Crime Data analysis using Hadoop

Crime Rate Analysis

**Ramdharan Donda (CIN: 305061502)**

**Sridhar reddy Puli (CIN: 305080001)**

**Raviteja Sontinani (CIN: 304988793)**

**Sri Chakra Sanjay Devabathini (CIN: 305067976)**

**Table of Contents**

Introduction and Prerequisites

Create a Cluster

Create table and load data into the table

Calculate the distance and save into a new table

View the Tables

Analyzing the data

Create new data source in ODBC administrator

Importing data into excel sheets

Graphical Representation

View crime locations on map

Crime Rate Data Analysis

*Overview*

In this tutorial, we will analyze the crime dataset using different types of Hive and import the output into the excel for displaying it in graphical view. Since we have latitude and longitude values of locations, we even use map functionality in excel to display the crime locations in map. Through this we can learn and apply the Hive concepts to analyze crime data of Los Angeles occurred in the year 2014.

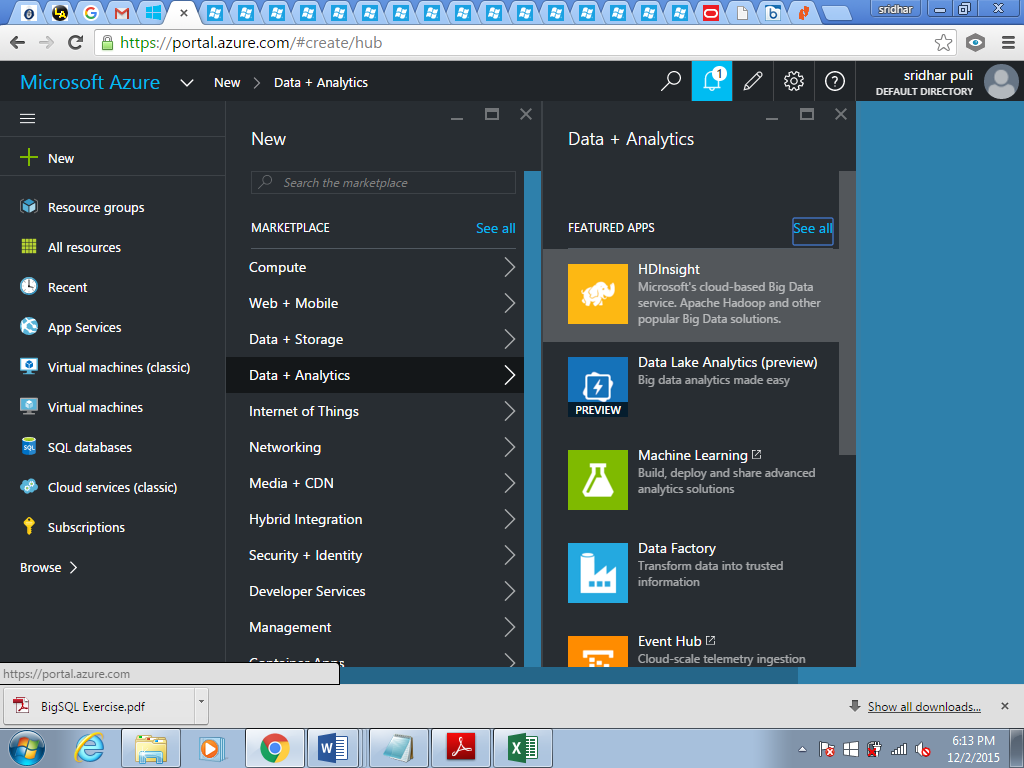
The crime data includes crime occurred date, crime reported date, place, crime type, latitude and longitude of location, etc. Using these data, we finds out number of crimes occurred per month, number of crime types and occurrence, crimes occurred at specific time intervals, crimes happened near CSULA, etc.

**Prerequisites**

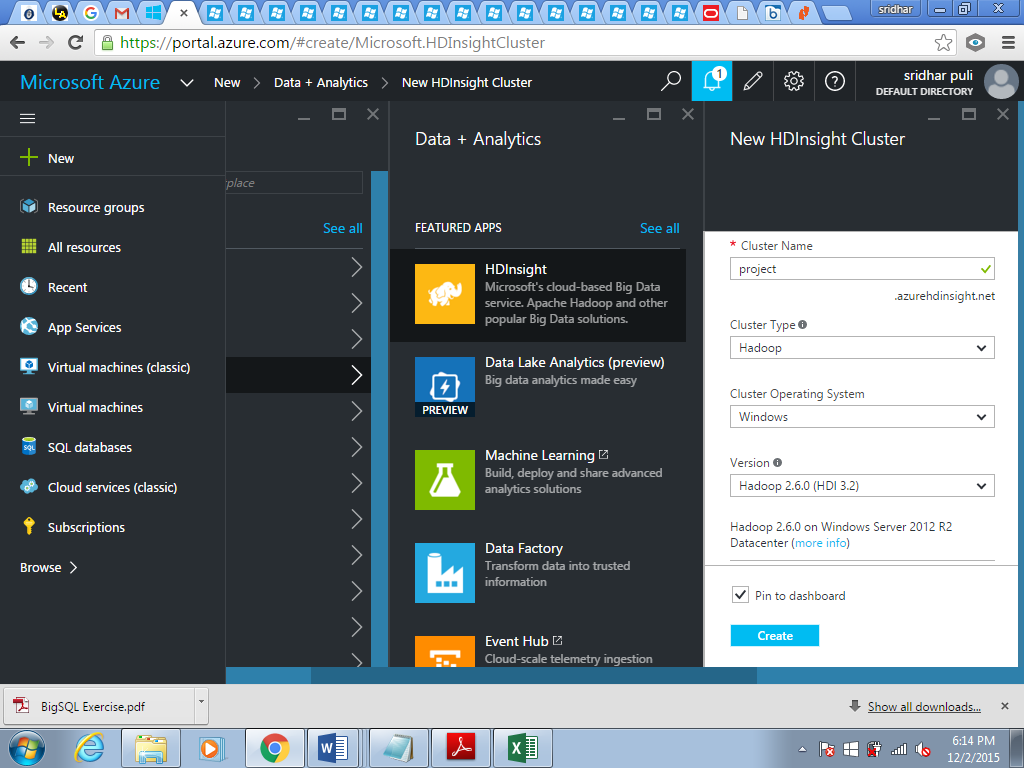
* Microsoft Azure
* Microsoft Excel
* Hive editor
* Odbc Driver

