GROUP 2

Crimes to Arrests

COURSE NUMBER: CIS 4400

COURSE SECTION: CMWA

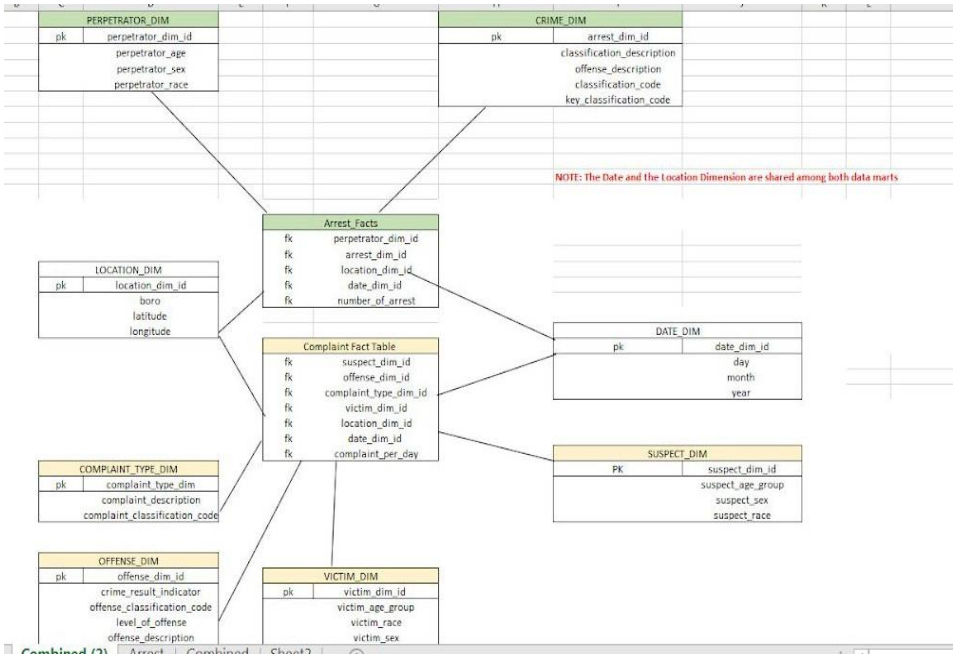
1. An introduction section similar to the proposal section including the narrative description of the business used for the data warehouse being created and description of the source data.

The business that we are conducting is a non-profit that is trying to see the most affected population who is arrested. We are trying to see if there are any attributes that contribute the most to the arrested population and tracking if all the crimes reported lead to arrests.

The data sources that were used for this project are derived from NYC Open Data. We found the NYPD Complaint Data set which contains all valid felonies, misdemeanors, and violation crimes reported to the NYPD; this data set includes all records beginning from 2016 and the granularity of this data is quarterly. This data set contains 35 columns but for the purpose of this project we only chose the important attributes which you will see in the dimensional model below.

Next up we have the NYPD Arrest Data set which is a breakdown of every arrest in NYC by the NYPD. Like the data set above, it is updated quarterly but contains those records beginning from 2018. The set has 18 columns but did not use all of them as shown below by our dimensional diagram.

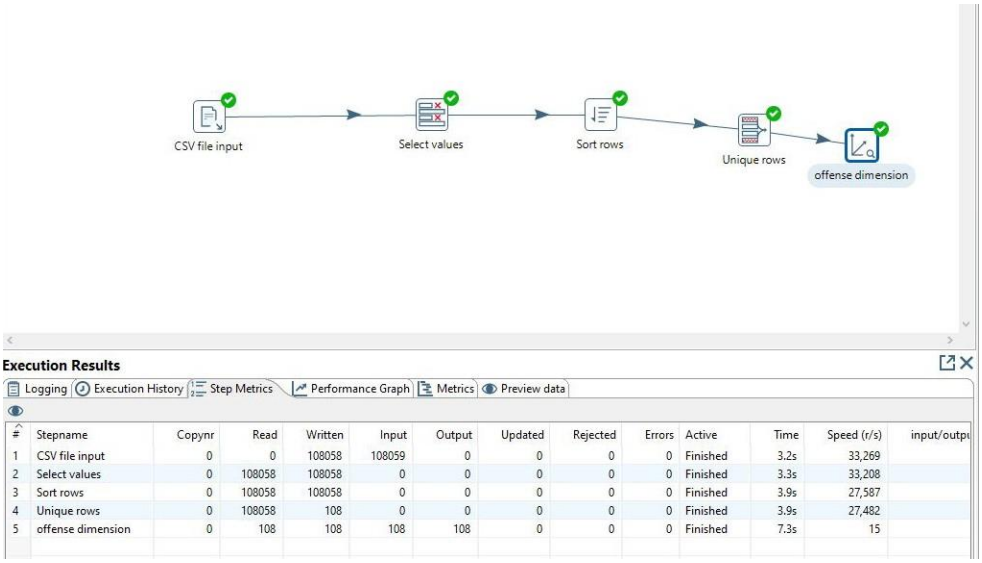
2. Dimensional Model Diagram



3. A description and screen pictures of the ETL process

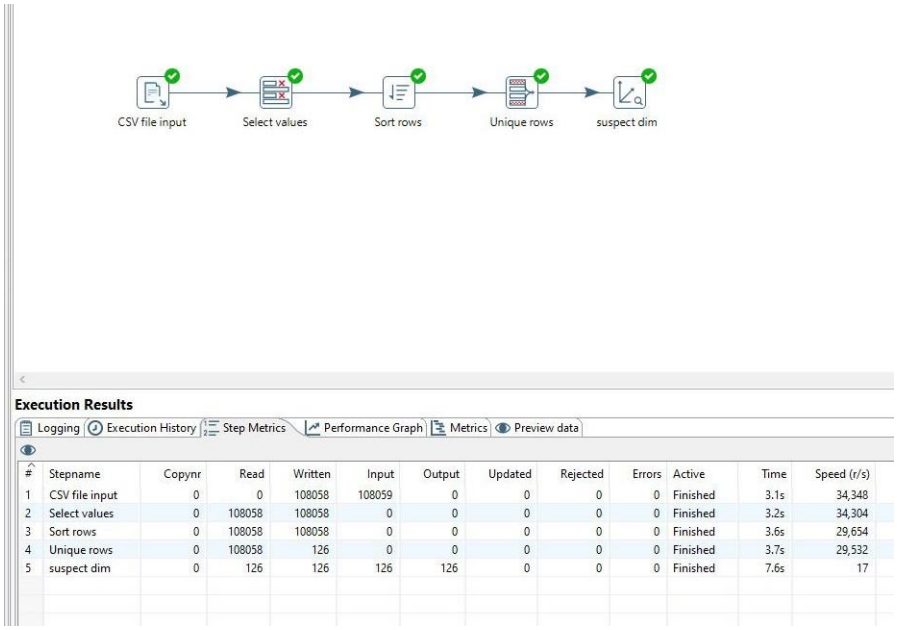
Description: This transformation started off with the CSV file. Then we selected off the values for the dimension, sorted and picked out the unique values from the set. Finally we created the **complaint** dimension

