



2/6/2022

FINAL PROJECT REPORT

BIG DATA ANALYTICS

Arslan Ahmed Shaikh
ERP 23388

Contents

Introduction	2
Dataset.....	2
APACHE DRILL	3
Introduction:	3
Working.....	3
Steps	3
Queries.....	7
Total Execution Time on Apache-Drill Queries Performed	9
APACHE SPARK	10
Introduction	10
Working.....	10
Steps	10
Queries.....	12
Total Execution Time on Apache-Spark Queries Performed.....	14
APACHE HIVE.....	15
Introduction	15
Working.....	15
Steps	15
Queries.....	17
Total Execution Time on Apache-Hive Queries Performed	20
Frontend	21
Total Time	21
Hive	22
Spark	23
Drill.....	23
Apache Spark Tutorial Link	25
Apache Hive Tutorial Link	25
Apache Drill Tutorial Link.....	25

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.

Chicago Crime Dataset (2001 - Present)

Data

Code (0)

Discussion (0)

Metadata

4

New Notebook

Detail	Compact	Column	10 of 22 columns											
	Primary Type		Description		Location Description		Arrest		Domestic					
8%	THEFT	21%	SIMPLE	12%	STREET	26%	<div><div></div></div>	true	1.85m	28%	<div><div></div></div>	true	872k	13%
8%	BATTERY	18%	\$500 AND UNDER	8%	RESIDENCE	17%	<div><div></div></div>	false	4.81m	72%	<div><div></div></div>	false	5.79m	87%
4%	Other (4048422)	61%	Other (5336021)	80%	Other (3781692)	57%								
	DECEPTIVE PRACTICE		FINANCIAL IDENTITY THEFT OVER \$ 300		RESIDENCE		false				false			
	OTHER OFFENSE		VIOLENT OFFENDER: ANNUAL REGISTRATION		JAIL / LOCK-UP FACILITY		true				false			
	ROBBERY		ARMED: HANDGUN		SIDEWALK		true				false			
	CRIM SEXUAL ASSAULT		NON-AGGRAVATED		RESIDENCE		false				false			
	BURGLARY		UNLAWFUL ENTRY		OTHER		false				false			
	THEFT		OVER \$500		RESIDENCE		false				false			
	ASSAULT		AGGRAVATED: HANDGUN		DEPARTMENT STORE		true				false			
	CRIM SEXUAL ASSAULT		NON-AGGRAVATED		HOTEL/MOTEL		false				false			

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 from 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

```
E:\MS-DS\BigDataAnalytics\docker-drill\docker-apache-drill>docker-compose up -d
Pulling journalnode-1 (smizy/hadoop-base:2.7.7-alpine)...
2.7.7-alpine: Pulling from smizy/hadoop-base
f4900964ff56: Pull complete
3cff98b53f39: Pull complete
cb83007c8208: Pull complete
3bbc69e9f311: Pull complete
f5580162dbe0: Pull complete
Digest: sha256:69c184024241a486a48c1c8fbed759ce796bb79744342b41350b018460a3631e
Status: Downloaded newer image for smizy/hadoop-base:2.7.7-alpine
Recreating zookeeper-1 ... done
Creating namenode-2 ... done
Creating journalnode-3 ... done
Creating journalnode-2 ... done
Creating zookeeper-2 ... done
Creating datanode-1 ... done
Creating datanode-3 ... done
Creating datanode-2 ... done
Creating namenode-1 ... done
Creating journalnode-1 ... done
Creating zookeeper-3 ... done
Recreating drillbit-1 ... done

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

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->38047/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
3b3b495e0cda	smizy/hadoop-base:2.7.7-alpine	"entrypoint.sh datan..."	3 minutes ago	Up 2 minutes	50010/tcp, 50020/tcp, 50075/tcp	datanode-2
433b349184e9	smizy/zookeeper:3.4-alpine	"entrypoint.sh -serv..."	3 minutes ago	Up 2 minutes	2181/tcp, 2888/tcp, 3888/tcp	zookeeper-3
c714be101231	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
4404212c7cd3	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
bb401d8342e7	smizy/hadoop-base:2.7.7-alpine	"entrypoint.sh namen..."	3 minutes ago	Exited (1) 2 minutes ago		namenode-2
48bb27db8932	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_dat
anode_1						
efc9e31a1f93	bde2020/hadoop-namenode:1.1.0-hadoop2.8-java8	"/entrypoint.sh /run..."	22 hours ago	Exited (137) 21 hours ago		namenode
8c3d1f22b6f6	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_spa
pk-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						
1c71c01129bd	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
56f21ae178b3	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
1a1b1b560f2f	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
21fcaecbe57f	shawnzhu/prestodb:0.101	"/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 interactive 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 -mkdir** 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
```

```
bash-4.3$ cd ..
```

```
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
```

