2/6/2022

FINAL PROJECT REPORT

BIG DATA ANALYTICS

Arslan Ahmed Shaikh
ERP 23388

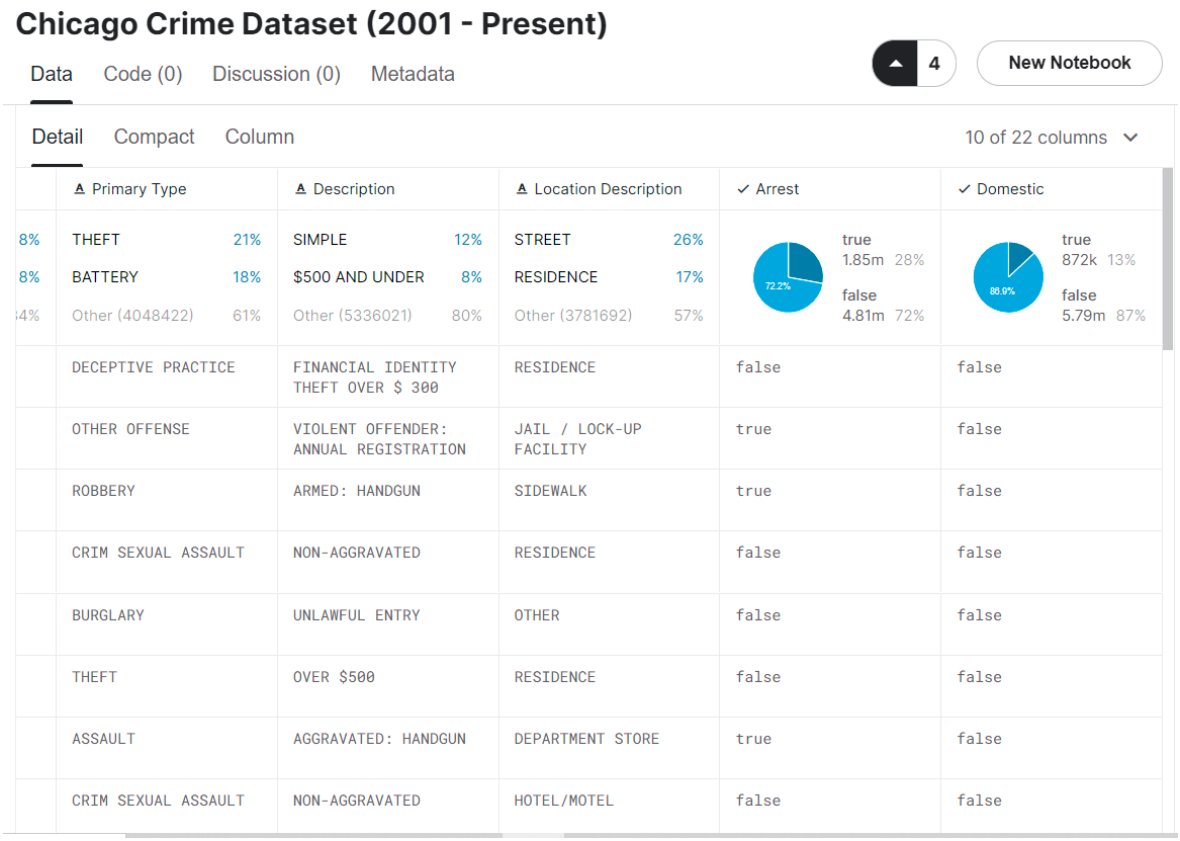
Contents

Introduction

In this project, I have performed a thorough analysis on the working of three popular querying platforms available to query big data i.e. Drill, Hive and Spark. All three are designed to query the modern day big data applications. In this project, I have mainly focused on determining the time taken to execute the query by all three mentioned technologies.

Dataset

In this project, I have used the **Chicago Crimes Data**, from 2001 to present. The dataset contains the data of different types of crimes being recorded at Chicago. The dataset contains more than seventy hundred thousand records and total size of dataset is above 1.5 GB. The dataset has total 22 columns e.g. Case Number, Date, Block, Arrest, Location Description etc.



The dataset is publically available on **kaggle** website, mentioned below is the link to dataset.

https://www.kaggle.com/datasets/armyaviator/chicago-crime-dataset-2001-present

APACHE DRILL

Introduction:

Drill is an apache open source SQL query engine for big data exploration. Drill is designed form the ground up to support high-performance analysis on the semi-structured and rapidly evolving data coming from modern big data applications.

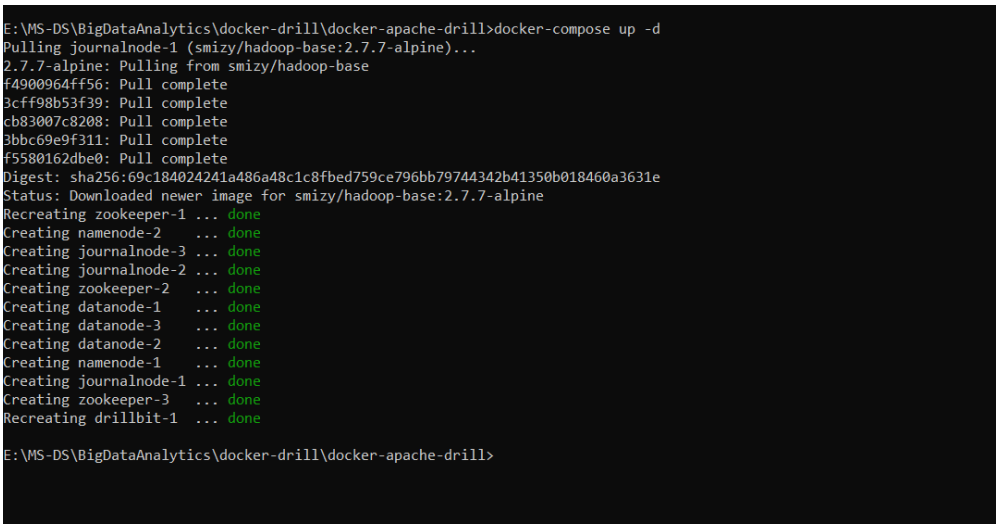
Working

In our **docker-compose.yml** file, we have added the zookeeper services, apache drill, data-node and name-node. These are the core features which we will use to perform queries on apache drill platform. We will be using the **Hadoop** as our storage mechanism where we will keep our dataset (Big Data) and we will load our data into apache drill from hdfs path.

Steps

1. **Run docker-compose.yml file:**

   You can run the docker-compose file using the below command:
   Docker-compose up –d



2. **Check the containers are running:**

You can check this using docker ps -a

```
E:\MS-DS\BigDataAnalytics\docker-drill\docker-apache-drill>docker ps -a
CONTAINER ID   IMAGE                                          COMMAND                  CREATED         STATUS                     PORTS                                       NAMES
2022464c07b6   smizy/apache-drill:1.16.0-alpine               "entrypoint.sh drill…"   2 minutes ago   Up 2 minutes               0.0.0.0:52810->8047/tcp                     drillbit-1
cff25fc17da0   smizy/zookeeper:3.4-alpine                     "entrypoint.sh -serv…"   3 minutes ago   Up 2 minutes               2181/tcp, 2888/tcp, 3888/tcp                zookeeper-1
3b3b405e0cda   smizy/hadoop-base:2.7.7-alpine                 "entrypoint.sh datan…"   3 minutes ago   Up 2 minutes               50010/tcp, 50020/tcp, 50075/tcp             datanode-2
433b349184c9   smizy/zookeeper:3.4-alpine                     "entrypoint.sh -serv…"   3 minutes ago   Up 2 minutes               2181/tcp, 2888/tcp, 3888/tcp                zookeeper-3
e714be101231   smizy/hadoop-base:2.7.7-alpine                 "entrypoint.sh datan…"   3 minutes ago   Up 2 minutes               50010/tcp, 50020/tcp, 50075/tcp             datanode-3
ee389c200385   smizy/hadoop-base:2.7.7-alpine                 "entrypoint.sh journ…"   3 minutes ago   Up 2 minutes               8480/tcp, 8485/tcp                          journalnode-2
6b68f04b77bd   smizy/hadoop-base:2.7.7-alpine                 "entrypoint.sh journ…"   3 minutes ago   Up 2 minutes               8480/tcp, 8485/tcp                          journalnode-1
27d1e9c48b85   smizy/hadoop-base:2.7.7-alpine                 "entrypoint.sh namen…"   3 minutes ago   Up 2 minutes               8020/tcp, 0.0.0.0:52803->50070/tcp          namenode-1
900ea6d978bc   smizy/hadoop-base:2.7.7-alpine                 "entrypoint.sh datan…"   3 minutes ago   Up 2 minutes               50010/tcp, 50020/tcp, 50075/tcp             datanode-1
d404212c7cd3   smizy/zookeeper:3.4-alpine                     "entrypoint.sh -serv…"   3 minutes ago   Up 2 minutes               2181/tcp, 2888/tcp, 3888/tcp                zookeeper-2
ddaaf1312d54   smizy/hadoop-base:2.7.7-alpine                 "entrypoint.sh journ…"   3 minutes ago   Up 2 minutes               8480/tcp, 8485/tcp                          journalnode-3
bb403d8342e7   smizy/hadoop-base:2.7.7-alpine                 "entrypoint.sh namen…"   3 minutes ago   Exited (1) 2 minutes ago                                               namenode-2
d8bb27db8932   bde2020/hadoop-datanode:1.1.0-hadoop2.8-java8  "/entrypoint.sh /run…"   22 hours ago    Exited (137) 21 hours ago                                              docker-hadoop-spark-workbench-master_datanode_1
efc9e31a1f93   bde2020/hadoop-namenode:1.1.0-hadoop2.8-java8  "/entrypoint.sh /run…"   22 hours ago    Exited (137) 21 hours ago                                              namenode
8c3d1f12b6f6   bde2020/spark-worker:2.1.0-hadoop2.8-hive-java8  "entrypoint.sh /bin/…"   22 hours ago    Exited (137) 21 hours ago                                              docker-hadoop-spark-workbench-master_spark-worker_1
16e20e129e35   bde2020/spark-notebook:2.1.0-hadoop2.8-hive    "/entrypoint.sh /run…"   22 hours ago    Exited (137) 21 hours ago                                              spark-notebook
bcb4e464e482   bde2020/spark-master:2.1.0-hadoop2.8-hive-java8  "entrypoint.sh /bin/…"   22 hours ago    Exited (137) 21 hours ago                                              spark-master
072d309b5855   bde2020/hdfs-filebrowser:3.11                  "/entrypoint.sh buil…"   22 hours ago    Exited (137) 21 hours ago                                              docker-hadoop-spark-workbench-master_hue_1
1c71c011290d   bde2020/hadoop-namenode:2.0.0-hadoop2.7.4-java8  "/entrypoint.sh /run…"   2 days ago      Exited (137) 23 hours ago                                              docker-hive_namenode_1
56f21aee1703   bde2020/hadoop-datanode:2.0.0-hadoop2.7.4-java8  "/entrypoint.sh /run…"   2 days ago      Exited (137) 23 hours ago                                              docker-hive_datanode_1
1a3b1b560f2f   bde2020/hive:2.3.2-postgresql-metastore        "entrypoint.sh /opt/…"   2 days ago      Exited (143) 23 hours ago                                              docker-hive_hive-metastore_1
021846ee18d5   bde2020/hive-metastore-postgresql:2.3.0        "/docker-entrypoint.…"   2 days ago      Exited (0) 23 hours ago                                                docker-hive_hive-metastore-postgresql_1
32755347790b   bde2020/hive:2.3.2-postgresql-metastore        "entrypoint.sh /bin/…"   2 days ago      Exited (137) 23 hours ago                                              docker-hive_hive-server_1
21fcaecba57f   shawnzhu/prestodb:0.181                        "./bin/launcher run"     2 days ago      Exited (143) 23 hours ago                                              docker-hive_presto-coordinator_1

E:\MS-DS\BigDataAnalytics\docker-drill\docker-apache-drill>
```

3. **Load data into HDFS:**

In order to load data onto HDFS, we firstly need to copy the data from our local file system in to our data-node container. Then, from our container, we will put the data onto hdfs so that it can be referenced or linked to drill while performing he queries. Now, in order to copy the dataset from our local system, we use **docker cp sourcePath destinationPath** command.