Description: Following the same steps as before we created the **offense** dimension which contained 4 of the original attributes from the dataset.



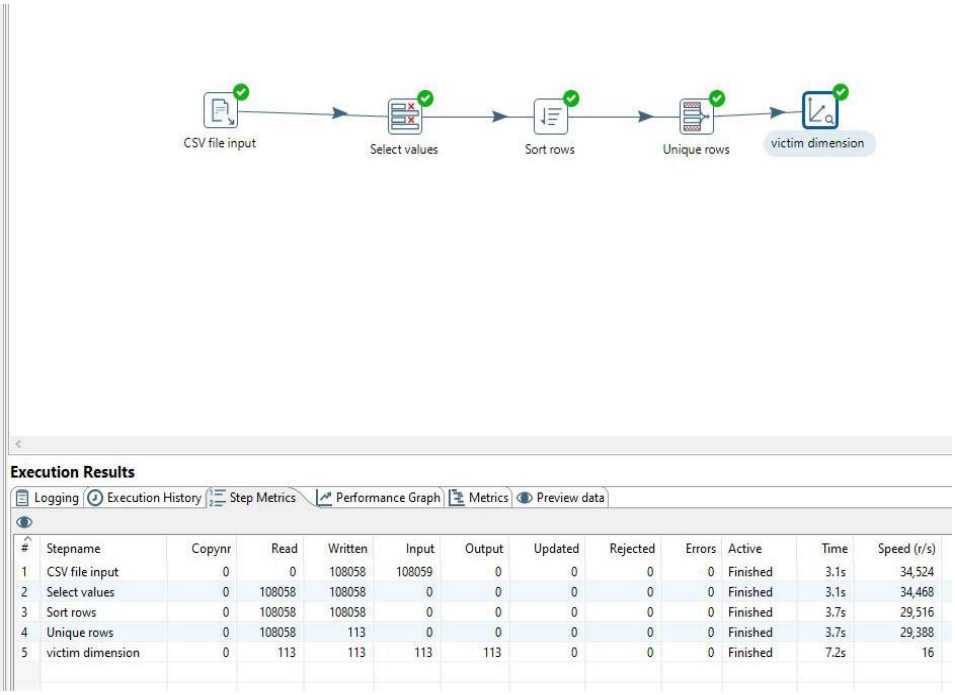
#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/outpu
1	CSV file input	0	0	108058	108059	0	0	0	0	Finished	3.2s	33,269	
2	Select values	0	108058	108058	0	0	0	0	0	Finished	3.3s	33,208	
3	Sort rows	0	108058	108058	0	0	0	0	0	Finished	3.9s	27,587	
4	Unique rows	0	108058	108	0	0	0	0	0	Finished	3.9s	27,482	
5	offense dimension	0	108	108	108	108	0	0	0	Finished	7.3s	15	

Description: This is the **suspect** dimension which contains the suspects age group, sex, and race.



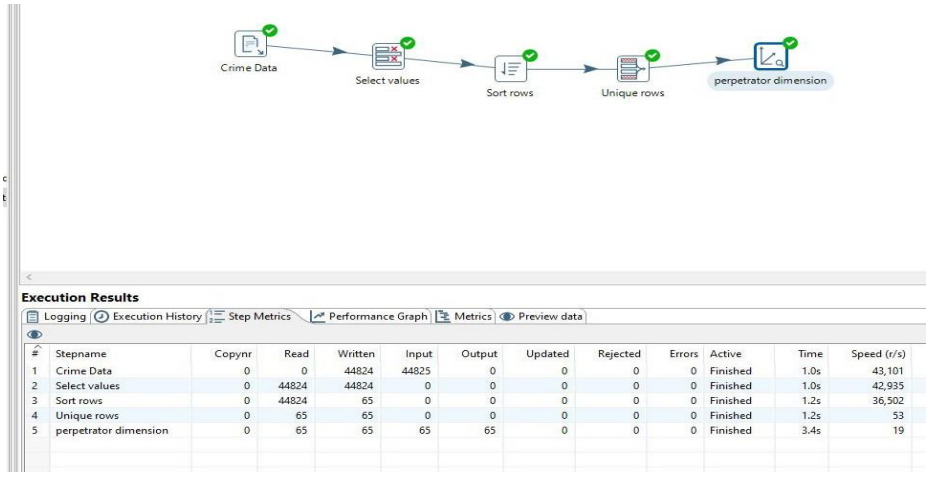
#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)
1	CSV file input	0	0	108058	108059	0	0	0	0	Finished	3.1s	34,348
2	Select values	0	108058	108058	0	0	0	0	0	Finished	3.2s	34,304
3	Sort rows	0	108058	108058	0	0	0	0	0	Finished	3.6s	29,654
4	Unique rows	0	108058	126	0	0	0	0	0	Finished	3.7s	29,532
5	suspect dim	0	126	126	126	126	0	0	0	Finished	7.6s	17

Description: This is the **victim** dimension and similar to the one above this one contains the victims age group, sex, and race



