For New York Area:

1. curl -c /tmp/cookies "https://drive.google.com/uc?export=download&id=1Pqk6x-W4MW4a0xVGOeLztVYkXnkiEauu"> /tmp/enriched\_ny.html
2. curl -L -b /tmp/cookies "https://drive.google.com$(cat /tmp/enriched\_ny.html | grep -Po 'uc-download-link" [^>]\* href="\K[^"]\*' | sed 's/\&amp;/\&/g')" > nycrimedata.zip
3. unzip nycrimedata.zip
4. Open hive terminal using: hive
5. Run Following commands on hive shell do:
   1. create database if not exists crime\_data;
   2. use crime\_data;
   3. CREATE TABLE IF NOT EXISTS ny\_crime\_data(

cmplnt\_num DECIMAL,

cmplnt\_fr\_dt DATE,

cmplnt\_fr\_tm STRING,

cmplnt\_to\_dt DATE,

cmplnt\_to\_tm STRING,

rpt\_dt DATE,

ky\_cd DECIMAL,

ofns\_desc STRING,

pd\_cd DECIMAL,

pd\_desc STRING,

crm\_atpt\_cptd\_cd STRING,

law\_cat\_cd STRING,

juris\_desc STRING,

boro\_nm STRING,

addr\_pct\_cd DECIMAL,

loc\_of\_occur\_desc STRING,

prem\_typ\_desc STRING,

parks\_nm STRING,

hadevelopt STRING,

x\_coord\_cd DECIMAL,

y\_coord\_cd DECIMAL,

latitude DECIMAL,

longitude DECIMAL,

lat\_lon STRING,

zipcode STRING,

city STRING,

state STRING) ROW FORMAT SERDE'org.apache.hadoop.hive.serde2.OpenCSVSerde'WITH SERDEPROPERTIES ("separatorChar" = ",","quoteChar" = "\"")TBLPROPERTIES ("skip.header.line.count"="1");

* 1. load data local inpath '/home/ngupta8/fdny\_final.csv' into table ny\_crime\_data;
  2. Select \* from ny\_crime\_data limit 10;

1. Run Following commands from the bash shell and your home directory for instance /home/ngupta8 (NOT from Hive shell).

**Query -1: Top 10 crimes:**

hive -e "use default ;Select count(\*) as crime\_count,ofns\_desc from ny\_crime\_data group by ofns\_desc order by crime\_count DESC limit 10;"| perl -lpe 's/"/\\"/g; s/^|$/"/g; s/\t/","/g' > Top\_10\_Crime\_NY.csv;

**Query -2 : Top 10 crimes per year :**

hive -e 'set hive.cli.print.header=true ; SET hive.groupby.orderby.position.alias=true ; select count(\*) as crime\_count,from\_unixtime(unix\_timestamp(rpt\_dt,"mm/dd/yyyy"),"Y") as year from ny\_crime\_data where ofns\_desc is not NULL and ofns\_desc != "" AND ofns\_desc IN(

"PETIT LARCENY",

"HARRASSMENT 2",

"ASSAULT 3 & RELATED OFFENSES",

"CRIMINAL MISCHIEF & RELATED OF",

"GRAND LARCENY",

"DANGEROUS DRUGS",

"OFF. AGNST PUB ORD SENSBLTY &",

"ROBBERY",

"FELONY ASSAULT",

"BURGLARY") group by 2'| perl -lpe 's/"/\\"/g; s/^|$/"/g; s/\t/","/g' >Top\_10\_Crime\_Per\_Year\_NY.csv ;

**Query -3 :TOP 10 crimes by day :**

hive -e 'set hive.cli.print.header=true ; SET hive.groupby.orderby.position.alias=true ; use default ; with query as (Select count(\*) as crime\_count,ofns\_desc,from\_unixtime(unix\_timestamp(cmplnt\_fr\_dt,"mm/dd/yyyy"),"E") as day ,zipcode,city,state from ny\_crime\_data where ofns\_desc is not NULL and ofns\_desc != "" AND ofns\_desc IN(

"PETIT LARCENY",

"HARRASSMENT 2",

"ASSAULT 3 & RELATED OFFENSES",

"CRIMINAL MISCHIEF & RELATED OF",

"GRAND LARCENY",

"DANGEROUS DRUGS",

"OFF. AGNST PUB ORD SENSBLTY &",

"ROBBERY",

"FELONY ASSAULT",

"BURGLARY") AND zipcode != "NONE" group by ofns\_desc,3, zipcode,city,state) select \* from query where day != "NULL";'| perl -lpe 's/"/\\"/g; s/^|$/"/g; s/\t/","/g' >Top\_10\_Crime\_In\_Day\_NY.csv;

**Query -4 : No of crimes by hour :**

hive -e 'set hive.cli.print.header=true ; SET hive.groupby.orderby.position.alias=true ; use default ; with query as (Select count(\*) as crime\_count,ofns\_desc,hour(cmplnt\_fr\_tm) as hour,zipcode,city,state from ny\_crime\_data where ofns\_desc is not NULL and ofns\_desc != "" AND ofns\_desc IN(

"PETIT LARCENY",

"HARRASSMENT 2",

"ASSAULT 3 & RELATED OFFENSES",

"CRIMINAL MISCHIEF & RELATED OF",

"GRAND LARCENY",

"DANGEROUS DRUGS",

"OFF. AGNST PUB ORD SENSBLTY &",

"ROBBERY",

"FELONY ASSAULT",

"BURGLARY") AND zipcode != "NONE" group by ofns\_desc,3, zipcode,city,state) select \* from query where hour is not NULL order by hour;'| perl -lpe 's/"/\\"/g; s/^|$/"/g; s/\t/","/g' >Top\_10\_Crime\_In\_Hour\_NY.csv;

**Query -5: Top 10 zipcodes:**

hive -e 'set hive.cli.print.header=true;use default ; select count(\*) as crime\_count,zipcode from ny\_crime\_data where zipcode is not NULL and zipcode !="NONE" and zipcode != "" group by zipcode order by crime\_count DESC limit 10;'| perl -lpe 's/"/\\"/g; s/^|$/"/g; s/\t/","/g' > Top\_10\_Crime\_Prone\_zipcode\_NY.csv;

**Download processed data on local:**

Scp ngupta8@bi-hadoop-prod-4126.bi.services.us-south.bluemix.net:\*\_NY.csv .