmental hell '       |
|Fired British vegan files landmark discrimination case                 |
|US places ban on Cuban defense chief                                   |
|Medicinal cannabis substitute can treat Parkinson's disease            |
|2020 trends to watch in US beverage                                    |
|Madonna rings in the New Year, ready for another decade of dreams      |
|Three Industrial Giants You Should Own In 2020                         |
|My experience of surviving cancer twice                                |
|Alcohol-free craft beer from Partake looks to fill whitespace of US market|
|Buonafide 0.0 imports alcohol-free wine from Italy                     |
|Oil prices soar as US kills top Iranian general, fans war fears        |
|On the road to AI adoption, execs grapple with expertise and data      |
+-----------------------------------------------------------------------+
only showing top 20 rows

Time taken: 87 ms
```

Mehak Jamil Khatri
23653

```
val scType7 = spark.sql("show columns from covid_dataset");
```

```
scala> val scType7 = spark.sql("show columns from covid_dataset");
scType7: org.apache.spark.sql.DataFrame = [col_name: string]

scala> spark.time(scType7.show(false));
+---------+
|col_name |
+---------+
|_c0      |
|author   |
|date     |
|domain   |
|title    |
|content  |
|datatype |
|companies|
|locations|
|sdgs     |
+---------+

Time taken: 84 ms
```

```
val scType8 = spark.sql("describe covid_dataset");
```

```
scala> val scType8 = spark.sql("describe covid_dataset");
scType8: org.apache.spark.sql.DataFrame = [col_name: string, data_type: string ... 1 more field]

scala> spark.time(scType8.show(false));
+---------+---------+-------+
|col_name |data_type|comment|
+---------+---------+-------+
|_c0      |string   |null   |
|author   |string   |null   |
|date     |string   |null   |
|domain   |string   |null   |
|title    |string   |null   |
|content  |string   |null   |
|datatype |string   |null   |
|companies|string   |null   |
|locations|string   |null   |
|sdgs     |string   |null   |
+---------+---------+-------+

Time taken: 26 ms
```

Mehak Jamil Khatri
23653

Apache Drill

Download git project from https://github.com/smizy/docker-apache-drill

```
D:\big_data\docker-apache-drill-master>docker network create vnet
6e5e769c2c3cbc1f5b6ed6442dd2a780efb493bda4b4d6a93b9f44eea4e36aaf

D:\big_data\docker-apache-drill-master>
```

Docker-compose up –d

```
D:\big_data\docker-apache-drill-master>Docker-compose up -d
Pulling zookeeper-1 (smizy/zookeeper:3.4-alpine)...
3.4-alpine: Pulling from smizy/zookeeper
Image docker.io/smizy/zookeeper:3.4-alpine uses outdated schema1 manifest format. Please upgrade to a schema2 image for
better future compatibility. More information at https://docs.docker.com/registry/spec/deprecated-schema-v1/
709515475419: Pull complete
a3ed95caeb02: Pull complete
99676f001172: Pull complete
b56553ecb14e: Pull complete
c9e594a31794: Pull complete
Digest: sha256:a6b5e73f3fbd31e50f205c020adf0974745265ace6b1173a30cd813f36619f5e
Status: Downloaded newer image for smizy/zookeeper:3.4-alpine
Pulling drillbit-1 (smizy/apache-drill:1.16.0-alpine)...
1.16.0-alpine: Pulling from smizy/apache-drill
5a3ea8efae5d: Pull complete
54c7015d798d: Pull complete
a0b94afdd1f3: Pull complete
1b12b77f6e9d: Pull complete
b3580b1e4cfc: Pull complete
Digest: sha256:a857b53a531b36cfbe2e9f6800f5672c1dd6a65471ceae2fa7118442b73655e1
Status: Downloaded newer image for smizy/apache-drill:1.16.0-alpine
Creating zookeeper-1 ... done
Creating drillbit-1  ... done

D:\big_data\docker-apache-drill-master>
```

Copy the data into the container

```
D:\big_data\docker-apache-drill-master>docker cp D:\big_data\data e6647e058d68:dataset

