databricks Big Data Lahari Project

(https://databricks.com)

Mounting to the bucket

```
%fs

ls /mnt/lahari-bigdata
```

Table

	path	name	size	modificationTime
1	dbfs:/mnt/lahari-bigdata/athena_output/	athena_output/	0	0
2	dbfs:/mnt/lahari-bigdata/fall2022/	fall2022/	0	0

2 rows

Mounting to the Project folder

```
%fs

ls /mnt/lahari-bigdata/fall2022/project/nypd-arrests-data
```

Table

	path	name	size
1	dbfs:/mnt/lahari-bigdata/fall2022/project/nypd-arrests-data/bronze/	bronze/	0
2	dbfs:/mnt/lahari-bigdata/fall2022/project/nypd-arrests-data/bronze_delta/	bronze_delta/	0
3	dbfs:/mnt/lahari-bigdata/fall2022/project/nypd-arrests-data/nypd-big-data-arrests-data.csv.bz2	nypd-big-data-arrests-data.csv.bz2	1502
4	dbfs:/mnt/lahari-bigdata/fall2022/project/nypd-arrests-data/silver/	silver/	0

4 rows

Creating Bronze Table

```
%python

#https://spark.apache.org/docs/latest/sql-data-sources-csv.html


df=spark.read.option("delimiter", ";").option("header", "true").csv("dbfs:/mnt/lahari-bigdata/fall2022/project/nypd-
arrests-data/nypd-big-data-arrests-data.csv.bz2/")
df.show(truncate=False)
```

```
+----------+-----------+-----+--------------------------------+-----+--------------------------+----------+----------
+-----------+---------------+-----------------+---------+--------+--------------+----------+----------+------------------
+------------------+---------------------------------------------+
|ARREST_KEY|ARREST_DATE|PD_CD|PD_DESC                          |KY_CD|OFNS_DESC                 |LAW_CODE  |LAW_CAT_CD
|ARREST_BORO|ARREST_PRECINCT|JURISDICTION_CODE|AGE_GROUP|PERP_SEX|PERP_RACE     |X_COORD_CD|Y_COORD_CD|Latitude
|Longitude          |Lon_Lat                                      |
+----------+-----------+-----+--------------------------------+-----+--------------------------+----------+----------
+-----------+---------------+-----------------+---------+--------+--------------+----------+----------+------------------
+------------------+---------------------------------------------+
|236791704 |11/22/2021 |581  |null                             |null |null                      |PL 2225001|M
|M          |28             |0                |45-64    |M       |BLACK         |997427    |230378    |40.799008797000056
|-73.95240854099995 |POINT (-73.95240854099995 40.799008797000056)|
```

```
|237354740 |12/04/2021 |153  |RAPE 3                              |104  |RAPE                          |PL 1302502|F
|B          |41             |0                |25-44    |M       |WHITE HISPANIC|1013232   |236725    |40.816391847000034
|-73.89529641399997 |POINT (-73.89529641399997 40.816391847000034)|
|236081433 |11/09/2021 |681  |CHILD, ENDANGERING WELFARE          |233  |SEX CRIMES                    |PL 2601001|M
|Q          |113            |0                |25-44    |M       |BLACK         |1046367   |186986    |40.67970040800003
|-73.77604736799998 |POINT (-73.77604736799998 40.67970040800003) |
|32311380  |06/18/2007 |511  |CONTROLLED SUBSTANCE, POSSESSION 7|235  |DANGEROUS DRUGS               |PL 2200300|M
|Q          |27             |1                |18-24    |M       |BLACK         |null      |null      |null
```

```
%py
spark.conf.set("spark.sql.legacy.allowCreatingManagedTableUsingNonemptyLocation","true")
```

```
%python


#https://ganeshchandrasekaran.com/how-to-save-the-spark-data-frame-as-an-external-table-in-databricks-6e7ce0ab12cc

df.write.format("csv").option("path","dbfs:/mnt/lahari-bigdata/fall2022/project/nypd-arrests-
data/bronze/").saveAsTable("projectbronze")
```

```
select * from projectbronze limit 5;
```

