Unsanitary Conditions in New York City

Rachel Abreu (rachel.abreu@baruchmail.cuny.edu)
Mei Chang (mei.chang@baruchmail.cuny.edu)
Shakir Mahabirsingh(shakir.mahabirsingh@baruchmail.cuny.edu)
Hoichun Ip (hoichun.ip@baruchmail.cuny.edu)

Description

In our project, our group hopes to address the unsanitary conditions in New York City. As New Yorkers ourselves, we see these conditions on a daily basis. Often, complaints made to 311 are not always acted upon. In efforts to create change, we hope that our database management system is able to be effective. Not only would this database management system help us implement change in our communities, but it would also help the future use this system as means of maintaining cleanliness. The 311 complaints we will be focusing on are **dirty conditions**, **rodents**, and **sewer**.

While these 311 complaints can be seen visibly on the streets, we can easily notice it on the subway as well. With the creation of our database management system, it should also facilitate the way in which subways are being cleaned. Subways are also victim to rodent infestations, dirty conditions, and overall uncleanliness. The Key Performance Indicators we will be using to address the problem of cleanliness in New York City are Dirty Conditions(35887), Rodent(35075), and Sewer(32529).

KPIs:

Complaint Types -

1. Total count of complaints
    a. Sum all complaints
2. % of total for each complaint type
    a. Complaint Type Total / Total Complaints
3. Frequency of each complaint type each month
    a. Complaint Type Total / Days in the Month

Location -

1. Total complaint status in each borough
    a. Sum all statuses per borough
2. Total of location type in each borough
    a. Sum all location types per borough
3. Frequency of complaints per month in each neighborhood
    a. (Complaint Type Total / Days in the Month) per neighborhood

Time -

1. Peak Season for complaints
    a. Sum complaints for each season
2. Avg complaints per month
    a. Sum all complaints / Days in the Month
3. Time until the resolved complaint
    a. Time Complaint Started - Time Complaint Resolved

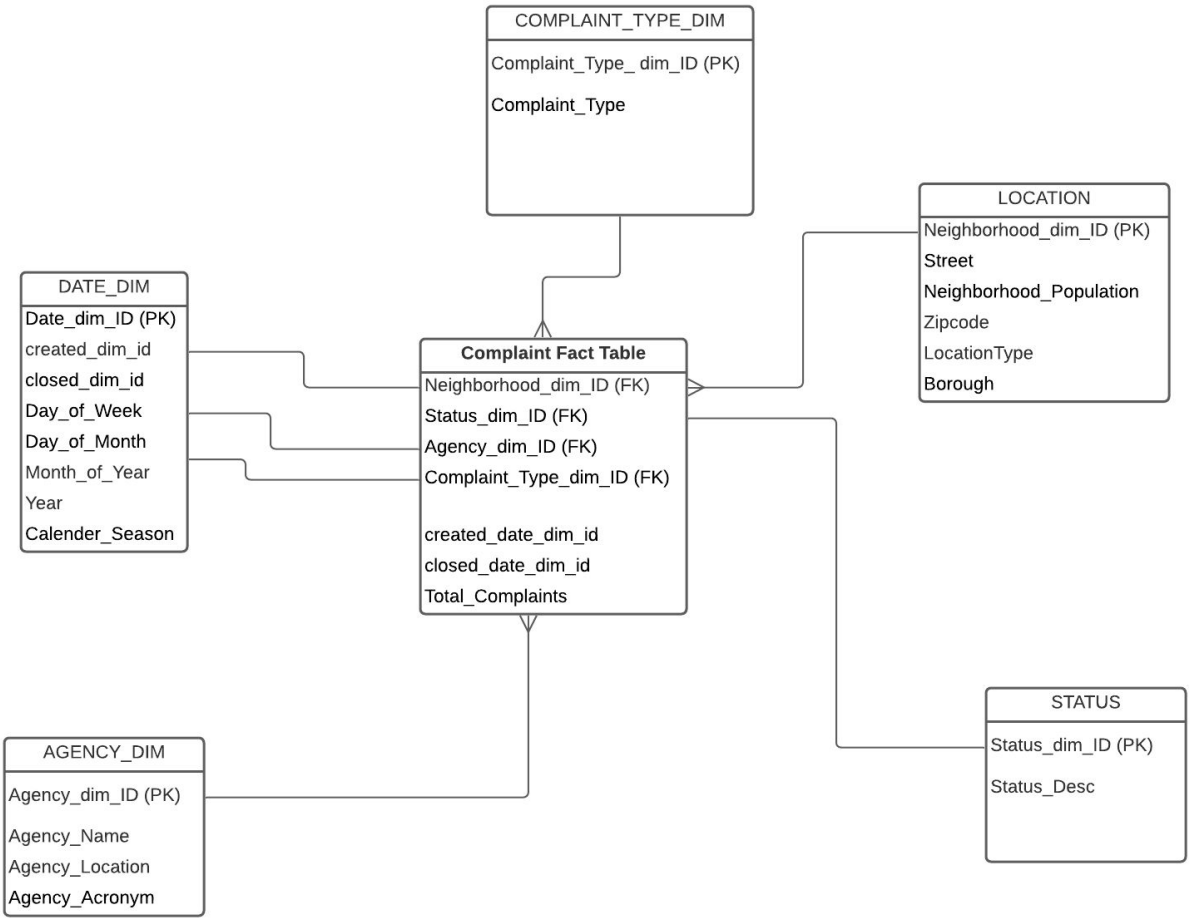
Meeting Log:

Meeting Times: various times done through GroupMe app

Attendees: Rachel, Mei, Shakir, Hoichun

Topic Discussed: Goals and ideas for the project. What we wanted to find out/show.