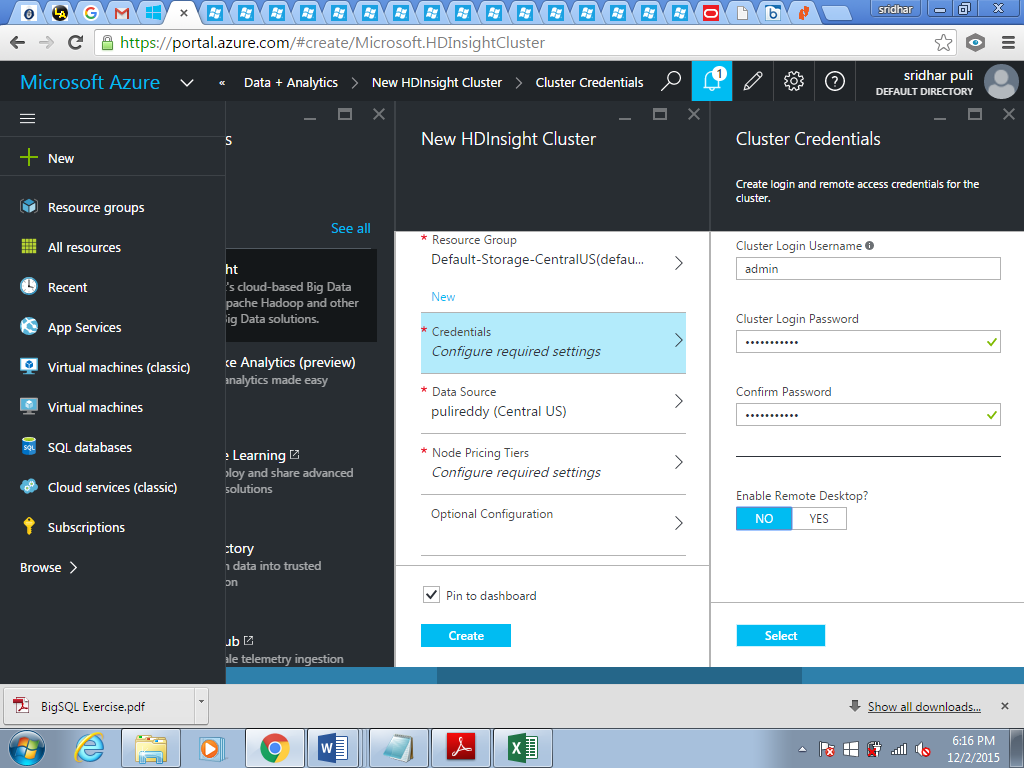
Create cluster

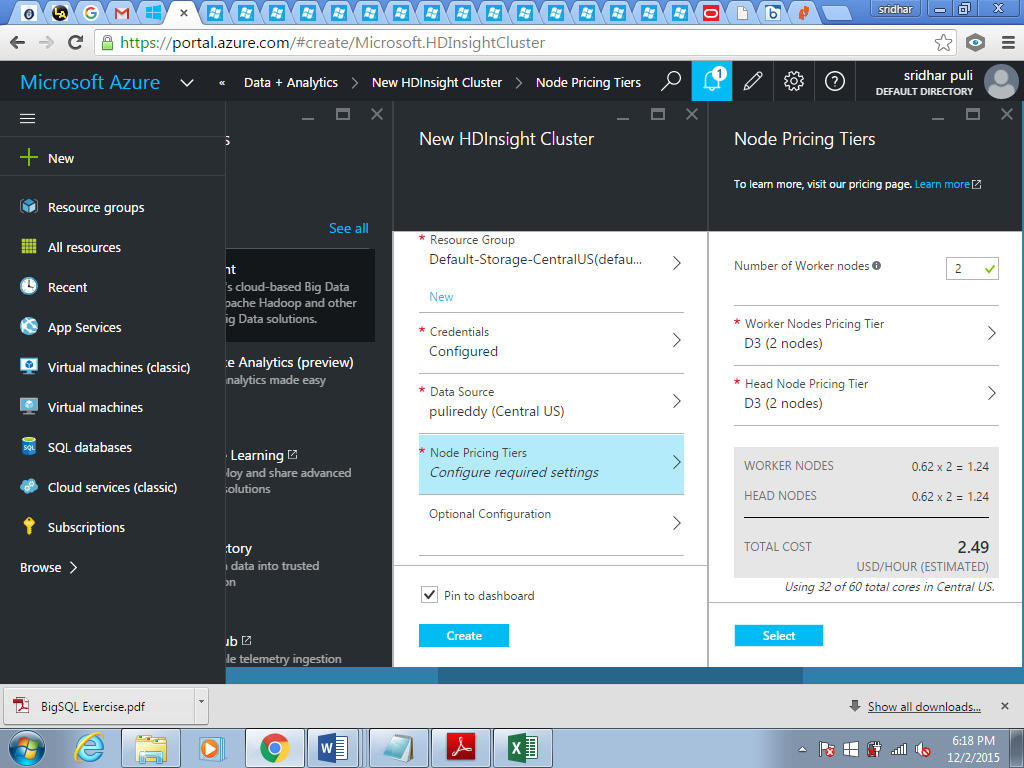
1. Open new Azure Portal, click on new and under Data + Analytics select HDInsight.



