WEATHER PREDICTION FOR UNITED STATES

CMPT -732

Programming for Big Data 1

Fall - 2017

INTRODUCTION

Predicting the Meteorological and Pollutant level for U.S. based upon the last 20 years data collected by EPA – Environmental Protection Agency, for computing the overall AQI of a particular region.

BACKGROUND AND BRIEF INFORMATION ABOUT DATA

US EPA collects the outside air quality data, the data is collected and reported across US till the county level. It is reported in the below timelines: -

a. Annual Summary
b. Daily Monitored Data by Counties
c. Hourly Data/8 Hour Data

The format and type of data reported is different for each of these reports. Please refer the below links for more information: -

a. Data Download Links - https://aqs.epa.gov/aqsweb/airdata/download_files.html
b. Data Description - https://aqs.epa.gov/aqsweb/airdata/FileFormats.html

The data is available since 1980 and it is collected for the below parameters:

Data for the criteria gases	Data for the particulate matters	Meteorological Data
Ozone (44201)	PM2.5 (88101)	Temperature
SO2 (42401)	PM2.5 nonFRM (88502)	Wind
CO (42101)	PM10 Mass (81102)	RH and Dewpoint
NO2 (42602)	PM2.5 Speciation	Barometric Pressure

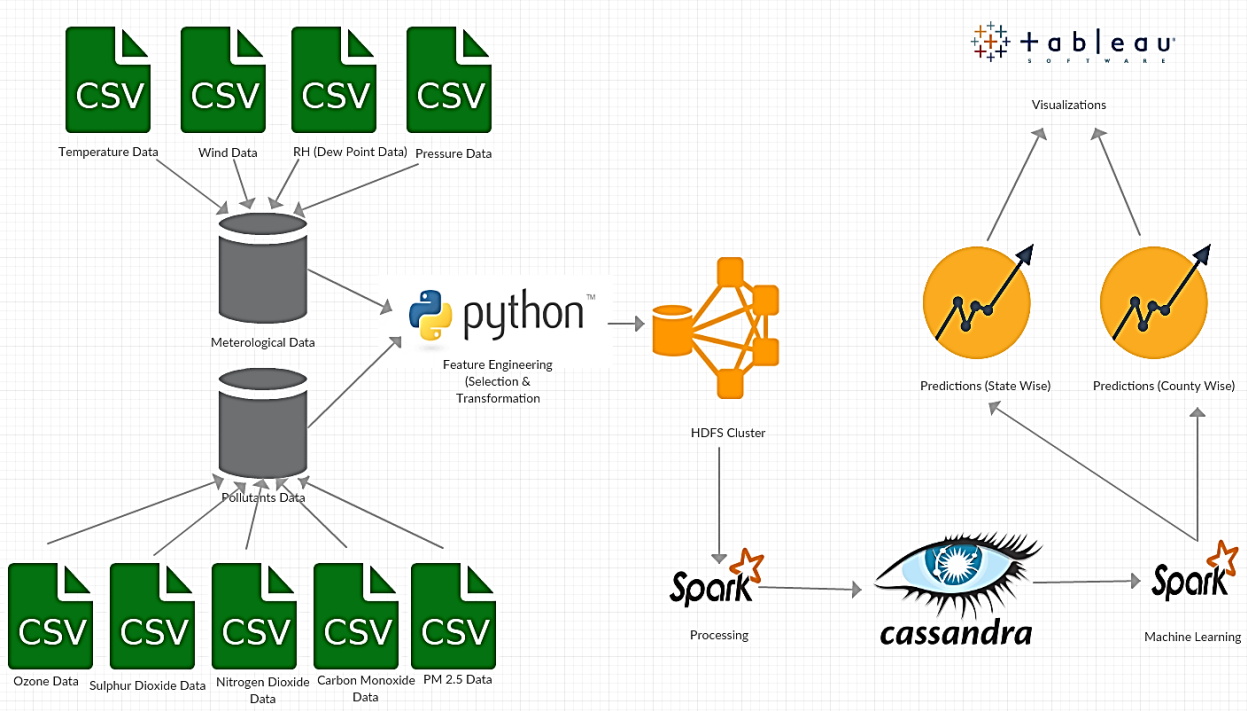
There are several other parameters as well but we have included these parameters in to our project scope.