D:\big_data\docker-apache-drill-master>
```

Bash into the container

```
D:\big_data\docker-apache-drill-master>docker exec -it -u hdfs datanode-1 /bin/bash
bash-4.3$ cd ..
bash-4.3$ cd ..
bash-4.3$ ls
bin      include  lib      libexec  local    sbin     share
bash-4.3$ cd ..
bash-4.3$ ls
bin      dataset  dev      etc      hadoop   home     lib      media    mnt      proc     root     run      sbin     srv      sys      tmp      usr      var
bash-4.3$ cd dataset
bash-4.3$ ls
covid19_articles.csv  data
bash-4.3$
```

Create directory and put data on hdfs

Mehak Jamil Khatri
23653

```
bash-4.3$ ls
covid19_articles.csv  data
bash-4.3$ hdfs dfs -mkdir -p /user/hdfs/output
bash-4.3$ hdfs dfs -ls /user/hdfs/output
bash-4.3$
```

```
bash-4.3$ hdfs dfs -mkdir -p /user/hdfs/output
bash-4.3$ hdfs dfs -put covid19_articles.csv /user/hdfs/output
bash-4.3$ hdfs dfs -ls /user/hdfs/output
Found 1 items
-rw-r--r--   3 hdfs hadoop 2802728054 2022-06-02 15:05 /user/hdfs/output/covid19_articles.csv
bash-4.3$
```

Execute drill container

```
bash-4.3$ exit
exit

D:\big_data\docker-apache-drill-master>docker exec -it drillbit-1 drill-conf
Apache Drill 1.16.0
"Drill never goes out of style."
apache drill>
```

Change the file extension to make it accessible with column name (optional)

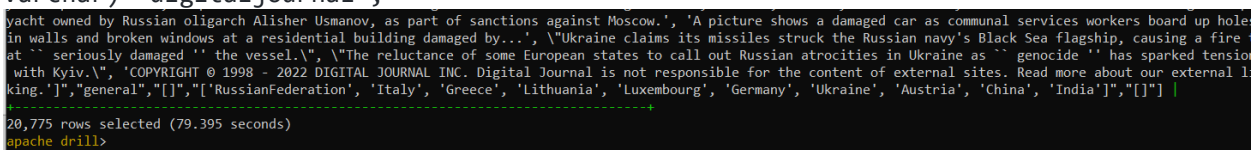
```
D:\big_data\docker-apache-drill-master>docker exec -it -u hdfs datanode-1 /bin/bash
bash-4.3$ cd ..
bash-4.3$ cd ..
bash-4.3$ ls
bin   include  lib   libexec  local  sbin  share
bash-4.3$ cd ..
bash-4.3$ ls
bin   dataset  dev   etc   hadoop  home  lib   media  mnt   proc  root  run   sbin  srv   sys   tmp   usr   var
bash-4.3$ cd dataset
bash-4.3$ ls
covid19_articles.csv  data
bash-4.3$ hdfs dfs -ls /user/hdfs/output
Found 1 items
-rw-r--r--   3 hdfs hadoop 2802728054 2022-06-02 15:05 /user/hdfs/output/covid19_articles.csv
bash-4.3$ hdfs dfs -mv /user/hdfs/output/covid19_articles.csv /user/hdfs/output/covid_dataset.csvh
bash-4.3$ hdfs dfs -ls /user/hdfs/output
Found 1 items
-rw-r--r--   3 hdfs hadoop 2802728054 2022-06-02 15:05 /user/hdfs/output/covid_dataset.csvh
bash-4.3$
```

Mehak Jamil Khatri
23653

Queries