Now run the data-node container using the **docker exec –it name /bin/bash** to start the container in the iterative mode. Once the container is up, you can locate your dataset into the directory you copied in. The, create the directory on Hadoop using **hdfs dfs –mkidr** command and put the dataset inside the **HDFS.**

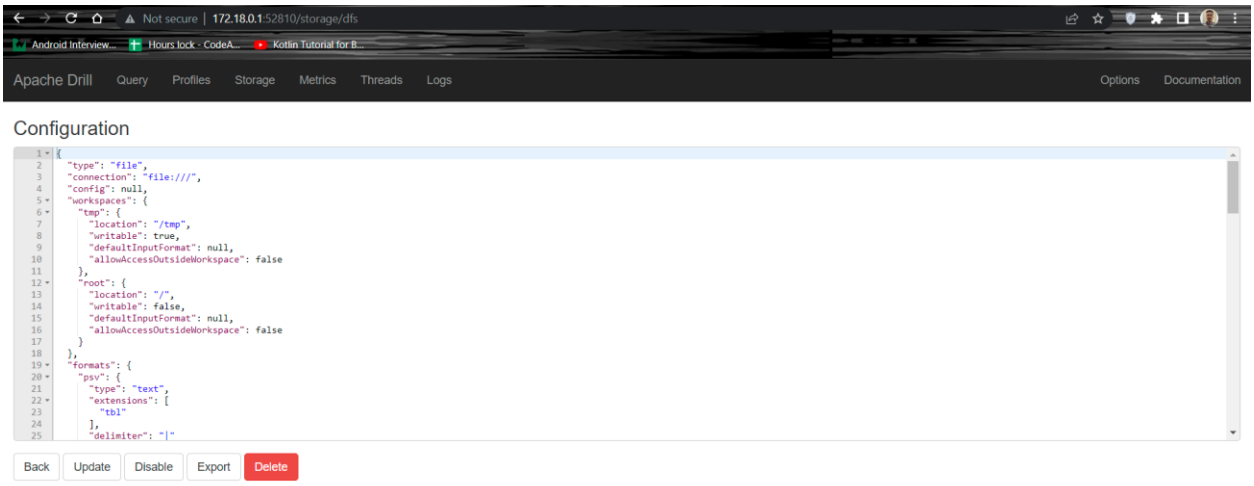
/user/hdfs/output created in HDFS

```
E:\MS-DS\BigDataAnalytics\docker-drill\docker-apache-drill>docker cp E:\MS-DS\BigDataAnalytics\data 900ea6d978bc:dataset

E:\MS-DS\BigDataAnalytics\docker-drill\docker-apache-drill>docker exec -it -u hdfs datanode-1 /bin/bash
bash-4.3$ cd ..
bash-4.3$ ls
bin          hadoop-2.7    hadoop-2.7.7  lib          share
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
ChicagoCrimesData.csv
bash-4.3$ hdfs dfs -mkdir -p /user/hdfs/output
bash-4.3$ hdfs dfs -put -f ChicagoCrimesData.csv /user/hdfs/output
bash-4.3$ hdfs dfs -ls  /user/hdfs/output
Found 2 items
-rw-r--r--   3 hdfs hadoop 1684766921 2022-06-01 20:48 /user/hdfs/output/ChicagoCrimesData.csv
drwxr-xr-x   - hdfs hadoop          0 2022-06-01 18:45 /user/hdfs/output/home
bash-4.3$
```

4. **Browse the UI of Apache-drill:**

Once the container is up and running, you can use the web interface of Apache-drill under the url, http://yourMachine'sIPV4Address:dockerImagePort/storage/dfs.



Once this web page is accessible, you need to change few configurations parameters:

Set connection parameter = "hdfs://namenode-1.vnet:8020/"

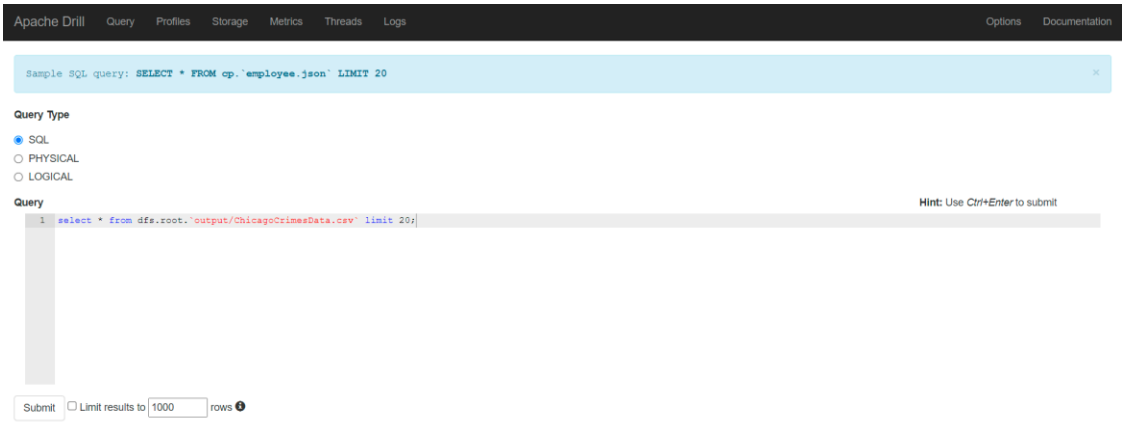
Add enabled = true

Set root.location = /user/hdfs

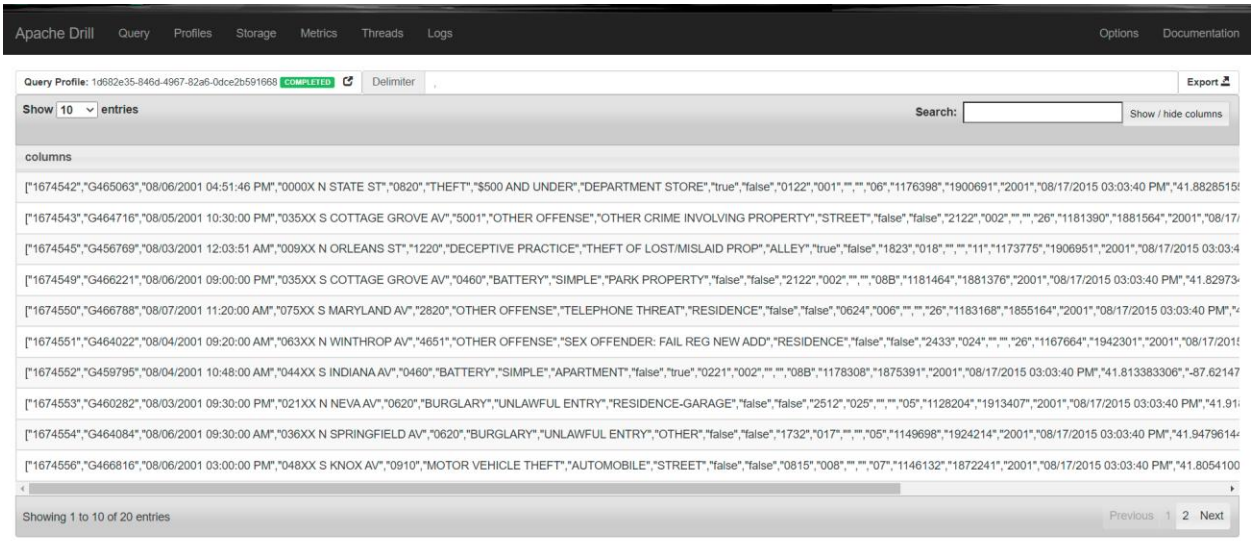
After changing all the configurations as mentioned above, click the update button to save the configuration.

5. **Query Apache-Drill on Web UI:**

Once the configurations are updated successfully, you can click the **Query** button from upper tab over the same web interface. As we had already uploaded the dataset file in the output directory in **HDFS**, we will now refer the same directory to pull the data.

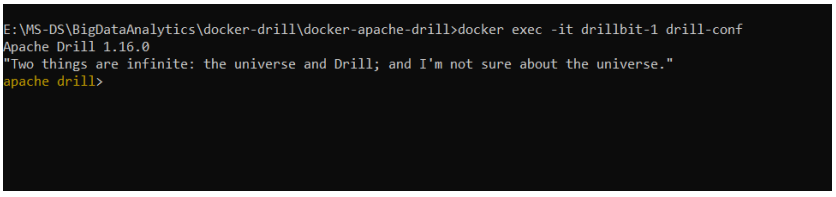


Output:



6. Query Apache-Drill on CLI:

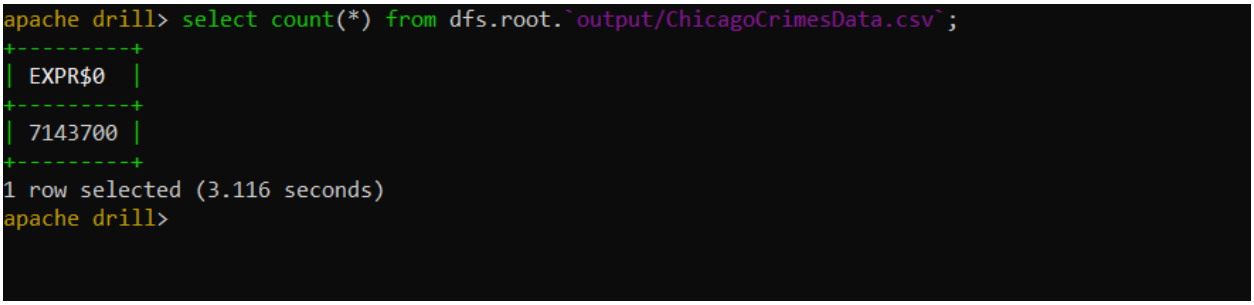
As we had performed the query on web UI, the same query can be processed on cli version as well. For that purpose, we need to start the drill container using the same **docker exec –it** command.



Queries

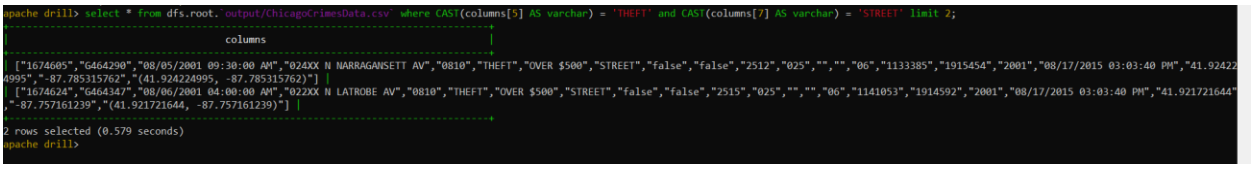
a. Get the first 20 rows from the data:

b. Get the total count of the data:



c. Filter data where Primary Type of crime is Theft

d. Filter data where Primary Type of crime is Theft and LocationDescription is Street



e. Filter data where LocationDescription is Street

```
apache drill> select * from dfs.root.`output/ChicagoCrimesData.csv` where CAST(columns[7] AS varchar) = 'STREET' limit 2;
+------------------------------------------------------------------------------+
|                                   columns                                    |
+------------------------------------------------------------------------------+
| ["1674543","G464716","08/05/2001 10:30:00 PM","035XX S COTTAGE GROVE AV","5001","OTHER OFFENSE","OTHER CRIME INVOLVING PROPERTY","STREET","false","false","2122","002","","","26","1181390","1881564","2001","08/
17/2015 03:03:40 PM","41.830251864","-87.609976054","(41.830251864, -87.609976054)"] |
| ["1674556","G466816","08/06/2001 03:00:00 PM","048XX S KNOX AV","0910","MOTOR VEHICLE THEFT","AUTOMOBILE","STREET","false","false","0815","008","","","07","1146132","1872241","2001","08/17/2015 03:03:40 PM","4
1.80541001","-87.739575644","(41.80541001, -87.739575644)"] |
+------------------------------------------------------------------------------+
2 rows selected (0.735 seconds)
apache drill>
```

f. Filter data where Arrest = false

```
apache drill> select * from dfs.root.`output/ChicagoCrimesData.csv` where CAST(columns[8] AS varchar) = 'false' limit 2;
+------------------------------------------------------------------------------+
|                                   columns                                    |
+------------------------------------------------------------------------------+
| ["11034701","JA366925","01/01/2001 11:00:00 AM","016XX E 86TH PL","1153","DECEPTIVE PRACTICE","FINANCIAL IDENTITY THEFT OVER $ 300","RESIDENCE","false","false","0412","004","8","45","11","","","2001","08/05/20
17 03:50:08 PM","","","",""] |
| ["11227287","JB147188","10/08/2017 03:00:00 AM","092XX S RACINE AVE","0281","CRIM SEXUAL ASSAULT","NON-AGGRAVATED","RESIDENCE","false","false","2222","022","21","73","02","","","2017","02/11/2018 03:57:41 PM",
"","","",""] |
+------------------------------------------------------------------------------+
2 rows selected (0.286 seconds)
apache drill>
```

g. Filter data where Primary Type of crime is Theft and LocationDescription is Street

```
apache drill> select * from dfs.root.`output/ChicagoCrimesData.csv` where CAST(columns[7] AS varchar) = 'STREET' and CAST(columns[5] AS varchar) = 'THEFT'  limit 4;
+------------------------------------------------------------------------------+
|                                   columns                                    |
+------------------------------------------------------------------------------+
| ["3127961","HK117644","01/10/2004 01:00:00 PM","050XX W JACKSON BLVD","0820","THEFT","$500 AND UNDER","STREET","false","false","1533","015","28","25","06","1142671","1898197","2004","02/10/2018 03:50:01 PM","4
1.876701895","-87.751624656","(41.876701895, -87.751624656)"] |
| ["3127989","HK117846","01/09/2004 07:00:00 PM","056XX S CARPENTER ST","0820","THEFT","$500 AND UNDER","STREET","false","false","0712","007","16","68","06","1170276","1867439","2004","02/28/2018 03:56:25 PM","4
1.79174096","-87.651164679","(41.79174096, -87.651164679)"] |
| ["3128011","HK117857","01/09/2004 07:00:00 PM","026XX W 68TH ST","0810","THEFT","OVER $500","STREET","false","false","0831","008","15","66","06","1160027","1859462","2004","02/28/2018 03:56:25 PM","41.77006797
8","-87.688965217","(41.770067978, -87.688965217)"] |
| ["3128034","HK117968","01/09/2004 09:00:00 AM","025XX W DIVISION ST","0820","THEFT","$500 AND UNDER","STREET","false","false","1312","012","26","24","06","1159389","1907896","2004","02/28/2018 03:56:25 PM","41
.902989475","-87.689974104","(41.902989475, -87.689974104)"] |
+------------------------------------------------------------------------------+
4 rows selected (0.346 seconds)
apache drill>
```

h. Filter data where Primary Type of crime is Battery and LocationDescription is Street