1. Give Cluster name, type, OS, Credentials and Node details and click create.

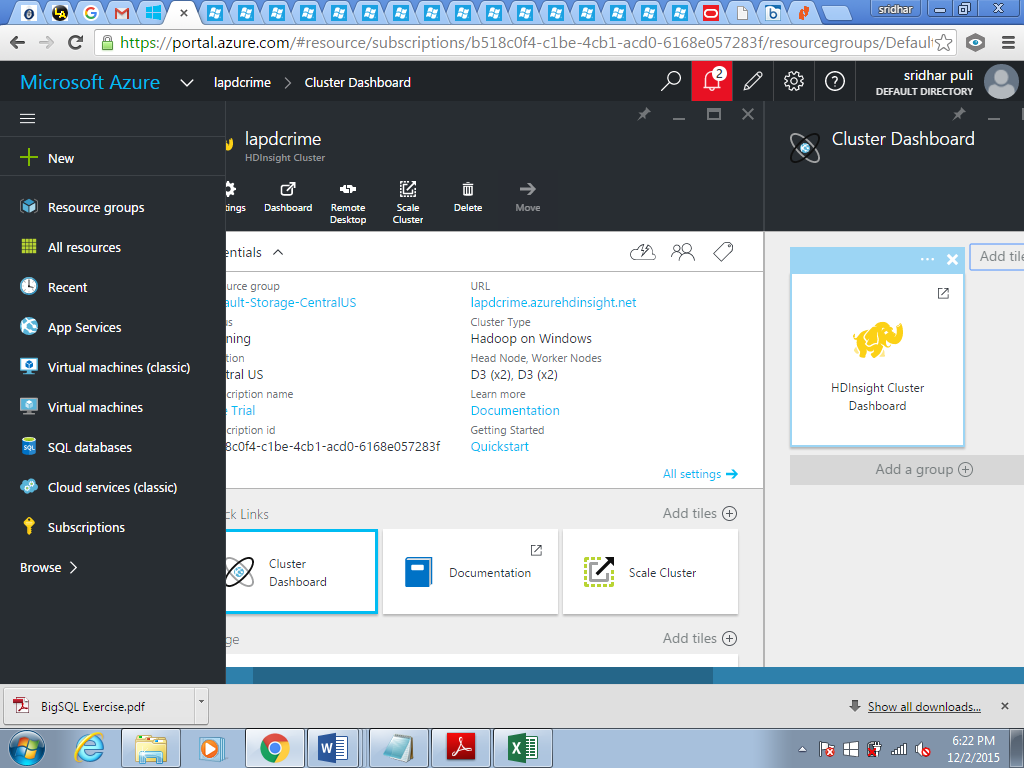






Create table and load data

1. After creation of cluster, click on the cluster at the dashboard. Under the quick links, click on the cluster dashboard and the HD Insight cluster dashboard appears, click on that.



1. Goto hive editor and Submit the following query to create crimedata table.

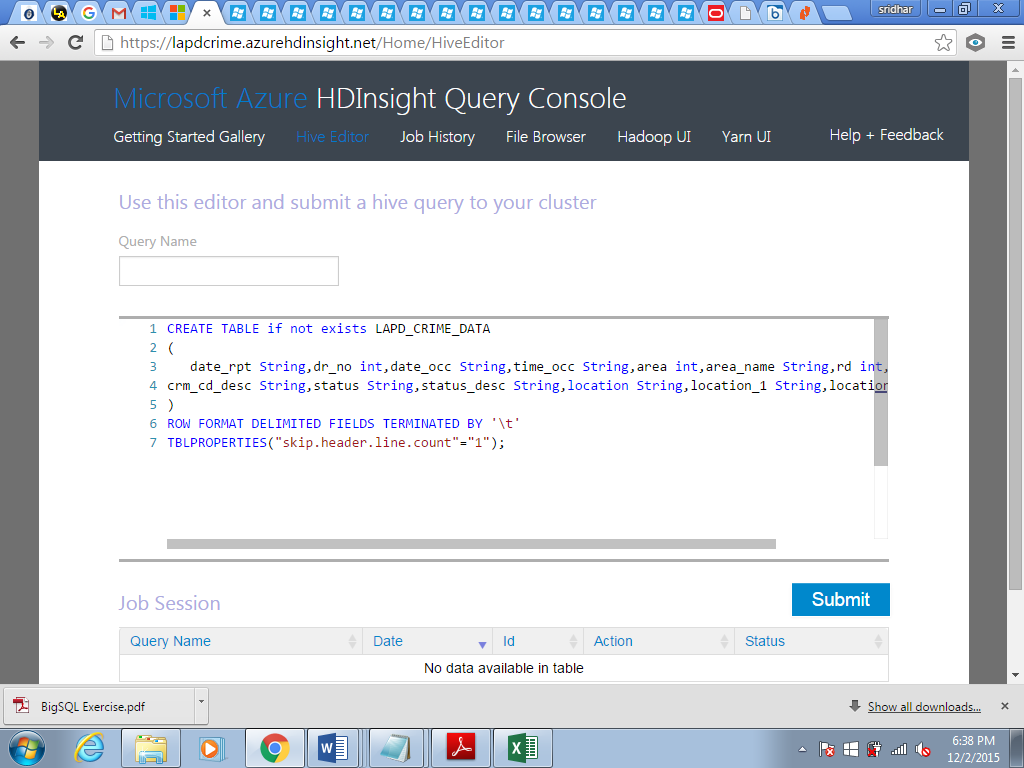
CREATE TABLE if not exists LAPD\_CRIME\_DATA