Unsanitary Conditions Dimensional Model



Since we are loading and using this data in a specific time frame, we will be using the ***Periodic Snapshot Grain.***

Meeting Log:

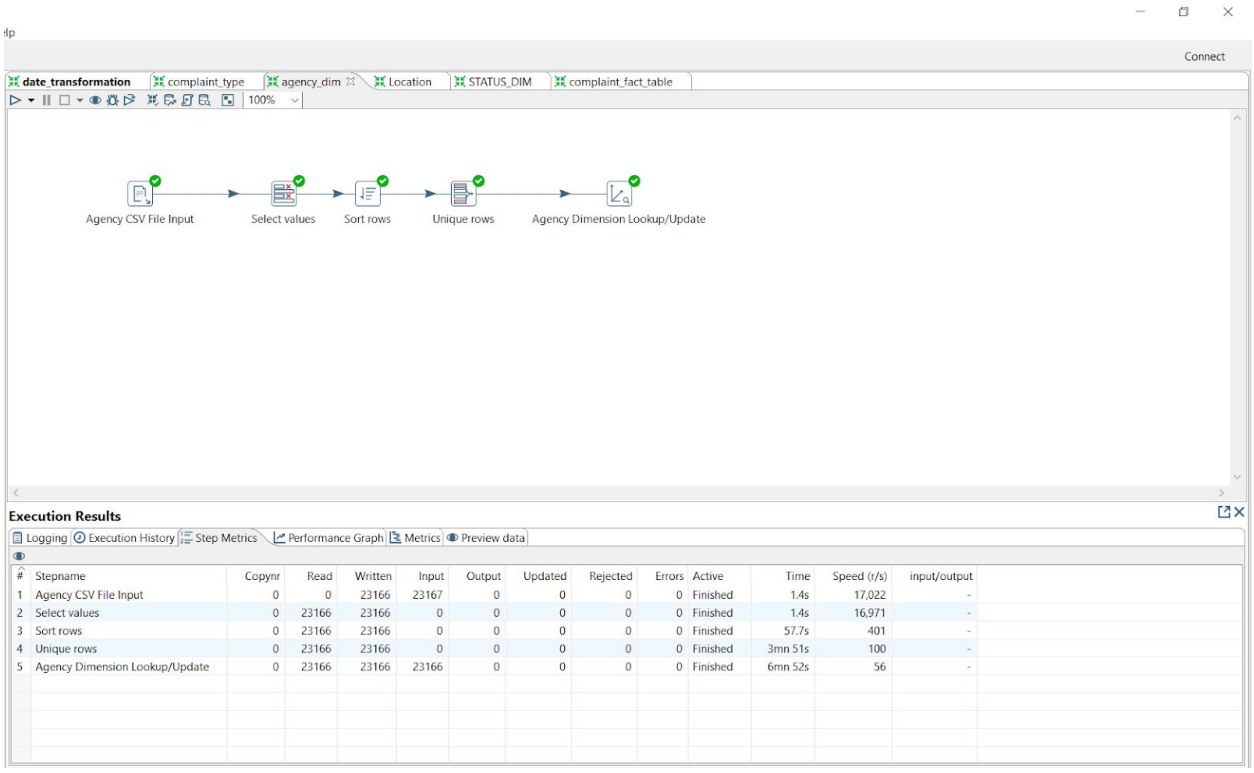
Meeting Times: various times done through GroupMe app

Attendees: Rachel, Mei, Shakir, Hoichun

Topic Discussed: Layout of our dimensional tables. Various columns, keys, attributes would populate each table. What FKs would be used for the central fact table.