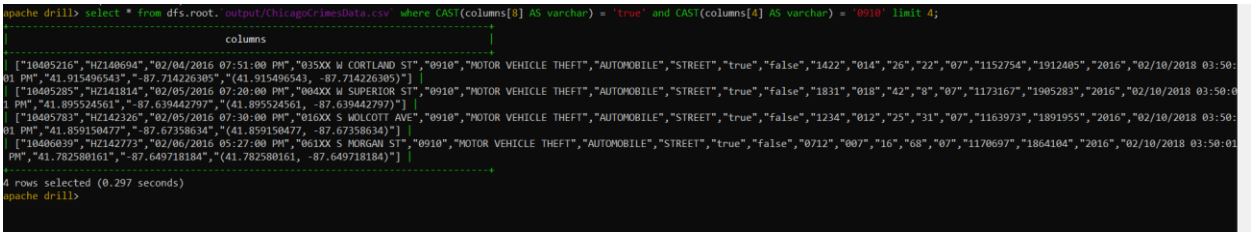
```
apache drill> select * from dfs.root.`output/ChicagoCrimesData.csv` where CAST(columns[5] AS varchar) = 'BATTERY' and CAST(columns[7] AS varchar) = 'STREET' limit 4;
+------------------------------------------------------------------------------+
|                                   columns                                    |
+------------------------------------------------------------------------------+
| ["1744168","G553545","09/15/2001 02:00:00 AM","013XX W POLK ST","0460","BATTERY","SIMPLE","STREET","false","false","1213","012","","","08B","","","2001","08/17/2015 03:03:40 PM","","","",""] |
| ["1896258","G749215","12/15/2001 02:00:00 AM","011XX N STATE ST","0460","BATTERY","SIMPLE","STREET","false","false","1824","018","","","08B","","","2001","08/17/2015 03:03:40 PM","","","",""] |
| ["2370646","HH676216","09/26/2002 01:30:00 PM","034XX W 79TH ST","0460","BATTERY","SIMPLE","STREET","false","false","0835","008","18","70","08B","","","2002","08/17/2015 03:03:40 PM","","","",""] |
| ["8070766","HR559439","09/27/2009 06:20:00 PM","002XX W 95TH ST","0460","BATTERY","SIMPLE","STREET","false","false","0634","006","21","49","08B","","","2009","08/17/2015 03:03:40 PM","","","",""] |
+------------------------------------------------------------------------------+
4 rows selected (0.289 seconds)
apache drill>
```

i. Get all data from the provided data (limit=10)

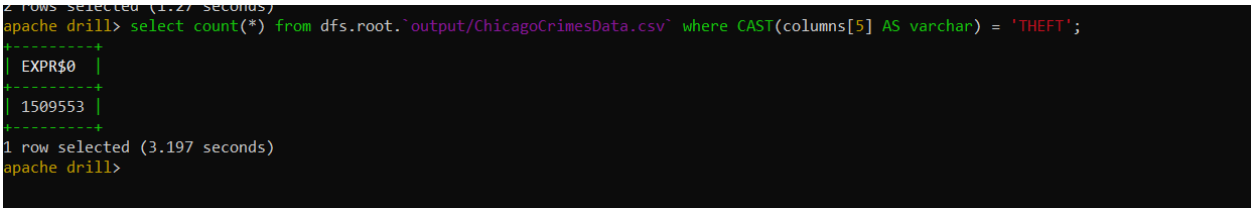
```
apache drill> select * from dfs.root.`output/ChicagoCrimesData.csv` limit 10;
+------------------------------------------------------------------------------+
|                                   columns                                    |
+------------------------------------------------------------------------------+
| ["11272723","JB207383","04/01/2018 10:45:00 AM","010XX W POLK ST","0820","THEFT","$500 AND UNDER","OTHER","false","false","1232","012","25","28","06","1169611","1896627","2018","05/04/2018 03:51:04 PM","41.871
850066","-87.65275518","(41.871850066, -87.65275518)"] |
| ["11272724","JB207215","04/01/2018 05:15:00 AM","077XX S ESSEX AVE","0460","BATTERY","SIMPLE","APARTMENT","false","false","0421","004","7","43","08B","1194202","1854229","2018","05/04/2018 03:51:04 PM","41.754
037343","-87.56386716","(41.754037343, -87.56386716)"] |
| ["11272725","JB207348","03/31/2018 03:00:00 PM","028XX E 130TH ST","0820","THEFT","$500 AND UNDER","VEHICLE NON-COMMERCIAL","false","false","0433","004","10","55","06","1196858","1819361","2018","05/04/2018 03
:51:04 PM","41.659190644","-87.555288394","(41.659190644, -87.555288394)"] |
| ["11272726","JB207224","04/01/2018 03:30:00 AM","104XX S EGGLESTON AVE","0810","THEFT","OVER $500","RESIDENCE","false","false","2233","022","34","49","06","1175131","1835642","2018","05/04/2018 03:51:04 PM","4
1.704378949","-87.634309567","(41.704378949, -87.634309567)"] |
| ["11272727","JB207322","04/01/2018 12:00:00 AM","095XX S EUCLID AVE","0910","MOTOR VEHICLE THEFT","AUTOMOBILE","STREET","false","false","0431","004","7","51","07","1190847","1842066","2018","05/04/2018 03:51:0
4 PM","41.721642614","-87.576554028","(41.721642614, -87.576554028)"] |
| ["11272728","JB207381","04/01/2018 12:00:00 AM","020XX W DIVISION ST","0486","BATTERY","DOMESTIC BATTERY SIMPLE","BAR OR TAVERN","false","true","1212","012","1","24","08B","1162670","1907990","2018","05/04/201
8 03:51:04 PM","41.903179264","-87.677919727","(41.903179264, -87.677919727)"] |
| ["11272729","JB207296","04/01/2018 09:15:00 AM","075XX S DORCHESTER AVE","0486","BATTERY","DOMESTIC BATTERY SIMPLE","RESIDENCE","true","true","0411","004","5","43","08B","1186849","1855270","2018","05/04/2018
03:51:04 PM","41.75797125","-87.59078048","(41.75797125, -87.59078048)"] |
| ["11272730","JB207368","04/01/2018 02:20:00 AM","040XX N DRAKE AVE","1310","CRIMINAL DAMAGE","TO PROPERTY","RESIDENCE","false","true","1723","017","33","16","14","1152001","1926688","2018","05/04/2018 03:51:04
PM","41.954705124","-87.716615295","(41.954705124, -87.716615295)"] |
| ["11272731","JB207406","03/31/2018 06:00:00 PM","080XX S VERNON AVE","0890","THEFT","FROM BUILDING","RESIDENCE","false","false","0631","006","6","44","06","1180642","1851698","2018","05/04/2018 03:51:04 PM","4
1.748314058","-87.613637728","(41.748314058, -87.613637728)"] |
| ["11272732","JB207344","04/01/2018 09:55:00 AM","115XX S LAFLIN ST","0486","BATTERY","DOMESTIC BATTERY SIMPLE","RESIDENCE","true","true","0524","005","34","53","08B","1168405","1828220","2018","05/04/2018 03:5
1:04 PM","41.68415894","-87.65915185","(41.68415894, -87.65915185)"] |
+------------------------------------------------------------------------------+
10 rows selected (0.191 seconds)
apache drill>
```

j. Filter data where Arrest = true and LocationDescription is Street

k. Filter data where Arrest = true and IUCR is 0910



l. Get count of data where Primary Type of crime is Theft



Total Execution Time on Apache-Drill Queries Performed

Query1	3.116 (Seconds)
Query2	0.378 (Seconds)
Query3	0.579 (Seconds)
Query4	0.735 (Seconds)
Query5	0.286 (Seconds)
Query6	0.346 (Seconds)
Query7	0.289 (Seconds)
Query8	0.191 (Seconds)
Query9	1.761 (Seconds)
Query10	0.297 (Seconds)
Query11	3.197 (Seconds)
Query12	0.445 (Seconds)

APACHE SPARK

Introduction

Apache spark is a lightning-fast cluster computing technology, desgined for fast computation. It is based on Hadoop map reduce working mechanism and it extends map reduce power to perform processing more efficiently. The main feature of spark is its in-memory cluster computing that increases the processing speed of an application.

Spark is designed to cover a wide range of workloads such as batch applications, iterative algorithms, interactive, interactive queries and streaming. Apart from supporting all these workload in a respective system, it reduces the management burden of maintaining separate tools.

Working

In our **docker-compose.yml** file, we have added all the required docker images e.g. Hadoop images (datanode and namenode), spark-master etc. These are the important docker images which we will use to perform queries on apache spark platform. We will be using the **Hadoop** as our storage mechanism where we will keep our dataset (Big Data) and we will load our data into apache spark from hdfs path.

Steps

a. **Run docker-compose file**

```
E:\MS-DS\docker-hadoop-spark-workbench-master>docker-compose up -d
spark-master is up-to-date
Recreating namenode ...
docker-hadoop-spark-workbench-master_hue_1 is up-to-date
spark-notebook is up-to-date
Recreating namenode ... done
Creating docker-hadoop-spark-workbench-master_datanode_1 ... done
```

b. **Check Containers**
Command: docker ps -a

```
E:\MS-DS\docker-hadoop-spark-workbench-master>docker ps -a
CONTAINER ID   IMAGE                                            COMMAND                   CREATED          STATUS                       PORTS                                                NAMES
d8bb27db8932   bde2020/hadoop-datanode:1.1.0-hadoop2.8-java8     "/entrypoint.sh /run…"    39 seconds ago   Up 36 seconds (healthy)      0.0.0.0:40075->40075/tcp, 50075/tcp                  docker-hadoo
p-spark-workbench-master_datanode_1
efc9e31a1f93   bde2020/hadoop-namenode:1.1.0-hadoop2.8-java8     "/entrypoint.sh /run…"    42 seconds ago   Up 40 seconds (healthy)      0.0.0.0:40070->40070/tcp, 50070/tcp                  namenode
8c3d1f12b6f6   bde2020/spark-worker:2.1.0-hadoop2.8-hive-java8   "entrypoint.sh /bin/…"    2 minutes ago    Up 2 minutes (healthy)       0.0.0.0:8081->8081/tcp                               docker-hadoo
p-spark-workbench-master_spark-worker_1
16e20e129e35   bde2020/spark-notebook:2.1.0-hadoop2.8-hive       "/entrypoint.sh /run…"    2 minutes ago    Up 2 minutes                 0.0.0.0:9001->9001/tcp                               spark-notebo
ok
bcb4e464e482   bde2020/spark-master:2.1.0-hadoop2.8-hive-java8   "entrypoint.sh /bin/…"    2 minutes ago    Up 2 minutes (healthy)       0.0.0.0:7077->7077/tcp, 6066/tcp, 0.0.0.0:8080->8080/tcp   spark-master

072d309b5855   bde2020/hdfs-filebrowser:3.11                     "/entrypoint.sh buil…"    2 minutes ago    Up 2 minutes                 0.0.0.0:8088->8088/tcp                               docker-hadoo
p-spark-workbench-master_hue_1
```

c. Copy Data from local into Container

```
presto-coordinator_1

E:\MS-DS\docker-hadoop-spark-workbench-master>docker cp E:\MS-DS\BigDataAnalytics\data bcb4e464e482:home

E:\MS-DS\docker-hadoop-spark-workbench-master>
```

d. Execute Spark Container

```
E:\MS-DS\docker-hadoop-spark-workbench-master>docker exec -it spark-master /bin/bash
```

e. Put data onto Hadoop

Command: -mkdir to make the directory on HDFS

-put to load data on HDFS

-ls to list the data from HDFS

```
root@bcb4e464e482:/home/data# hadoop fs -mkdir /user
root@bcb4e464e482:/home/data#  hadoop fs -put -f /home/data /user
root@bcb4e464e482:/home/data# hadoop fs -ls /user
Found 1 items
drwxr-xr-x   - root supergroup          0 2022-05-31 20:55 /user/data
root@bcb4e464e482:/home/data#
```