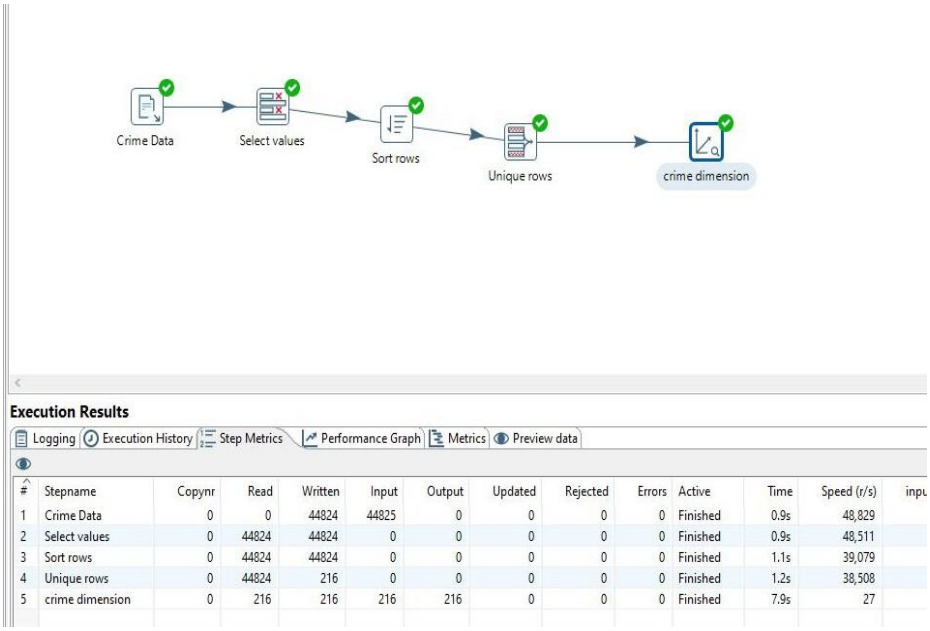
#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)
1	CSV file input	0	0	108058	108059	0	0	0	0	Finished	3.1s	34,524
2	Select values	0	108058	108058	0	0	0	0	0	Finished	3.1s	34,468
3	Sort rows	0	108058	108058	0	0	0	0	0	Finished	3.7s	29,516
4	Unique rows	0	108058	113	0	0	0	0	0	Finished	3.7s	29,388
5	victim dimension	0	113	113	113	113	0	0	0	Finished	7.2s	16

Description: This is the **perpretrator** dimension and similar to the one above it contains the perpretrators age group, sex, and race



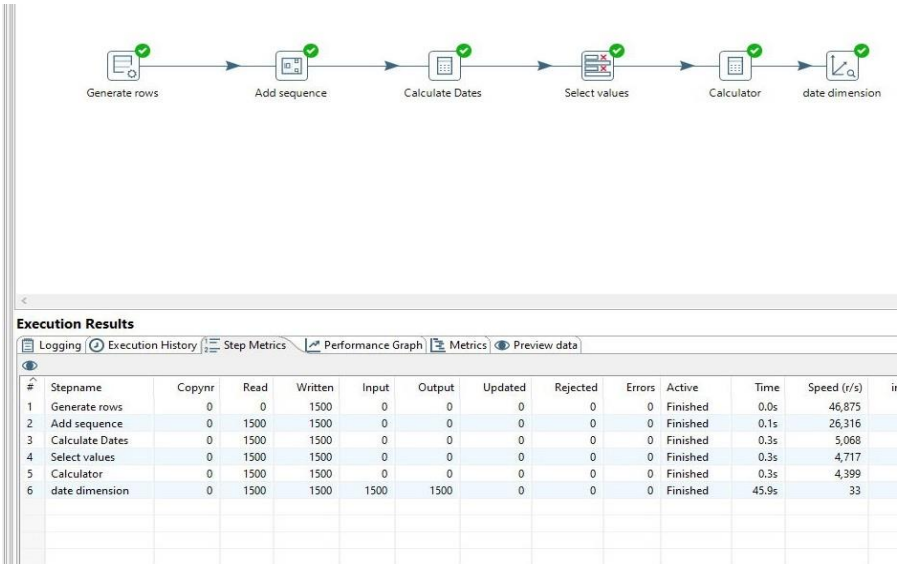
#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)
1	Crime Data	0	0	44824	44825	0	0	0	0	Finished	1.0s	43,101
2	Select values	0	44824	44824	0	0	0	0	0	Finished	1.0s	42,935
3	Sort rows	0	44824	65	0	0	0	0	0	Finished	1.2s	36,502
4	Unique rows	0	65	65	0	0	0	0	0	Finished	1.2s	53
5	perpetrator dimension	0	65	65	65	65	0	0	0	Finished	3.4s	19

Description: This is the **crime** dimension which contains original attributes from the data set including the classification code, classification description, key classification code and the offense description.



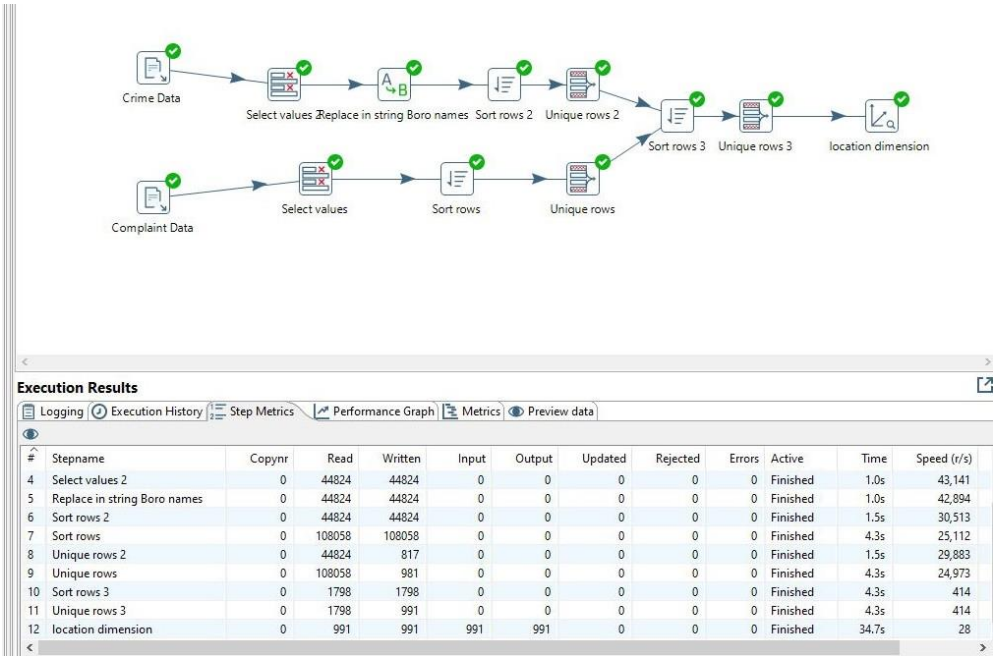
#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	inpu
1	Crime Data	0	0	44824	44825	0	0	0	0	Finished	0.9s	48,829	
2	Select values	0	44824	44824	0	0	0	0	0	Finished	0.9s	48,511	
3	Sort rows	0	44824	44824	0	0	0	0	0	Finished	1.1s	39,079	
4	Unique rows	0	44824	216	0	0	0	0	0	Finished	1.2s	38,508	
5	crime dimension	0	216	216	216	216	0	0	0	Finished	7.9s	27	