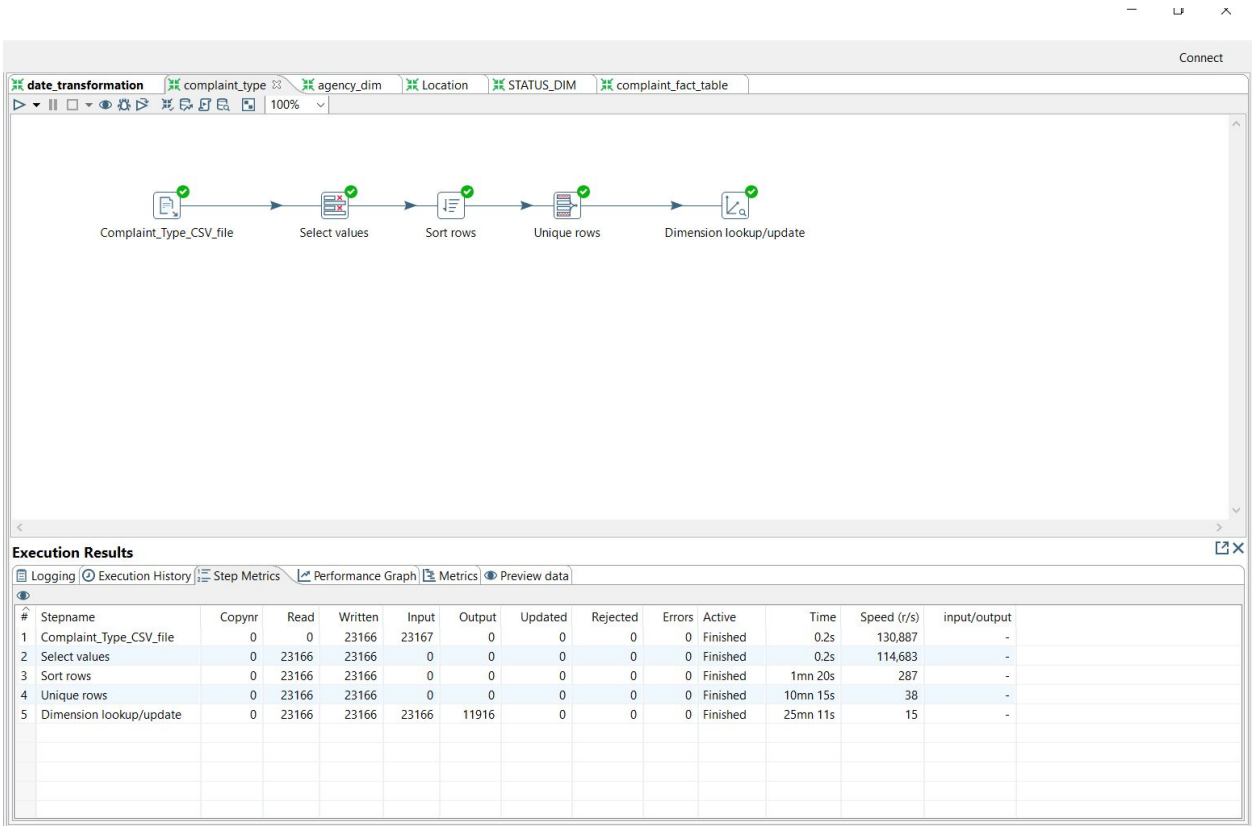
ETL Transformations

1. Agency Dimension



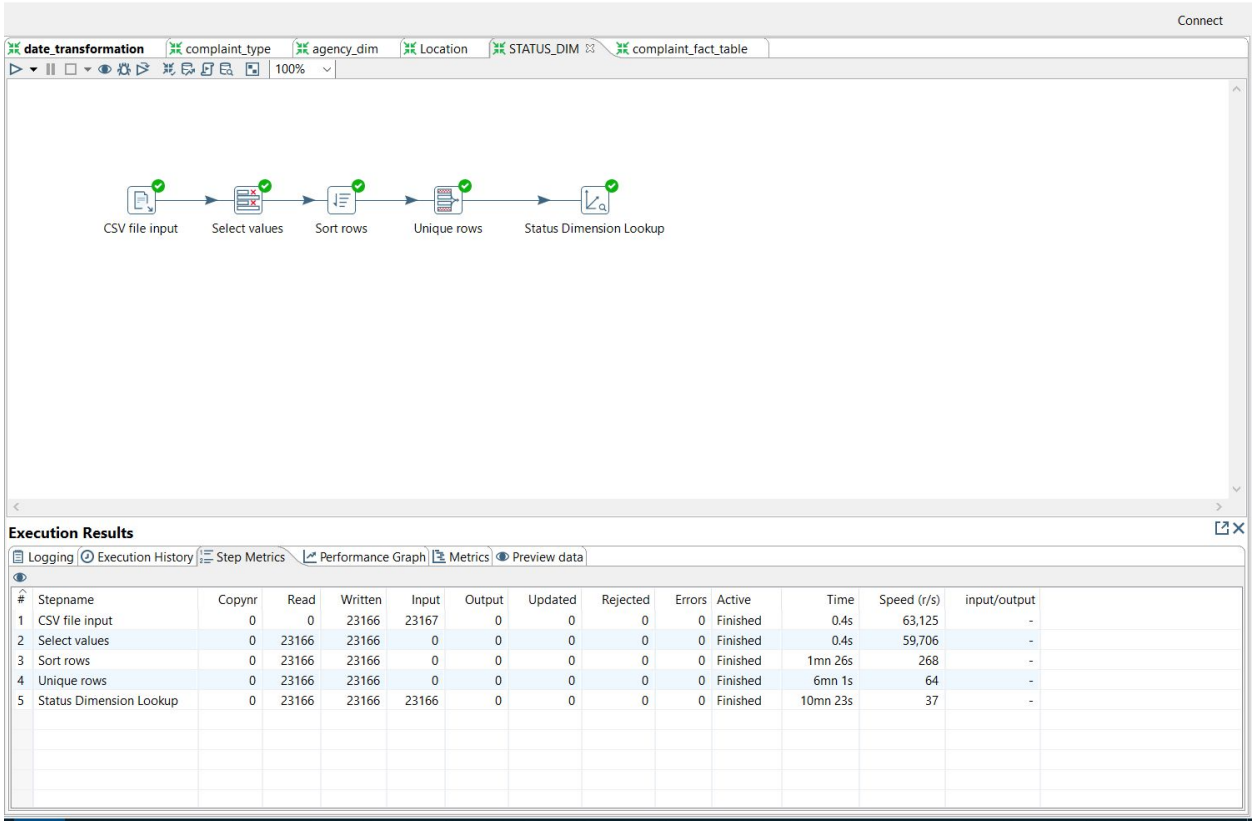
#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Agency CSV File Input	0	0	23166	23167	0	0	0	0	Finished	1.4s	17,022	-
2	Select values	0	23166	23166	0	0	0	0	0	Finished	1.4s	16,971	-
3	Sort rows	0	23166	23166	0	0	0	0	0	Finished	57.7s	401	-
4	Unique rows	0	23166	23166	0	0	0	0	0	Finished	3mn 51s	100	-
5	Agency Dimension Lookup/Update	0	23166	23166	23166	0	0	0	0	Finished	6mn 52s	56	-

