DBMS PM5

SafeStay is a comprehensive analysis application that determines the safety index for various tenants and buyers of property in Boston.

Data Warehouse:

We have combined the data aggregated from the following links with our custom designed databases.

Data sources that we have employed are as follows:

1) This dataset contains the locations of Hospitals within the city:
https://data.boston.gov/dataset/hospital-locations/resource/6222085d-ee88-45c6-ae40-0c7464620d64

2) This dataset contains property, or parcel, ownership together with value information, which ensures fair assessment of Boston taxable and non-taxable property of all types and classifications.
https://data.boston.gov/dataset/property-assessment

3) This dataset is our primary dataset that contains the incidents that have occurred and specific details about the incidents:
https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system

4) This dataset contains the information about various police stations in Boston.
http://bostonopendata-boston.opendata.arcgis.com/datasets/e5a0066d38ac4e2abbc7918197a4f6af_6

5) This dataset contains information about various schools in Boston.
https://data.boston.gov/dataset/public-schools/resource/16c8f02c-e32a-423a-8e27-0080117e6845

Combining these data sources:
To compute the safety index of a particular area, we had to accumulate the information gathered from all these tables.

1) We have studied the safety index of an area (more specifically, a street) by mapping the number of incidents that have occurred on the street to the number of police stations.

2) We have studied the safety index of an area (more specifically, a street) by mapping the number of incidents that have occurred on the street to the number of hospitals.

3) We have studied the property rates for each street based on its safety by mapping the number of incidents that have occurred on the street to the property rates on that street.

4) We have studied our dataset to map the number of incidents to every month.

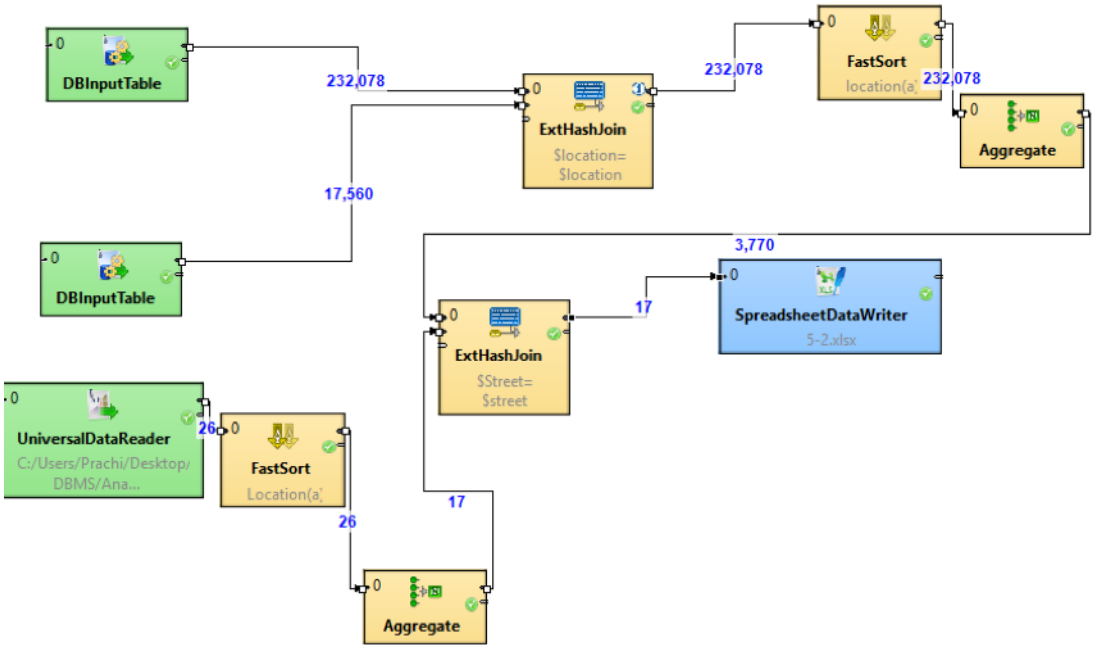
5) We have studied the number of schools for each street based on its safety by mapping the number of incidents that have occurred on the street to the number of schools on that street.