Date and Location Dimensions



#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	i
1	Generate rows	0	0	1500	0	0	0	0	0	Finished	0.0s	46,875	
2	Add sequence	0	1500	1500	0	0	0	0	0	Finished	0.1s	26,316	
3	Calculate Dates	0	1500	1500	0	0	0	0	0	Finished	0.3s	5,068	
4	Select values	0	1500	1500	0	0	0	0	0	Finished	0.3s	4,717	
5	Calculator	0	1500	1500	0	0	0	0	0	Finished	0.3s	4,399	
6	date dimension	0	1500	1500	1500	1500	0	0	0	Finished	45.9s	33	

Description: This is the date dimension for both sets of data. Because one of them starts off in 2016 and the other in 2018 we needed to make sure that they have enough dates to measure up to the last time both sets were updated. We calculated the dates so that it shows up for four years. Also, since our model wants only the date, month, and year we did another calculated step where we could see the month and year only



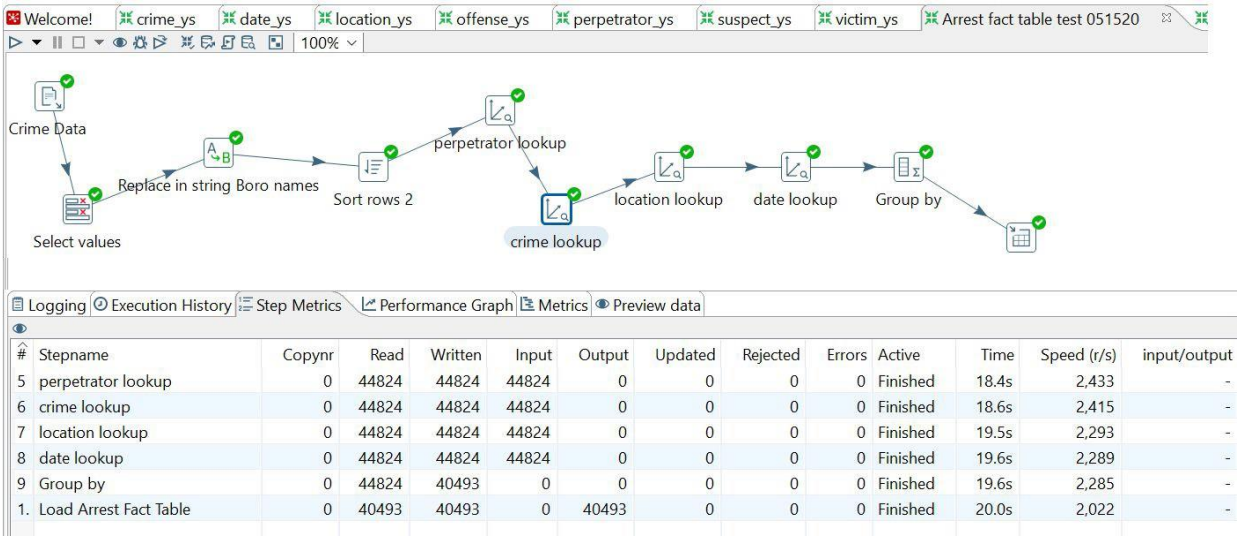
Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