Table

	ARREST_KEY	ARREST_DATE	PD_CD	PD_DESC	KY_CD	OFNS_DESC	LAW
1	236791704	11/22/2021	581	null	null	null	PL 22
2	237354740	12/04/2021	153	RAPE 3	104	RAPE	PL 13
3	236081433	11/09/2021	681	CHILD, ENDANGERING WELFARE	233	SEX CRIMES	PL 26
4	32311380	06/18/2007	511	CONTROLLED SUBSTANCE, POSSESSION 7	235	DANGEROUS DRUGS	PL 22
5	192799737	01/26/2019	177	SEXUAL ABUSE	116	SEX CRIMES	PL 13

5 rows

```
select count(*) from projectbronze;
```

Table

	count(1)
1	5308876

1 row

```
describe extended projectbronze;
```

Table

	col_name	data_type	comment
1	ARREST_KEY	string	null
2	ARREST_DATE	string	null
3	PD_CD	string	null
4	PD_DESC	string	null
5	KY_CD	string	null
6	OFNS_DESC	string	null
7	LAW_CODE	string	null

33 rows

Creating Bronze Delta

Removed two columns-Lon_Lat and Law_CT_CD(because similiar ones exist already)

Changed the datatypes of 9 columns during the delta creation itself

```
CREATE TABLE projectdelta (
arrest_id STRING,
arrest_date date,
arrest_boro string,
pd_cd int,
ky_cd int,
pd_desc string ,
offense_desc string,
law_code string,
arrest_precinct int,
jurisdiction int,
perp_age string,
perp_sex string,
perp_race string,
x_coord double,
y_coord double,
latitude double,
longitude double
)
LOCATION 'dbfs:/mnt/lahari-bigdata/fall2022/project/nypd-arrests-data/bronze_delta/'
```

OK

```
INSERT INTO projectdelta select
arrest_key,
to_date(ARREST_DATE,'M/d/y'),
arrest_boro,
pd_cd ,
ky_cd,
PD_DESC,
OFNS_DESC,
LAW_CODE,
ARREST_PRECINCT,
JURISDICTION_CODE,
AGE_GROUP,
PERP_SEX,
PERP_RACE,
X_COORD_CD,
Y_COORD_CD,
Latitude,
Longitude
from projectbronze;
```

Table

	num_affected_rows	num_inserted_rows
1	5308876	5308876

1 row

```
select * from projectdelta limit 5;
```

Table

	arrest_id	arrest_date	arrest_boro	pd_cd	ky_cd	pd_desc	offense_desc
1	236791704	2021-11-22	M	581	null	null	null
2	237354740	2021-12-04	B	153	104	RAPE 3	RAPE
3	236081433	2021-11-09	Q	681	233	CHILD, ENDANGERING WELFARE	SEX CRIMES

4	32311380	2007-06-18	Q	511	235	CONTROLLED SUBSTANCE, POSSESSION 7	DANGEROUS DRU
5	192799737	2019-01-26	M	177	116	SEXUAL ABUSE	SEX CRIMES

5 rows

```
select count(*) from projectdelta;
```

Table

	count(1)
1	5308876

1 row

```
describe extended projectdelta;
```

Table

	col_name	data_type	comment
1	arrest_id	string	
2	arrest_date	date	
3	arrest_boro	string	
4	pd_cd	int	
5	ky_cd	int	
6	pd_desc	string	
7	offense_desc	string	

30 rows

Validation

Checking the number of null values in each column

```
select
    sum(case when arrest_id is null then 1 else 0 end) as arrest_id,
    sum(case when arrest_date is null then 1 else 0 end) as arrest_date,
    sum(case when arrest_boro is null then 1 else 0 end) as arrest_boro,
    sum(case when pd_cd is null then 1 else 0 end) as pd_cd,
    sum(case when ky_cd is null then 1 else 0 end) as ky_cd,
    sum(case when pd_desc is null then 1 else 0 end) as pd_desc,
    sum(case when offense_desc is null then 1 else 0 end) as offense_desc,
    sum(case when law_code is null then 1 else 0 end) as law_code,
    sum(case when arrest_precinct is null then 1 else 0 end) as arrest_precinct ,
    sum(case when jurisdiction is null then 1 else 0 end) as jurisdiction,
    sum(case when perp_age is null then 1 else 0 end) as perp_age,
    sum(case when perp_race is null then 1 else 0 end) as perp_race,
    sum(case when x_coord is null then 1 else 0 end) as x_coord,
    sum(case when y_coord is null then 1 else 0 end) as y_coord,
    sum(case when latitude is null then 1 else 0 end) as latitude,
    sum(case when longitude is null then 1 else 0 end) as longitude

from projectdelta;
```

