Task List

Anuj Bora and Gowtham Goli

Task	Who	Start Date	End Date	Comments
Identify data sources	Anuj & Gowtham	18th March	20th March	• *Identified relevant data sources from NYC OpenData*
Plan where data will reside	Anuj & Gowtham	21st March	25th March	• *Roughly around 12GB* • *The data will fit in QuickStart VM*
Write code to ingest data source 1, 2, 3 and 4	Anuj & Gowtham	26th March	31st March	• *The data is historical and one-time. So, all the data will be transferred only once at the beginning to HDFS.*
Write code to clean (ETL) data source 1 and 2	Gowtham	1st April	3rd April	• *Cleaning and filtering of data to remove unwanted columns*
Write code to clean (ETL) data source 3 and 4	Anuj	1st April	3rd April	• *Cleaning and filtering of data to remove unwanted columns*
Write code to profile data source 1 and 2	Gowtham	4th April	14th April	• *Check if the boroughs names are consistent in the two datasets. If not, make them consistent*
Write code to profile data source 3 and 4	Anuj	4th April	14th April	• *In the air quality dataset, map the measurement values to whether the air is good, moderate, unhealthy or hazardous* • *Some of the values of years, boroughs are erroneous which needs to be profiled*

Task	Who	Start Date	End Date	Comments
Design the analytic(s)	Anuj & Gowtham	15th April	20th April	• *Classify health complaints by water, air and rodents and compare them against the registered complaints on water, air and garbage and infer the results obtained*
Code the analytic(s)	Anuj & Gowtham	21st April	24th April	• *The above discussed analytic will be coded using MapReduce*
Test the analytic(s)	Anuj & Gowtham	25th April	28th April	• *Plot the results obtained*
Analyze results of analytic(s)	Anuj & Gowtham	29th April	2nd May	• *We will check the results to find out whether there is a direct correlation between the registered water and garbage complaints, air quality values to the health conditions in each of the five boroughs*

Legend:

Data Source 1 - NYC OpenData: 311 Service Request Dataset

Data Source 2 - NYC OpenData: Water Quality Complaints Dataset

Data Source 3 - NYC OpenData: Air Quality Dataset

Data Source 4 - NYC OpenData: OATH ECB Hearings Case Status Dataset