```
select * from dfs.`/user/hdfs/output/covid19_articles.csv` limit 50;
```


```
select * from dfs.`/user/hdfs/output/covid19_articles.csv` where cast(columns[3] as varchar)='digitaljournal';
```



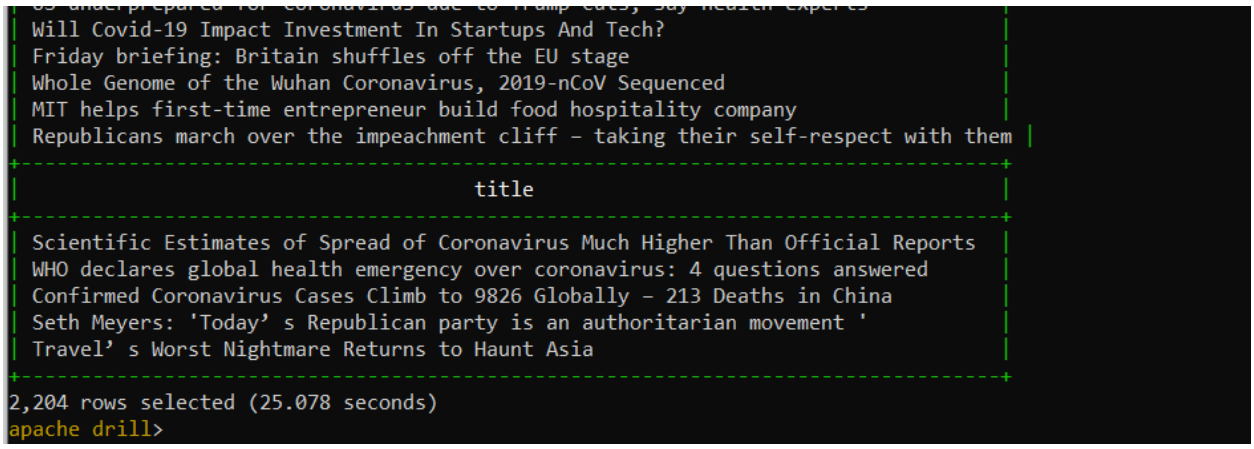

```
select columns[1] as author from dfs.`/user/hdfs/output/covid19_articles.csv` where cast(columns[1] as varchar)='' limit 5;
```


```
select columns[1] as author from dfs.`/user/hdfs/output/covid19_articles.csv` where cast(columns[1] as varchar) is not null limit 5;
```


```
select columns[4] as title from dfs.`/user/hdfs/output/covid19_articles.csv` where cast(columns[2] as varchar) between '2020-01-02' and '2020-02-02';
```

select columns[4] as title from dfs.`/user/hdfs/output/covid19_articles.csv` where cast(columns[2] as varchar) like '2020-01-%';

```
apache drill> select columns[4] as title from dfs.`/user/hdfs/output/covid19_articles.csv` where cast(columns[2] as varchar) like '2020-01-%';
+-------------------------------------------------------------------------------+
|                                     title                                     |
+-------------------------------------------------------------------------------+
| Looking into the truth about modern workplace environments                    |
| Hexo refiles financial statements                                             |
| Japan raid, Turkey arrests in widening Ghosn probe                            |
| Pope's bodyguards criticised over slapping incident                           |
```



```
| Will Covid-19 Impact Investment In Startups And Tech?                         |
| Friday briefing: Britain shuffles off the EU stage                            |
| Whole Genome of the Wuhan Coronavirus, 2019-nCoV Sequenced                    |
| MIT helps first-time entrepreneur build food hospitality company              |
| Republicans march over the impeachment cliff – taking their self-respect with them |
+-------------------------------------------------------------------------------+
|                                     title                                     |
+-------------------------------------------------------------------------------+
| Scientific Estimates of Spread of Coronavirus Much Higher Than Official Reports |
| WHO declares global health emergency over coronavirus: 4 questions answered   |
| Confirmed Coronavirus Cases Climb to 9826 Globally – 213 Deaths in China      |
| Seth Meyers: 'Today' s Republican party is an authoritarian movement '        |
| Travel' s Worst Nightmare Returns to Haunt Asia                               |
+-------------------------------------------------------------------------------+
2,204 rows selected (25.078 seconds)
apache drill>
```

- describe is not available here, reference: https://stackoverflow.com/questions/56825619/describe-table-returns-nothing
- show is not available here, reference: https://drill.apache.org/docs/show-tables/

Browse Web UI

http://yourMachine'sIPV4Address:dockerImagePort/storage/dfs

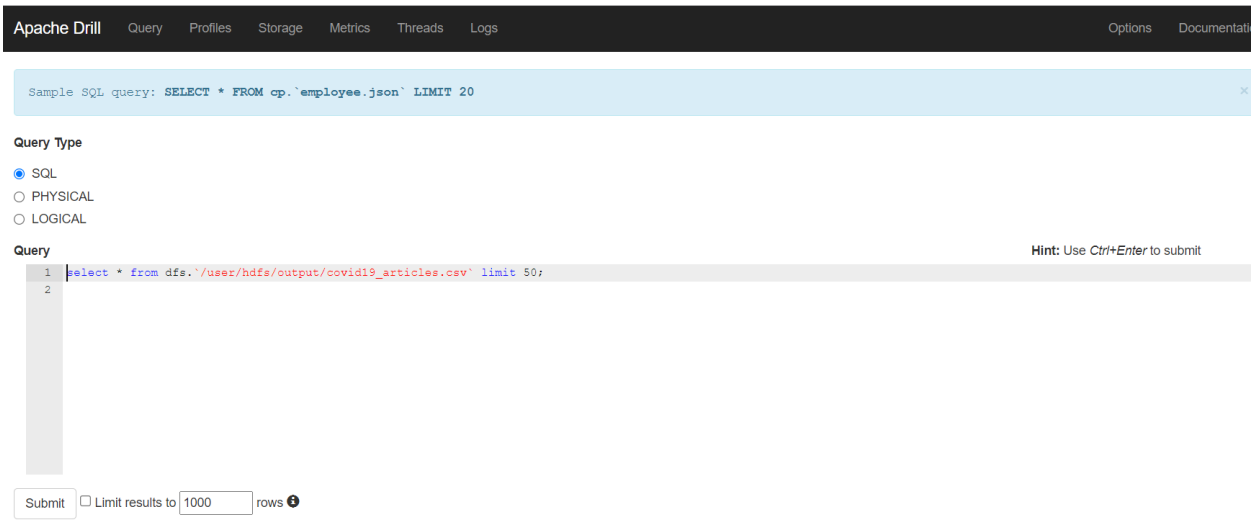
Add "enabled":true and update the connection path and then click update