2. Complaint Type Dimension



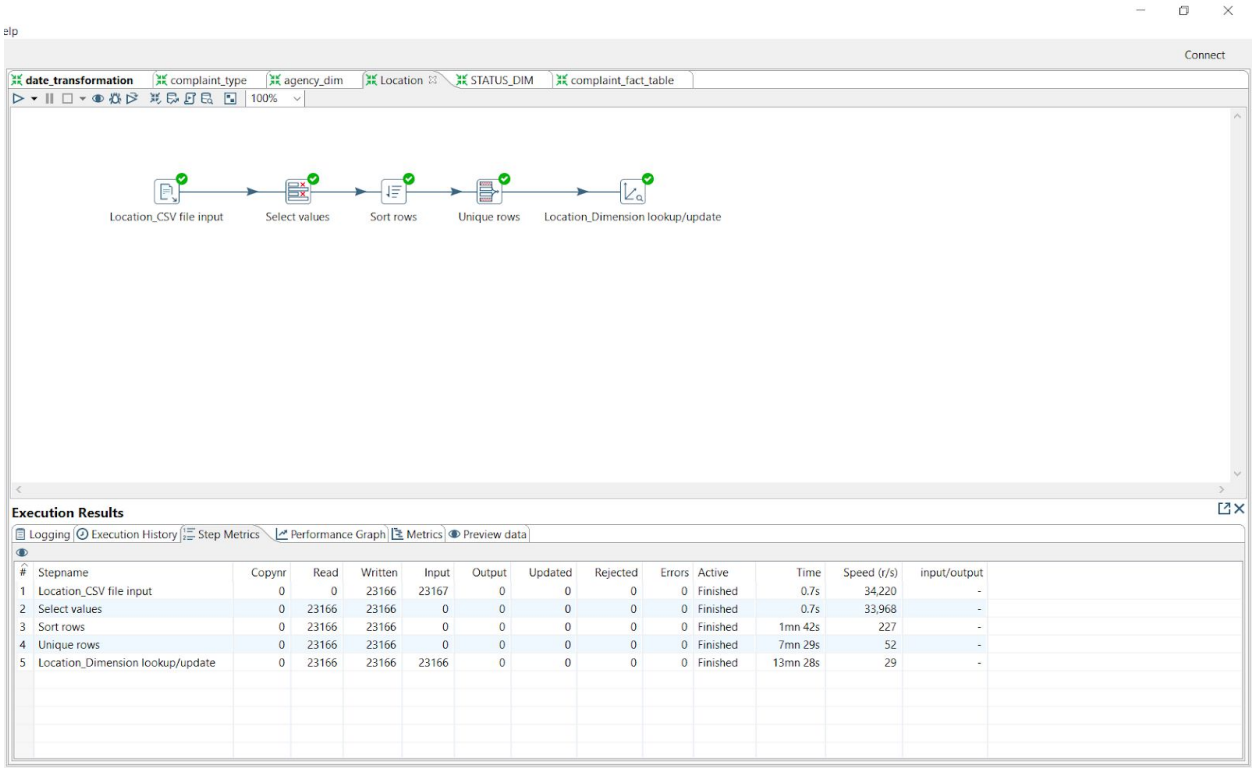
#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Complaint_Type_CSV_file	0	0	23166	23167	0	0	0	0	Finished	0.2s	130,887	-
2	Select values	0	23166	23166	0	0	0	0	0	Finished	0.2s	114,683	-
3	Sort rows	0	23166	23166	0	0	0	0	0	Finished	1mn 20s	287	-
4	Unique rows	0	23166	23166	0	0	0	0	0	Finished	10mn 15s	38	-
5	Dimension lookup/update	0	23166	23166	23166	11916	0	0	0	Finished	25mn 11s	15	-