f. Execute Spark

Command: /spark/bin/spark-shell --master=local spark://spark-master:7077

```
root@bcb4e464e482:/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).
22/05/31 20:59:52 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
22/05/31 20:59:59 WARN metastore.ObjectStore: Version information not found in metastore. hive.metastore.schema.verification is not enabled so recording the schema version 1.2.0
22/05/31 20:59:59 WARN metastore.ObjectStore: Failed to get database default, returning NoSuchObjectException
22/05/31 21:00:00 WARN metastore.ObjectStore: Failed to get database global_temp, returning NoSuchObjectException
Spark context Web UI available at http://172.19.0.3:4040
Spark context available as 'sc' (master = local[*], app id = local-1654030793647).
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>
```

g. Read data from HDFS

```
scala> val df = spark.read.format("csv").option("header", "true").load("hdfs://namenode:8020/user/data/ChicagoCrimesData");
df: org.apache.spark.sql.DataFrame = [ID: string, Case Number: string ... 20 more fields]

scala>
```

h. Create Table

```
scala> df.createOrReplaceTempView("ChicagoDataTable");

scala> spark.time(df.show(false));
+--------+-----------+----------------------+-------------+----+---------------------+--------+-------+--------+
|ID      |Case Number|Date                  |Block        |IUCR|Primary Type         |Description                      |Location Description|Arrest|Domestic|Beat|District|Ward|Community Area|FBI Code|X Coordinate|Y Coordinate|Year|Updated On            |Latitude|Longitude|Location|
+--------+-----------+----------------------+-------------+----+---------------------+---------------------------------+--------------------+------+--------+----+--------+----+--------------+--------+------------+------------+----+----------------------+--------+---------+--------+
|11034701|JA366925   |01/01/2001 11:00:00 AM|016XX E 86TH PL|1153|DECEPTIVE PRACTICE   |FINANCIAL IDENTITY THEFT OVER $ 300|RESIDENCE           |false |false   |0412|004     |8   |45            |11      |null        |null        |2001|08/05/2017 03:50:08 PM|null    |null     |null    |
|11227287|JB147188   |10/08/2017 03:00:00 AM|092XX S RACINE AVE|0281|CRIM SEXUAL ASSAULT  |NON-AGGRAVATED                   |RESIDENCE           |false |false   |2222|022     |21  |73            |02      |null        |null        |2017|02/11/2018 03:57:41 PM|null    |null     |null    |
|11227583|JB147595   |03/28/2017 02:00:00 PM|026XX W 79TH ST|0620|BURGLARY             |UNLAWFUL ENTRY                   |OTHER               |false |false   |0835|008     |18  |70            |05      |null        |null        |2017|02/11/2018 03:57:41 PM|null    |null     |null    |
|11227293|JB147230   |09/09/2017 08:17:00 PM|060XX S EBERHART AVE|0810|THEFT                |OVER $500                        |RESIDENCE           |false |false   |0313|003     |20  |42            |06      |null        |null        |2017|02/11/2018 03:57:41 PM|null    |null     |null    |
|11227634|JB147599   |08/26/2017 10:00:00 AM|001XX W RANDOLPH ST|0281|CRIM SEXUAL ASSAULT  |NON-AGGRAVATED                   |HOTEL/MOTEL         |false |false   |0122|001     |42  |32            |02      |null        |null        |2017|02/11/2018 03:57:41 PM|null    |null     |null    |
|11227517|JB138481   |02/10/2013 12:00:00 AM|071XX S LAFAYETTE AVE|0266|CRIM SEXUAL ASSAULT  |PREDATORY                        |RESIDENCE           |false |false   |0731|007     |6   |69            |02      |null        |null        |2013|02/11/2018 03:57:41 PM|null    |null     |null    |
|11227503|JB146383   |01/01/2015 12:01:00 AM|061XX S KILBOURN AVE|1751|OFFENSE INVOLVING CHILDREN|CRIM SEX ABUSE BY FAM MEMBER|RESIDENCE           |false |true    |0813|008     |13  |65            |17      |null        |null        |2015|04/12/2019 04:00:15 PM|null    |null     |null    |
|11227508|JB146365   |01/01/2017 12:01:00 AM|027XX S WHIPPLE ST|1754|OFFENSE INVOLVING CHILDREN|AGG SEX ASSLT OF CHILD FAM MBR|RESIDENCE           |false |false   |1033|010     |12  |30            |02      |null        |null        |2017|02/11/2018 03:57:41 PM|null    |null     |null    |
|11022695|JA353568   |07/17/2017 10:10:00 AM|021XX W MC LEAN AVE|0810|THEFT                |OVER $500                        |RESIDENCE           |false |false   |1432|014     |32  |22            |06      |null        |null        |2017|07/24/2017 03:54:23 PM|null    |null     |null    |
|11227633|JB147500   |12/28/2017 03:55:00 PM|011XX S MICHIGAN AVE|1153|DECEPTIVE PRACTICE   |FINANCIAL IDENTITY THEFT OVER $ 300|null                |false |false   |0123|001     |2   |32            |11      |null        |null        |2017|02/11/2018 03:57:41 PM|null    |null     |null    |
|11227586|JB147613   |02/10/2017 12:00:00 PM|089XX S COTTAGE GROVE AVE|1310|CRIMINAL DAMAGE      |TO PROPERTY                      |APARTMENT           |false |false   |0633|006     |8   |44            |14      |null        |null        |2017|02/11/2018 03:57:41 PM|null    |null     |null    |
|11227475|JB147314   |11/22/2017 02:42:00 AM|056XX N CHRISTIANA AVE|2826|OTHER OFFENSE        |HARASSMENT BY ELECTRONIC MEANS   |APARTMENT           |false |true    |1711|017     |39  |13            |26      |null        |null        |2017|02/11/2018 03:57:41 PM|null    |null     |null    |
|11227247|JB147078   |01/01/2012 09:00:00 AM|105XX S INDIANAPOLIS AVE|1153|DECEPTIVE PRACTICE   |FINANCIAL IDENTITY THEFT OVER $ 300|RESIDENCE           |false |false   |0432|004     |10  |52            |11      |null        |null        |2012|02/11/2018 03:57:41 PM|null    |null     |null    |
|11036284|JA370282   |07/29/2017 03:40:00 PM|014XX W Devon Ave|0820|THEFT                |$500 AND UNDER                   |SIDEWALK            |false |false   |2432|024     |40  |1             |06      |null        |null        |2017|08/05/2017 03:50:08 PM|null    |null     |null    |
|11227509|JB146413   |01/22/2017 12:01:00 AM|079XX S JEFFERY BLVD|1752|OFFENSE INVOLVING CHILDREN|AGG CRIM SEX ABUSE FAM MEMBER|APARTMENT           |false |false   |0414|004     |8   |46            |20      |null        |null        |2017|02/11/2018 03:57:41 PM|null    |null     |null    |
```

Queries

a. Get 10 records from the provided data

```
scala> spark.time(spark.sql("select * from ChicagoDataTable limit 10"))
Time taken: 43 ms
res4: org.apache.spark.sql.DataFrame = [ID: string, Case Number: string ... 20 more fields]
```

b. Get total count of provided data

```
scala> spark.time(spark.sql("select count(*) from ChicagoDataTable"))
Time taken: 36 ms
res3: org.apache.spark.sql.DataFrame = [count(1): bigint]
```

c. Get count of all the crimes where no arrest happened

```
scala> spark.time(spark.sql("select count(*) from ChicagoDataTable where Arrest = 'false'"))
Time taken: 38 ms
res8: org.apache.spark.sql.DataFrame = [count(1): bigint]
```

d. Get count of the crimes where no arrest happened in 0910 ICUR

```
scala> spark.time(spark.sql("select count(*) from ChicagoDataTable where IUCR = '0910' and Arrest='false'"))
Time taken: 16 ms
res14: org.apache.spark.sql.DataFrame = [count(1): bigint]
```

e. Get the all the records which were not domestic and happened in 0910 ICUR

```
scala> spark.time(spark.sql("select * from ChicagoDataTable where IUCR = '0910' and Domestic='false'"))
Time taken: 14 ms
res17: org.apache.spark.sql.DataFrame = [ID: string, Case Number: string ... 20 more fields]
```

f. Get the all the records which were not domestic and order by ICUR

```
scala> spark.time(spark.sql("select * from ChicagoDataTable where Domestic='false' order by IUCR"))
Time taken: 25 ms
res18: org.apache.spark.sql.DataFrame = [ID: string, Case Number: string ... 20 more fields]
```

g. Get all the records which were domestic, and the arrest happened as well, order by District

```
scala> spark.time(spark.sql("select * from ChicagoDataTable where Domestic='true' and Arrest ='true' order by District"))
Time taken: 17 ms
res19: org.apache.spark.sql.DataFrame = [ID: string, Case Number: string ... 20 more fields]
```

h. Get all the records of crimes happened in 022 district and in which arrest happened as well order by district

```
scala> spark.time(spark.sql("select * from ChicagoDataTable where District ='022' and Arrest ='true' order by District"))
Time taken: 18 ms
res20: org.apache.spark.sql.DataFrame = [ID: string, Case Number: string ... 20 more fields]
```

i. Get all the records of crimes happened in 21 Ward and in which arrest happened as well order by district

```
scala> spark.time(spark.sql("select * from ChicagoDataTable where Ward='21' and Arrest ='true' order by District"))
Time taken: 12 ms
res21: org.apache.spark.sql.DataFrame = [ID: string, Case Number: string ... 20 more fields]

scala>
```

j. Get all the records where no arrest happened order by district

```
scala> spark.time(spark.sql("select * from ChicagoDataTable where Arrest ='false' order by District"))
Time taken: 10 ms
res22: org.apache.spark.sql.DataFrame = [ID: string, Case Number: string ... 20 more fields]
```

k. Get all the records of crimes which were domestic and the arrest happened as well

```
scala> spark.time(spark.sql("select * from ChicagoDataTable where Domestic='true' and Arrest ='true'" ))
Time taken: 10 ms
res23: org.apache.spark.sql.DataFrame = [ID: string, Case Number: string ... 20 more fields]
```

l. Get all records of crimes happened in 021 District and 21 Ward

```
scala> spark.time(spark.sql("select * from ChicagoDataTable where District ='021' and Ward='21'" ))
Time taken: 13 ms
res25: org.apache.spark.sql.DataFrame = [ID: string, Case Number: string ... 20 more fields]
```

Total Execution Time on Apache-Spark Queries Performed

Query1	0.043 (Seconds)
Query2	0.036 (Seconds)
Query3	0.038 (Seconds)
Query4	0.016 (Seconds)
Query5	0.014 (Seconds)
Query6	0.025 (Seconds)
Query7	0.017 (Seconds)
Query8	0.018 (Seconds)
Query9	0.012 (Seconds)
Query10	0.010 (Seconds)
Query11	0.010 (Seconds)
Query12	0.013 (Seconds)

APACHE HIVE

Introduction

Apache Hive is a data warehousing package built on top of Hadoop and is used for data analysis. Hive is targeted towards users who are comfortable with SQL. It is similar to SQL and call HiveQL, used for managing and querying structured data. Apache Hive is used to abstract complexity of Hadoop.

Working

In our **docker-compose.yml** file, we have added all the required docker images e.g. Hadoop images (datanode and namenode), hive etc. These are the important docker images which we will use to perform queries on apache spark platform. We will be using the **Hadoop** as our storage mechanism where we will keep our dataset (Big Data) and we will load our data into apache spark from hdfs path.