```
{
  "type": "file",
  "enabled": true,
  "connection": "hdfs://namenode-1.vnet:8020",
  "config": null,
  "workspaces": {
    "tmp": {
      "location": "/tmp",
      "writable": true,
      "defaultInputFormat": null,
      "allowAccessOutsideWorkspace": false
    },
    "root": {
      "location": "/usr/hdfs",
      "writable": false,
      "defaultInputFormat": null,
      "allowAccessOutsideWorkspace": false
    }
  },
  "formats": {
    "psv": {
      "type": "text",
      "extensions": [
        "tbl"
      ],
```

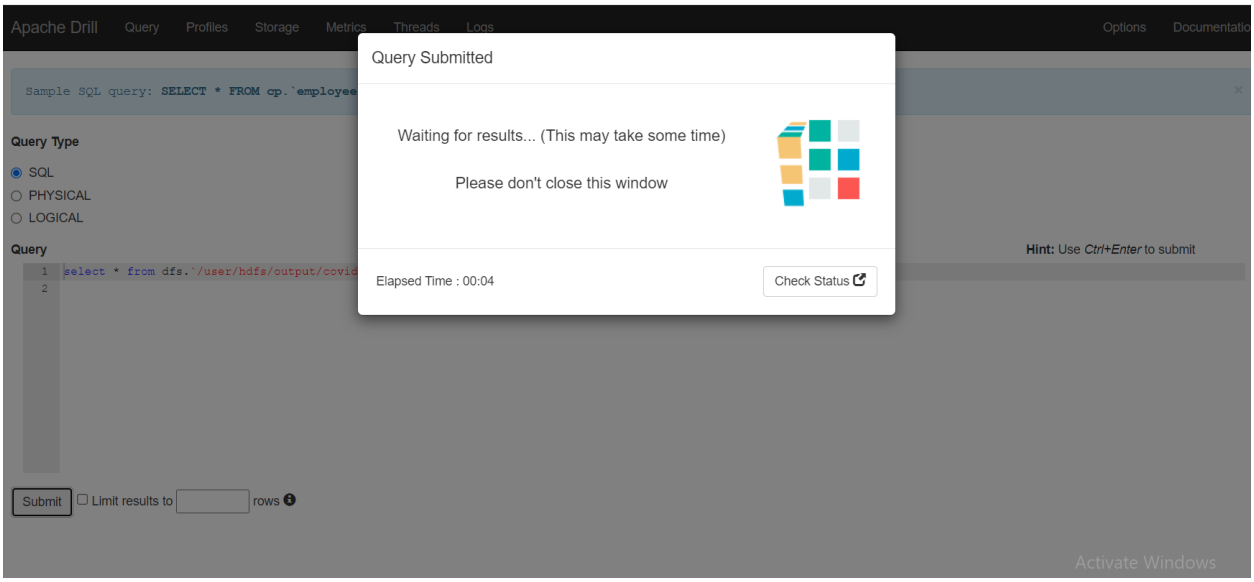
Back | Update | Disable | Export | Delete

You can execute the same queries on browser

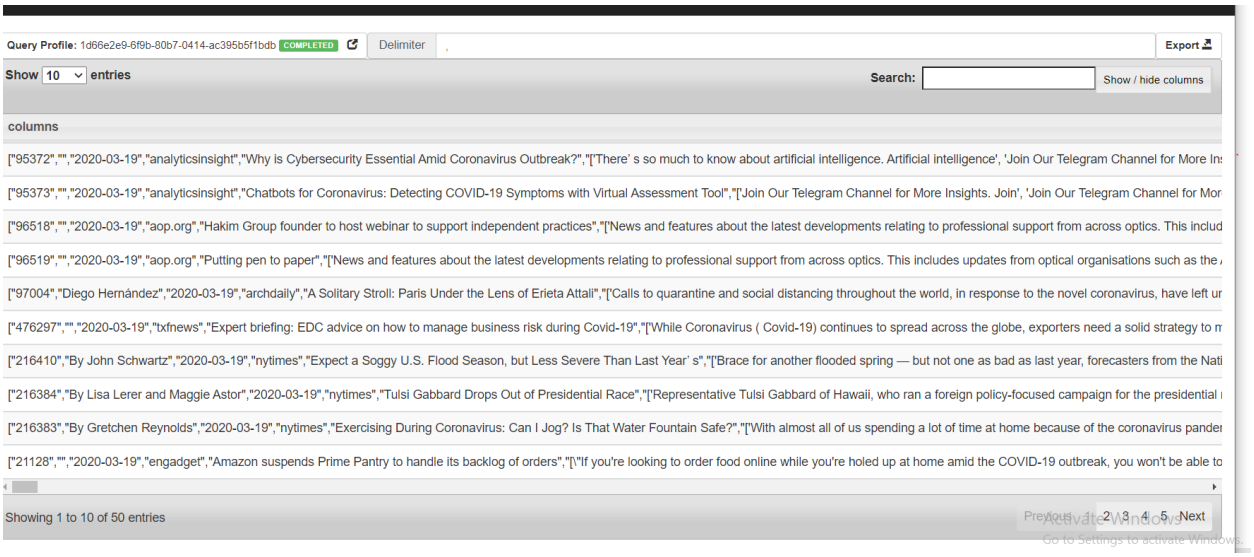
Mehak Jamil Khatri
23653



Hit submit



Mehak Jamil Khatri
23653



Query Profile: 1d66e2e9-6f9b-80b7-0414-ac395b5f1bdb COMPLETED | Delimiter , | Export

Show 10 entries | Search: | Show / hide columns

columns

["95372","","2020-03-19","analyticsinsight","Why is Cybersecurity Essential Amid Coronavirus Outbreak?","['There' s so much to know about artificial intelligence. Artificial intelligence', 'Join Our Telegram Channel for More In

["95373","","2020-03-19","analyticsinsight","Chatbots for Coronavirus: Detecting COVID-19 Symptoms with Virtual Assessment Tool","['Join Our Telegram Channel for More Insights. Join', 'Join Our Telegram Channel for Mor

["96518","","2020-03-19","aop.org","Hakim Group founder to host webinar to support independent practices","['News and features about the latest developments relating to professional support from across optics. This includ