( date\_rpt String,dr\_no int,date\_occ String,time\_occ String,area int,area\_name String,rd int,crm\_cd int,

crm\_cd\_desc String,status String,status\_desc String,location String,location\_1 String,location\_2 String)

ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'

TBLPROPERTIES("skip.header.line.count"="1");



1. Now load the data into table from tsv file.

LOAD DATA INPATH '\HdiSamples\LAPD\_Crime\_and\_Collision\_Raw\_Data\_-\_2014.tsv'OVERWRITE INTO TABLE LAPD\_CRIME\_DATA;

Create table LOCATION\_VALUES

1. Create a table “LOCATION\_VALUES” to store latitude,longitude and distance to that locations from CSULA.

CREATE table location\_values(latVal double,longVal double,dr\_no int,distance double);

1. Now, calculate the distance and insert the data into the LOCATION\_VALUES table.

INSERT OVERWRITE TABLE LOCATION\_VALUES Select cast(regexp\_replace(split(location\_2,',')[0],'\\(','') as double),

cast(regexp\_replace(split(location\_2,',')[1],'\\)','') as double),

dr\_no,

2 \* asin(

sqrt(cos(radians(34.0667)) \*

cos(radians(cast(regexp\_replace(split(location\_2,',')[0],'\\(','') as double))) \*

pow(sin(radians((-118.1678 - cast(regexp\_replace(split(location\_2,',')[1],'\\)','') as double))/2)), 2)

+pow(sin(radians((34.0667 - cast(regexp\_replace(split(location\_2,',')[0],'\\(','') as double))/2)), 2)) )\*3956 from LAPD\_CRIME\_DATA;

View Tables

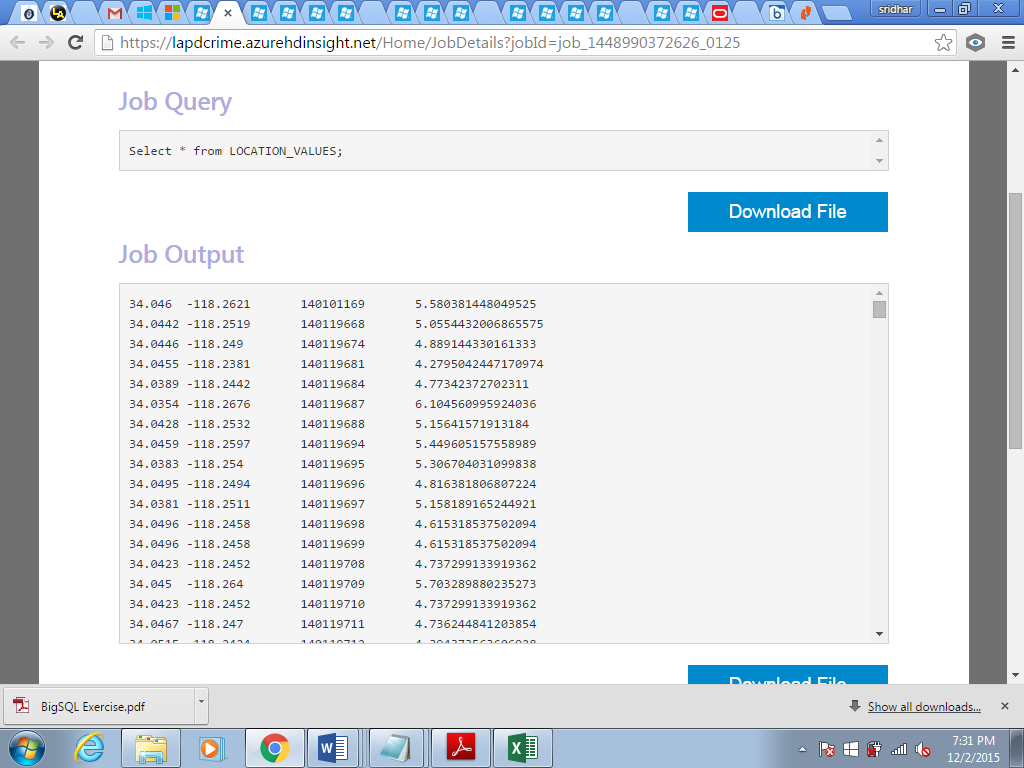
Check the data in the tables.

Select \* from LAPD\_CRIME\_DATA;

Click on “view details” to see the query result.