Hypothesis:

1) If the number of incidents in an area are more, then the number of police stations in that area should be more.
2) If the number of incidents in an area are more, then the number of hospitals in those areas should be more.
3) If the number of incidents in an area are more, then the cost of living in that area should be less.
4) Number of Incidents will be more in the month of November because of the holiday season starting in the month of December.
5) If the number of incidents in an area are more, then the number of schools in that area should be less.

ETL WORKFLOWS:

ETL WORKFLOW 1:



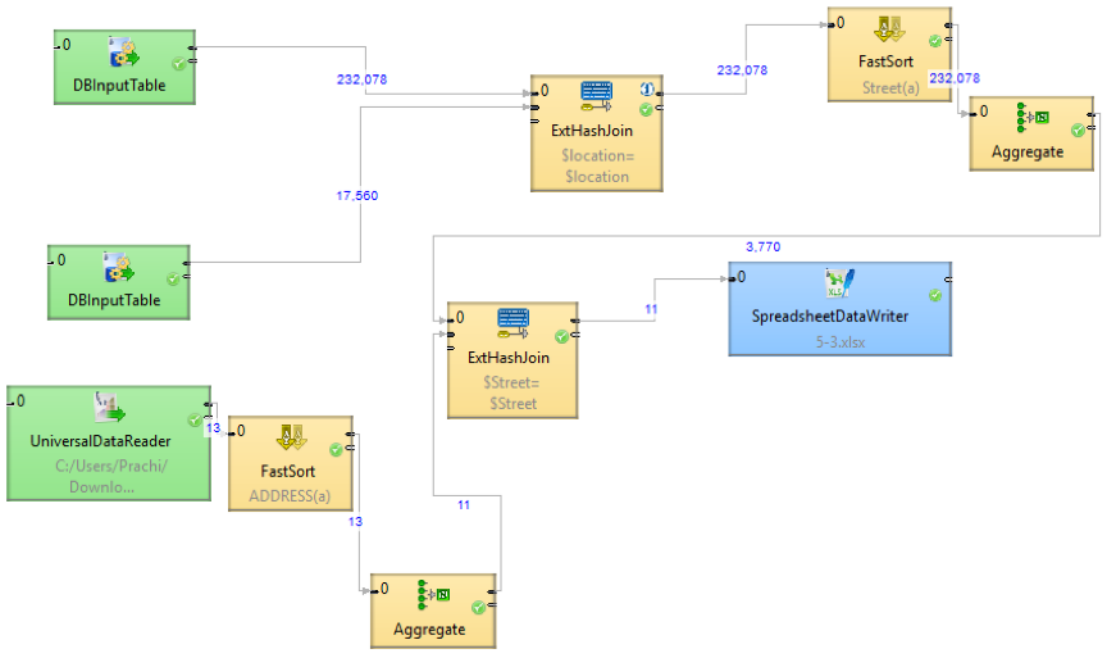
In this workflow, we are taking number of incidents and location from the Incidents table, and location and street from the Address table, and performing an inner join on the same, and after sorting it, we are using a group by street and counting the number of incidents per street.

We are using the csv file that has the data regarding the hospitals on every street, and we are performing an aggregate function and group by street and counting the number of hospitals on every street.

We are joining the results of these two by mapping the streets and exporting it to a excel file where we insert a chart based on the output generated.

The excel file contains three columns, one shows the street, the other shows number of incidents on that street, and the last one displays the number of hospitals on that street.