Table

	arrest_id	arrest_date	arrest_boro	pd_cd	ky_cd	pd_desc	offense_desc	law_code	arrest_pre
1	0	0	8	313	9169	9169	9169	196	0

1 row

Cleaning

```
select * from projectdelta where x_coord is NULL;
```

Table

	arrest_id	arrest_date	arrest_boro	pd_cd	ky_cd	pd_desc	offense_desc
1	32311380	2007-06-18	Q	511	235	CONTROLLED SUBSTANCE, POSSESSION 7	DANGEROUS DRU

1 row

We can see that there exists only one row where x_coord, y_coord, latitude,longitude are null.

```
delete from projectdelta where x_coord is NULL;
```

Table

	num_affected_rows
1	1

1 row

```
select * from projectdelta limit 5;
```

Table

	arrest_id	arrest_date	arrest_boro	pd_cd	ky_cd	pd_desc	offense
1	183601677	2018-06-07	K	109	106	ASSAULT 2,1,UNCLASSIFIED	FELONY
2	182980759	2018-05-22	K	507	117	CONTROLLED SUBSTANCE, POSSESSION 5	DANGEF
3	191458539	2018-12-23	Q	779	126	PUBLIC ADMINISTRATION,UNCLASSIFIED FELONY	MISCELL
4	183321230	2018-05-31	K	494	111	STOLEN PROPERTY 2,1,POSSESSION,UNCLASSIFIED	POSSES
5	191448370	2018-12-22	B	339	341	LARCENY,PETIT FROM OPEN AREAS,UNCLASSIFIED	PETIT LA

5 rows

```
select count(*) from projectdelta where ky_cd is NULL and pd_desc is NULL and offense_desc is NULL;
```

Table

	count(1)
1	9169

1 row

Above we can see that we 9169 rows where ky_cd , offense_desc, and pd_desc are null so we drop them in below

```
delete from projectdelta where ky_cd is null and pd_desc is null and offense_desc is null;
```

Table

	num_affected_rows
1	9169

1 row

```
select
    sum(case when arrest_id is null then 1 else 0 end) as arrest_id,
    sum(case when arrest_date is null then 1 else 0 end) as arrest_date,
    sum(case when arrest_boro is null then 1 else 0 end) as arrest_boro,
    sum(case when pd_cd is null then 1 else 0 end) as pd_cd,
    sum(case when ky_cd is null then 1 else 0 end) as ky_cd,
    sum(case when pd_desc is null then 1 else 0 end) as pd_desc,
    sum(case when offense_desc is null then 1 else 0 end) as offense_desc,
    sum(case when law_code is null then 1 else 0 end) as law_code,
    sum(case when arrest_precinct is null then 1 else 0 end) as arrest_precinct ,
    sum(case when jurisdiction is null then 1 else 0 end) as jurisdiction,
    sum(case when perp_age is null then 1 else 0 end) as perp_age,
    sum(case when perp_race is null then 1 else 0 end) as perp_race,
    sum(case when x_coord is null then 1 else 0 end) as x_coord,
    sum(case when y_coord is null then 1 else 0 end) as y_coord,
    sum(case when latitude is null then 1 else 0 end) as latitude,
    sum(case when longitude is null then 1 else 0 end) as longitude

from projectdelta;
```

Table

	arrest_id	arrest_date	arrest_boro	pd_cd	ky_cd	pd_desc	offense_desc	law_code	arrest_pre
1	0	0	8	0	0	0	0	0	0

1 row

By deleting nulls in ky_cd, pd_desc , offense_desc we can see that pd_cd,law_code nulls are also removed

Deleting nulls in other columns(arrest_boro, jurisdiction, perp_age)

```
delete from projectdelta where arrest_boro is NULL;
```