["96519","","2020-03-19","aop.org","Putting pen to paper","['News and features about the latest developments relating to professional support from across optics. This includes updates from optical organisations such as the

["97004","Diego Hernández","2020-03-19","archdaily","A Solitary Stroll: Paris Under the Lens of Erieta Attali","['Calls to quarantine and social distancing throughout the world, in response to the novel coronavirus, have left ur

["476297","","2020-03-19","txfnews","Expert briefing: EDC advice on how to manage business risk during Covid-19","['While Coronavirus (Covid-19) continues to spread across the globe, exporters need a solid strategy to m

["216410","By John Schwartz","2020-03-19","nytimes","Expect a Soggy U.S. Flood Season, but Less Severe Than Last Year' s","['Brace for another flooded spring — but not one as bad as last year, forecasters from the Nati

["216384","By Lisa Lerer and Maggie Astor","2020-03-19","nytimes","Tulsi Gabbard Drops Out of Presidential Race","['Representative Tulsi Gabbard of Hawaii, who ran a foreign policy-focused campaign for the presidential

["216383","By Gretchen Reynolds","2020-03-19","nytimes","Exercising During Coronavirus: Can I Jog? Is That Water Fountain Safe?","['With almost all of us spending a lot of time at home because of the coronavirus pande

["21128","","2020-03-19","engadget","Amazon suspends Prime Pantry to handle its backlog of orders","[\"If you're looking to order food online while you're holed up at home amid the COVID-19 outbreak, you won't be able to

Showing 1 to 10 of 50 entries | Previous 1 2 3 4 5 Next

Mehak Jamil Khatri
23653

HIVEQL

Downloaded project from big data Europe

https://github.com/big-data-europe/docker-hive

copy data file to the container, then on the Hadoop, and then run hive