3. Status Dimension



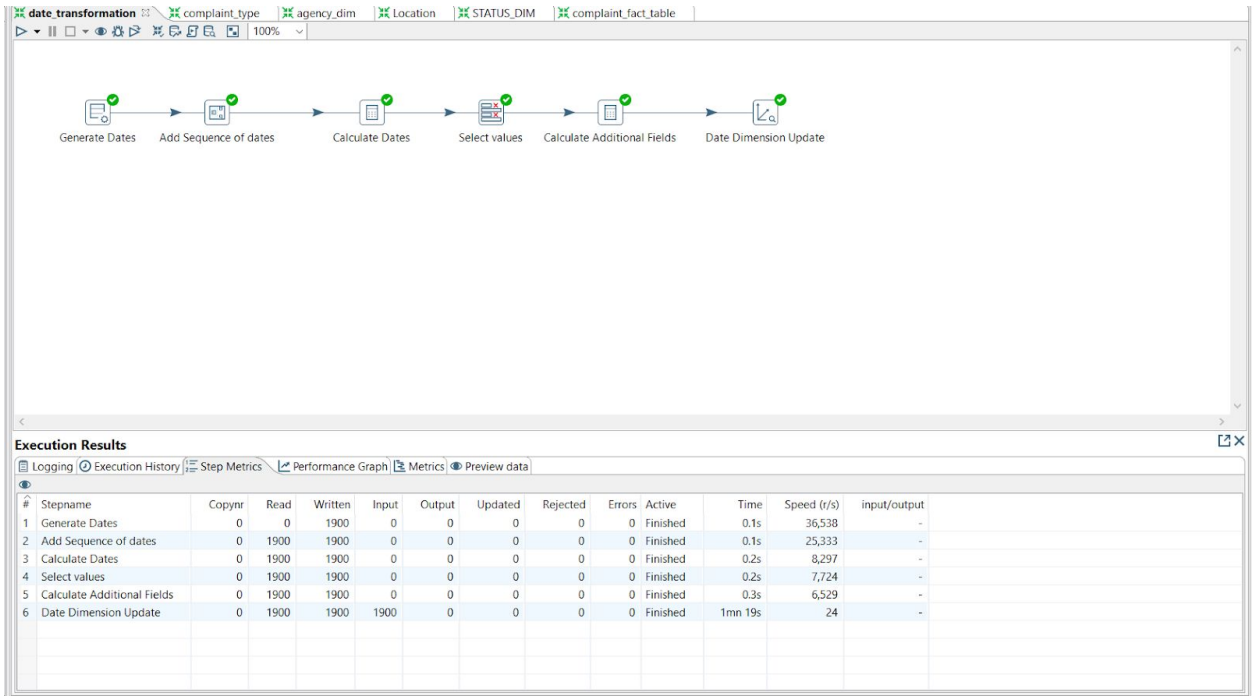
#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	CSV file input	0	0	23166	23167	0	0	0	0	Finished	0.4s	63,125	-
2	Select values	0	23166	23166	0	0	0	0	0	Finished	0.4s	59,706	-
3	Sort rows	0	23166	23166	0	0	0	0	0	Finished	1mn 26s	268	-
4	Unique rows	0	23166	23166	0	0	0	0	0	Finished	6mn 1s	64	-
5	Status Dimension Lookup	0	23166	23166	23166	0	0	0	0	Finished	10mn 23s	37	-