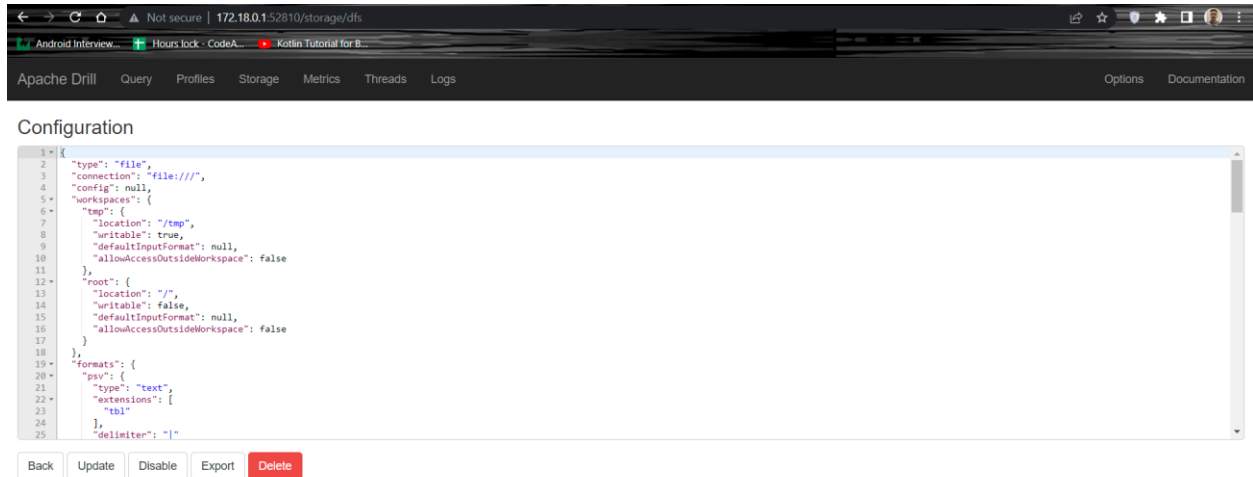
```
-rw-r--r-- 3 hdfs hadoop 1684766921 2022-06-01 20:48 /user/hdfs/output/ChicagoCrimesData.csv
```

```
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.

Apache Drill

Query

Profiles

Storage

Metrics

Threads

Logs

Options

Documentation

Sample SQL query: `SELECT * FROM cp.`employee.json` LIMIT 20`

Query Type

SQL

PHYSICAL

LOGICAL

Query

1

`select * from dfs.root.`output/ChicagoCrimesData.csv` limit 20;`

Hint: Use `Ctrl+Enter` to submit

Submit

Limit results to

1000

rows

Output:

Apache Drill

Query

Profiles

Storage

Metrics

Threads

Logs

Options

Documentation

Query Profile: 19582e35-846d-4967-82a6-0dce2b591668 COMPLETED

Delimiter

Export

Show 10 entries

Search:

Show / hide columns

columns

[{"1674542","G465063","08/06/2001 04:51:46 PM","0000X N STATE ST","0820","THEFT","\$500 AND UNDER","DEPARTMENT STORE","true","false","0122","001","","06","1176398","1900691","2001","08/17/2015 03:03:40 PM","41.88285151

[{"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/

[{"1674545","G456769","08/03/2001 12:03:51 AM","009XX N ORLEANS ST","1220","DECEPTIVE PRACTICE","THEFT OF LOST/MISLAID PROP","ALLEY","true","false","1823","018","","11","1173775","1906951","2001","08/17/2015 03:03:4

[{"1674549","G466221","08/06/2001 09:00:00 PM","035XX S COTTAGE GROVE AV","0460","BATTERY","SIMPLE","PARK PROPERTY","false","false","2122","002","","08B","1181464","1881376","2001","08/17/2015 03:03:40 PM","41.82973-

[{"1674550","G466788","08/07/2001 11:20:00 AM","075XX S MARYLAND AV","2820","OTHER OFFENSE","TELEPHONE THREAT","RESIDENCE","false","false","0624","006","","26","1183168","1855164","2001","08/17/2015 03:03:40 PM","4

[{"1674551","G464022","08/04/2001 09:20:00 AM","063XX N WINTHROP AV","4651","OTHER OFFENSE","SEX OFFENDER: FAIL REG NEW ADD","RESIDENCE","false","false","2433","024","","26","1167664","1942301","2001","08/17/201

[{"1674552","G459795","08/04/2001 10:48:00 AM","044XX S INDIANA AV","0460","BATTERY","SIMPLE","APARTMENT","false","true","0221","002","","08B","1178308","1875391","2001","08/17/2015 03:03:40 PM","41.813383306","-87.62147

[{"1674553","G460282","08/03/2001 09:30:00 PM","021XX N NEVA AV","0620","BURGLARY","UNLAWFUL ENTRY","RESIDENCE-GARAGE","false","false","2512","025","","05","1128204","1913407","2001","08/17/2015 03:03:40 PM","41.91

[{"1674554","G464084","08/06/2001 09:30:00 AM","036XX N SPRINGFIELD AV","0620","BURGLARY","UNLAWFUL ENTRY","OTHER","false","false","1732","017","","05","1149698","1924214","2001","08/17/2015 03:03:40 PM","41.9479614

[{"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","41.8054100

Showing 1 to 10 of 20 entries

Previous

1

2

Next

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.

```
E:\MS-DS\BigDataAnalytics\docker-drill\docker-apache-drill>docker exec -it drillbit-1 drill-conf
Apache Drill 1.16.0
"Two things are infinite: the universe and Drill; and I'm not sure about the universe."
apache drill>
```

Queries

- a. Get the first 20 rows from the data:

```
apache drill> select * from dfs.root.`output/ChicagoCrimesData.csv` limit 20;

+-----+
| columns |
+-----+
| ["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","","","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","08/20/2017 10:00:00 AM","001XX W RANDOLPH ST","0201","CRIM SEXUAL ASSAULT","NON-AGGRAVATED","HOTEL/HOTEL","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","0260","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","02/11/2018 03:57:41 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","JA333580","07/17/2017 10:10:00 AM","021XX W MC LEAN AVE","0818","THEFT","OVER $500","RESIDENCE","false","false","1432","014","32","22","06","","","2017","07/24/2017 03:54:23 PM","","",""] |
| ["11227633","JB147500","12/28/2017 03:55:00 PM","011XX S MICHIGAN AVE","1153","DECEPTIVE PRACTICE","FINANCIAL IDENTITY THEFT OVER $ 300","RESIDENCE","false","false","0123","001","2","32","11","","","2017","02/11/2018 03:57:41 PM","","",""] |
| ["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","","","2017","02/11/2018 03:57:41 PM","","",""] |
| ["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","","","2017","02/11/2018 03:57:41 PM","","",""] |
| ["11227247","JB147078","01/01/2012 00:00:00 AM","105XX S INDIANAPOLIS AVE","1153","DECEPTIVE PRACTICE","FINANCIAL IDENTITY THEFT OVER $ 300","RESIDENCE","false","false","0432","004","10","52","11","","","2012","02/11/2018 03:57:41 PM","","",""] |
| ["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","","","2017","08/05/2017 03:50:08 PM","","",""] |
| ["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","","","2017","02/11/2018 03:57:41 PM","","",""] |
| ["11227407","JB147329","10/14/2017 12:01:00 AM","037XX N SOUTHPORT AVE","1150","DECEPTIVE PRACTICE","CREDIT CARD FRAUD","OTHER","false","false","1922","019","44","6","11","","","2017","02/22/2018 03:56:47 PM","","",""] |
| ["11028056","JA350834","10/15/2014 03:00:00 PM","047XX S PULASKI RD","1153","DECEPTIVE PRACTICE","FINANCIAL IDENTITY THEFT OVER $ 300","PARKING LOT/GARAGE(NON-RESID.)","false","false","0821","008","14","57","","","2014","02/11/2018 03:57:41 PM","","",""] |
| ["523983","JB065450","10/17/2006 01:05:00 PM","052XX W HURON ST","2017","HARCOITICS","MAMA/DELIVER:CRACK","STREET","true","false","1524","015","28","25","18","","","2006","08/17/2015 03:03:40 PM","","",""] |
| ["11028299","JA360073","05/30/2015 12:00:00 AM","074XX S HARVARD AVE","1753","OFFENSE INVOLVING CHILDREN","SEX ASSLT OF CHILD BY FAM MBR","RESIDENCE","false","true","0731","007","17","69","02","","","2015","02/11/2018 03:57:41 PM","","",""] |
20 rows selected (0.444 seconds)
apache drill>
```

- b. Get the total count of the data:

```
apache drill> select count(*) from dfs.root.`output/ChicagoCrimesData.csv`;

+-----+
| EXPR$0 |
+-----+
| 7143700 |
+-----+

1 row selected (3.116 seconds)
apache drill>
```

- c. Filter data where Primary Type of crime is Theft

```
apache drill> select * from dfs.root.`output/ChicagoCrimesData.csv` where CAST(columns[5] AS varchar) = 'THEFT' limit 2;

+-----+
| columns |
+-----+
| ["3127060","JBK117658","01/10/2004 04:20:00 PM","005XX N STATE ST","0890","THEFT","FROM BUILDING","RESTAURANT","false","false","1832","018","42","8","06","1176262","1904268","2004","02/28/2018 03:56:25 PM","41.8926701","-87.628106353","(41.8926701, -87.628106353)"] |
| ["3127961","JBK117644","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","41.876701895","-87.751624656","(41.876701895, -87.751624656)"] |
2 rows selected (0.378 seconds)
apache drill>
```

- d. 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[5] AS varchar) = 'THEFT' and CAST(columns[7] AS varchar) = 'STREET' limit 2;

+-----+
| columns |
+-----+
| ["1674605","G464290","08/05/2001 09:30:00 AM","024XX N HARRAGANSETT AV","0810","THEFT","OVER $500","STREET","false","false","2512","025","","","06","1133385","1915454","2001","08/17/2015 03:03:40 PM","41.924224995","-87.785315762","(41.924224995, -87.785315762)"] |
| ["1674624","G464347","08/06/2001 04:00:00 AM","022XX N LATROBE AV","0810","THEFT","OVER $500","STREET","false","false","2515","025","","","06","1141053","1914592","2001","08/17/2015 03:03:40 PM","41.921721644","-87.757161239","(41.921721644, -87.757161239)"] |
2 rows selected (0.579 seconds)
apache drill>
```

- 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 |
+-----+
| ["174543","G464716","08/05/2001 10:30:00 PM","055XX 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","41.80541001","-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","010XX 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","J0147188","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","41.876701895","-87.751624656","(41.876701895,-87.751624656)"] |
| ["3127989","HK117846","01/09/2004 07:00:00 PM","050XX 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","41.79174086","-87.651164679","(41.79174086,-87.651164679)"] |
| ["1128011","HK117857","01/09/2004 07:00:00 PM","020XX W 68TH ST","0810","THEFT","OVER $500","STREET","false","false","0831","000","15","66","06","1160027","1859462","2004","02/28/2018 03:56:25 PM","41.770067978","-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","","","088","","","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","","","088","","","2001","08/17/2015 03:03:40 PM","","",""] |
| ["2370646","H0676216","09/26/2002 01:30:00 PM","034XX W 79TH ST","0460","BATTERY","SIMPLE","STREET","false","false","0835","008","18","70","088","","","2002","08/17/2015 03:03:40 PM","","",""] |
| ["0070760","HK559439","09/22/2009 06:10:00 PM","002XX W 95TH ST","0460","BATTERY","SIMPLE","STREET","false","false","0634","006","21","49","080","","","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","J0207383","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","41.871050866","-87.65275518","(41.871050866,-87.65275518)"] |
| ["11272724","J0207215","04/01/2018 05:15:00 AM","077XX S ESSEX AVE","0460","BATTERY","SIMPLE","APARTMENT","false","false","0421","004","7","43","088","1194202","1854229","2018","05/04/2018 03:51:04 PM","41.754937343","-87.56386716","(41.754937343,-87.56386716)"] |
| ["11272725","J0207348","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","J0207224","04/01/2018 03:30:00 AM","104XX S EGLESTON AVE","0810","THEFT","OVER $500","RESIDENCE","false","false","2233","022","34","49","06","1175131","1835642","2018","05/04/2018 03:51:04 PM","41.704378049","-87.634309567","(41.704378049,-87.634309567)"] |
| ["11272727","J0207322","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 03:51:04 PM","41.721642614","-87.576554028","(41.721642614,-87.576554028)"] |
| ["11272728","J0207381","04/01/2018 12:00:00 AM","020XX W DIVISION ST","0480","BATTERY","DOMESTIC BATTERY SIMPLE","BAR OR TAVERN","false","true","1212","012","1","24","088","1162670","1907990","2018","05/04/2018 03:51:04 PM","41.903179264","-87.677919727","(41.903179264,-87.677919727)"] |
| ["11272729","J0207290","04/01/2018 00:15:00 AM","095XX S DORCHESTER AVE","0480","BATTERY","DOMESTIC BATTERY SIMPLE","RESIDENCE","true","true","0411","004","5","43","088","1186849","1855270","2018","05/04/2018 03:51:04 PM","41.75797125","-87.59078048","(41.75797125,-87.59078048)"] |
| ["11272730","J0207368","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","J0207486","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","41.748314058","-87.613637728","(41.748314058,-87.613637728)"] |
| ["11272732","J0207344","04/01/2018 00:55:00 AM","115XX S LAFLIN ST","0480","BATTERY","DOMESTIC BATTERY SIMPLE","RESIDENCE","true","true","0524","005","34","53","088","1168405","1828220","2018","05/04/2018 03:51:04 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

```
10 rows selected (0.191 seconds)
apache drill> select * from dfs.root.`output/ChicagoCrimesData.csv` where CAST(columns[8] AS varchar) = 'true' and CAST(columns[7] AS varchar) = 'STREET' limit 4;
+-----+
| columns |
+-----+
| ["10405216","H2140694","02/04/2016 07:51:00 PM","035XX W CORTLAND ST","0910","MOTOR VEHICLE THEFT","AUTOMOBILE","STREET","true","false","1422","014","26","22","07","1152754","1912405","2016","02/10/2018 03:50:01 PM","41.915496543","-87.714226305","(41.915496543, -87.714226305)"] |
| ["10405222","H2141590","02/05/2016 03:28:00 PM","091XX S EXCHANGE AVE","502P","OTHER OFFENSE","FALSE/STOLEN/ALTERED TRD","STREET","true","false","0423","004","10","46","26","1197355","1844922","2016","02/10/2018 03:50:01 PM","41.729320257","-87.552621993","(41.729320257, -87.552621993)"] |
| ["10405226","H2141668","02/05/2016 04:12:00 PM","001XX W DAMEN AVE","0520","ASSAULT","AGGRAVATED:KNIFE/CUTTING INSTR","STREET","true","true","1223","012","27","28","04A","1163053","1900944","2016","02/10/2018 03:50:01 PM","41.883836447","-87.676710959","(41.883836447, -87.676710959)"] |
| ["10405242","H2141525","02/05/2016 04:14:00 PM","079XX S ESCANABA AVE","051A","ASSAULT","AGGRAVATED: HANDGUN","STREET","true","false","0422","004","7","46","04A","1196879","1852912","2016","02/10/2018 03:50:01 PM","41.751257279","-87.554100638","(41.751257279, -87.554100638)"] |
+-----+
4 rows selected (1.761 seconds)
apache drill>
```