```
D:\big_data\docker-hive-master>docker cp D:\big_data\data 8ffc3f107c87:home

D:\big_data\docker-hive-master>docker compose exec hive-server bash
root@8ffc3f107c87:/opt# ls
hadoop-2.7.4  hive
root@8ffc3f107c87:/opt# cd ..
root@8ffc3f107c87:/# ls
bin  boot  dev  entrypoint.sh  etc  hadoop-data  home  lib  lib64  media  mnt  opt  proc  root  run  sbin  srv  sys  tmp  usr  var
root@8ffc3f107c87:/# cd home
root@8ffc3f107c87:/home# ls
data
root@8ffc3f107c87:/home# cd data
root@8ffc3f107c87:/home/data# ls
covid19_articles.csv
root@8ffc3f107c87:/home/data# exit
exit

D:\big_data\docker-hive-master>docker compose exec hive-server bash
root@8ffc3f107c87:/opt# hadoop fs -mkdir /user/data
mkdir: `/user/data': File exists
root@8ffc3f107c87:/opt# hadoop fs -ls /user/data
Found 1 items
drwxr-xr-x   - root supergroup          0 2022-06-01 18:52 /user/data/home
root@8ffc3f107c87:/opt# hadoop fs -ls /user/data/home
Found 3 items
-rw-r--r--   3 root supergroup 2802728054 2022-05-31 04:24 /user/data/home/covid19_articles.csv
drwxr-xr-x   - root supergroup          0 2022-06-01 18:53 /user/data/home/data
-rw-r--r--   3 root supergroup 1224216058 2022-05-31 04:24 /user/data/home/geolocation
root@8ffc3f107c87:/opt#
```



```
root@8ffc3f107c87:/# ls
bin  boot  dev  entrypoint.sh  etc  hadoop-data  home  lib  lib64  media  mnt  opt  proc  root  run  sbin  srv  sys  tmp  usr  var
root@8ffc3f107c87:/# cd home
root@8ffc3f107c87:/home# ls
data
root@8ffc3f107c87:/home# cd data
root@8ffc3f107c87:/home/data# ls
covid19_articles.csv
root@8ffc3f107c87:/home/data# /opt/hive/bin/beeline -u jdbc:hive2://localhost:10000
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hadoop-2.7.4/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Connecting to jdbc:hive2://localhost:10000
Connected to: Apache Hive (version 2.3.2)
Driver: Hive JDBC (version 2.3.2)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 2.3.2 by Apache Hive
0: jdbc:hive2://localhost:10000>
```

Create table to load data in it

```
0: jdbc:hive2://localhost:10000> create table covid_data (c0 string, author string, datee string, domain string, title
string, content string, datatype string, companies string, locations string, sdgs string) row format delimited fields t
erminated by ',';
No rows affected (0.696 seconds)
0: jdbc:hive2://localhost:10000>
```

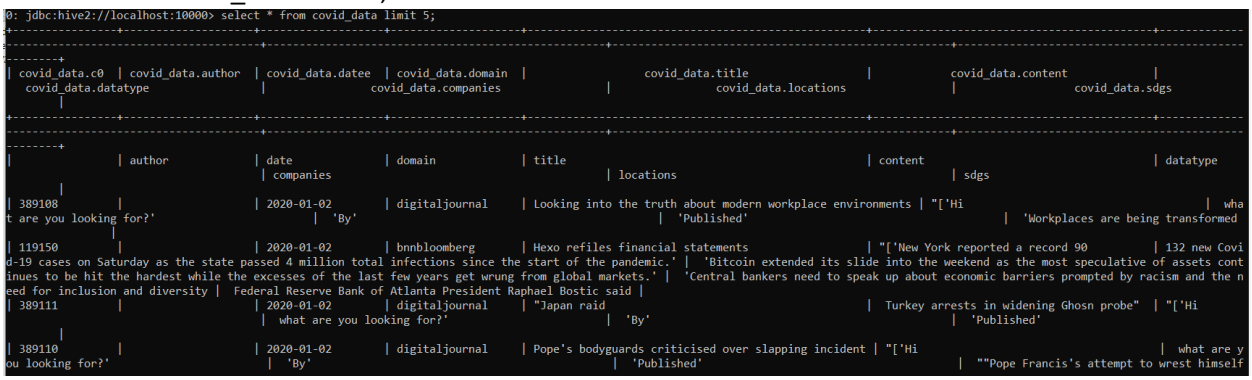
Load data