4. Location Dimension



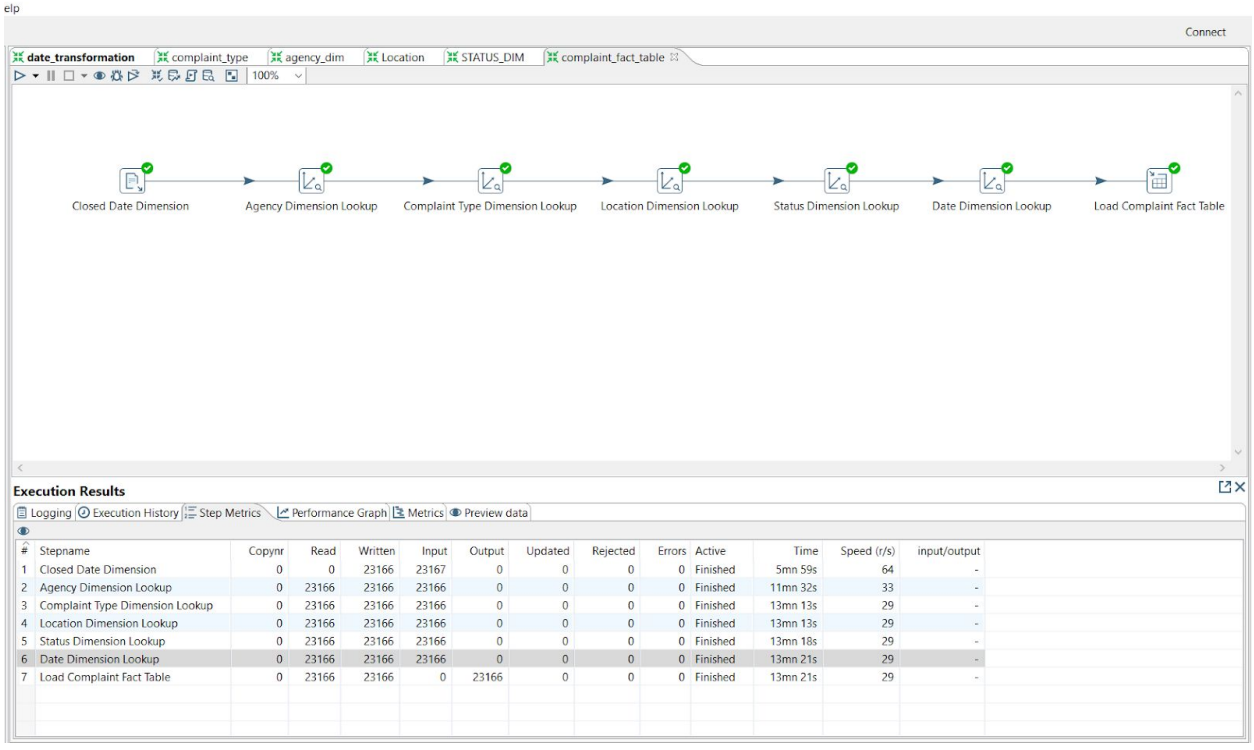
#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Location_CSV file input	0	0	23166	23167	0	0	0	0	Finished	0.7s	34,220	-
2	Select values	0	23166	23166	0	0	0	0	0	Finished	0.7s	33,968	-
3	Sort rows	0	23166	23166	0	0	0	0	0	Finished	1mn 42s	227	-
4	Unique rows	0	23166	23166	0	0	0	0	0	Finished	7mn 29s	52	-
5	Location_Dimension lookup/update	0	23166	23166	23166	0	0	0	0	Finished	13mn 28s	29	-

5. Date Dimension



#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Generate Dates	0	0	1900	0	0	0	0	0	Finished	0.1s	36,538	-
2	Add Sequence of dates	0	1900	1900	0	0	0	0	0	Finished	0.1s	25,333	-
3	Calculate Dates	0	1900	1900	0	0	0	0	0	Finished	0.2s	8,297	-
4	Select values	0	1900	1900	0	0	0	0	0	Finished	0.2s	7,724	-
5	Calculate Additional Fields	0	1900	1900	0	0	0	0	0	Finished	0.3s	6,529	-
6	Date Dimension Update	0	1900	1900	1900	0	0	0	0	Finished	1mn 19s	24	-