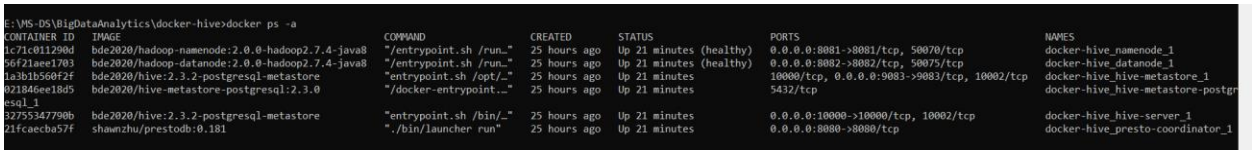
Steps

a. Run docker-compose file

```
E:\MS-DS\BigDataAnalytics\docker-hive>docker-compose up -d
Starting docker-hive_datanode_1                ... done
Starting docker-hive_hive-metastore_1          ... done
Starting docker-hive_presto-coordinator_1      ... done
Starting docker-hive_hive-server_1             ... done
Starting docker-hive_namenode_1                ... done
Starting docker-hive_hive-metastore-postgresql_1 ... done
```

b. Check containers are up and running

Command: docker ps -a



```
E:\MS-DS\BigDataAnalytics\docker-hive>docker ps -a
CONTAINER ID   IMAGE                                          COMMAND                  CREATED       STATUS                   PORTS                                              NAMES
1c71c011290d   bde2020/hadoop-namenode:2.0.0-hadoop2.7.4-java8   "/entrypoint.sh /run…"   25 hours ago   Up 21 minutes (healthy)   0.0.0.0:8081->8081/tcp, 50070/tcp                  docker-hive_namenode_1
56f21aee1703   bde2020/hadoop-datanode:2.0.0-hadoop2.7.4-java8   "/entrypoint.sh /run…"   25 hours ago   Up 21 minutes (healthy)   0.0.0.0:8082->8082/tcp, 50075/tcp                  docker-hive_datanode_1
1a3b1b560f2f   bde2020/hive:2.3.2-postgresql-metastore        "entrypoint.sh /opt/…"   25 hours ago   Up 21 minutes             10000/tcp, 0.0.0.0:9083->9083/tcp, 10002/tcp       docker-hive_hive-metastore_1
021846ee18d5   bde2020/hive-metastore-postgresql:2.3.0        "/docker-entrypoint.…"   25 hours ago   Up 21 minutes             5432/tcp                                           docker-hive_hive-metastore-postgr
esql_1
32755347790b   bde2020/hive:2.3.2-postgresql-metastore        "entrypoint.sh /bin/…"   25 hours ago   Up 21 minutes             0.0.0.0:10000->10000/tcp, 10002/tcp                docker-hive_hive-server_1
21fcaecba57f   shawnzhu/prestodb:0.181                        "./bin/launcher run"     25 hours ago   Up 21 minutes             0.0.0.0:8080->8080/tcp                             docker-hive_presto-coordinator_1
```

c. Copy data from local machine into container

```
E:\MS-DS\BigDataAnalytics\docker-hive>docker ps -a
CONTAINER ID   IMAGE                                          COMMAND                  CREATED       STATUS                    PORTS                                                NAMES
1c71c011290d   bde2020/hadoop-namenode:2.0.0-hadoop2.7.4-java8   "/entrypoint.sh /run…"   25 hours ago   Up 21 minutes (healthy)   0.0.0.0:8081->8081/tcp, 50070/tcp                    docker-hive_namenode_1
56f21aee1703   bde2020/hadoop-datanode:2.0.0-hadoop2.7.4-java8   "/entrypoint.sh /run…"   25 hours ago   Up 21 minutes (healthy)   0.0.0.0:8082->8082/tcp, 50075/tcp                    docker-hive_datanode_1
1a3b1b560f2f   bde2020/hive:2.3.2-postgresql-metastore        "entrypoint.sh /opt/…"   25 hours ago   Up 21 minutes             10000/tcp, 0.0.0.0:9083->9083/tcp, 10002/tcp         docker-hive_hive-metastore_1
021846ee18d5   bde2020/hive-metastore-postgresql:2.3.0        "/docker-entrypoint.…"   25 hours ago   Up 21 minutes             5432/tcp                                             docker-hive_hive-metastore-postgr
esql_1
32755347790b   bde2020/hive:2.3.2-postgresql-metastore        "entrypoint.sh /bin/…"   25 hours ago   Up 21 minutes             0.0.0.0:10000->10000/tcp, 10002/tcp                  docker-hive_hive-server_1
21fcaecba57f   shawnzhu/prestodb:0.181                        "./bin/launcher run"     25 hours ago   Up 21 minutes             0.0.0.0:8080->8080/tcp                               docker-hive_presto-coordinator_1

E:\MS-DS\BigDataAnalytics\docker-hive>docker cp E:\MS-DS\BigDataAnalytics\data 32755347790b:home

E:\MS-DS\BigDataAnalytics\docker-hive>
```

d. Execute hive container

```
E:\MS-DS\BigDataAnalytics\docker-hive>docker cp E:\MS-DS\BigDataAnalytics\data 32755347790b:home

E:\MS-DS\BigDataAnalytics\docker-hive>docker compose exec hive-server bash
root@32755347790b:/opt#
```

e. Locate your data into container

```
E:\MS-DS\BigDataAnalytics\docker-hive>docker cp E:\MS-DS\BigDataAnalytics\data 32755347790b:home

E:\MS-DS\BigDataAnalytics\docker-hive>docker compose exec hive-server bash
root@32755347790b:/opt# cd ..
root@32755347790b:/# cd home/
root@32755347790b:/home# cd data/
root@32755347790b:/home/data# ls
Chicago Crimes_-_2001_to_Present.csv
root@32755347790b:/home/data#
```

f. Rename your file (Remove spaces)

```
E:\MS-DS\BigDataAnalytics\docker-hive>docker compose exec hive-server bash
root@32755347790b:/opt# cd ..
root@32755347790b:/# cd home/
root@32755347790b:/home# cd data/
root@32755347790b:/home/data# ls
Chicago Crimes_-_2001_to_Present.csv
root@32755347790b:/home/data# mv Chicago\ Crimes_-_2001_to_Present.csv ChicagoCrimesData
root@32755347790b:/home/data# ls
ChicagoCrimesData
root@32755347790b:/home/data#
```

g. Put data onto Hadoop

```
root@32755347790b:/home/data# hadoop fs -mkdir /user/dataset
root@32755347790b:/home/data# hadoop fs -put -f /home/data /user/dataset
root@32755347790b:/home/data#
```

h. Run Apache within container

```
root@32755347790b:/home/data# hadoop fs -put -f /home/data /user/dataset
root@32755347790b:/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>
```

i. Create Table

```
0: jdbc:hive2://localhost:10000> CREATE TABLE ChicagoCrimesTable (ID String,CaseNumber String,Block String,IUCR String,PrimaryType String,Description String,Arrest String,Domestic String,Beat String,District Str
ing,Ward String,CommunityArea String,FBICode String,XCoordinate String,YCoordinate String,Year String,UpdatedOn String,Latitude String,Longitude String,Location String) row format delimited fields terminated by
',';
No rows affected (1.538 seconds)
0: jdbc:hive2://localhost:10000>
```

Queries

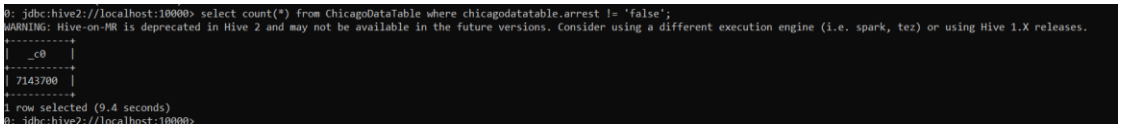
a. Get 10 records from the provided data

```
0: jdbc:hive2://localhost:10000> select * from ChicagoDataTable limit 10;
+-----------------------+-----------------------------+-----------------------+-----------------------+-------------------------------+-----------------------------------+-----------------------+-----------------------+-----------------------+-----------------------+-----------------------+-----------------------+
| chicagodatatable.id  | chicagodatatable.casenumber  | chicagodatatable.block  | chicagodatatable.iucr  | chicagodatatable.primarytype  | chicagodatatable.description  |       chicagodatatable.arrest       | chicagodatatable.domestic  | chicagodatatable.beat  | chicagodatatable.district  | chicagodatatable.ward  | chicagodatatable.communityarea  | chicagodatatable.fbicode  | chicagodatatable.xcoordinate  | chicagodatatable.ycoordinate  | chicagodatatable.year  | chicagodatatable.updatedon  | chicagodatatable.latitude  | chicagodatatable.longitude  | chicagodatatable.location  |
+-----------------------+-----------------------------+-----------------------+-----------------------+-------------------------------+-----------------------------------+-----------------------+-----------------------+-----------------------+-----------------------+-----------------------+-----------------------+
+-----------------------+-----------------------------+-----------------------+-----------------------+-------------------------------+-----------------------------------+-----------------------+-----------------------+-----------------------+-----------------------+-----------------------+-----------------------+
| ID                    | Case Number                 | Date                  | Block                 | IUCR                          | Primary Type                      | Description           | Location Description  | Arrest                | Domestic              | Beat                  | District              | Ward                  | Community Area        | FBI Code              | X Coordinate          | Y Coordinate          | Year                  | Updated On            | Latitude              |
| 11034701              | JA366925                    | 01/01/2001 11:00:00 AM | 016XX E 86TH PL       | 1153                          | DECEPTIVE PRACTICE                | FINANCIAL IDENTITY THEFT OVER $ 300 | RESIDENCE             | false                 | false                 | 0412                  | 004                   | 8                     | 45                    | 11                    |                       |                       | 2001                  | 08/05/2017 03:50:08 PM |                       |
| 11227287              | JB147188                    | 10/08/2017 03:00:00 AM | 092XX S RACINE AVE    | 0281                          | CRIM SEXUAL ASSAULT               | NON-AGGRAVATED        | RESIDENCE             | false                 | false                 | 2222                  | 022                   | 21                    | 73                    | 02                    |                       |                       | 2017                  | 02/11/2018 03:57:41 PM |                       |
| 11227583              | JB147595                    | 03/28/2017 02:00:00 PM | 026XX W 79TH ST       | 0620                          | BURGLARY                          | UNLAWFUL ENTRY        | OTHER                 | false                 | false                 | 0835                  | 008                   | 18                    | 70                    | 05                    |                       |                       | 2017                  | 02/11/2018 03:57:41 PM |                       |
| 11227293              | JB147230                    | 09/09/2017 08:17:00 PM | 060XX S EBERHART AVE  | 0810                          | THEFT                             | OVER $500             | RESIDENCE             | false                 | false                 | 0313                  | 003                   | 20                    | 42                    | 06                    |                       |                       | 2017                  | 02/11/2018 03:57:41 PM |                       |
| 11227634              | JB147599                    | 08/26/2017 10:00:00 AM | 001XX W RANDOLPH ST   | 0281                          | CRIM SEXUAL ASSAULT               | NON-AGGRAVATED        | HOTEL/MOTEL           | false                 | false                 | 0122                  | 001                   | 42                    | 32                    | 02                    |                       |                       | 2017                  | 02/11/2018 03:57:41 PM |                       |
| 11227517              | JB138481                    | 02/10/2013 12:00:00 AM | 071XX S LAFAYETTE AVE | 0266                          | CRIM SEXUAL ASSAULT               | PREDATORY             | RESIDENCE             | false                 | false                 | 0731                  | 007                   | 6                     | 69                    | 02                    |                       |                       | 2013                  | 02/11/2018 03:57:41 PM |                       |
| 11227503              | JB146383                    | 01/01/2015 12:01:00 AM | 061XX S KILBOURN AVE  | 1751                          | OFFENSE INVOLVING CHILDREN        | CRIM SEX ABUSE BY FAM MEMBER | RESIDENCE             | false                 | true                  | 0813                  | 008                   | 13                    | 65                    | 17                    |                       |                       | 2015                  | 04/12/2019 04:00:15 PM |                       |
| 11227508              | JB146365                    | 01/01/2017 12:01:00 AM | 027XX S WHIPPLE ST    | 1754                          | OFFENSE INVOLVING CHILDREN        | AGG SEX ASSLT OF CHILD FAM MBR | RESIDENCE             | false                 | false                 | 1033                  | 010                   | 12                    | 30                    | 02                    |                       |                       | 2017                  | 02/11/2018 03:57:41 PM |                       |
| 11022695              | JA353568                    | 07/17/2017 10:10:00 AM | 021XX W MC LEAN AVE   | 0810                          | THEFT                             | OVER $500             | RESIDENCE             | false                 | false                 | 1432                  | 014                   | 32                    | 22                    | 06                    |                       |                       | 2017                  | 07/24/2017 03:54:23 PM |                       |
+-----------------------+-----------------------------+-----------------------+-----------------------+-------------------------------+-----------------------------------+-----------------------+-----------------------+-----------------------+-----------------------+-----------------------+-----------------------+
10 rows selected (0.12 seconds)
0: jdbc:hive2://localhost:10000>
```

b. Get count of all records available in the dataset

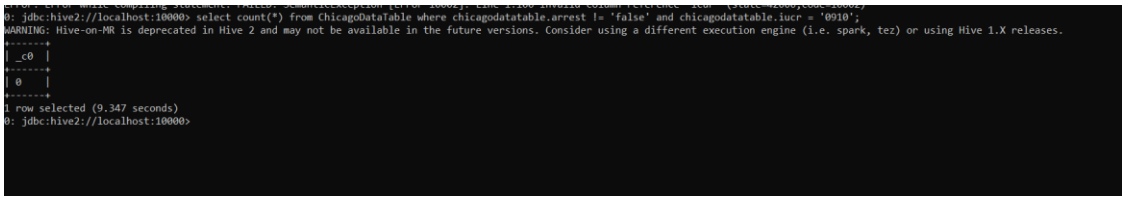
```
0: jdbc:hive2://localhost:10000> select count(*) from ChicagoDataTable;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
+----------+
|   _c0    |
+----------+
| 7143700  |
+----------+
1 row selected (5.434 seconds)
0: jdbc:hive2://localhost:10000>
```

c. Get count of all the crimes where no arrest happened



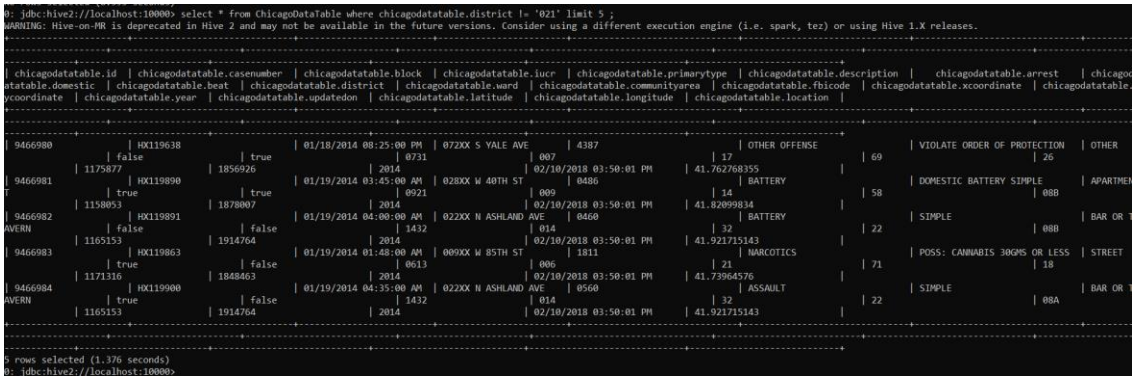
```
0: jdbc:hive2://localhost:10000> select count(*) from ChicagoDataTable where chicagodatatable.arrest != 'false';
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
+----------+
|   _c0    |
+----------+
| 7143700  |
+----------+
1 row selected (9.4 seconds)
0: jdbc:hive2://localhost:10000>
```

d. Get the count of all the crimes where no arrest happened in the icur 0910.