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

```
apache drill> select * from dfs.root.`output/ChicagoCrimesData.csv` where CAST(columns[8] AS varchar) = 'true' and CAST(columns[4] AS varchar) = '0910' limit 4;
+-----+
| columns |
+-----+
| ["10405216","H2140694","02/04/2016 07:51:00 PM","035XX W CORTLAND ST","0910","MOTOR VEHICLE THEFT","AUTOMOBILE","STREET","true","false","1422","014","26","22","07","1152754","1912405","2016","02/10/2018 03:50:01 PM","41.915496543","-87.714226305","(41.915496543, -87.714226305)"] |
| ["10405285","H2141814","02/05/2016 07:20:00 PM","004XX W SUPERIOR ST","0910","MOTOR VEHICLE THEFT","AUTOMOBILE","STREET","true","false","1831","018","42","8","07","1173167","1905283","2016","02/10/2018 03:50:01 PM","41.895524561","-87.639442297","(41.895524561, -87.639442297)"] |
| ["10405783","H2142326","02/05/2016 07:30:00 PM","016XX S WOLCOTT AVE","0910","MOTOR VEHICLE THEFT","AUTOMOBILE","STREET","true","false","1234","012","25","31","07","1163973","1891955","2016","02/10/2018 03:50:01 PM","41.859150477","-87.67358634","(41.859150477, -87.67358634)"] |
| ["10406039","H2142773","02/06/2016 05:27:00 PM","061XX S MORGAN ST","0910","MOTOR VEHICLE THEFT","AUTOMOBILE","STREET","true","false","0712","007","16","68","07","1170697","1864104","2016","02/10/2018 03:50:01 PM","41.782580161","-87.649718184","(41.782580161, -87.649718184)"] |
+-----+
4 rows selected (0.297 seconds)
apache drill>
```

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

```
2 rows selected (1.27 seconds)
apache drill> select count(*) from dfs.root.`output/ChicagoCrimesData.csv` where CAST(columns[5] AS varchar) = 'THEFT';
+-----+
| EXPR$0 |
+-----+
| 1509553 |
+-----+
1 row selected (3.197 seconds)
apache drill>
```

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, designed 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`

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	ICR	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	1A366925	01/01/2001 11:00:00 AM	010100X E 86TH PL	1153	DECEPTIVE PRACTICE	[FINANCIAL IDENTITY THEFT OVER \$ 300]	RESIDENCE	[false]	[false]	0412	004	18	45
11227287	1B147188	10/08/2017 03:00:00 AM	09200X S RACINE AVE	0281	CRIM SEXUAL ASSAULT	[NON-AGGRAVATED]	RESIDENCE	[false]	[false]	2222	022	21	73
11227583	1B147595	03/28/2017 02:00:00 PM	02052X W 79TH ST	0620	BURGLARY	[UNLAWFUL ENTRY]	OTHER	[false]	[false]	0835	008	18	70
11227293	1B147230	09/09/2017 08:17:00 PM	06000X S EBERHART AVE	0810	THEFT	[OVER \$500]	RESIDENCE	[false]	[false]	0313	003	20	42
11227634	1B147599	08/26/2017 10:00:00 AM	06100X S KILBOURN AVE	1751	OFFENSE INVOLVING CHILDREN	[CRIM SEX ABUSE BY FAM MEMBER]	RESIDENCE	[false]	[true]	0813	008	13	65
11227517	1B138481	02/10/2013 12:00:00 AM	07100X S LAFAYETTE AVE	0266	CRIM SEXUAL ASSAULT	[PREDATORY]	RESIDENCE	[false]	[false]	0731	007	16	69
11227508	1B146365	01/01/2017 12:01:00 AM	02700X S WHIPPLE ST	1754	OFFENSE INVOLVING CHILDREN	[AGG SEX ASLT OF CHILD FAM MBR]	RESIDENCE	[false]	[false]	1033	010	12	30
11022695	1A353568	07/17/2017 10:10:00 AM	02100X W MC LEAN AVE	0810	THEFT	[OVER \$500]	RESIDENCE	[false]	[false]	1432	014	32	22
11227633	1B147500	12/28/2017 03:55:00 PM	01100X S MICHIGAN AVE	1153	DECEPTIVE PRACTICE	[FINANCIAL IDENTITY THEFT OVER \$ 300]	null	[false]	[false]	0123	001	12	32
11227586	1B147613	02/10/2017 12:00:00 PM	0800X S COTTAGE GROVE AVE	1310	CRIMINAL DAMAGE	[TO PROPERTY]	APARTMENT	[false]	[false]	0633	006	18	44
11227475	1B147314	11/22/2017 02:42:00 AM	05000X W CHRISTIANA AVE	2826	OTHER OFFENSE	[HARASSMENT BY ELECTRONIC MEANS]	APARTMENT	[false]	[true]	1711	017	39	13
11227247	1B147078	01/01/2012 09:00:00 AM	18500X S INDIANAPOLIS AVE	1153	DECEPTIVE PRACTICE	[FINANCIAL IDENTITY THEFT OVER \$ 300]	RESIDENCE	[false]	[false]	0432	004	10	52
11036284	1A370282	07/29/2017 03:40:00 PM	01000X W DECON AVE	0820	THEFT	[\$500 AND UNDER]	SIDEWALK	[false]	[false]	2432	024	40	11
11227509	1B146413	01/22/2017 12:01:00 AM	07000X S JEFFERY BLVD	1752	OFFENSE INVOLVING CHILDREN	[AGG CRIM SEX ABUSE FAM MEMBER]	APARTMENT	[false]	[false]	0414	004	18	46
120	null	null	2017/02/11/2018 03:57:41 PM	null	null	null	null	null	null	null	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]
```

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**

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)	0.0.0.0:8001->8001/tcp, 50070/tcp	docker-hive_namenode_1
56f21ee1703	bde2020/hadoop-datanode:2.0.0-hadoop2.7.4-java8	"/entrypoint.sh /run..."	25 hours ago	Up 21 minutes (healthy)	0.0.0.0:8002->8002/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	shanzhu/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
1c71ce112904        bde2020/hadoop-namenode:2.0.0-hadoop2.7.4-java8        "/entrypoint.sh /run..."   25 hours ago       Up 21 minutes (healthy)   0.0.0.0:8001->8001/tcp, 50070/tcp   docker-hive_namenode_1
56f21aeef703        bde2020/hadoop-datanode:2.0.0-hadoop2.7.4-java8        "/entrypoint.sh /run..."   25 hours ago       Up 21 minutes (healthy)   0.0.0.0:8002->8002/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-postgresql_1
e9a1_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
21fcaeb57f        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,CasNumber String,Block String,IUCR String,PrimaryType String,Description String,Arrest String,Domestic String,Beat String,District String,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;
```