ETL WORKFLOW 2:



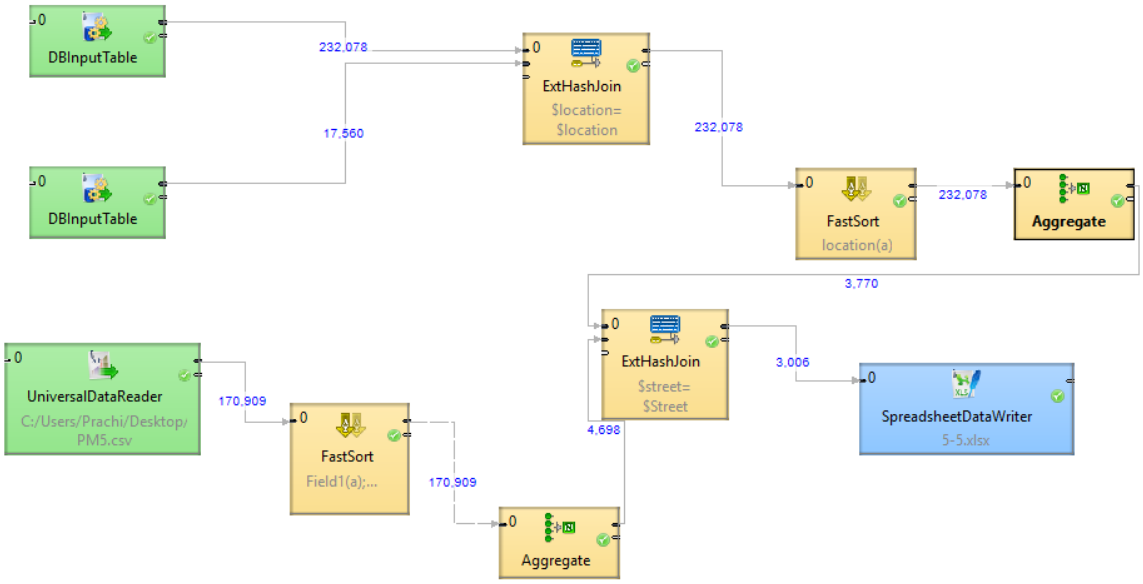
In this workflow, we are taking number of incidents and location from the Incidents table, and location and street from the Address table, and performing an inner join on the same, and after sorting it, we are using a group by street and counting the number of incidents per street.

We are using the csv file that has the data regarding the police stations on every street, and we are performing an aggregate function and group by street and counting the number of hospitals on every street.

We are joining the results of these two by mapping the streets and exporting it to a excel file where we insert a chart based on the output generated.

The excel file contains three columns, one shows the street, the other shows number of incidents on that street, and the last one displays the number of police stations on that street.

ETL WORKFLOW 3:



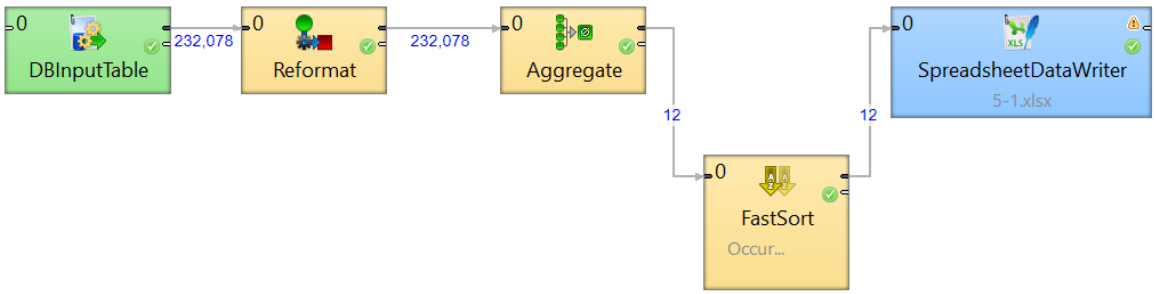
In this workflow, we are taking number of incidents and location from the Incidents table, and location and street from the Address table, and performing an inner join on the same, and after sorting it, we are using a group by street and counting the number of incidents per street.

We are using the csv file that has the data regarding the cost of apartments on every street, and we are performing an aggregate function and group by street and averaging the cost of property on each street.

We are joining the results of these two by mapping the streets and exporting it to a excel file where we insert a chart based on the output generated.