DATA ANALYSIS AND PREPROCESSING STEPS

As per the project requirement, we had choosen the "Daily Summary" dataset for the required parameters. Below is the structure of the dataset. The file is comma separated variables (CSV) with a header row: -

Below are the set of operations we have applied on the Dataset in order to use it for the computation:

a. **Data Pipeline**: Below is the high-level view of the complete pipeline for our project:



b. **Data Aggregation**: As have ran the prediction across states and on a monthly basis but the initial challenge was that EPA doesn't report the data on a monthly basis. So, due to this we had used the Daily Summary data collected across U.S. at the county level. So the transformation is:

State, County, Daily -> State, Monthly

While doing this transformation, we have applied Statistical function on the aggregated data by keeping them consistent with the actual range of the monitored values.

c. Feature Selection: We have done **Feature engineering** in order to get efficient prediction. This was done based upon research about how the Meteorological data effects the pollutants level of a particular place and also analysis in python locally like applying **Corr**() and **VarianceThreshold()** functionality of Sklearn which enabled us to identify the most prominent features to train our model on. Below are some articles we referred before including the additional weather/meteorological feature for predicting the Max value of a pollutant at a particular region: [4][5][6]

d. **Data Joining**: As mentioned above, we have used the meteorological data with the daily gases and particulate data for making their predictions.

   a. But all of this data is reported in separate files.
   b. So, we have extracted the required features from the meteorological data and clubbed them with the gases and particulate specific data as one of the independent values. (Columns)

Please refer the link [3] in references for feature description.

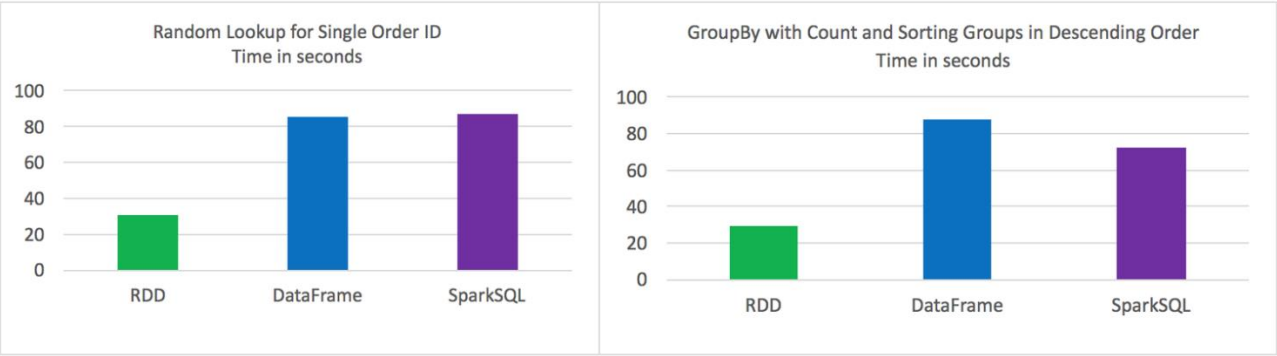
IMPLEMENTATION

There are 2 parallel implementations of this project:

a. **On Cluster**: On the cluster we have used Spark and Cassandra for all the implementation related tasks. We have defined the schema for the required table keeping in mind efficiency of reading and writing back to Cassandra.
b. **On Local Machine**: We have also implemented a bigger version of this project on Spark installed on the Local Machine with 4 cores. Here we are saving the output in the .csv format.

Methodology: Below are list of important steps implemented in this project to scale the solution for **18.5GB** of data:

a. By using the SQL-like functionality of the Spark Dataframes, we have clubbed the required features from Cassandra tables.
b. On the clubbed data, we have trained multiple models on training data as mentioned below and evaluated the trained models using the test data. The list of models are:
    - **Linear Regression**
    - **Gradient Boosting**
    - **Random forest**
c. Based upon our analysis, we have identified that the gradient boosting is the best model; hence we have used the same model for predicting all the 2018 parameters including the meteorological data.
d. Important/Interesting Observation:
    - **Dataframe Functions vs Spark.Sql:** We have observed that for the aggregating tasks especially when the data size is large as our initial dataset was around 18.5GB. The SQL queries inside spark.sql perform better than the Dataframe functions. Please refer the below explanation:



Ref: https://community.hortonworks.com/articles/42027/rdd-vs-dataframe-vs-sparksql.html

- **Checkpoints:** As the model was getting trained on a **big dataset**, if we were writing the program output at the end, then the program was throwing "out of memory error". In order to execute the program, we added the insertions during the execution of the program i.e. the Cassandra table was updating during the program execution. Hence, we were able to resolve the "Out of Memory" issue.

e. **Scalability of Code**: We have also implemented our code to predict the trends at the county levels for the state of California. All the trends and predictions are consistent with the type of predictions done on the country level. This was to check the scalability of the code and checking on how it deals with more granular data.

f. **AQI Calculation:** We have understood EPA methodology for calculating the AQI at a particular place. Refer the table below:

EPA's table of breakpoints is:

O3 (ppb)	**O3 (ppb)**	**PM2.5 (µg/m3)**	**PM10 (µg/m3)**	**CO (ppm)**	**SO2 (ppb)**	**NO2 (ppb)**	**AQI**	**AQI**
Clow - Chigh (avg)	*Clow - Chigh (avg)*	*Clow- Chigh (avg)*	*Clow - Chigh (avg)*	*Clow - Chigh (avg)*	*Clow - Chigh (avg)*	*Clow - Chigh (avg)*	*Ilow - Ihigh*	**Category**
0-54 (8-hr)	-	0.0-12.0 (24-hr)	0-54 (24-hr)	0.0-4.4 (8-hr)	0-35 (1-hr)	0-53 (1-hr)	0-50	Good
55-70 (8-hr)	-	12.1-35.4 (24-hr)	55-154 (24-hr)	4.5-9.4 (8-hr)	36-75 (1-hr)	54-100 (1-hr)	51-100	Moderate
71-85 (8-hr)	125-164 (1-hr)	35.5-55.4 (24-hr)	155-254 (24-hr)	9.5-12.4 (8-hr)	76-185 (1-hr)	101-360 (1-hr)	101-150	Unhealthy (Sensitive Groups)
86-105 (8-hr)	165-204 (1-hr)	55.5-150.4 (24-hr)	255-354 (24-hr)	12.5-15.4 (8-hr)	186-304 (1-hr)	361-649 (1-hr)	151-200	Unhealthy
106-200 (8-hr)	205-404 (1-hr)	150.5-250.4 (24-hr)	355-424 (24-hr)	15.5-30.4 (8-hr)	305-604 (24-hr)	650-1249 (1-hr)	201-300	Very Unhealthy
-	405-504 (1-hr)	250.5-350.4 (24-hr)	425-504 (24-hr)	30.5-40.4 (8-hr)	605-804 (24-hr)	1250-1649 (1-hr)	301-400	Hazardous
-	505-604 (1-hr)	350.5-500.4 (24-hr)	505-604 (24-hr)	40.5-50.4 (8-hr)	805-1004 (24-hr)	1650-2049 (1-hr)	401-500	