chicagodatatatable.id	chicagodatatatable.casnumber	chicagodatatatable.block	chicagodatatatable.iucr	chicagodatatatable.primarytype	chicagodatatatable.description	chicagodatatatable.arrest	chicagodatatatable.domestic	chicagodatatatable.beat	chicagodatatatable.district	chicagodatatatable.ward	chicagodatatatable.communityarea	chicagodatatatable.fbi	chicagodatatatable.xcoordinate	chicagodatatatable.ycoordinate	chicagodatatatable.year	chicagodatatatable.updatedon	chicagodatatatable.latitude	chicagodatatatable.longitude	chicagodatatatable.location
11034701	J366925	01/01/2001 11:00:00 AM	0160X E 86TH PL	0412	004	08/05/2017 03:50:08 PM	0281	022	02/11/2018 03:57:41 PM	0620	008	003	02/11/2018 03:57:41 PM	007	008	04/12/2019 04:00:15 PM	1754	010	02/11/2018 03:57:41 PM
11227287	JB147188	10/08/2017 03:00:00 AM	0920X S RACINE AVE	2222	004	08/05/2017 03:50:08 PM	0281	022	02/11/2018 03:57:41 PM	0620	008	003	02/11/2018 03:57:41 PM	007	008	04/12/2019 04:00:15 PM	1754	010	02/11/2018 03:57:41 PM
11227583	JB147595	03/28/2017 02:00:00 PM	0260X W 79TH ST	0835	008	08/05/2017 03:50:08 PM	0281	022	02/11/2018 03:57:41 PM	0620	008	003	02/11/2018 03:57:41 PM	007	008	04/12/2019 04:00:15 PM	1754	010	02/11/2018 03:57:41 PM
11227293	JB147230	09/09/2017 08:17:00 PM	0600X S EBERHART AVE	0313	001	08/05/2017 03:50:08 PM	0281	022	02/11/2018 03:57:41 PM	0620	008	003	02/11/2018 03:57:41 PM	007	008	04/12/2019 04:00:15 PM	1754	010	02/11/2018 03:57:41 PM
11227634	JB147599	08/26/2017 10:00:00 AM	0010X W RANDOLPH ST	0122	001	08/05/2017 03:50:08 PM	0281	022	02/11/2018 03:57:41 PM	0620	008	003	02/11/2018 03:57:41 PM	007	008	04/12/2019 04:00:15 PM	1754	010	02/11/2018 03:57:41 PM
11227517	JB138481	02/10/2013 12:00:00 AM	0710X S LAFAYETTE AVE	0731	007	08/05/2017 03:50:08 PM	0281	022	02/11/2018 03:57:41 PM	0620	008	003	02/11/2018 03:57:41 PM	007	008	04/12/2019 04:00:15 PM	1754	010	02/11/2018 03:57:41 PM
11227503	JB146383	01/01/2015 12:01:00 AM	0610X S KILBOURN AVE	0813	008	08/05/2017 03:50:08 PM	0281	022	02/11/2018 03:57:41 PM	0620	008	003	02/11/2018 03:57:41 PM	007	008	04/12/2019 04:00:15 PM	1754	010	02/11/2018 03:57:41 PM
11227508	JB146365	01/01/2017 12:01:00 AM	0270X S WHIPPLE ST	1033	010	08/05/2017 03:50:08 PM	0281	022	02/11/2018 03:57:41 PM	0620	008	003	02/11/2018 03:57:41 PM	007	008	04/12/2019 04:00:15 PM	1754	010	02/11/2018 03:57:41 PM
11022695	J353568	07/17/2017 10:10:00 AM	0210X W MC LEAN AVE	1432	014	08/05/2017 03:50:08 PM	0281	022	02/11/2018 03:57:41 PM	0620	008	003	02/11/2018 03:57:41 PM	007	008	04/12/2019 04:00:15 PM	1754	010	02/11/2018 03:57:41 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 chicanodatable.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 chicanodatable.arrest = 'false' and chicanodatable.icur = '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 chicanodatable.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.
+-----+
|_c0_|
+-----+
| 9466980 |
+-----+
| 9466981 |
+-----+
| 9466982 |
+-----+
| 9466983 |
+-----+
| 9466984 |
+-----+
5 rows selected (1.376 seconds)
0: jdbc:hive2://localhost:10000:
```

chicanodatable.id	chicanodatable.casenumbr	chicanodatable.block	chicanodatable.icur	chicanodatable.primarytype	chicanodatable.description	chicanodatable.arrest	chicanodatable.domestic	chicanodatable.beat	chicanodatable.district	chicanodatable.ward	chicanodatable.communityarea	chicanodatable.fbcode	chicanodatable.xcoordinate	chicanodatable.ycoordinate	chicanodatable.year	chicanodatable.updatedon	chicanodatable.latitude	chicanodatable.longitude	chicanodatable.location
9466980	1175877	1856926	0720X S YALE AVE	007	OTHER OFFENSE	69	VIOLATE ORDER OF PROTECTION	26	OTHER										
9466981	1175877	1856926	0720X S YALE AVE	007	BATTERY	58	DOMESTIC BATTERY SIMPLE	08B	APARTMENT										
9466982	1175877	1856926	0720X S YALE AVE	007	BATTERY	22	SIMPLE	08B	BAR OR										
9466983	1175877	1856926	0720X S YALE AVE	007	POSSESS: CANNABIS 30GMS OR LESS	71	SIMPLE	08A	BAR OR										
9466984	1175877	1856926	0720X S YALE AVE	007	ASSAULT	22	SIMPLE	08A	BAR OR										

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 chicanodatable.ward != '021' order by chicanodatable.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.
+-----+
|_c0_|
+-----+
| 58388414 |
+-----+
| 5838845 |
+-----+
| 5838590 |
+-----+
| 7668253 |
+-----+
| 5837984 |
+-----+
5 rows selected (18.35 seconds)
0: jdbc:hive2://localhost:10000:
```