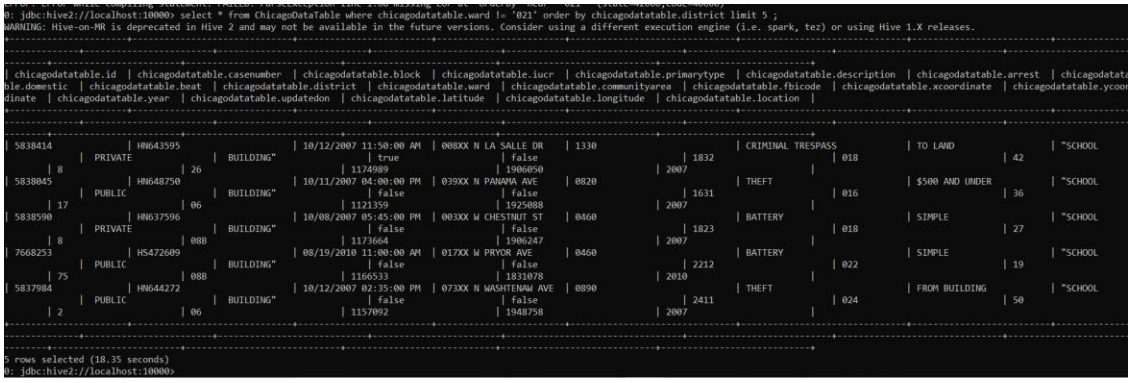
```
0: jdbc:hive2://localhost:10000> select count(*) from ChicagoDataTable where chicagodatatable.arrest != 'false' and chicagodatatable.iucr = '0910';
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
+------+
| _c0  |
+------+
| 0    |
+------+
1 row selected (9.347 seconds)
0: jdbc:hive2://localhost:10000>
```

e. Get five records of crimes which happened in districts other than 021



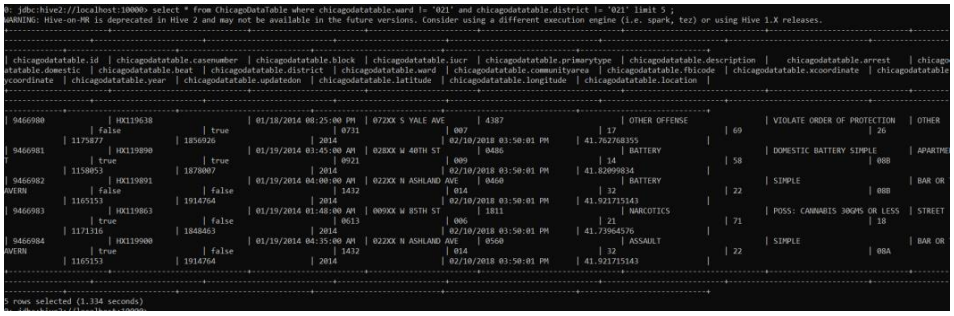
```
0: jdbc:hive2://localhost:10000> select * from ChicagoDataTable where chicagodatatable.district != '021' limit 5 ;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
+-----------------------------+-----------------------------------+-----------------------------+-----------------------------+-------------------------------+-----------------------------------+-----------------------------+
| chicagodatatable.id | chicagodatatable.casenumber | chicagodatatable.block | chicagodatatable.iucr | chicagodatatable.primarytype | chicagodatatable.description | chicagodatatable.arrest | chicagodatatable.domestic | chicagodatatable.beat | chicagodatatable.district | chicagodatatable.ward | chicagodatatable.communityarea | chicagodatatable.fbicode | chicagodatatable.xcoordinate | chicagodatatable.ycoordinate | chicagodatatable.year | chicagodatatable.updatedon | chicagodatatable.latitude | chicagodatatable.longitude | chicagodatatable.location |
+-----------------------------+-----------------------------------+-----------------------------+-----------------------------+-------------------------------+-----------------------------------+-----------------------------+
| 9466980 | HX119638 | 01/18/2014 08:25:00 PM | 072XX S YALE AVE | 4387 | OTHER OFFENSE | VIOLATE ORDER OF PROTECTION | OTHER | false | true | 0731 | 007 | 17 | 69 | 26 | 1175877 | 1856926 | 2014 | 02/10/2018 03:50:01 PM | 41.762768355 |
| 9466981 | HX119890 | 01/19/2014 03:45:00 AM | 028XX W 40TH ST | 0486 | BATTERY | DOMESTIC BATTERY SIMPLE | APARTMENT | true | true | 0921 | 009 | 14 | 58 | 08B | 1158053 | 1878007 | 2014 | 02/10/2018 03:50:01 PM | 41.82099834 |
| 9466982 | HX119891 | 01/19/2014 04:00:00 AM | 022XX N ASHLAND AVE | 0460 | BATTERY | SIMPLE | BAR OR TAVERN | false | false | 1432 | 014 | 32 | 22 | 08B | 1165153 | 1914764 | 2014 | 02/10/2018 03:50:01 PM | 41.921715143 |
| 9466983 | HX119863 | 01/19/2014 01:48:00 AM | 009XX W 85TH ST | 1811 | NARCOTICS | POSS: CANNABIS 30GMS OR LESS | STREET | true | false | 0613 | 006 | 21 | 71 | 18 | 1171316 | 1848463 | 2014 | 02/10/2018 03:50:01 PM | 41.73964576 |
| 9466984 | HX119900 | 01/19/2014 04:35:00 AM | 022XX N ASHLAND AVE | 0560 | ASSAULT | SIMPLE | BAR OR TAVERN | true | false | 1432 | 014 | 32 | 22 | 08A | 1165153 | 1914764 | 2014 | 02/10/2018 03:50:01 PM | 41.921715143 |
+-----------------------------+-----------------------------------+-----------------------------+-----------------------------+-------------------------------+-----------------------------------+-----------------------------+
5 rows selected (1.376 seconds)
0: jdbc:hive2://localhost:10000>
```

f. Get five records of crimes which happened in ward other than 021 and fetch order by district

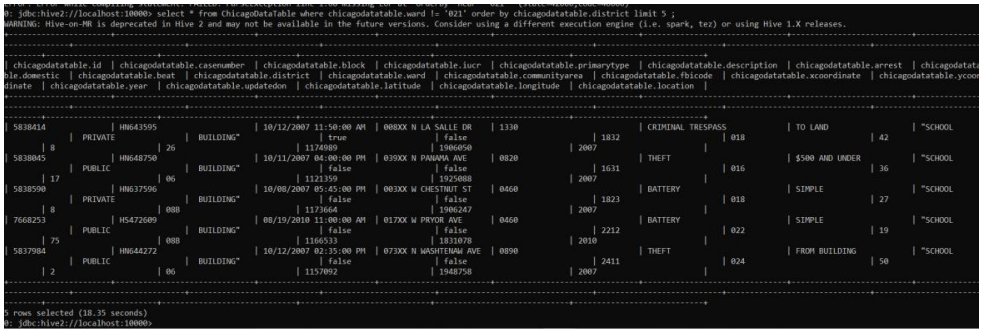


```
0: jdbc:hive2://localhost:10000> select * from ChicagoDataTable where chicagodatatable.ward != '021' order by chicagodatatable.district limit 5 ;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
+-----------------------------+-----------------------------------+-----------------------------+-----------------------------+-------------------------------+-----------------------------------+-----------------------------+
| chicagodatatable.id | chicagodatatable.casenumber | chicagodatatable.block | chicagodatatable.iucr | chicagodatatable.primarytype | chicagodatatable.description | chicagodatatable.arrest | chicagodatatable.domestic | chicagodatatable.beat | chicagodatatable.district | chicagodatatable.ward | chicagodatatable.communityarea | chicagodatatable.fbicode | chicagodatatable.xcoordinate | chicagodatatable.ycoordinate | chicagodatatable.year | chicagodatatable.updatedon | chicagodatatable.latitude | chicagodatatable.longitude | chicagodatatable.location |
+-----------------------------+-----------------------------------+-----------------------------+-----------------------------+-------------------------------+-----------------------------------+-----------------------------+
| 5838414 | HN643595 | 10/12/2007 11:50:00 AM | 008XX N LA SALLE DR | 1330 | CRIMINAL TRESPASS | TO LAND | "SCHOOL | PRIVATE | BUILDING" | true | false | 1832 | 018 | 42 | 8 | 26 | 1174989 | 1906050 | 2007 |
| 5838045 | HN648750 | 10/11/2007 04:00:00 PM | 039XX N PANAMA AVE | 0820 | THEFT | $500 AND UNDER | "SCHOOL | PUBLIC | BUILDING" | false | false | 1631 | 016 | 36 | 17 | 06 | 1121359 | 1925088 | 2007 |
| 5838590 | HN637596 | 10/08/2007 05:45:00 PM | 003XX W CHESTNUT ST | 0460 | BATTERY | SIMPLE | "SCHOOL | PRIVATE | BUILDING" | false | false | 1823 | 018 | 27 | 8 | 08B | 1173664 | 1906247 | 2007 |
| 7668253 | HS472609 | 08/19/2010 11:00:00 AM | 017XX W PRYOR AVE | 0460 | BATTERY | SIMPLE | "SCHOOL | PUBLIC | BUILDING" | false | false | 2212 | 022 | 19 | 75 | 08B | 1166533 | 1831078 | 2010 |
| 5837984 | HN644272 | 10/12/2007 02:35:00 PM | 073XX N WASHTENAW AVE | 0890 | THEFT | FROM BUILDING | "SCHOOL | PUBLIC | BUILDING" | false | false | 2411 | 024 | 50 | 2 | 06 | 1157092 | 1948758 | 2007 |
+-----------------------------+-----------------------------------+-----------------------------+-----------------------------+-------------------------------+-----------------------------------+-----------------------------+
5 rows selected (18.35 seconds)
0: jdbc:hive2://localhost:10000>
```

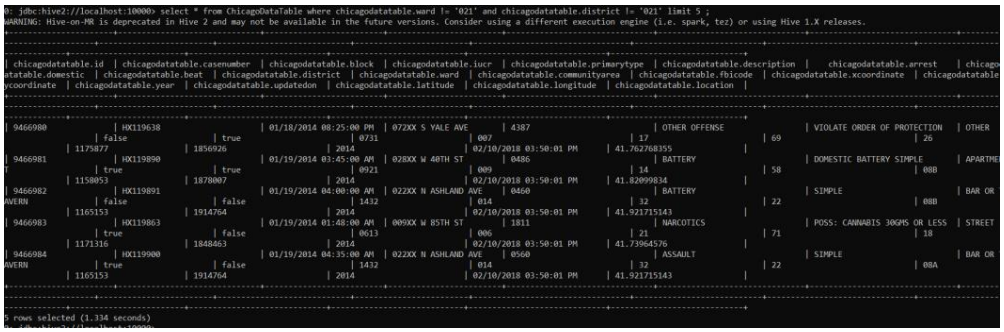
g. Get the five records of same ward and district



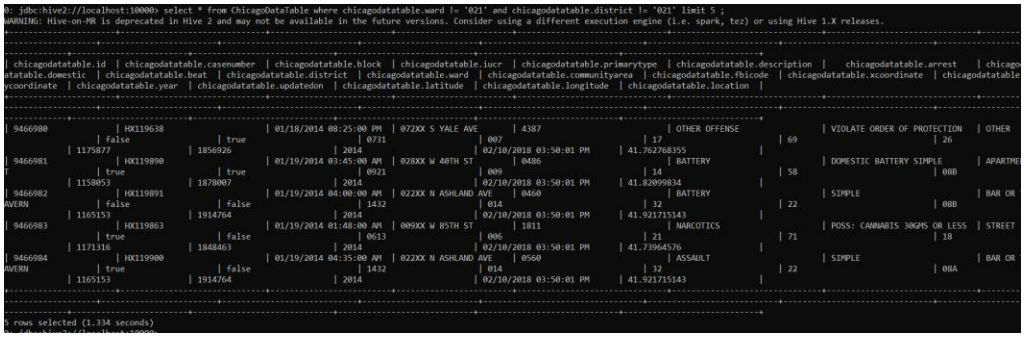
h. Get records where the arrest happened in the districts order by iucr