```
0: jdbc:hive2://localhost:10000> load data inpath '/user/data/home/data' overwrite into table covid_data;
No rows affected (0.761 seconds)
```

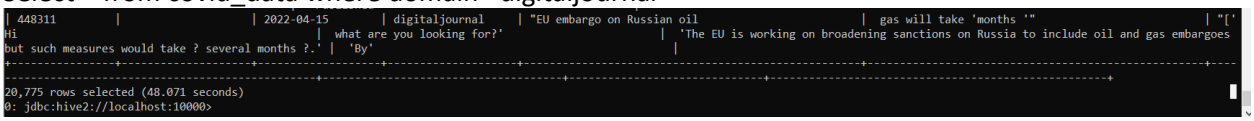
Queries

Mehak Jamil Khatri
23653

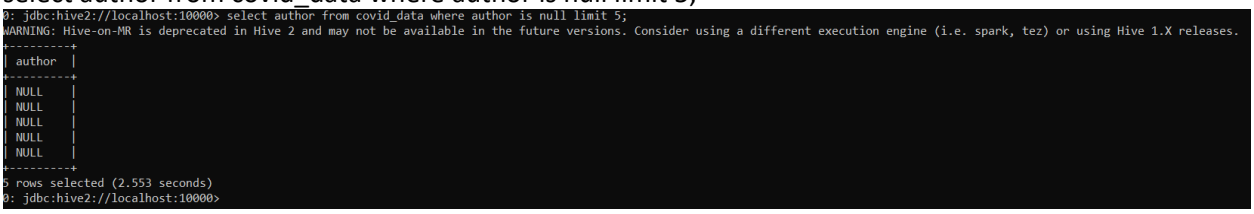
Select * from covid_data limit 5;



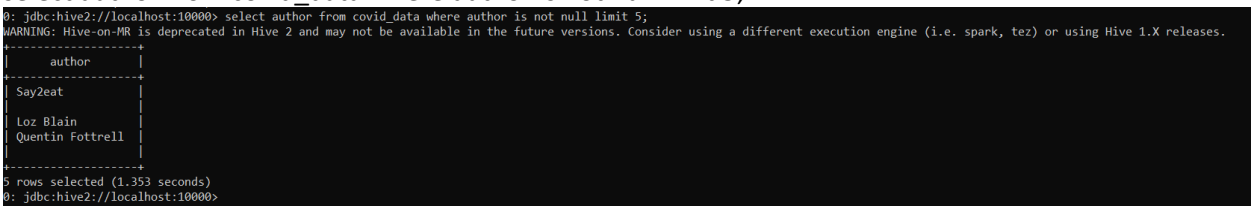
Select * from covid_data where domain='digitaljournal'



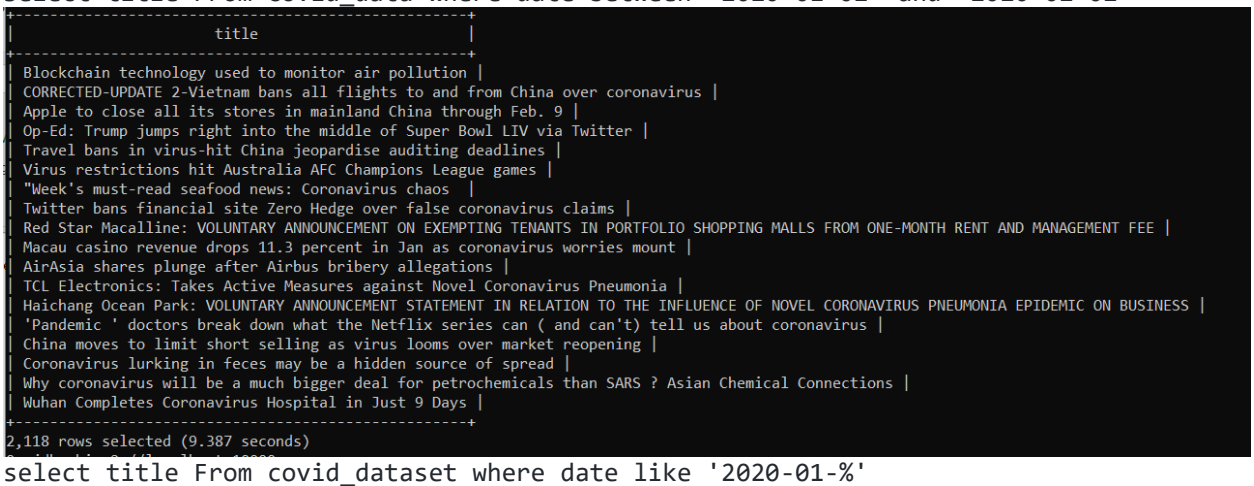
select author from covid_data where author is null limit 5;



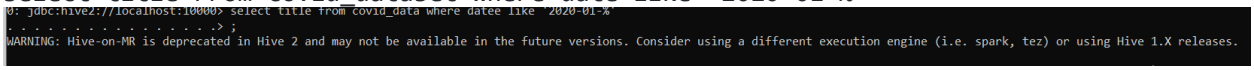
select author from covid_data where author is not null limit 5;



select title From covid_data where date between '2020-01-02' and '2020-02-02'



select title From covid_dataset where date like '2020-01-%'