chicanodatable.id	chicanodatable.casenumbr	chicanodatable.block	chicanodatable.icur	chicanodatable.primarytype	chicanodatable.description	chicanodatable.arrest	chicanodatable.domestic	chicanodatable.beat	chicanodatable.district	chicanodatable.ward	chicanodatable.communityarea	chicanodatable.fbcode	chicanodatable.xcoordinate	chicanodatable.ycoordinate	chicanodatable.year	chicanodatable.updatedon	chicanodatable.latitude	chicanodatable.longitude	chicanodatable.location
58388414	1175877	1856926	0720X S YALE AVE	007	CRIMINAL TRESPASS	018	TO LAND	42	SCHOOL										
5838845	1175877	1856926	0720X S YALE AVE	007	THEFT	016	\$500 AND UNDER	36	SCHOOL										
5838590	1175877	1856926	0720X S YALE AVE	007	BATTERY	018	SIMPLE	27	SCHOOL										
7668253	1175877	1856926	0720X S YALE AVE	007	BATTERY	022	SIMPLE	19	SCHOOL										
5837984	1175877	1856926	0720X S YALE AVE	007	THEFT	024	FROM BUILDING	50	SCHOOL										

g. Get the five records of same ward and district

```
0: jdbc:hive2://localhost:10000: select * from ChicagoDataTable where chicagodatatable.ward = '021' and chicagodatatable.district = '021' limit 5 ;
WARNING: Hive-on-PK 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.case_number	chicagodatatable.block	chicagodatatable.iucr	chicagodatatable.primarytype	chicagodatatable.description	chicagodatatable.arrest	chicagodatatable.domestic	chicagodatatable.beat	chicagodatatable.district	chicagodatatable.ward	chicagodatatable.communityarea	chicagodatatable.fbi_code	chicagodatatable.xcoordinate	chicagodatatable.ycoordinate	chicagodatatable.year	chicagodatatable.updatedon	chicagodatatable.latitude	chicagodatatable.longitude	chicagodatatable.location
9466980	HCI19638																		
		false	01/18/2014 08:25:00 PM	072XX S YALE AVE	007	4387													OTHER OFFENSE
		1175877	1856926	2014	02/18/2018 03:50:01 PM	41.762768355		69											VIOLATE ORDER OF PROTECTION
9466981	HCI19890																		OTHER
		true	01/19/2014 03:45:00 AM	028XX W 40TH ST	009	0486													BATTERY
		1158053	1878007	2014	02/18/2018 03:50:01 PM	41.82099834		58											DOMESTIC BATTERY SIMPLE
9466982	HCI19891																		APARTHE
		false	01/19/2014 04:00:00 AM	022XX N ASHLAND AVE	009	0460													BATTERY
		1165153	1914764	2014	02/18/2018 03:50:01 PM	41.921753143		22											SIMPLE
9466983	HCI19863																		BAR OR
		true	01/19/2014 01:48:00 AM	009XX W 85TH ST	006	1811													NARCOTICS
		1171316	1848463	2014	02/18/2018 03:50:01 PM	41.739645376		71											POSS: CANNABIS 30GMS OR LESS
9466984	HCI19900																		STREET
		true	01/19/2014 04:35:00 AM	022XX N ASHLAND AVE	009	0560													ASSAULT
		1165153	1914764	2014	02/18/2018 03:50:01 PM	41.921753143		22											SIMPLE
		false																	BAR OR

5 rows selected (1.334 seconds)

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

```
0: jdbc:hive2://localhost:10000: select * from ChicagoDataTable where chicagodatatable.ward = '021' order by chicagodatatable.district limit 5 ;
WARNING: Hive-on-PK 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.case_number	chicagodatatable.block	chicagodatatable.iucr	chicagodatatable.primarytype	chicagodatatable.description	chicagodatatable.arrest	chicagodatatable.domestic	chicagodatatable.beat	chicagodatatable.district	chicagodatatable.ward	chicagodatatable.communityarea	chicagodatatable.fbi_code	chicagodatatable.xcoordinate	chicagodatatable.ycoordinate	chicagodatatable.year	chicagodatatable.updatedon	chicagodatatable.latitude	chicagodatatable.longitude	chicagodatatable.location
5838414	HNI643595																		
		PRIVATE	10/12/2007 11:50:00 AM	080XX N LA SALLE DR	1330														CRIMINAL TRESPASS
		26	1117009	080650	1832														TO LAND
5838045	HNI64750																		"SCHOOL
		PUBLIC	10/11/2007 04:00:00 PM	039XX N PANAMA AVE	0820														THEFT
		17	1121359	1525400	1631														\$500 AND UNDER
5838590	HNI637596																		"SCHOOL
		PRIVATE	10/08/2007 05:45:00 PM	003XX W CHESTNUT ST	0460														BATTERY
		088	1175654	106247	1823														SIMPLE
7668253	HS472009																		BAR OR
		PUBLIC	08/19/2018 11:00:00 AM	017XX W PLYOR AVE	0460														BATTERY
		088	1166533	1813078	2212														SIMPLE
5837984	HNI644272																		"SCHOOL
		PUBLIC	10/12/2007 02:35:00 PM	073XX N WASHINGTON AVE	0890														FROM BUILDING
		2	1157909	1348758	2411														THEFT
		06			2007														024