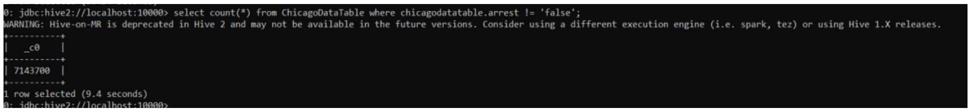
i. Get the records by the year where the ratio of crimes was found highest order by districts



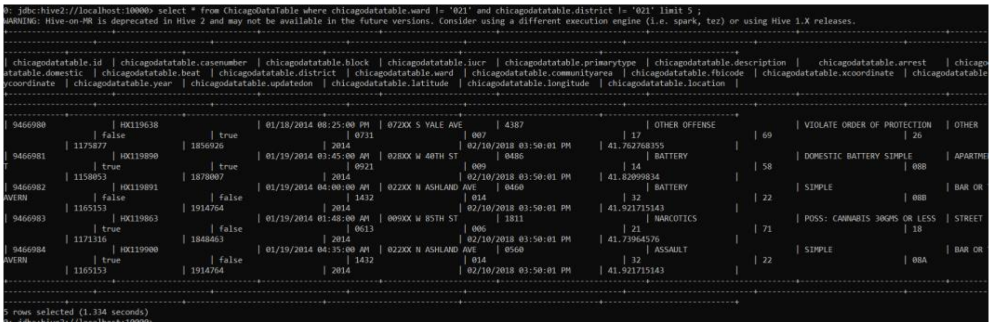
j. Get the records where the icur is not 021 and the ward is 21 order by location



k. Get the count of all arrest that happened in the district 038


0: jdbc:hive2://localhost:10000> select count(*) from ChicagoDataTable where chicagodatatable.arrest != 'false';
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
| _c0 |
| 7143700 |
1 row selected (9.4 seconds)

l. Get the records where the crimes were domestic and the ward is 021

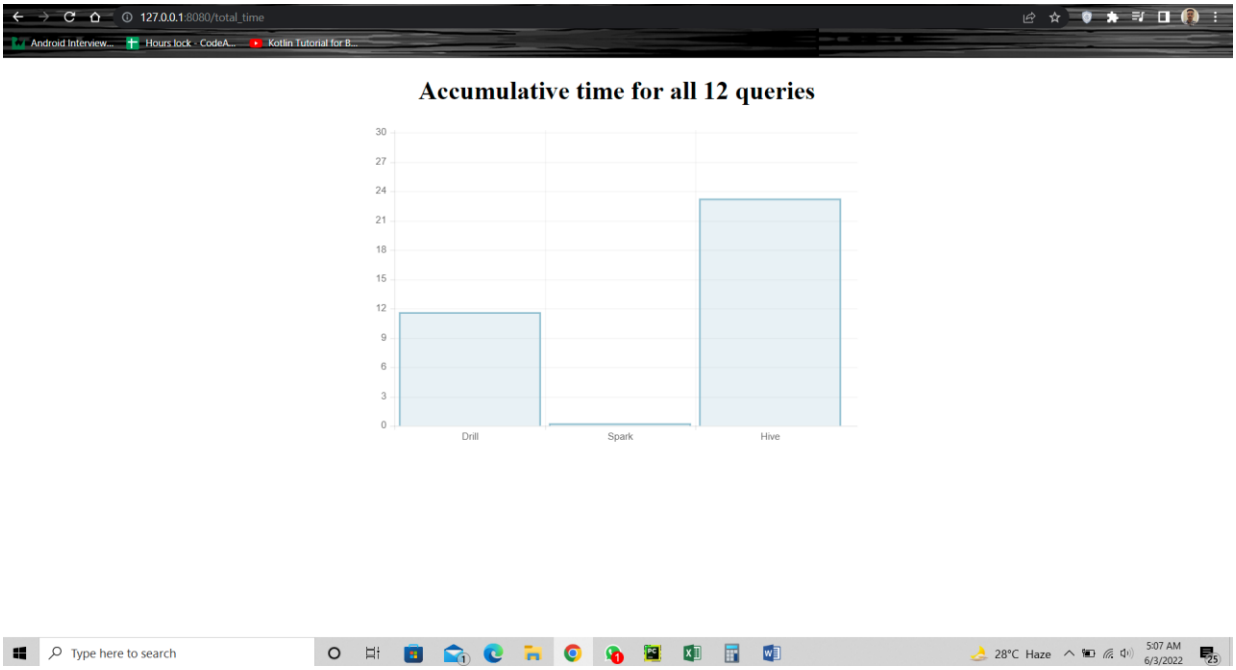

0: jdbc:hive2://localhost:10000> select * from ChicagoDataTable where chicagodatatable.ward != '021' and chicagodatatable.district != '021' limit 5 ;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
| chicagodatatable.id | chicagodatatable.casenumber | chicagodatatable.block | chicagodatatable.iucr | chicagodatatable.primarytype | chicagodatatable.description | chicagodatatable.arrest |
atatable.domestic | chicagodatatable.beat | chicagodatatable.district | chicagodatatable.ward | chicagodatatable.communityarea | chicagodatatable.fbicode | chicagodatatable.xcoordinate |
ycoordinate | chicagodatatable.year | chicagodatatable.updatedon | chicagodatatable.latitude | chicagodatatable.longitude | chicagodatatable.location |
| 9466980 | HX119638 | 01/18/2014 08:25:00 PM | 072XX S YALE AVE | 4387 | OTHER OFFENSE | VIOLATE ORDER OF PROTECTION | OTHER
| false | true | 0731 | 007 | 17 | 69 | 26
| 1175877 | 1856926 | 2014 | 02/10/2018 03:50:01 PM | 41.762768355 |
| 9466981 | HX119890 | 01/19/2014 03:45:00 AM | 028XX W 40TH ST | 0486 | BATTERY | DOMESTIC BATTERY SIMPLE
T | true | true | 0921 | 009 | 14 | 58 | 08B
| 1158053 | 1878007 | 2014 | 02/10/2018 03:50:01 PM | 41.82099834 |
| 9466982 | HX119891 | 01/19/2014 04:00:00 AM | 022XX N ASHLAND AVE | 0460 | BATTERY | SIMPLE
AVERN | false | false | 1432 | 014 | 32 | 22 | 08B
| 1165153 | 1914764 | 2014 | 02/10/2018 03:50:01 PM | 41.921715143 |
| 9466983 | HX119863 | 01/19/2014 01:48:00 AM | 009XX W 85TH ST | 1811 | NARCOTICS | POSS: CANNABIS 30GMS OR LESS | STREET
| true | false | 0613 | 006 | 21 | 71 | 18
| 1171316 | 1848463 | 2014 | 02/10/2018 03:50:01 PM | 41.73964576 |
| 9466984 | HX119900 | 01/19/2014 04:35:00 AM | 022XX N ASHLAND AVE | 0560 | ASSAULT | SIMPLE
AVERN | true | false | 1432 | 014 | 32 | 22 | 08A
| 1165153 | 1914764 | 2014 | 02/10/2018 03:50:01 PM | 41.921715143 |
5 rows selected (1.334 seconds)

Total Execution Time on Apache-Hive Queries Performed

Query1	1.043 (Seconds)
Query2	4.037 (Seconds)
Query3	2.038 (Seconds)
Query4	1.091 (Seconds)
Query5	0.914 (Seconds)
Query6	1.025 (Seconds)
Query7	1.317 (Seconds)
Query8	1.818 (Seconds)
Query9	1.512 (Seconds)
Query10	1.810 (Seconds)
Query11	3.810 (Seconds)
Query12	2.013 (Seconds)

Frontend

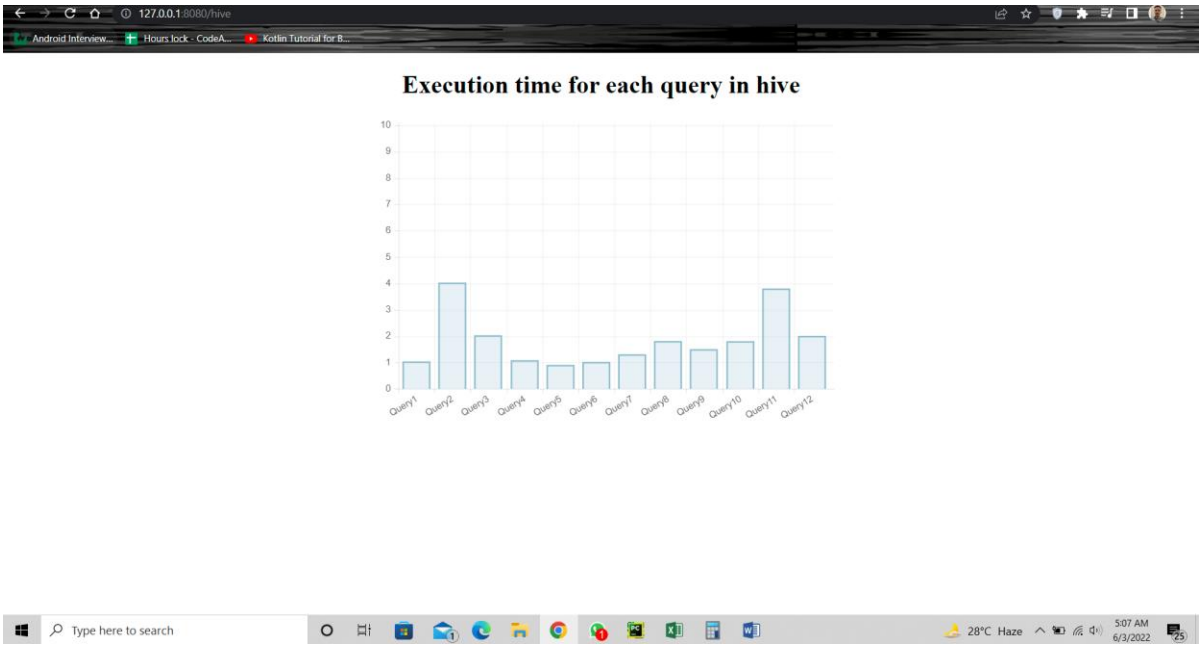
Total Time

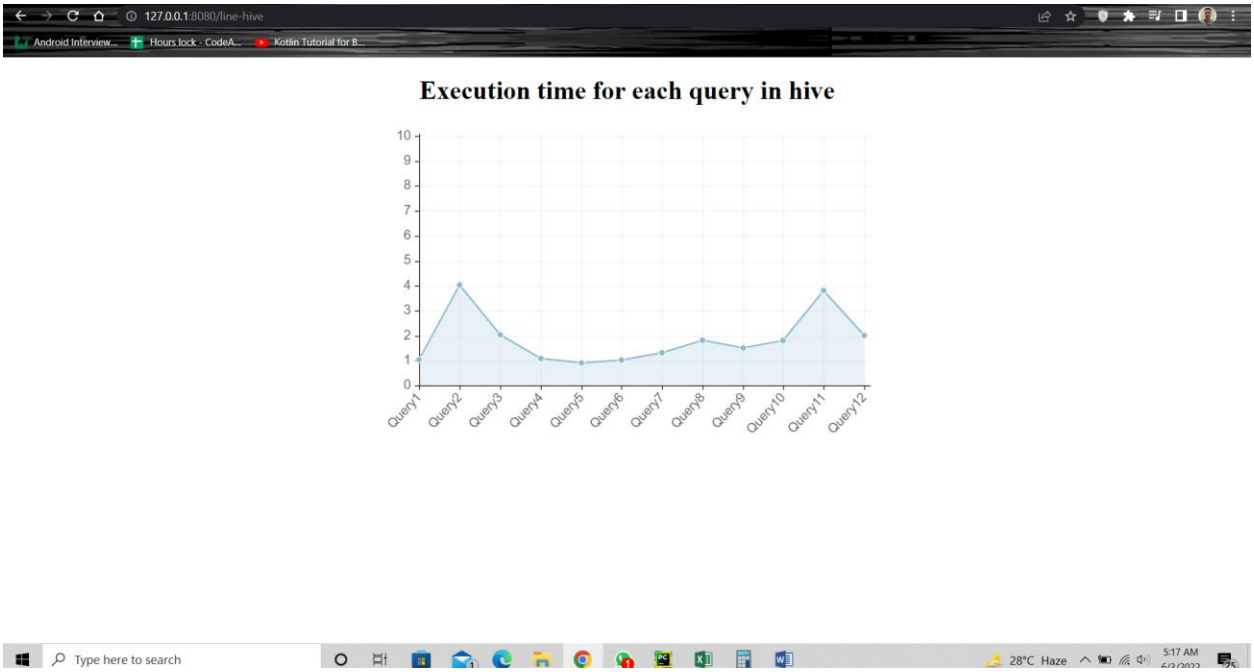


One can easily see the difference between the execution time of spark vs hive and drill. Spark has executed each query in milliseconds. Whereas, for some of the complex queries hive and drill execution time went to two to five seconds respectively.