Table

	num_affected_rows
1	8

1 row

```
delete from projectdelta where jurisdiction is NULL;
```

Table

	num_affected_rows
1	10

1 row

```
delete from projectdelta where perp_age is NULL;
```

Table

	num_affected_rows

1	13

1 row

Final Validation once again

```
select
    sum(case when arrest_id is null then 1 else 0 end) as arrest_id,
    sum(case when arrest_date is null then 1 else 0 end) as arrest_date,
    sum(case when arrest_boro is null then 1 else 0 end) as arrest_boro,
    sum(case when pd_cd is null then 1 else 0 end) as pd_cd,
    sum(case when ky_cd is null then 1 else 0 end) as ky_cd,
    sum(case when pd_desc is null then 1 else 0 end) as pd_desc,
    sum(case when offense_desc is null then 1 else 0 end) as offense_desc,
    sum(case when law_code is null then 1 else 0 end) as law_code,
    sum(case when arrest_precinct is null then 1 else 0 end) as arrest_precinct ,
    sum(case when jurisdiction is null then 1 else 0 end) as jurisdiction,
    sum(case when perp_age is null then 1 else 0 end) as perp_age,
    sum(case when perp_race is null then 1 else 0 end) as perp_race,
    sum(case when x_coord is null then 1 else 0 end) as x_coord,
    sum(case when y_coord is null then 1 else 0 end) as y_coord,
    sum(case when latitude is null then 1 else 0 end) as latitude,
    sum(case when longitude is null then 1 else 0 end) as longitude

from projectdelta;
```

Table

	arrest_id	arrest_date	arrest_boro	pd_cd	ky_cd	pd_desc	offense_desc	law_code	arrest_pre
1	0	0	0	0	0	0	0	0	0

1 row

Following command shows the count of rows after clearing NULL values

```
select count(*) from projectdelta;
```

Table

	count(1)
1	5299675

1 row

Total 9201 null records are cleaned out of 5.2 million rows

Creation of Silver Table

```
CREATE TABLE projectsilver (
arrest_id STRING,
arrest_date date,
arrest_boro string,
pd_cd int,
ky_cd int,
pd_desc string ,
offense_desc string,
law_code string,
arrest_precinct int,
jurisdiction int,
perp_age string,
perp_sex string,
perp_race string,
x_coord double,
y_coord double,
latitude double,
longitude double
)
LOCATION 'dbfs:/mnt/lahari-bigdata/fall2022/project/nypd-arrests-data/silver/'
```

OK

```
insert into projectsilver select * from projectdelta;
```

Table

	num_affected_rows	num_inserted_rows
1	5299675	5299675

1 row

```
select count(*) from projectsilver;
```

Table

	count(1)
1	5299675

1 row

```
describe extended projectsilver;
```

Table

	col_name	data_type	comment
1	arrest_id	string	
2	arrest_date	date	
3	arrest_boro	string	
4	pd_cd	int	
5	ky_cd	int	
6	pd_desc	string	
7	offense_desc	string	

30 rows

```
select * from projectsilver limit 5;
```

Table

	arrest_id	arrest_date	arrest_boro	pd_cd	ky_cd	pd_desc	offense_desc	law_
1	237354740	2021-12-04	B	153	104	RAPE 3	RAPE	PL 1
2	236081433	2021-11-09	Q	681	233	CHILD, ENDANGERING WELFARE	SEX CRIMES	PL 2
3	192799737	2019-01-26	M	177	116	SEXUAL ABUSE	SEX CRIMES	PL 1
4	236106641	2021-11-10	B	263	114	ARSON 2,3,4	ARSON	PL 1
5	238383628	2021-12-28	Q	729	113	FORGERY,ETC.,UNCLASSIFIED-FELO	FORGERY	PL 1

5 rows

5 Queries on Silver table

Subquery

The following is the subquery to retrieve description of crime and other details of MinorMale(under 18 years age) Arrests whose borough of arrest is Brooklyn('K') and arrest date is between September 2020 and December 2020

```
SELECT arrest_id,arrest_date,arrest_boro,perp_age,perp_sex,pd_desc FROM projectsilver WHERE arrest_date between '2020-11-01' and '2020-12-31' and arrest_id IN (SELECT arrest_id FROM projectsilver WHERE perp_age = '<18' and perp_sex='M' and arrest_boro='K');
```