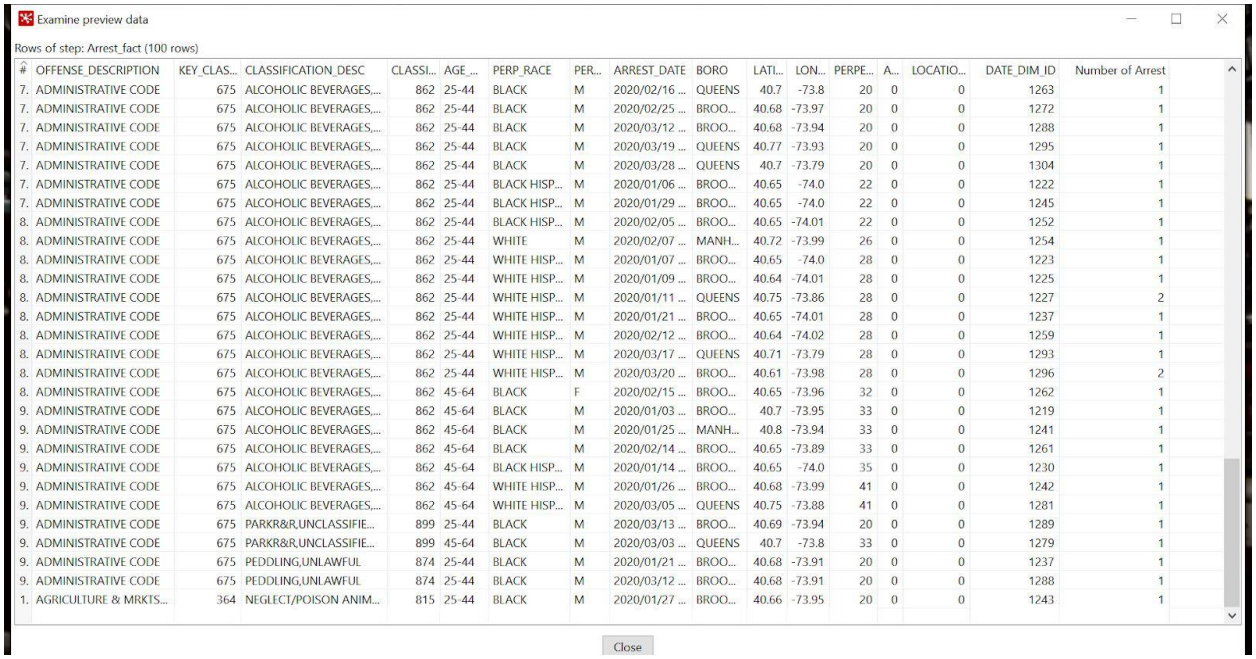
#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)
4	Select values 2	0	44824	44824	0	0	0	0	0	Finished	1.0s	43,141
5	Replace in string Boro names	0	44824	44824	0	0	0	0	0	Finished	1.0s	42,894
6	Sort rows 2	0	44824	44824	0	0	0	0	0	Finished	1.5s	30,513
7	Sort rows	0	108058	108058	0	0	0	0	0	Finished	4.3s	25,112
8	Unique rows 2	0	44824	817	0	0	0	0	0	Finished	1.5s	29,883
9	Unique rows	0	108058	981	0	0	0	0	0	Finished	4.3s	24,973
10	Sort rows 3	0	1798	1798	0	0	0	0	0	Finished	4.3s	414
11	Unique rows 3	0	1798	991	0	0	0	0	0	Finished	4.3s	414
12	location dimension	0	991	991	991	991	0	0	0	Finished	34.7s	28

Description: This is the location dimension and this is what will bind together the entire data set in addition to the date dimension. We took the CSV files from both data sets and filtered through them to get the values for location (borough name, longitude, latitude). For one of the inputs, the borough name was limited to one character so we needed to incorporate the ‘replace with string’ step to make sure that they both aligned with each other. We also rounded two decimal points for this dimension. After filtering and sorting through the records we ended up with 991 unique locations.

4. Screen shots and brief descriptions of the final schema that the business analytics tools are working with.



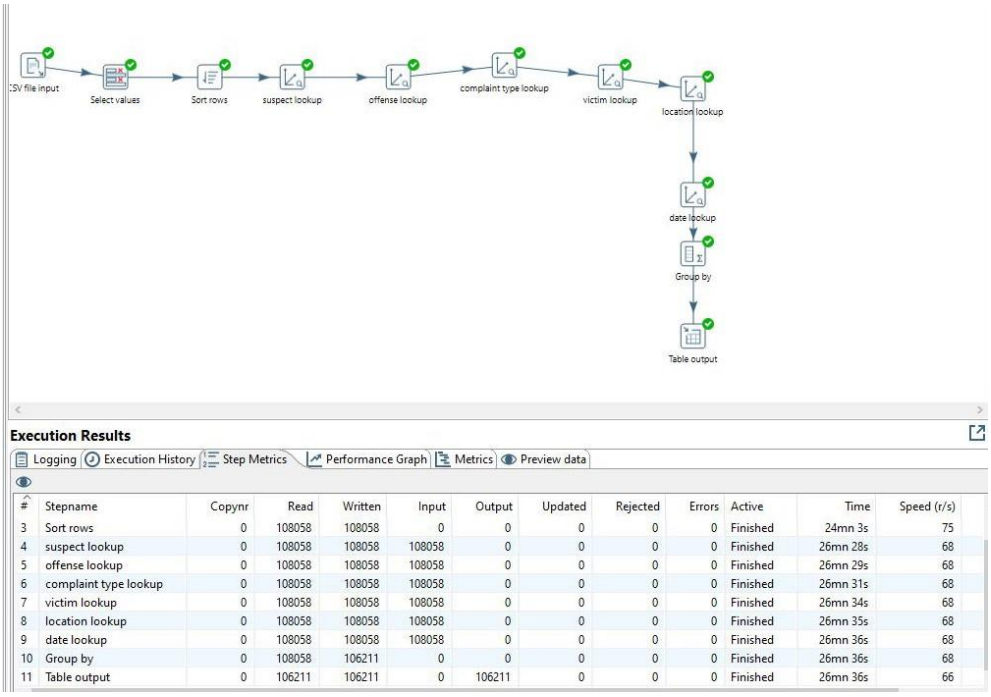
#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
5	perpetrator lookup	0	44824	44824	44824	0	0	0	0	Finished	18.4s	2,433	-
6	crime lookup	0	44824	44824	44824	0	0	0	0	Finished	18.6s	2,415	-
7	location lookup	0	44824	44824	44824	0	0	0	0	Finished	19.5s	2,293	-
8	date lookup	0	44824	44824	44824	0	0	0	0	Finished	19.6s	2,289	-
9	Group by	0	44824	40493	0	0	0	0	0	Finished	19.6s	2,285	-
1.	Load Arrest Fact Table	0	40493	40493	0	40493	0	0	0	Finished	20.0s	2,022	-