Select \* from LOCATION\_VALUES;



Analyzing the data

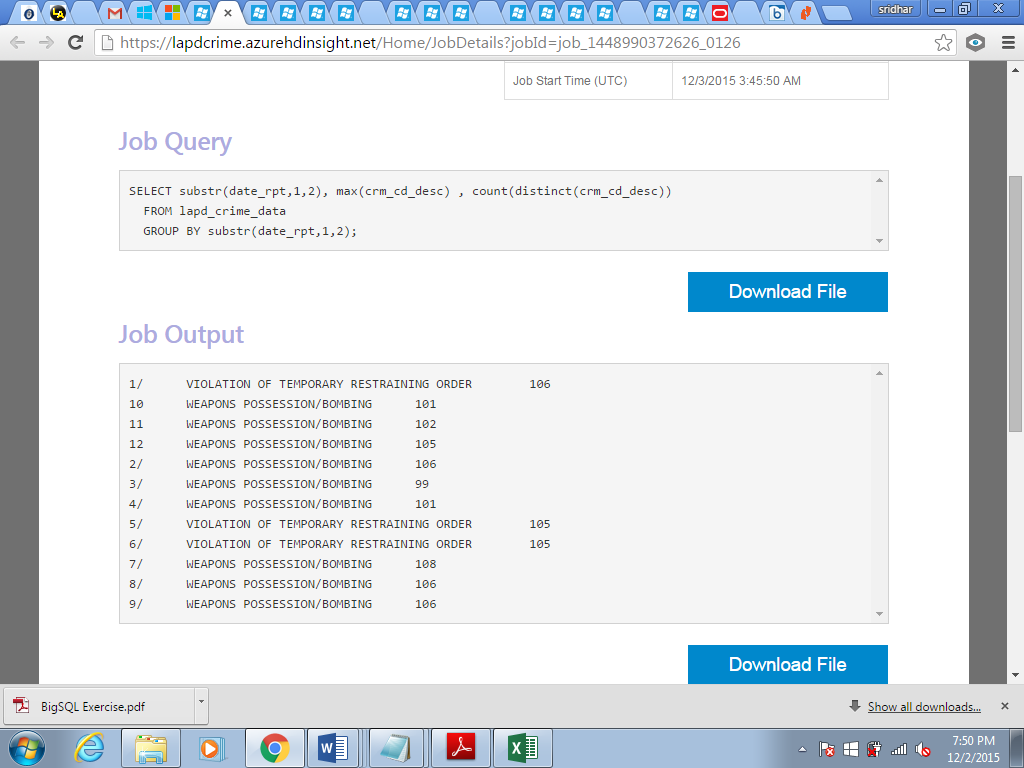
* Queries to analyze the data:

Crime types and number of crimes per month

SELECT substr(date\_rpt,1,2), max(crm\_cd\_desc) , count(distinct(crm\_cd\_desc))

FROM lapd\_crime\_data

GROUP BY substr(date\_rpt,1,2);



Similarly, run the following queries.

Total number of crimes per each day:

select date\_rpt,count(area) from LAPD\_CRIME\_DATA group by date\_rpt;

Distinct number of crimes performed:

select distinct crm\_cd\_desc from LAPD\_CRIME\_DATA;

Number of crimes performed based on radius:

select location\_values.distance as miles, count(\*) as totalcrimes

from (

select case

when distance between 0 and 5 then ' 0- 5'

when distance between 5 and 10 then '5-10'

when distance between 10 and 15 then '10-15'

when distance between 15 and 20 then '15-20'

when distance between 20 and 25 then '20-25'

when distance between 25 and 30 then '25-30'

when distance between 30 and 35 then '30-35'

when distance between 35 and 40 then '35-40'

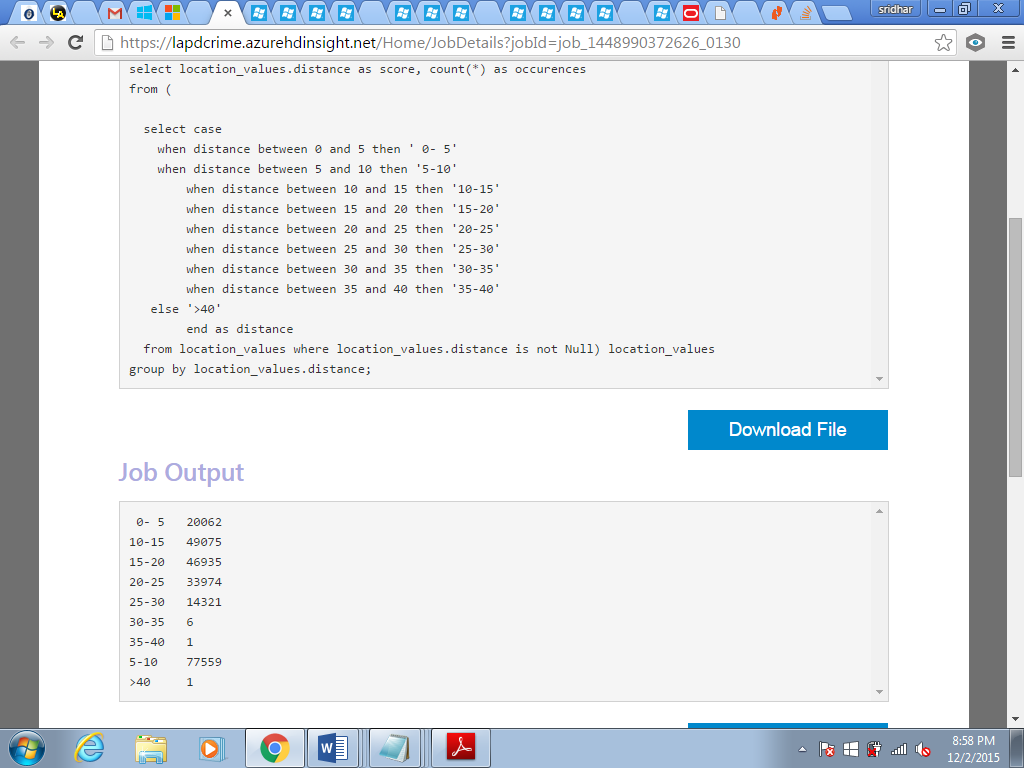
when distance between 40 and 50 then '40-50'

