## **⊗** databricksBig Data Lahari Project

(https://databricks.com)

## Mounting to the bucket

%fs

ls /mnt/lahari-bigdata

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

## **Mounting to the Project folder**

%fs

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

	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	150
4	dbfs:/mnt/lahari-bigdata/fall2022/project/nypd-arrests-data/silver/	silver/	0

## **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/nypdarrests-data/nypd-big-data-arrests-data.csv.bz2/") df.show(truncate=False)

|237354740 |12/04/2021 |153 |RAPE 3 |104 |RAPE |PL 1302502|F |41 0 | 25-44 M |WHITE HISPANIC|1013232 |236725 40.816391847000034 |236081433 |11/09/2021 |681 |CHILD, ENDANGERING WELFARE 233 | SEX CRIMES |PL 2601001|M 186986 40.67970040800003 |25-44 |M BLACK |1046367 |-73.77604736799998 | POINT (-73.77604736799998 40.67970040800003) | |32311380 |06/18/2007 |511 |CONTROLLED SUBSTANCE, POSSESSION 7|235 |DANGEROUS DRUGS |PL 2200300|M 18-24 null |1 IBLACK Inull Inull |Q 127 M

vq%

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-arrestsdata/bronze/").saveAsTable("projectbronze")

select \* from projectbronze limit 5;



select count(\*) from projectbronze;



describe extended projectbronze;



## **Creating Bronze Delta**

10

### Removed two columns-Lon\_Lat and Law\_CT\_CD(because similiar ones exist already)

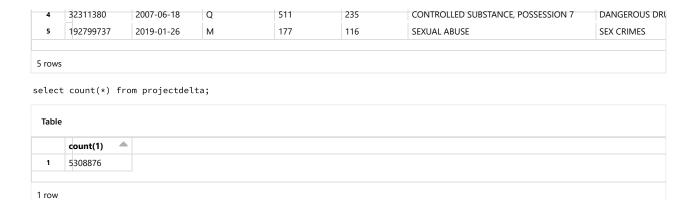
### Changed the datatypes of 9 columns during the delta creation itself

```
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
ΟK
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	Table					
	num_affected_rows	num_inserted_rows				
1	5308876	5308876				
1 row						

select \* from projectdelta limit 5;

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



describe extended projectdelta;



#### **Validation**

### Checking the number of null values in each column

```
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 ,
\operatorname{sum}(\operatorname{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,
\operatorname{sum}(\operatorname{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

## **Cleaning**

select \* from projectdelta where x\_coord is NULL;

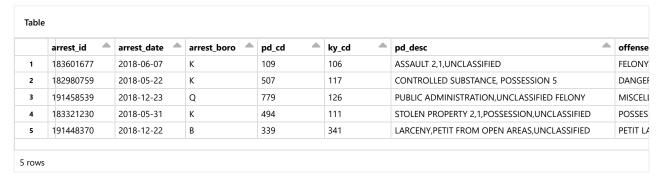


# We can see that there exists only one row where x\_coord, y\_coord, latitude,longitude are null.

delete from projectdelta where x\_coord is NULL;



select \* from projectdelta limit 5;



 $\texttt{select count(*) from projectdelta where ky\_cd is NULL and pd\_desc is NULL and offense\_desc is NULL;}$ 



# 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;

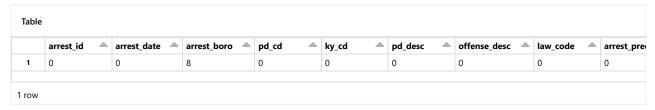




select

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,  $\operatorname{\mathsf{sum}}(\operatorname{\mathsf{case}}\ \operatorname{\mathsf{when}}\ \operatorname{\mathsf{arrest\_prec}}\ \operatorname{\mathsf{incl}}\ \operatorname{\mathsf{tis}}\ \operatorname{\mathsf{null}}\ \operatorname{\mathsf{then}}\ 1\ \operatorname{\mathsf{else}}\ 0\ \operatorname{\mathsf{end}})$  as  $\operatorname{\mathsf{arrest\_prec}}\ \operatorname{\mathsf{incl}}\ \operatorname{\mathsf{inc$ 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;



# 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;

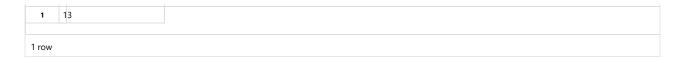


delete from projectdelta where jurisdiction is NULL;



delete from projectdelta where perp\_age is NULL;

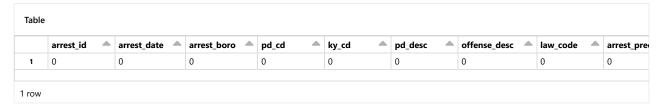




## **Final Validation once again**

```
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 ,
    \operatorname{sum}(\operatorname{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,
    \mathsf{sum}(\mathsf{case}\ \mathsf{when}\ \mathsf{latitude}\ \mathsf{is}\ \mathsf{null}\ \mathsf{then}\ \mathsf{1}\ \mathsf{else}\ \mathsf{0}\ \mathsf{end}) as latitude,
    sum(case when longitude is null then 1 else 0 end) as longitude
```

from projectdelta;



## Following command shows the count of rows after clearing NULL values

select count(\*) from projectdelta;



#### 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	ected_rows   num_inserted_rows
1	5299675	5299675
1 row		

select count(\*) from projectsilver;



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	strina	
30 rows	5		

select \* from projectsilver limit 5;

Table										

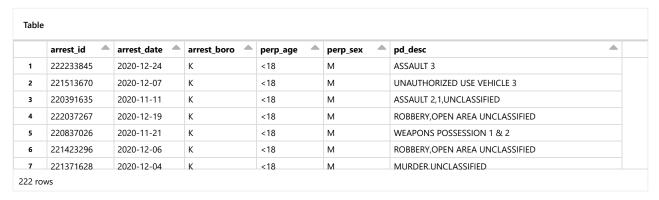


#### **5 Queries on Silver table**

#### 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

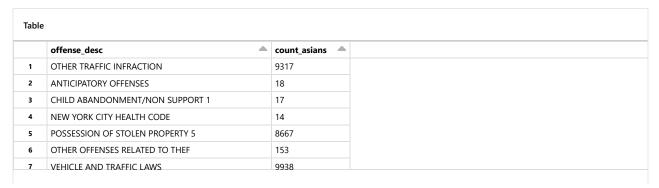
SELECT arrest\_id,arrest\_date,arrest\_boro,perp\_age,perp\_sex,pd\_desc FROM projectsilver WHERE arrest\_date between '2020-11-01' and '2020-12-31' and arrest\_id IN (SELECT arrest\_id FROM projectsilver WHERE perp\_age = '<18' and perp\_sex='M' and arrest\_boro='K');



### **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;



81 rows

#### **CTE**

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



#### **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

	arrest_id 🌰	offense_desc	perp_race	perp_age 📤	perp_sex	arrest_date 🌰	next_arr
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-

### 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)

	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+