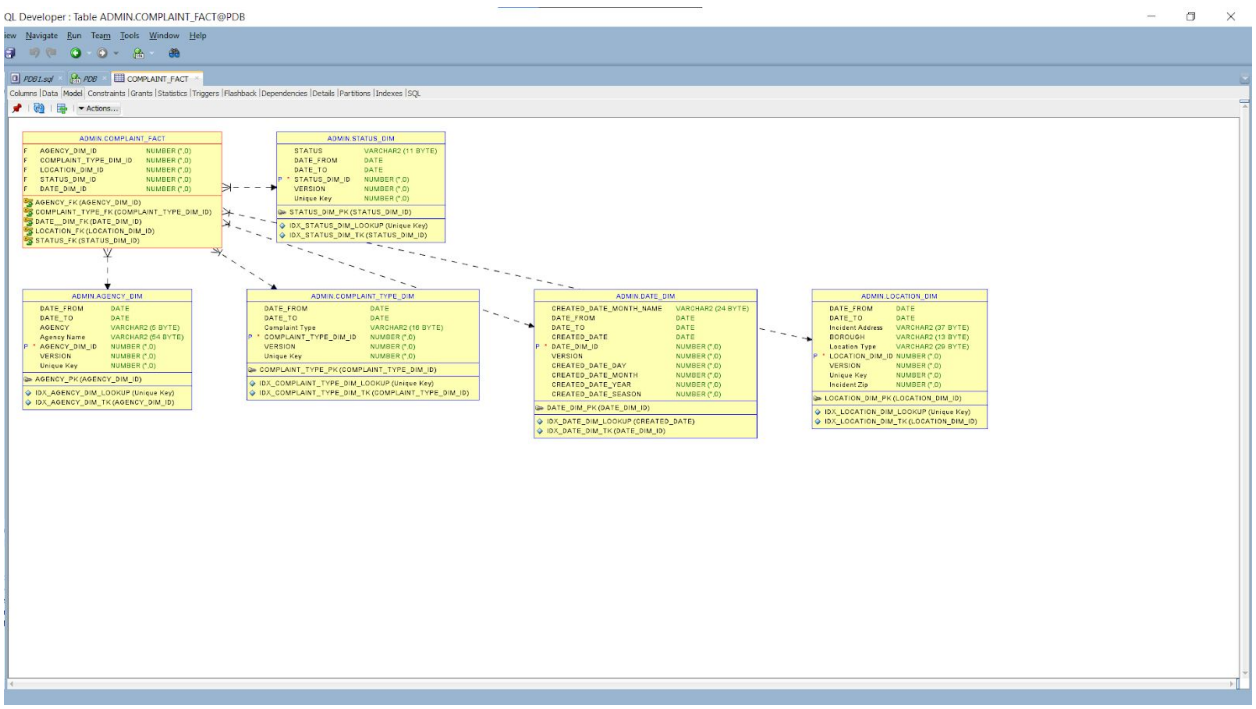
6. Complaint Fact Table



#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Closed Date Dimension	0	0	23166	23167	0	0	0	0	Finished	5mn 59s	64	-
2	Agency Dimension Lookup	0	23166	23166	23166	0	0	0	0	Finished	11mn 32s	33	-
3	Complaint Type Dimension Lookup	0	23166	23166	23166	0	0	0	0	Finished	13mn 13s	29	-
4	Location Dimension Lookup	0	23166	23166	23166	0	0	0	0	Finished	13mn 13s	29	-
5	Status Dimension Lookup	0	23166	23166	23166	0	0	0	0	Finished	13mn 18s	29	-
6	Date Dimension Lookup	0	23166	23166	23166	0	0	0	0	Finished	13mn 21s	29	-
7	Load Complaint Fact Table	0	23166	23166	0	23166	0	0	0	Finished	13mn 21s	29	-

Adding Keys

```
/* Adding the PKs in the respective Dims */
ALTER TABLE complaint_type_dim
ADD CONSTRAINT complaint_type_pk PRIMARY KEY (complaint_type_dim_id);

ALTER TABLE date_dim
ADD CONSTRAINT date_dim_pk PRIMARY KEY (date_dim_id);

ALTER TABLE location_dim
ADD CONSTRAINT location_dim_pk PRIMARY KEY (location_dim_id);

ALTER TABLE status_dim
ADD CONSTRAINT status_dim_pk PRIMARY KEY (status_dim_id);


/* Adding the FKs to the Sales Fact Table */
ALTER TABLE complaint_fact
    ADD CONSTRAINT agency_fk FOREIGN KEY (agency_dim_id)
    REFERENCES agency_dim(agency_dim_id);

ALTER TABLE complaint_fact
    ADD CONSTRAINT complaint_type_fk FOREIGN KEY (complaint_type_dim_id)
    REFERENCES complaint_type_dim(complaint_type_dim_id);

ALTER TABLE complaint_fact
    ADD CONSTRAINT date__dim_fk FOREIGN KEY (date_dim_id)
    REFERENCES date_dim(date_dim_id);

ALTER TABLE complaint_fact
    ADD CONSTRAINT location_fk FOREIGN KEY (location_dim_id)
    REFERENCES location_dim(location_dim_id);

ALTER TABLE complaint_fact
    ADD CONSTRAINT status_fk FOREIGN KEY (status_dim_id)
    REFERENCES status_dim(status_dim_id);
```



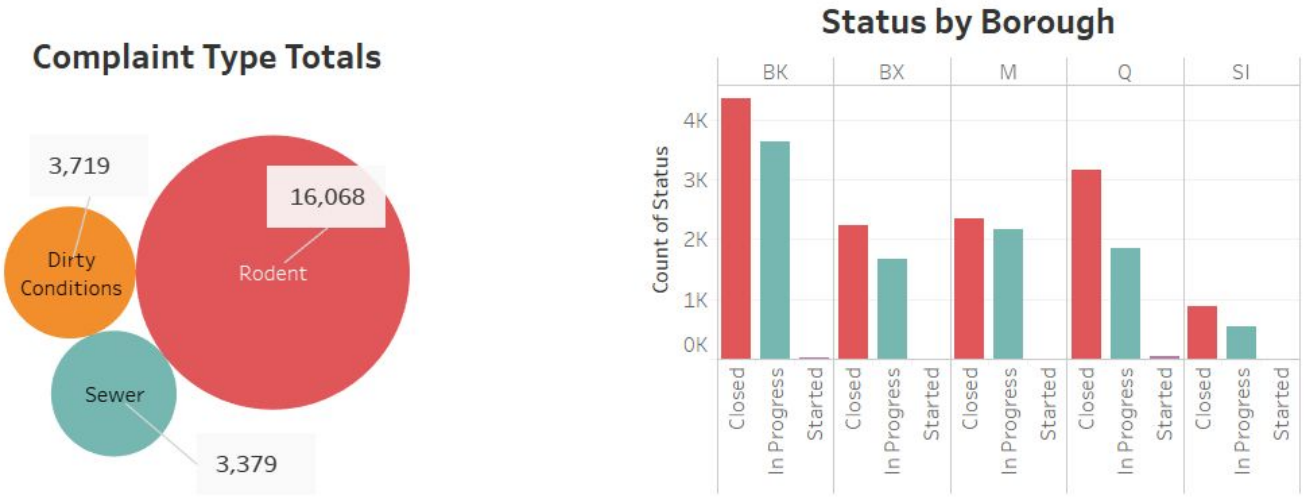
Meeting Log:

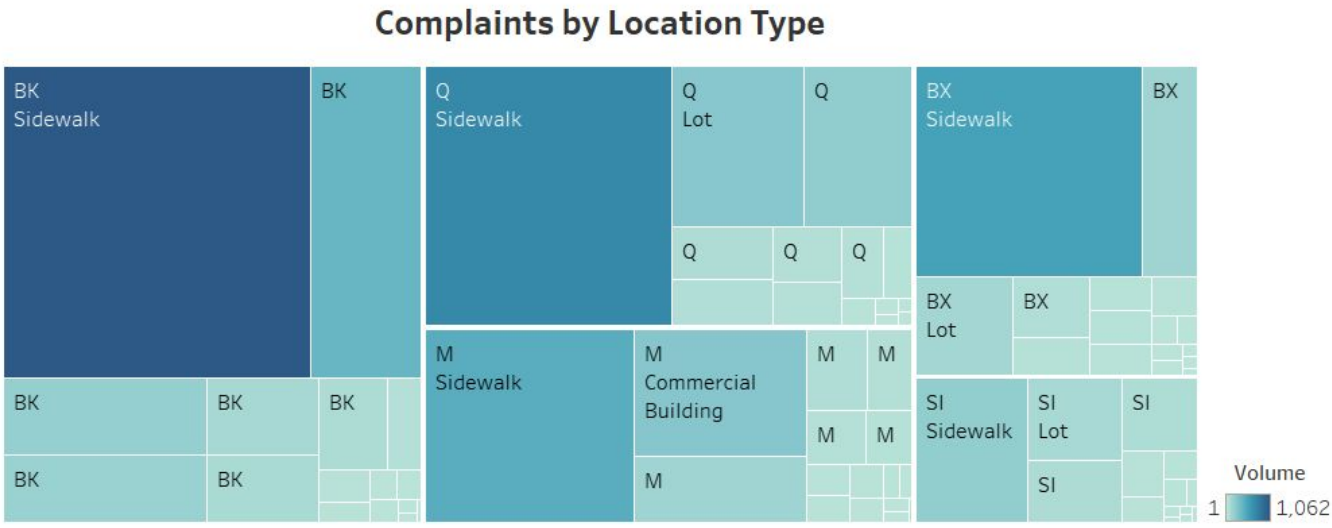
Meeting Times: various times done through GroupMe app

Attendees: Rachel, Mei, Shakir, Hoichun

Topic Discussed: How to set up the proper ETL workflow. Which tools/SQL to use to populate the database and create FKs.

Tableau Dashboards





Meeting Log:

Meeting Times: various times done through GroupMe app

Attendees: Rachel, Mei, Shakir, Hoichun

Topic Discussed: What KPIs could be shown and which visualizations to use. We wanted different graphs for respective KPIs.

a) the software and database tools the group used to coordinate and manage the project as well as carry out the programming tasks (list of bullet points with software or service and one sentence of what it was used for)

- MS Excel: used to create the CSV file.
- Pentaho(Spoon): used for ETL Transformations.
- Tableau: used to create the dashboard to visualize data.
- Oracle: add the value keys.

b) the group's experience with the project (which steps were the most difficult? Which were the easiest? what did you learn that you did not imagine you would have? if you had to do it all over again, what would you have done differently?)

The most challenging step of the project was creating the dimensions table. The way the excel spreadsheet was structured, that's hard to find the attributes of entities for different dimension tables.

The easiest step of the project was choosing the key performance indicators. Also, creating the visualizations.

What we learned that we did not imagine we would have was how to use a periodic snapshot grain.

If we were to do this project all over again, we would have added more interactive visualizations that allow the user to change filters as they please.

c) If the proposed benefits can be realized by the new system, we could centralize the recourse to implement the problems of dirty conditions, rodents, and sewer in our community. It is our main goal to change the unsanitary condition and maintain cleanliness in New York City.

d) any final comments and conclusions

None.

References

Cloud Storage for Work and Home. (n.d.). Retrieved September 08, 2020, from https://www.google.com/intl/en_in/drive/

Data.cityofnewyork.us. 2020. [online] Available at: <https://data.cityofnewyork.us/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9> [Accessed 27 September 2020].

GroupMe [Online software, App]. (Jared Hecht, Steve Martocci). Retrieved September 08, 2020, from groupme.com

Microsoft Excel [Computer software]. (n.d.).

ORACLE [Computer software]. (n.d.). Retrieved from https://www.oracle.com/index.html

Roldán, M. C. (n.d.). Pentaho Data Integration: Beginner's guide: Get up and running with the Pentaho Data Integration tool using this hands-on, easy-to-read guide [Computer software].

Tableau [Computer software]. (n.d.). Retrieved from https://www.tableau.com/