#	OFFENSE_DESCRIPTION	KEY_CLAS...	CLASSIFICATION_DESC	CLASSI...	AGE_...	PERP_RACE	PER...	ARREST_DATE	BORO	LATI...	LON...	PERPE...	A...	LOCATIO...	DATE_DIM_ID	Number of Arrest
7.	ADMINISTRATIVE CODE	675	ALCOHOLIC BEVERAGES,...	862	25-44	BLACK	M	2020/02/16 ...	QUEENS	40.7	-73.8	20	0	0	1263	1
7.	ADMINISTRATIVE CODE	675	ALCOHOLIC BEVERAGES,...	862	25-44	BLACK	M	2020/02/25 ...	BROO...	40.68	-73.97	20	0	0	1272	1
7.	ADMINISTRATIVE CODE	675	ALCOHOLIC BEVERAGES,...	862	25-44	BLACK	M	2020/03/12 ...	BROO...	40.68	-73.94	20	0	0	1288	1
7.	ADMINISTRATIVE CODE	675	ALCOHOLIC BEVERAGES,...	862	25-44	BLACK	M	2020/03/19 ...	QUEENS	40.77	-73.93	20	0	0	1295	1
7.	ADMINISTRATIVE CODE	675	ALCOHOLIC BEVERAGES,...	862	25-44	BLACK	M	2020/03/28 ...	QUEENS	40.7	-73.79	20	0	0	1304	1
7.	ADMINISTRATIVE CODE	675	ALCOHOLIC BEVERAGES,...	862	25-44	BLACK HISP...	M	2020/01/06 ...	BROO...	40.65	-74.0	22	0	0	1222	1
7.	ADMINISTRATIVE CODE	675	ALCOHOLIC BEVERAGES,...	862	25-44	BLACK HISP...	M	2020/01/29 ...	BROO...	40.65	-74.0	22	0	0	1245	1
8.	ADMINISTRATIVE CODE	675	ALCOHOLIC BEVERAGES,...	862	25-44	BLACK HISP...	M	2020/02/05 ...	BROO...	40.65	-74.01	22	0	0	1252	1
8.	ADMINISTRATIVE CODE	675	ALCOHOLIC BEVERAGES,...	862	25-44	WHITE	M	2020/02/07 ...	MANH...	40.72	-73.99	26	0	0	1254	1
8.	ADMINISTRATIVE CODE	675	ALCOHOLIC BEVERAGES,...	862	25-44	WHITE HISP...	M	2020/01/07 ...	BROO...	40.65	-74.0	28	0	0	1223	1
8.	ADMINISTRATIVE CODE	675	ALCOHOLIC BEVERAGES,...	862	25-44	WHITE HISP...	M	2020/01/09 ...	BROO...	40.64	-74.01	28	0	0	1225	1
8.	ADMINISTRATIVE CODE	675	ALCOHOLIC BEVERAGES,...	862	25-44	WHITE HISP...	M	2020/01/11 ...	QUEENS	40.75	-73.86	28	0	0	1227	2
8.	ADMINISTRATIVE CODE	675	ALCOHOLIC BEVERAGES,...	862	25-44	WHITE HISP...	M	2020/01/21 ...	BROO...	40.65	-74.01	28	0	0	1237	1
8.	ADMINISTRATIVE CODE	675	ALCOHOLIC BEVERAGES,...	862	25-44	WHITE HISP...	M	2020/02/12 ...	BROO...	40.64	-74.02	28	0	0	1259	1
8.	ADMINISTRATIVE CODE	675	ALCOHOLIC BEVERAGES,...	862	25-44	WHITE HISP...	M	2020/03/17 ...	QUEENS	40.71	-73.79	28	0	0	1293	1
8.	ADMINISTRATIVE CODE	675	ALCOHOLIC BEVERAGES,...	862	25-44	WHITE HISP...	M	2020/03/20 ...	BROO...	40.61	-73.98	28	0	0	1296	2
8.	ADMINISTRATIVE CODE	675	ALCOHOLIC BEVERAGES,...	862	45-64	BLACK	F	2020/02/15 ...	BROO...	40.65	-73.96	32	0	0	1262	1
9.	ADMINISTRATIVE CODE	675	ALCOHOLIC BEVERAGES,...	862	45-64	BLACK	M	2020/01/03 ...	BROO...	40.7	-73.95	33	0	0	1219	1
9.	ADMINISTRATIVE CODE	675	ALCOHOLIC BEVERAGES,...	862	45-64	BLACK	M	2020/01/25 ...	MANH...	40.8	-73.94	33	0	0	1241	1
9.	ADMINISTRATIVE CODE	675	ALCOHOLIC BEVERAGES,...	862	45-64	BLACK	M	2020/02/14 ...	BROO...	40.65	-73.89	33	0	0	1261	1
9.	ADMINISTRATIVE CODE	675	ALCOHOLIC BEVERAGES,...	862	45-64	BLACK HISP...	M	2020/01/14 ...	BROO...	40.65	-74.0	35	0	0	1230	1
9.	ADMINISTRATIVE CODE	675	ALCOHOLIC BEVERAGES,...	862	45-64	WHITE HISP...	M	2020/01/26 ...	BROO...	40.68	-73.99	41	0	0	1242	1
9.	ADMINISTRATIVE CODE	675	ALCOHOLIC BEVERAGES,...	862	45-64	WHITE HISP...	M	2020/03/05 ...	QUEENS	40.75	-73.88	41	0	0	1281	1
9.	ADMINISTRATIVE CODE	675	PARKR&R,UNCLASSIFIE...	899	25-44	BLACK	M	2020/03/13 ...	BROO...	40.69	-73.94	20	0	0	1289	1
9.	ADMINISTRATIVE CODE	675	PARKR&R,UNCLASSIFIE...	899	45-64	BLACK	M	2020/03/03 ...	QUEENS	40.7	-73.8	33	0	0	1279	1
9.	ADMINISTRATIVE CODE	675	PEDDLING,UNLAWFUL	874	25-44	BLACK	M	2020/01/21 ...	BROO...	40.68	-73.91	20	0	0	1237	1
9.	ADMINISTRATIVE CODE	675	PEDDLING,UNLAWFUL	874	25-44	BLACK	M	2020/03/12 ...	BROO...	40.68	-73.91	20	0	0	1288	1
1.	AGRICULTURE & MRKTS...	364	NEGLECT/POISON ANIM...	815	25-44	BLACK	M	2020/01/27 ...	BROO...	40.66	-73.95	20	0	0	1243	1

Arrest Fact Table: Reads a CSV file from NYPD_Arrest_Data_Year_to_Date (2020 from January 1 to March 31, 2020) data file. The total number of rows turns out to be 44824. "Group By" on Pentaho enables a column to be duplicated but counted by the number of N. It will represent "1" per perpetrator_dim_number, so that it can be calculated for KPI (arrests per day, arrests per age group and etc).