5 rows selected (18.35 seconds)

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

```
0: jdbc:hive2://localhost:10000: select * from ChicagoDataTable where chicagodatatable.ward = '021' and chicagodatatable.district = '021' limit 5 ;
WARNING: Hive-on-PK 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.case_number	chicagodatatable.block	chicagodatatable.iucr	chicagodatatable.primarytype	chicagodatatable.description	chicagodatatable.arrest	chicagodatatable.domestic	chicagodatatable.beat	chicagodatatable.district	chicagodatatable.ward	chicagodatatable.communityarea	chicagodatatable.fbi_code	chicagodatatable.xcoordinate	chicagodatatable.ycoordinate	chicagodatatable.year	chicagodatatable.updatedon	chicagodatatable.latitude	chicagodatatable.longitude	chicagodatatable.location
9466980	HCI19638																		
		false	01/18/2014 08:25:00 PM	072XX S YALE AVE	007	4387													OTHER OFFENSE
		1175877	1856926	2014	02/18/2018 03:50:01 PM	41.762768355		69											VIOLATE ORDER OF PROTECTION
9466981	HCI19890																		OTHER
		true	01/19/2014 03:45:00 AM	028XX W 40TH ST	009	0486													BATTERY
		1158053	1878007	2014	02/18/2018 03:50:01 PM	41.82099834		58											DOMESTIC BATTERY SIMPLE
9466982	HCI19891																		APARTHE
		false	01/19/2014 04:00:00 AM	022XX N ASHLAND AVE	009	0460													BATTERY
		1165153	1914764	2014	02/18/2018 03:50:01 PM	41.921753143		22											SIMPLE
9466983	HCI19863																		BAR OR
		true	01/19/2014 01:48:00 AM	009XX W 85TH ST	006	1811													NARCOTICS
		1171316	1848463	2014	02/18/2018 03:50:01 PM	41.739645376		71											POSS: CANNABIS 30GMS OR LESS
9466984	HCI19900																		STREET
		true	01/19/2014 04:35:00 AM	022XX N ASHLAND AVE	009	0560													ASSAULT
		1165153	1914764	2014	02/18/2018 03:50:01 PM	41.921753143		22											SIMPLE
		false																	BAR OR

5 rows selected (1.334 seconds)

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

```
0: jdbc:hive2://localhost:10000: select * from ChicagoDataTable where chicagodatatable.ward = '021' and chicagodatatable.district = '021' limit 5 ;
WARNING: Hive-on-PK 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.case_number	chicagodatatable.block	chicagodatatable.iucr	chicagodatatable.primarytype	chicagodatatable.description	chicagodatatable.arrest	chicagodatatable.domestic	chicagodatatable.beat	chicagodatatable.district	chicagodatatable.ward	chicagodatatable.communityarea	chicagodatatable.fbi_code	chicagodatatable.xcoordinate	chicagodatatable.ycoordinate	chicagodatatable.year	chicagodatatable.updatedon	chicagodatatable.latitude	chicagodatatable.longitude	chicagodatatable.location
9466980	HCI19638																		
		false	01/18/2014 08:25:00 PM	072XX S YALE AVE	007	4387													OTHER OFFENSE
		1175877	1856926	2014	02/18/2018 03:50:01 PM	41.762768355		69											VIOLATE ORDER OF PROTECTION
9466981	HCI19890																		OTHER
		true	01/19/2014 03:45:00 AM	028XX W 40TH ST	009	0486													BATTERY
		1158053	1878007	2014	02/18/2018 03:50:01 PM	41.82099834		58											DOMESTIC BATTERY SIMPLE
9466982	HCI19891																		APARTHE
		false	01/19/2014 04:00:00 AM	022XX N ASHLAND AVE	009	0460													BATTERY
		1165153	1914764	2014	02/18/2018 03:50:01 PM	41.921753143		22											SIMPLE
9466983	HCI19863																		BAR OR
		true	01/19/2014 01:48:00 AM	009XX W 85TH ST	006	1811													NARCOTICS
		1171316	1848463	2014	02/18/2018 03:50:01 PM	41.739645376		71											POSS: CANNABIS 30GMS OR LESS
9466984	HCI19900																		STREET
		true	01/19/2014 04:35:00 AM	022XX N ASHLAND AVE	009	0560													ASSAULT
		1165153	1914764	2014	02/18/2018 03:50:01 PM	41.921753143		22											SIMPLE
		false																	BAR OR

5 rows selected (1.334 seconds)

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

```
0: jdbc:hive2://localhost:10000 select count(*) from ChicagoData table where chicagoData table.arrest != 'false';
WARNING: Hive-on-PH 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
```