else '> 50'

end as distance

from location\_values where location\_values.distance is not Null) location\_values

group by location\_values.distance ;



Total no of crimes for every 2hrs time interval under 5 miles from university:

Select lapd\_crime\_data.time\_occ,count(\*) as crimes

from (

select case

when time\_occ between 0 and 200 then '00:00-02:00'

when time\_occ between 200 and 400 then '02:00-04:00'

when time\_occ between 400 and 600 then '04:00-06:00'

when time\_occ between 600 and 800 then '06:00-08:00'

when time\_occ between 800 and 1000 then '08:00-10:00'

when time\_occ between 1000 and 1200 then '10:00-12:00'

when time\_occ between 1200 and 1400 then '12:00-14:00'

when time\_occ between 1400 and 1600 then '14:00-16:00'

when time\_occ between 1600 and 1800 then '16:00-18:00'

when time\_occ between 1800 and 2000 then '18:00-20:00'

when time\_occ between 2000 and 2200 then '20:00-22:00'

when time\_occ between 2200 and 2400 then '22:00-24:00'

else 'others'

end as time\_occ

from lapd\_crime\_data

join location\_values on lapd\_crime\_data.dr\_no=location\_values.dr\_no

where location\_values.distance between 0 and 5

) lapd\_crime\_data

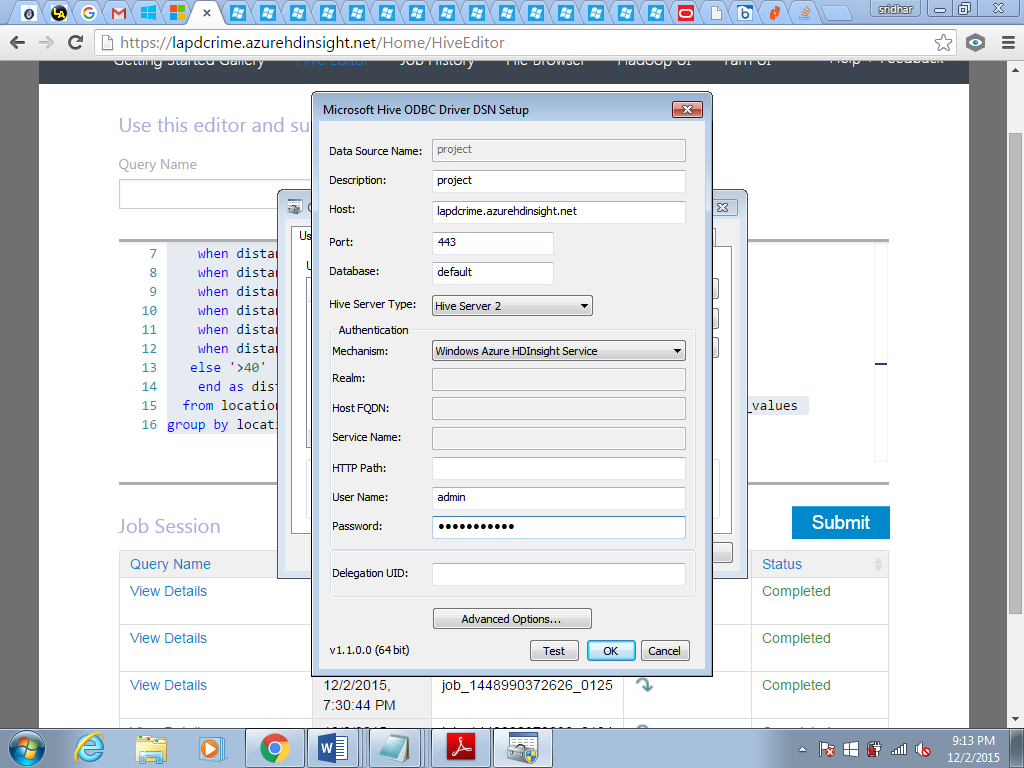
group by lapd\_crime\_data.time\_occ;

Similarly, we executed few queries to analyze the data in different way and these queries are saved in **“queries”** folder.