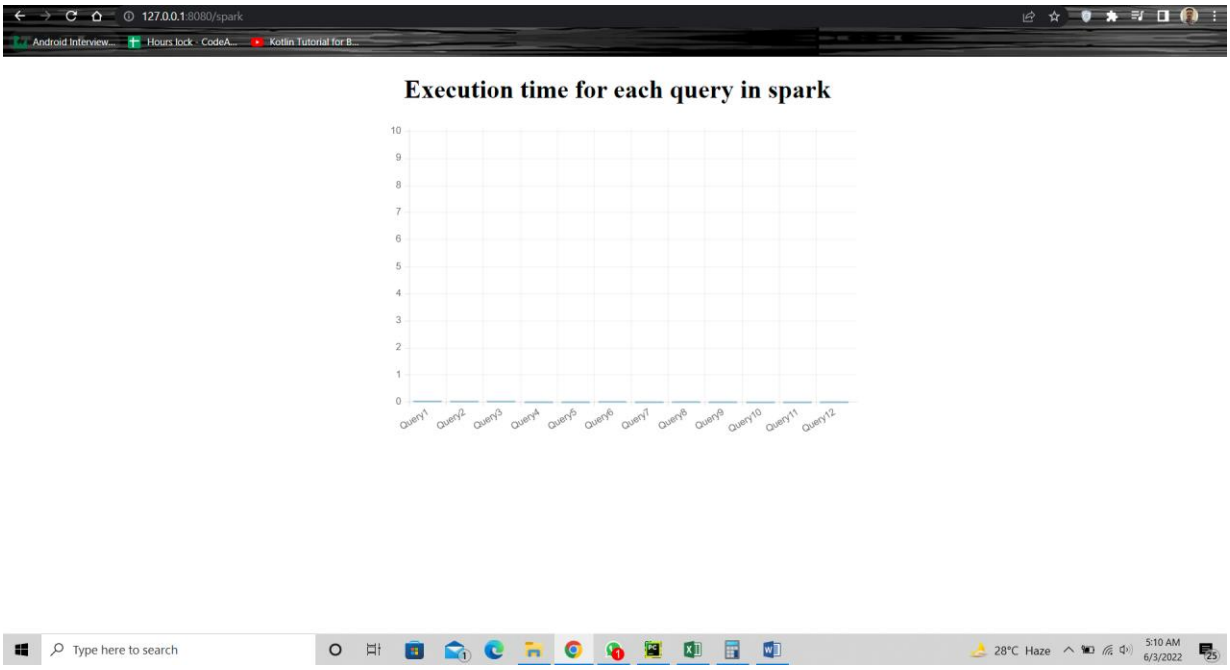
Some of the charts from the frontend project are pasted below as well to mark the query execution power of spark which has been proved to be a lightning fast query engine.

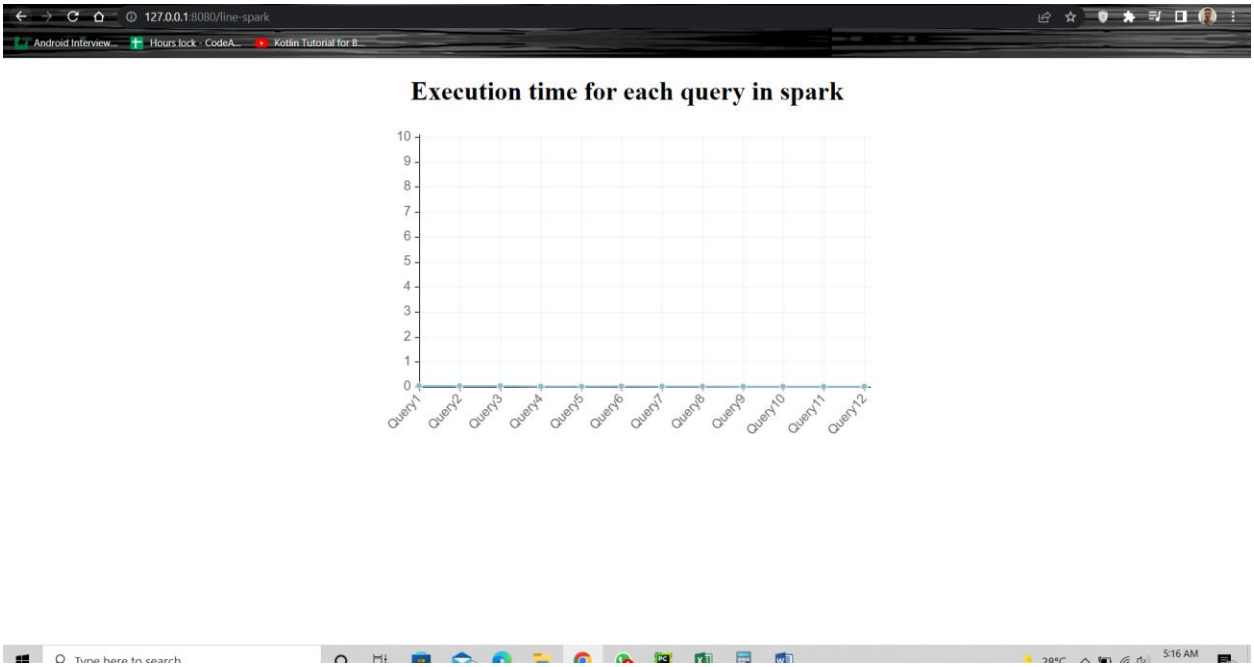
Hive


127.0.0.1:8080/hive
Android Interview...
Hours lock - CodeA...
Kotlin Tutorial for B...
Execution time for each query in hive
10
9
8
7
6
5
4
3
2
1
0
Query1
Query2
Query3
Query4
Query5
Query6
Query7
Query8
Query9
Query10
Query11
Query12
Type here to search
28°C Haze
5:07 AM
6/3/2022

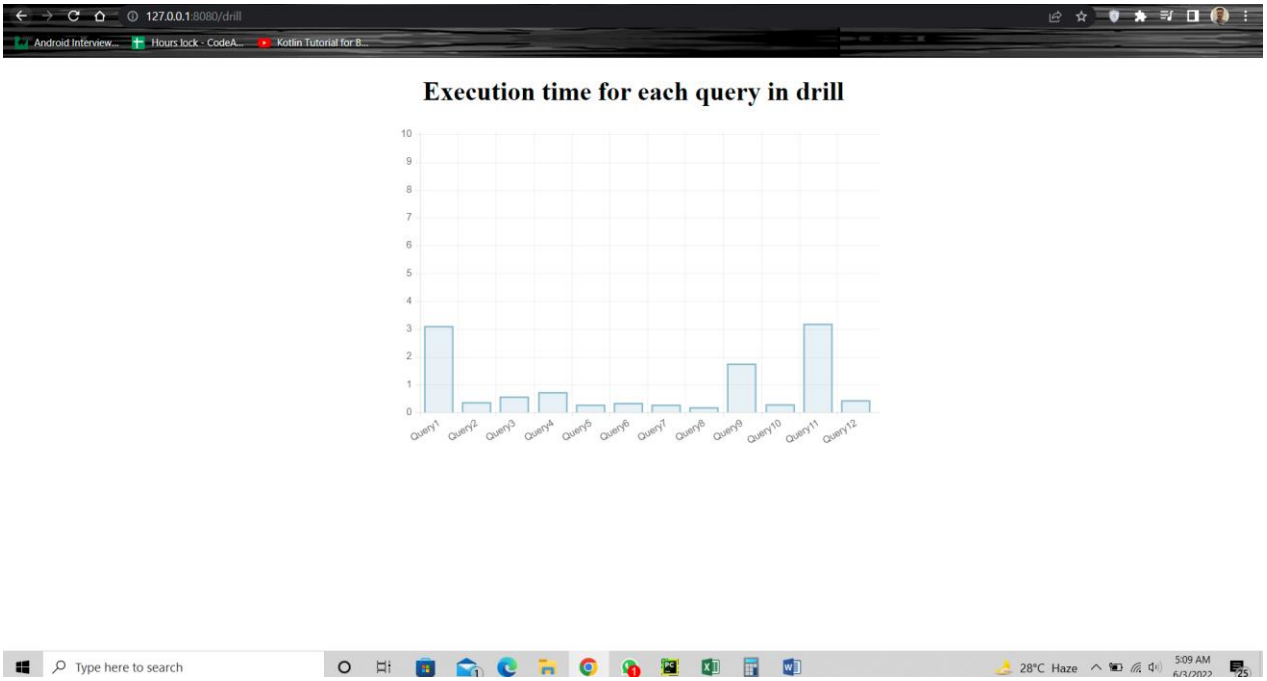

127.0.0.1:8080/line-hive
Android Interview...
Hours lock - CodeA...
Kotlin Tutorial for B...
Execution time for each query in hive
10
9
8
7
6
5
4
3
2
1
0
Query1
Query2
Query3
Query4
Query5
Query6
Query7
Query8
Query9
Query10
Query11
Query12
Type here to search
28°C Haze
5:17 AM

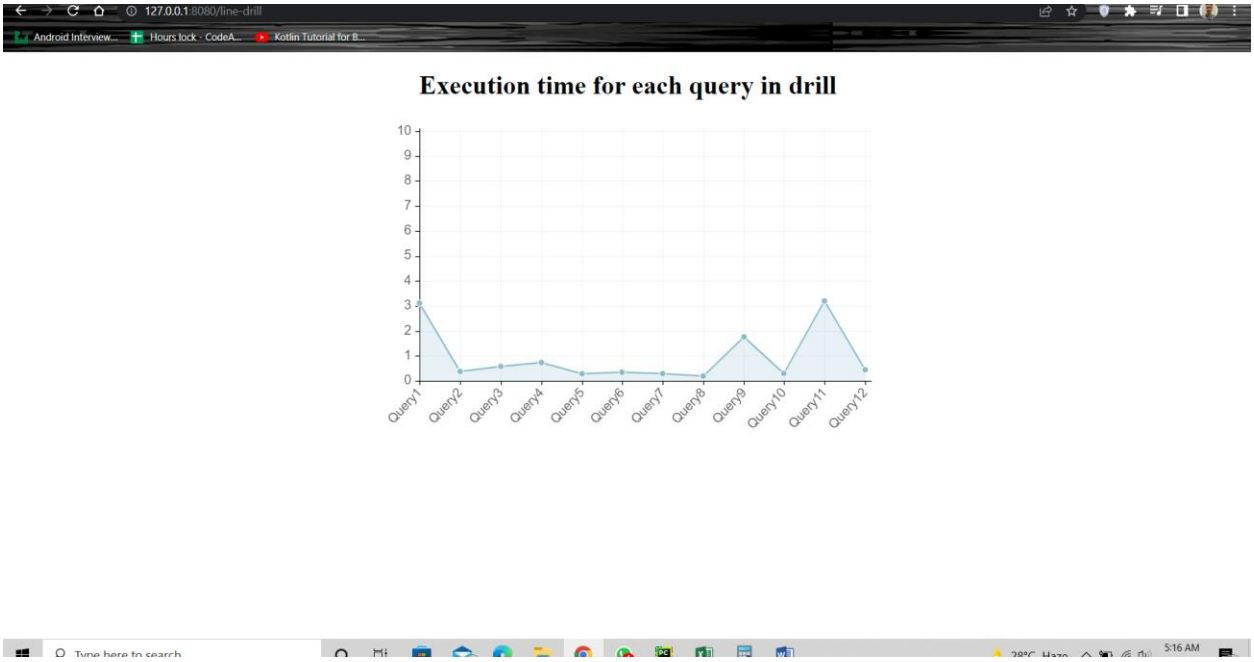
Spark


127.0.0.1:8080/spark
Android Interview...
Hours lock · CodeA...
Kotlin Tutorial for B...
Execution time for each query in spark
10
9
8
7
6
5
4
3
2
1
0
Query1
Query2
Query3
Query4
Query5
Query6
Query7
Query8
Query9
Query10
Query11
Query12
Type here to search
28°C Haze
5:10 AM
6/3/2022


127.0.0.1:8080/line-spark
Android Interview...
Hours lock · CodeA...
Kotlin Tutorial for B...
Execution time for each query in spark
10
9
8
7
6
5
4
3
2
1
0
Query1
Query2
Query3
Query4
Query5
Query6
Query7
Query8
Query9
Query10
Query11
Query12
5:16 AM

Drill





Few more charts are also integrated in the frontend project as well.

Apache Spark Tutorial Link

https://www.youtube.com/watch?v=2Vp27NADslw

(Available on my personal YouTube Channel)

Apache Hive Tutorial Link

https://www.youtube.com/watch?v=o99PJZstHdY

(Available on my personal YouTube Channel)

Apache Drill Tutorial Link

https://www.youtube.com/watch?v=mbT35-HN3bU

(Available on my personal YouTube Channel)