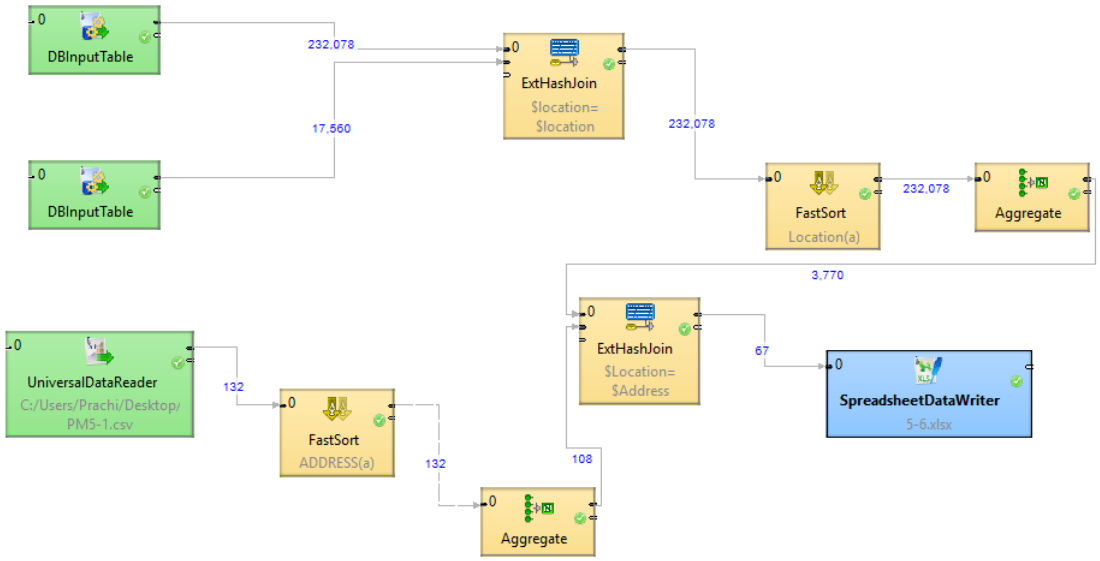
The excel file contains three columns, one shows the street, the other shows number of incidents on that street, and the last one displays the cost of property on that street.

ETL WORKFLOW 4:



In this workflow, we are taking the incidents from the dataset, and formatting the date of occurrence to the month, aggregating (group by month), sorting the results and writing the results to a csv file.

ETL WORKFLOW 5:



In this workflow, we are taking number of incidents and location from the Incidents table, and location and street from the Address table, and performing an inner join on the same, and after sorting it, we are using a group by street and counting the number of incidents per street.

We are using the csv file that has the data regarding the number of schools on every street, and we are performing an aggregate function and group by street and counting the number of schools on each street.

We are joining the results of these two by mapping the streets and exporting it to a excel file where we insert a chart based on the output generated.

The excel file contains three columns, one shows the street, the other shows number of incidents on that street, and the last one displays the number of schools on that street.

Insights:

1) The hypothesis is false in most cases, and the number of hospitals are more in the areas having the maximum incidents. This means that the hospitals are more in regions having less number of incidents (From 5-2.xlsx).
2) The hypothesis is true in most cases, and the number of police stations are more in the areas having the maximum incidents (From 5-3.xlsx).
3) The hypothesis is true in most cases, and the cost of living is more in the areas having the minimum incidents (From 5-5.xlsx).
4) The hypothesis is wrong in most cases, the maximum number of incidents have occured in the month of August, right before universities begin.
5) The hypothesis is true in most cases, and the number of schools are less in the areas having the maximum incidents (From 5-6.xlsx) with one exception.

Actions:

1) We can suggest having more hospitals where the number of incidents occur the most.
4) We can suggest having more patrolling in the month of August.
5) We can suggest having less schools in the areas where most incidents occur. If there is a new school plan in the city, reviews can be given in favor of having it built in the safe areas.

Grishma Thakkar
Oneil Contracctor
Prachi Ved
Sonal Singh