Create new Data Source

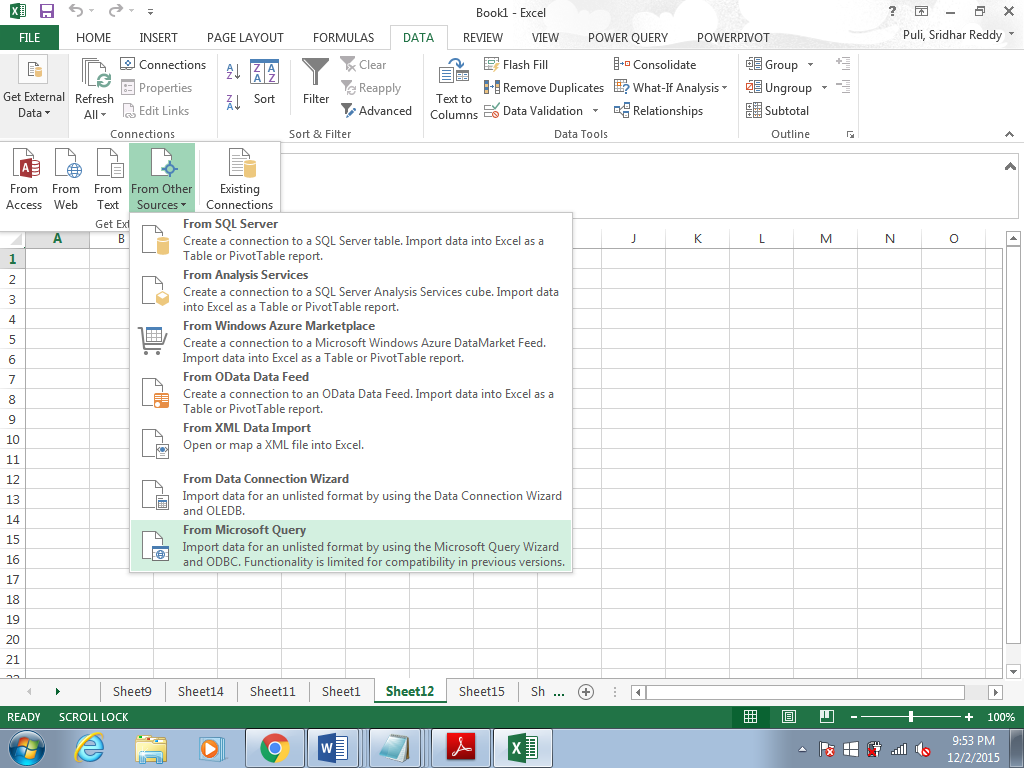
Steps to create DataSource:

* Install ODBC administrator
* Open ODBC administrator and click add under “**userdsn**”
* Select “Microsoft ODBC Hive Driver” and click finish.
* Enter data source name,host name which is nothing but clustername.azurehdinsight.net and also enter username,password.
* Click ‘ok’ when finished.