Table

	arrest_id	arrest_date	arrest_boro	perp_age	perp_sex	pd_desc
1	222233845	2020-12-24	K	<18	M	ASSAULT 3
2	221513670	2020-12-07	K	<18	M	UNAUTHORIZED USE VEHICLE 3
3	220391635	2020-11-11	K	<18	M	ASSAULT 2,1,UNCLASSIFIED
4	222037267	2020-12-19	K	<18	M	ROBBERY,OPEN AREA UNCLASSIFIED
5	220837026	2020-11-21	K	<18	M	WEAPONS POSSESSION 1 & 2
6	221423296	2020-12-06	K	<18	M	ROBBERY,OPEN AREA UNCLASSIFIED
7	221371628	2020-12-04	K	<18	M	MURDER.UNCLASSIFIED

222 rows

SQL Functions (Count and groupby)

The following is a query to get the count of Asians/Pacific Islanders involved in each offense.

```
select offense_desc , count(perp_race) as count_asians From projectsilver where perp_race='ASIAN / PACIFIC ISLANDER' group by offense_desc;
```

Table

	offense_desc	count_asians
1	OTHER TRAFFIC INFRACTION	9317
2	ANTICIPATORY OFFENSES	18
3	CHILD ABANDONMENT/NON SUPPORT 1	17
4	NEW YORK CITY HEALTH CODE	14
5	POSSESSION OF STOLEN PROPERTY 5	8667
6	OTHER OFFENSES RELATED TO THEF	153
7	VEHICLE AND TRAFFIC LAWS	9938

81 rows

CTE

The following is a query for printing the number of Females who have committed the Rape Crime

```
WITH my_cte AS (
  SELECT offense_desc, perp_sex
  FROM projectsilver
)
SELECT count(perp_sex) as count_female_rapists
FROM my_cte
WHERE offense_desc="RAPE" and perp_sex="F";
```

Table

	count_female_rapists
1	235

1 row

LEAD

The following query is for retrieving the information of Female UNKNOWN race crimes whose age is greater than 65 ordered by crime arrest dates

```
SELECT
        arrest_id,
        offense_desc,
    perp_race,
        perp_age,
    perp_sex,
    arrest_date,
        LEAD(arrest_date, 1) OVER (ORDER BY arrest_date) AS next_arrest_date
FROM
        projectsilver
where
   perp_race="UNKNOWN" and perp_sex="F" and perp_age="65+";
```

Table

	arrest_id	offense_desc	perp_race	perp_age	perp_sex	arrest_date	next_arre
1	10278158	ASSAULT 3 & RELATED OFFENSES	UNKNOWN	65+	F	2006-02-02	2006-09-
2	24347884	POSSESSION OF STOLEN PROPERTY 5	UNKNOWN	65+	F	2006-09-17	2006-12-
3	25889961	OFFENSES AGAINST PUBLIC ADMINISTRATION	UNKNOWN	65+	F	2006-12-21	2007-10-
4	35285698	DANGEROUS WEAPONS	UNKNOWN	65+	F	2007-10-25	2009-10-
5	66841991	CRIMINAL TRESPASS	UNKNOWN	65+	F	2009-10-15	2010-01-
6	69547685	OFF. AGNST PUB ORD SENSBLTY & RGHTS TO PRIV	UNKNOWN	65+	F	2010-01-04	2010-03-
7	71560738	OTHER OFFENSES RELATED TO THEFT	UNKNOWN	65+	F	2010-03-02	2010-03-

43 rows

Row_number

The following is a query to get data of above 65 age Female arrests in the order of their internal classification code(pd_cd)

Table

	arrest_id	pd_desc	offense_desc	pd_cd	perp_age
1	145169885	PEDDLING,UNLAWFUL	ADMINISTRATIVE CODE	874	65+
2	220951320	ADM.CODE,UNCLASSIFIED MISDEMEA	ADMINISTRATIVE CODE	878	65+
3	192427787	ADM.CODE,UNCLASSIFIED MISDEMEA	ADMINISTRATIVE CODE	878	65+
4	208321751	ADM.CODE,UNCLASSIFIED MISDEMEA	ADMINISTRATIVE CODE	878	65+
5	203796574	ADM.CODE,UNCLASSIFIED VIOLATIO	ADMINISTRATIVE CODE	879	65+
6	190135460	ADM.CODE,UNCLASSIFIED VIOLATION	ADMINISTRATIVE CODE	879	65+
7	189307240	ADM.CODE,UNCLASSIFIED VIOLATION	ADMINISTRATIVE CODE	879	65+

1,000 rows | Truncated data