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

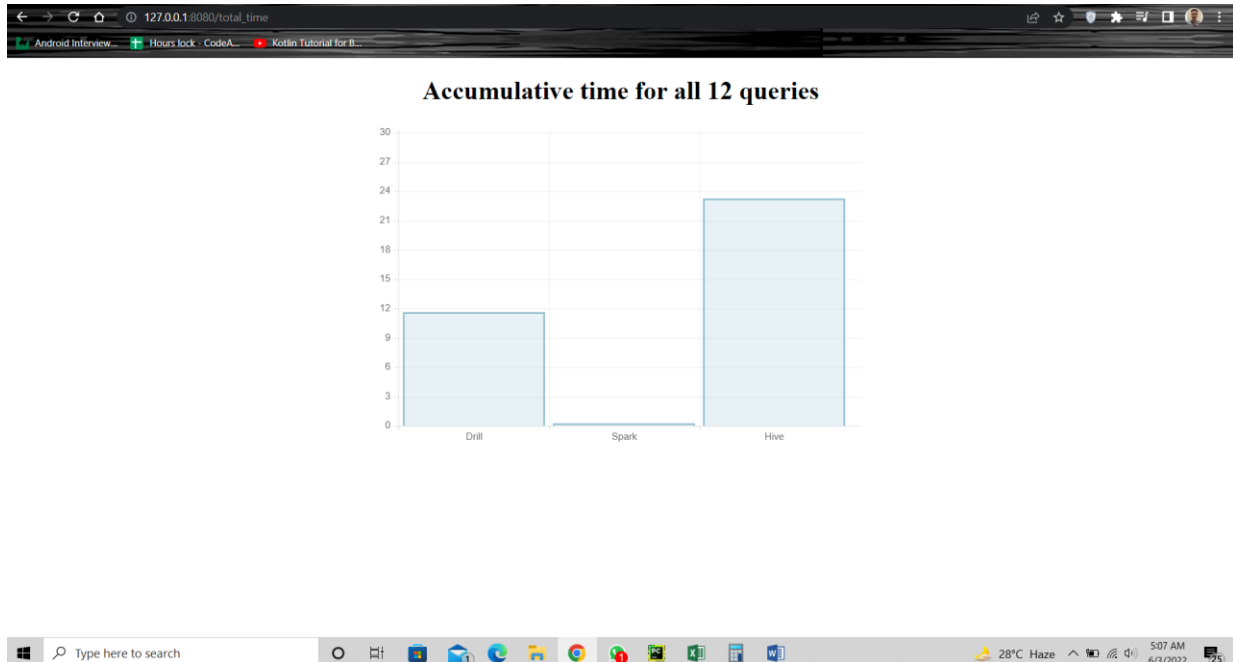
```
0: jdbc:hive2://localhost:10000 select * from ChicagoData table where chicagoData table.ward != '021' and chicagoData table.district != '021' limit 5 ;
WARNING: Hive-on-PH is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
+-----+
| chicagoData table.id | chicagoData table.case number | chicagoData table.block | chicagoData table.lur | chicagoData table.primary type | chicagoData table.description | chicagoData table.arrest | chicagoData table.domestic |
+-----+
| 9466980 | HK119638 | 1856926 | 0731 | 007 | 17 | OTHER OFFENSE | VIOLATE ORDER OF PROTECTION |
+-----+
| 9466981 | HK119890 | 1878907 | 0921 | 000 | 14 | BATTERY | DOMESTIC BATTERY SIMPLE |
+-----+
| 9466982 | HK119891 | 1878907 | 0921 | 000 | 14 | BATTERY | DOMESTIC BATTERY SIMPLE |
+-----+
| 9466983 | HK119863 | 1914764 | 1412 | 014 | 32 | BATTERY | SIMPLE |
+-----+
| 9466984 | HK119900 | 1848463 | 0613 | 000 | 21 | NARCOTICS | POSS: CANNABIS 30GHS OR LESS |
+-----+
| 9466985 | HK119900 | 1848463 | 0613 | 000 | 21 | NARCOTICS | POSS: CANNABIS 30GHS OR LESS |
+-----+
5 rows selected (1.334 seconds)
0: jdbc:hive2://localhost:10000
```

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



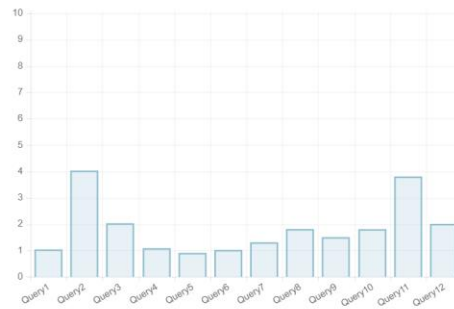
One can easily see the difference between the execution time of spark vs hive and drill. Spark has executed each query in milliseconds. Whereas, for some of the complex queries hive and drill execution time went to two to five seconds respectively.

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.

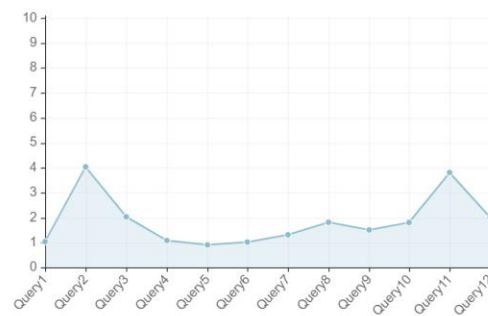
Hive



Execution time for each query in hive



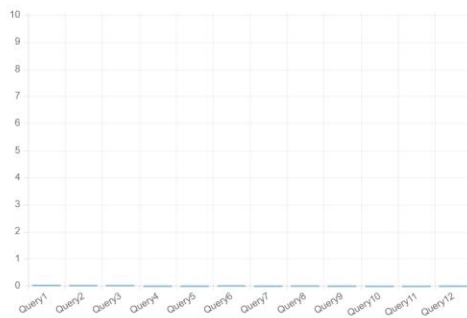
Execution time for each query in hive



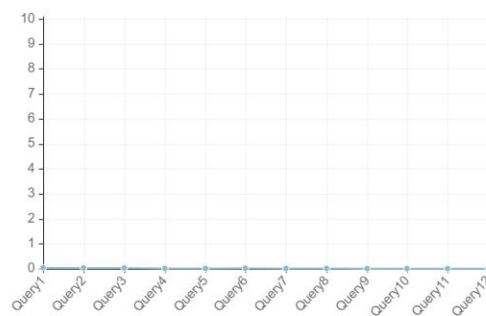
Spark



Execution time for each query in spark



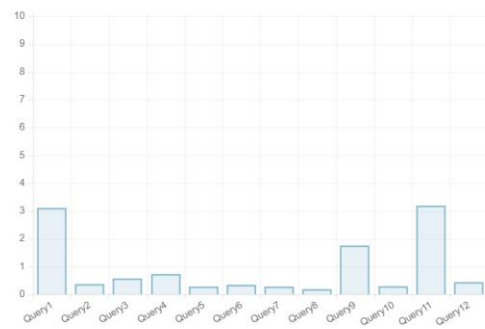
Execution time for each query in spark



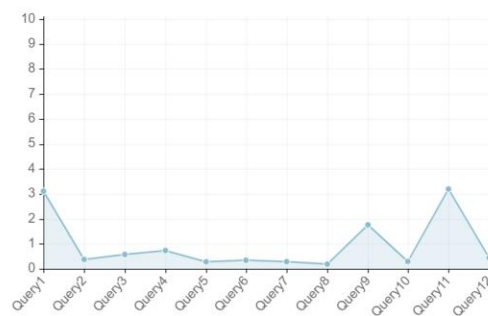
Drill



Execution time for each query in drill



Execution time for each query in 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)