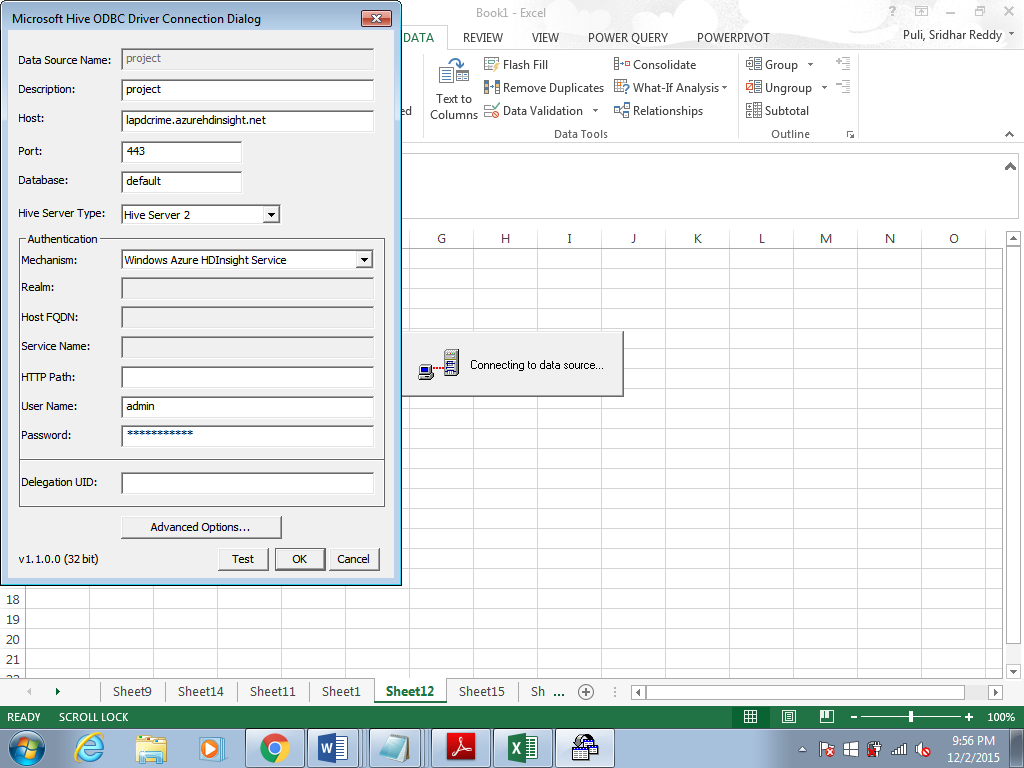


Import Data into Excel

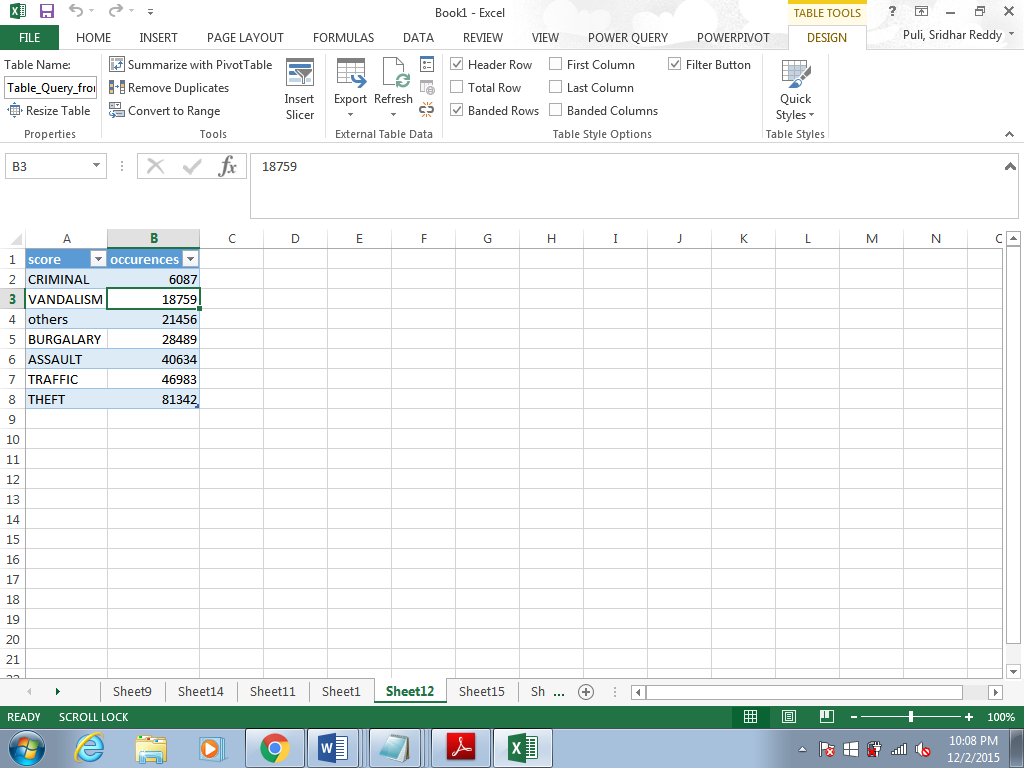
Open Excel sheet, under ‘Data’ select get external data->from other sources->from Microsoft query.



Under “databases” select datasource which we created and click ‘ok’. Enter password and click ‘ok’.

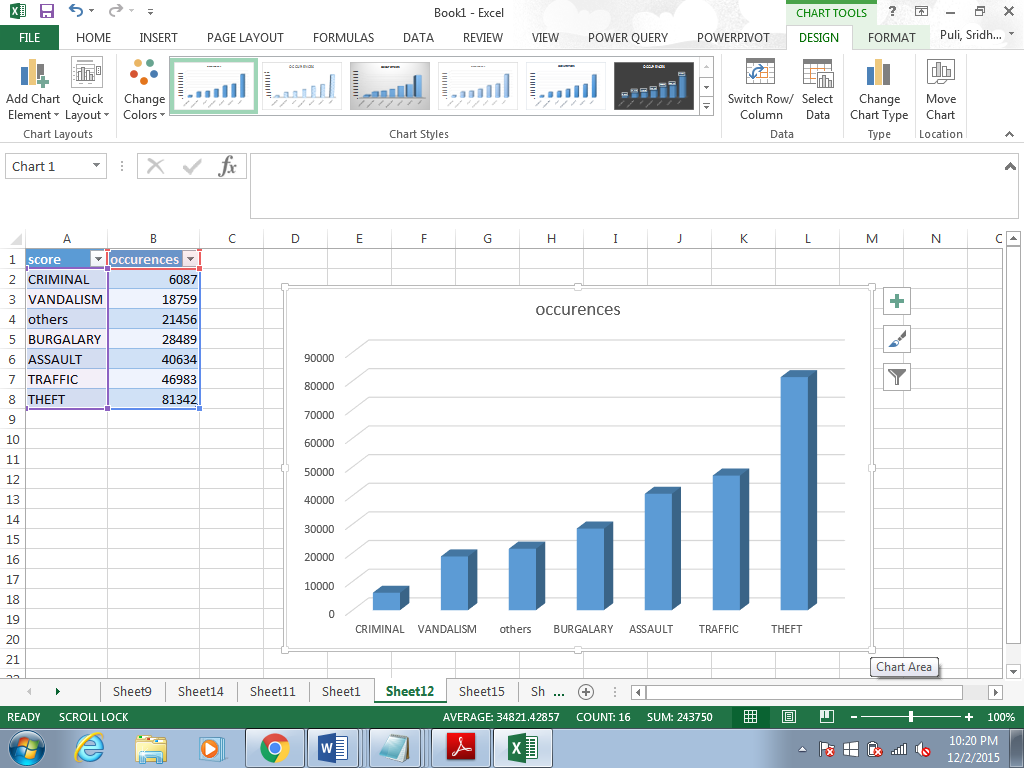


Select Table and click next and finish. When a Window named ‘import data’ appears select properties. Under ‘definition’ tab in ‘command text’ paste the hive query and click ‘ok’.



Graphical representation

Under insert tab, select the required type of chart and the chart appears on sheets.



View Crime locations on map

Select map under insert tab. Tick the latval and longval check boxes. When latval and longval appears in the bottom section, in the drop down beside these elements select latitude and longitude respectively.