Complaint Fact Table:



Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

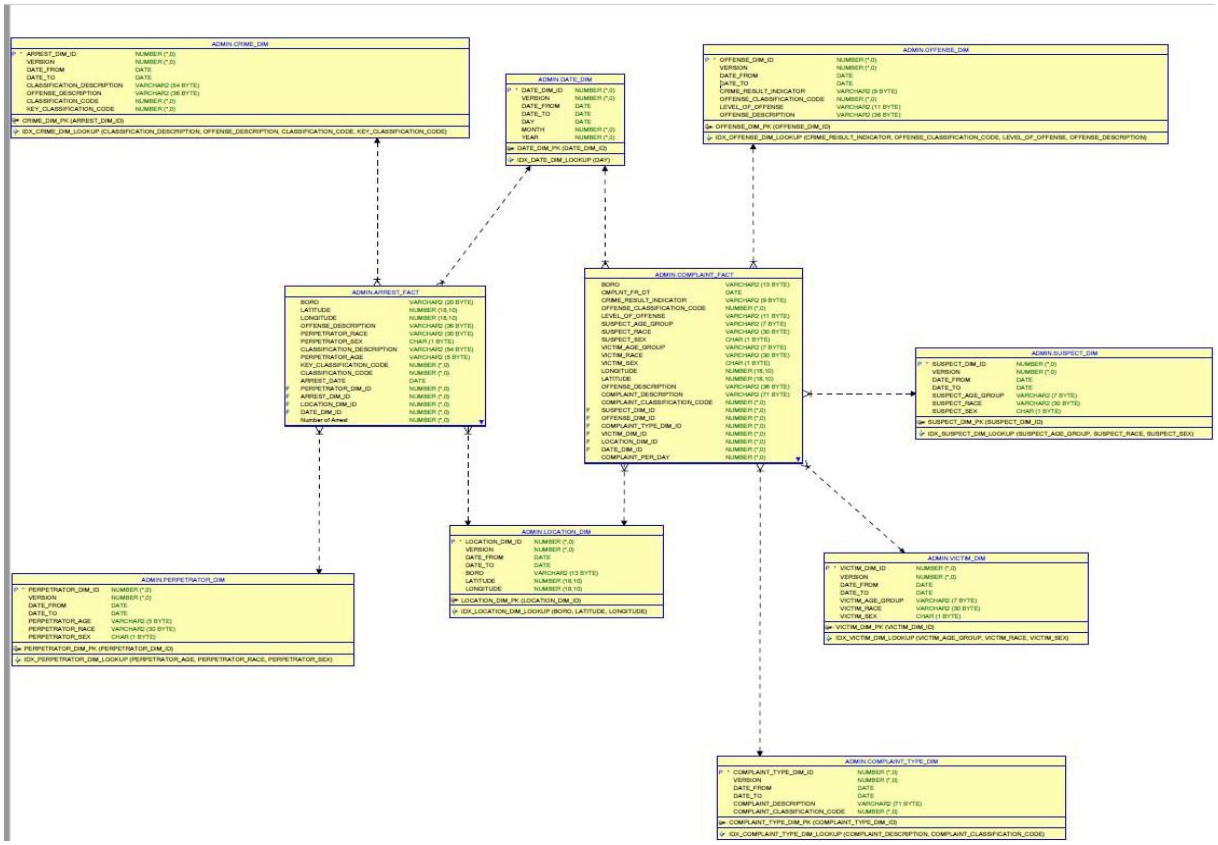
#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)
3	Sort rows	0	108058	108058	0	0	0	0	0	Finished	24mn 3s	75
4	suspect lookup	0	108058	108058	108058	0	0	0	0	Finished	26mn 28s	68
5	offense lookup	0	108058	108058	108058	0	0	0	0	Finished	26mn 29s	68
6	complaint type lookup	0	108058	108058	108058	0	0	0	0	Finished	26mn 31s	68
7	victim lookup	0	108058	108058	108058	0	0	0	0	Finished	26mn 34s	68
8	location lookup	0	108058	108058	108058	0	0	0	0	Finished	26mn 35s	68
9	date lookup	0	108058	108058	108058	0	0	0	0	Finished	26mn 36s	68
10	Group by	0	108058	106211	0	0	0	0	0	Finished	26mn 36s	68
11	Table output	0	106211	106211	0	106211	0	0	0	Finished	26mn 36s	66

Rows of step: COMPLAINT_FACT (100 rows)

	COMPLAINT_CLASSIFICATION_CODE	SUSPECT_DIM_ID	OFFENSE_DIM_ID	COMPLAINT_TYPE_DIM_ID	VICTIM_DIM_ID	LOCATION_DIM_ID	DATE_DIM_ID	COMPLAINT_PER_DAY
	<null>	48.0	42.0	0.0	38.0	0.0	0.0	1.0
	<null>	12.0	42.0	0.0	79.0	0.0	0.0	1.0
	<null>	36.0	42.0	0.0	52.0	0.0	0.0	1.0
	<null>	56.0	42.0	0.0	63.0	0.0	0.0	1.0
	<null>	0.0	42.0	0.0	28.0	0.0	1040.0	1.0
	<null>	36.0	42.0	0.0	28.0	0.0	1179.0	1.0
	<null>	56.0	42.0	0.0	53.0	0.0	1188.0	1.0
	<null>	45.0	42.0	0.0	54.0	0.0	1193.0	1.0
	<null>	48.0	42.0	0.0	43.0	0.0	1216.0	1.0
	<null>	35.0	42.0	0.0	28.0	0.0	1217.0	1.0
	<null>	36.0	42.0	0.0	28.0	0.0	1217.0	1.0
	729.0	39.0	53.0	106.0	103.0	0.0	1217.0	1.0
	661.0	113.0	95.0	212.0	45.0	0.0	1219.0	1.0
	<null>	9.0	42.0	0.0	71.0	0.0	1222.0	1.0
	175.0	66.0	71.0	321.0	24.0	0.0	1224.0	1.0
	<null>	36.0	42.0	0.0	28.0	0.0	1225.0	1.0
	<null>	12.0	42.0	0.0	28.0	0.0	1227.0	1.0
	<null>	24.0	42.0	0.0	72.0	0.0	1228.0	1.0
	<null>	36.0	42.0	0.0	28.0	0.0	1228.0	1.0
	<null>	0.0	42.0	0.0	46.0	0.0	1229.0	1.0
	<null>	36.0	42.0	0.0	28.0	0.0	1229.0	1.0
	<null>	36.0	42.0	0.0	64.0	0.0	1229.0	1.0
	101.0	56.0	80.0	19.0	37.0	0.0	1233.0	1.0
	<null>	47.0	42.0	0.0	11.0	0.0	1235.0	1.0
	<null>	36.0	42.0	0.0	45.0	0.0	1236.0	1.0
	<null>	36.0	42.0	0.0	46.0	0.0	1236.0	1.0
	<null>	36.0	42.0	0.0	57.0	0.0	1236.0	1.0
	<null>	0.0	42.0	0.0	28.0	0.0	1237.0	1.0
	<null>	15.0	42.0	0.0	30.0	0.0	1239.0	1.0
	<null>	55.0	42.0	0.0	64.0	0.0	1239.0	1.0
	<null>	104.0	42.0	0.0	28.0	0.0	1239.0	1.0
	<null>	36.0	42.0	0.0	28.0	0.0	1241.0	1.0