```
| Coronavirus: How Worried Should I Be About the Shortage of Face Masks? Or Can I Just Use a Scarf? |
| Kenya Airways: Kenyan Students Among Foreigners Stuck in Coronavirus-Hit Chinese City |
| "Novel coronavirus receptors show similarities to SARS-CoV |
| An Outbreak of Racist Sentiment as Coronavirus Reaches Australia |
| "Sabre                                                  |
| 195 Quarantined in California After Fleeing Coronavirus Epicenter |
| Diversified Communications puts coronavirus plan in place for Seafood Expo North America |
| American Airlines Pilots Sue to Stop Flights to China |
| Modeling study estimates spread of 2019 novel coronavirus -- ScienceDaily |
| Shock over virus drops the salmon price down once again |
| Near Miss of the Day 369: Taxi close pass ? but police claimed not to have received footage ( + video) |
| "US underprepared for coronavirus due to Trump cuts |
| Will Covid-19 Impact Investment In Startups And Tech? |
| Friday briefing: Britain shuffles off the EU stage |
| "Whole Genome of the Wuhan Coronavirus                  |
| MIT helps first-time entrepreneur build food hospitality company |
| Republicans march over the impeachment cliff ? taking their self-respect with them |
| Scientific Estimates of Spread of Coronavirus Much Higher Than Official Reports |
| WHO declares global health emergency over coronavirus: 4 questions answered |
| Confirmed Coronavirus Cases Climb to 9826 Globally ? 213 Deaths in China |
| Seth Meyers: 'Today? s Republican party is an authoritarian movement ' |
| Travel? s Worst Nightmare Returns to Haunt Asia     |
+-----------------------------------------------------+
1,954 rows selected (9.404 seconds)
0: jdbc:hive2://localhost:10000>
```

Show columns from covid_data;

```
0: jdbc:hive2://localhost:10000> show columns from covid_data
. . . . . .> ;
+------------+
|   field    |
+------------+
| c0         |
| author     |
| datee      |
| domain     |
| title      |
| content    |
| datatype   |
| companies  |
| locations  |
| sdgs       |
+------------+
10 rows selected (0.039 seconds)
0: jdbc:hive2://localhost:10000>
```

describe covid_data;

```
0: jdbc:hive2://localhost:10000> describe covid_data;
+------------+------------+----------+
|  col_name  | data_type  | comment  |
+------------+------------+----------+
| c0         | string     |          |
| author     | string     |          |
| datee      | string     |          |
| domain     | string     |          |
| title      | string     |          |
| content    | string     |          |
| datatype   | string     |          |
| companies  | string     |          |
| locations  | string     |          |
| sdgs       | string     |          |
+------------+------------+----------+
10 rows selected (0.037 seconds)
0: jdbc:hive2://localhost:10000>
```