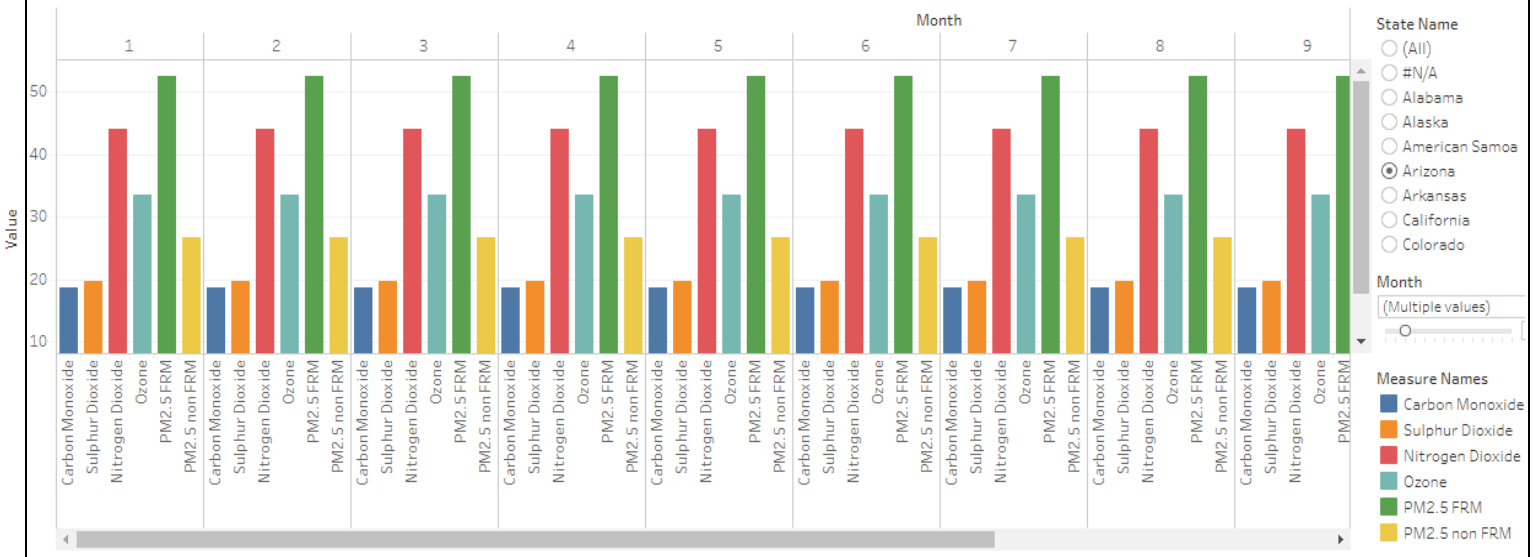
The formula used for calculating the AQI for all the contributing pollutants is:

$$I = \frac{I_{high} - I_{low}}{C_{high} - C_{low}}(C - C_{low}) + I_{low}$$

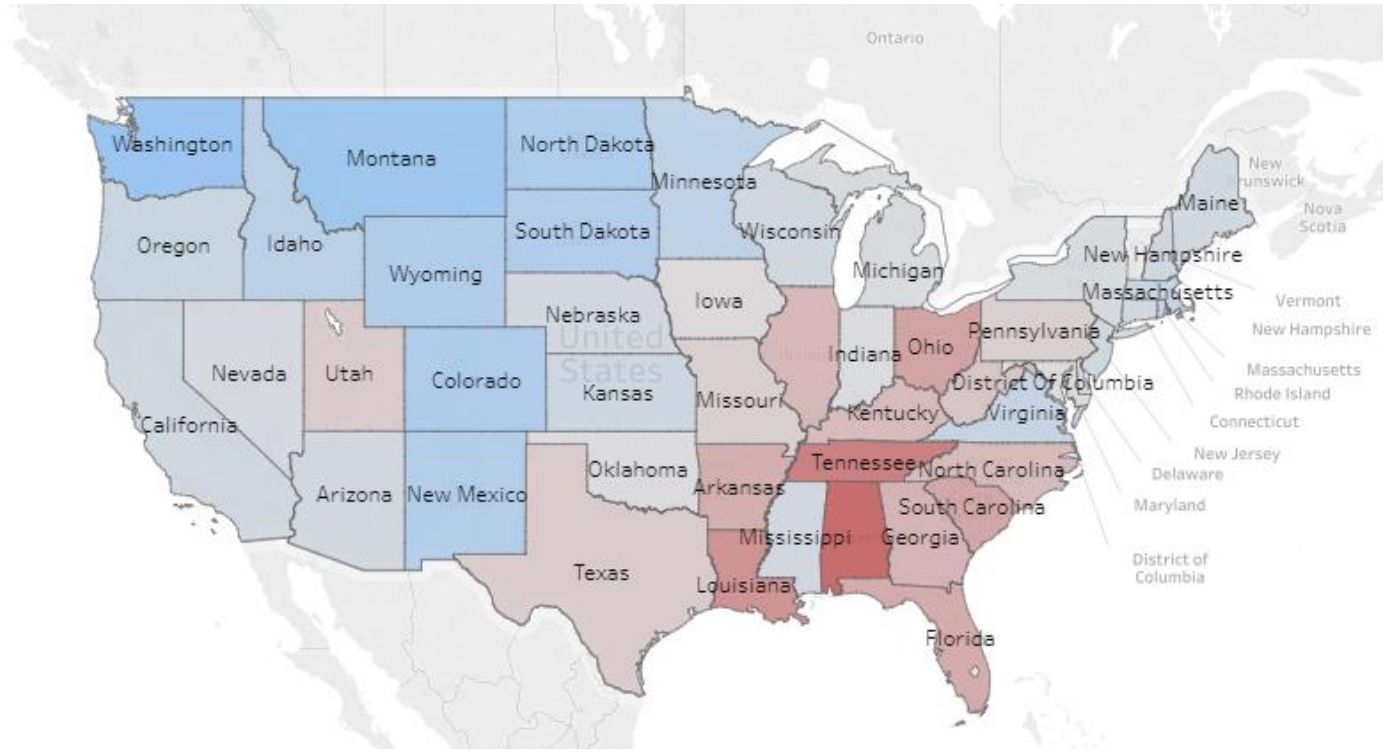
where:

I = the (Air Quality) index,
C = the pollutant concentration,
C_{low} = the concentration breakpoint that is ≤ C,
C_{high} = the concentration breakpoint that is ≥ C,
I_{low} = the index breakpoint corresponding to C_{low},
I_{high} = the index breakpoint corresponding to C_{high}.

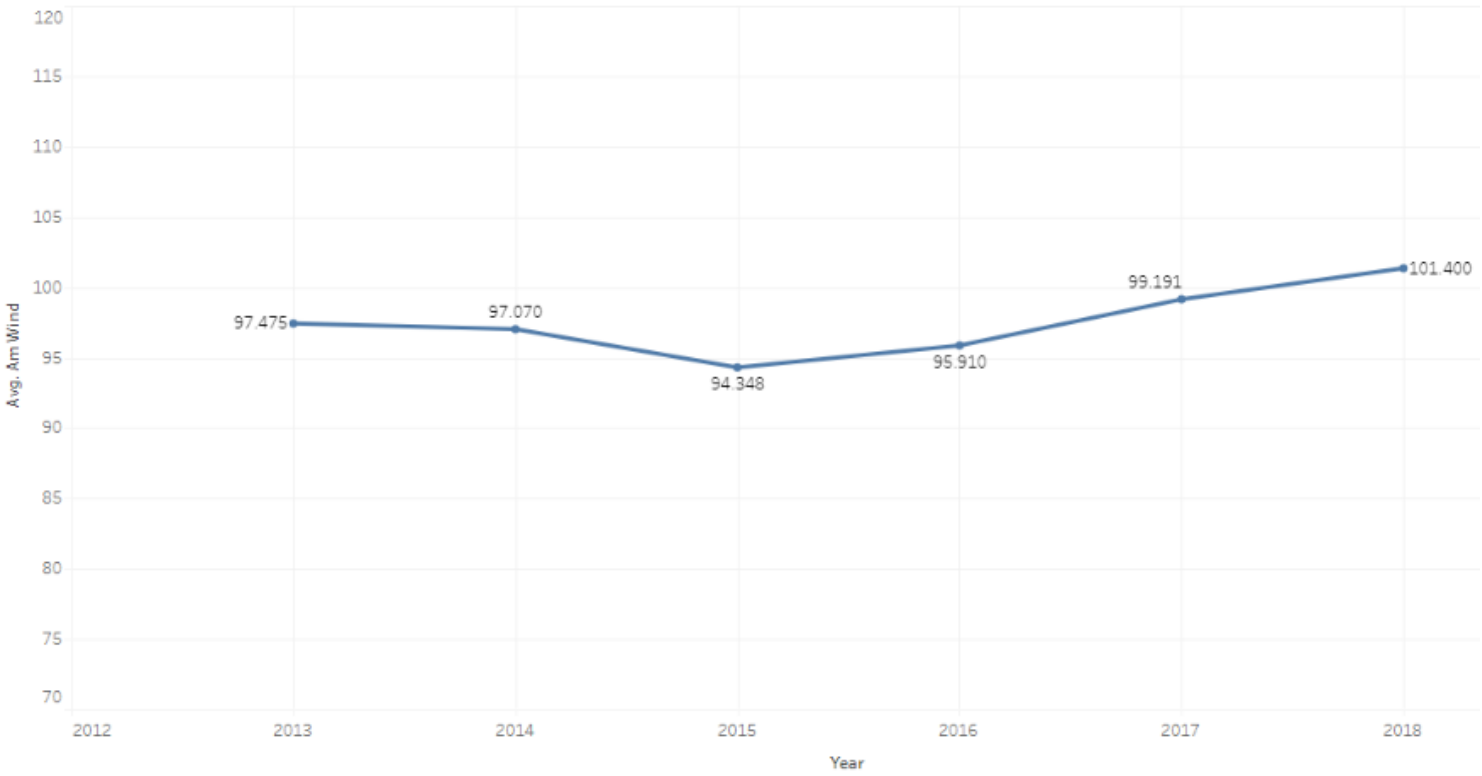
Using the above parameters, we would be getting the AQI of every pollutant and then the global AQI would be the AQI of the dominating pollutant. Below is the graph that we had plotted:



g. **Global Warming Trends**: In addition to predicting the meteorological data for the states across U.S., we have also plotted the trends of the temperature across all the state i.e. whether the next year would be colder or warmer for the respective states next year. Below is the diagram:



h. **Temporal Graph**: We have also show the pattern of the meteorological data for the last 5 years. This was done to showcase how consistent our prediction was from the past trends. Below is the example for wind: -



i. **Overall Accuracy:** As we have aggregated the data for the state and month i.e. taking away the granular details which will give better prediction (We have included those in the county wise implementation for California). We were able to achieve good accuracy for the Meteorological parameters such as Temperature, Wind, Rh/Dewpoint etc. And this can also be understood from the temporal graph above

VISUALIZATION / CONCLUSION

Please refer the below link for all the insights gathered from the data analysis and weather predictions, it is hosted on Tableau Public and explained by story:

https://public.tableau.com/profile/lakshayy.dua#!/vizhome/U_S_WeatherForecastingusingBigData/Story1?publish=yes

PROJECT SUMMARY

Initially, we proposed to predict the state-wise global AQI for every month across U.S. but after further discussion we expanded our scope to include the Meteorological Data and its prediction for 2018 for every state across U.S. i.e. Temperature, Pressure, Rh/Dewpoint.

Also, we researched more about the effects of Meteorological data on global AQI. Hence, we included these data points as feature to make predictions around AQI for a particular place.

In addition to this we also went till the county level as well as predicting Global Warming trends as per our discussion with Steven Bergner.

- Getting the data: Acquiring/gathering/downloading.
  5 – As we scaled our solution from initial dataset size of 4 Gb to final dataset size 18.5 Gb.

- ETL: Extract-Transform-Load work and cleaning the data set.
  3- performed transform load work on spark cluster for this huge dataset.

- Problem: Work on defining problem itself and motivation for the analysis.
  5-extended the scope of the project beyond it's proposal.

- Algorithmic work: Work on the algorithms needed to work with the data, including integrating data mining and machine learning techniques.
  3- Used feature selection, feature engineering concepts along with ensemble learning models (Random Forests, Gradient-Boosted Tress) to get precise results.

- Bigness/parallelization: Efficiency of the analysis on a cluster, and scalability to larger data sets.
  2- Scalable solution with use of checkpoints and digging deep up to county level on this huge dataset.

- UI: User interface to the results, possibly including web or data exploration frontends.
  0, no website made for this project

- Visualization: Visualization of analysis results.
  2 – Done various visualizations on tableau.

- Technologies: New technologies learned as part of doing the project.
  0 – But utilized already learned technologies and applied them to achieve scalable solution.

REFERENCES

[1] The AQI table was taken from Wikipedia, refer below
link: https://en.wikipedia.org/wiki/Air_quality_index#Computing_the_AQI
[2] Dataset was downloaded from https://aqs.epa.gov/aqsweb/airdata/download_files.html
[3] Table with Features https://aqs.epa.gov/aqsweb/airdata/FileFormats.html#_daily_summary_files
[4] https://www.kullabs.com/classes/subjects/units/lessons/notes/note-detail/7543
[5] https://www.waikatoregion.govt.nz/environment/natural-resources/air/weather-affects-air-quality/
[6] http://epa.tas.gov.au/Pages/How-Weather-Affects-Air-Quality.aspx