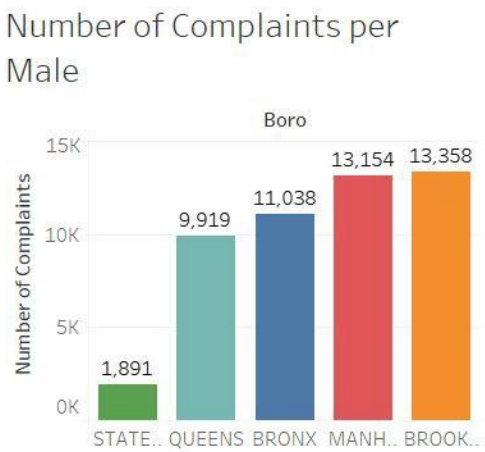
Close

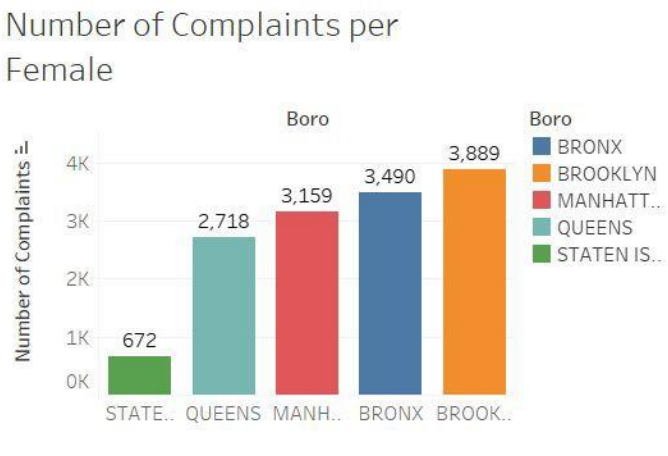
Description: Reads a CSV file from NYPD Complaint Data Current. The total number of rows turns out to be 108,058 because of its quarterly update. Similarly to the previous fact table, Group By was used to group all the rows so that we could do our aggregations..



Description: This is the final schema that we reverse engineered in Oracle SQL Developer. The relationships were made by altering the table to create the foreign key relationships.

5. Screen shots and descriptions of the analytics (at least 3) on the dashboard application developed based on the data warehouse data.





Number of Male Arrests per Age Group

Perpetrator Age	
<18	1,474
18-24	6,419
25-44	18,179
45-64	6,864
65+	464

Number of Female Arrests per Age Group

Perpetrator Age	
<18	341
18-24	1,471
25-44	3,970
45-64	1,205
65+	73

Description: As most of our KPIs were based on qualitative values we decided to stick with bar graphs and basic charts to reflect the number that we were looking for in our KPI's. These KPIs tell us a lot about the data set. The first being that the complaints per male are larger than that of females. We can also tell that both female and male complaints take place in the borough of Brooklyn more than any other borough. When looking at the arrests, we can see that there are more male arrests than female arrests which makes sense as they received more complaints than females. Looking from the basic count of arrests we can determine that the age group most likely to be arrested is from 25-44 and from this we can infer the same for the number or complaints per age group. Below are the KPIs we looked at and the code for each.

Number of Complaints per Male

```
SELECT complaint_description, suspect_sex, boro
FROM complaint_fact
WHERE complaint_description  IS NOT NULL AND boro IS NOT NULL AND suspect_sex =
'M';
```

Number of complaints per female

```
SELECT complaint_description, suspect_sex, boro
FROM complaint_fact
WHERE complaint_description IS NOT NULL AND boro IS NOT NULL AND suspect_sex =
'F';
```

Number of arrests per age group and are female

```
SELECT offense_description, perpetrator_age, boro, perpetrator_sex
FROM arrest_fact
WHERE offense_description  IS NOT NULL AND boro is NOT NULL AND perpetrator_sex
= 'F'
```

Number of arrests per age group and are male

```
SELECT offense_description, perpetrator_age, boro, perpetrator_sex
FROM arrest_fact
WHERE offense_description  IS NOT NULL AND boro is NOT NULL AND perpetrator_sex
= 'M'
```

6. A narrative conclusion section that describes:

After concluding our project there were many challenges that came along the way. The first one that we came across was loading all of the data from each dimension. After figuring out which schema to finally put everything on, we decided to work on the fact tables. It was difficult as we kept getting errors everytime we ran the fact tables which prevented us from getting to the next step. The errors mostly came from our Oracle cloud schema since it does not allow multiple users at the sametime when using the free version. However, we were able to eventually figure out that by separating the work on localhost (H2 database) and sharing the kettle files we would be able to implement working on dimensions or fact tables into the Oracle database so that everyone could stay and catch up to the same step of the project. The new system shows users the number of complaints or arrests per day across New York City, and can also be observed by their location /area. It will contribute to the police and fire department in looking at

the big picture to increase their operational efficiency in which areas they should place focus their efforts on and place more of their staff on.