# Clean air analytics and comparison of Big Data Technologies

**Arjun Chaudhary**

**Supervisor: Richard Sinnott**

**Melbourne School of Engineering**

**University of Melbourne**

This report is submitted for

*Computing Project (25 credit points)*

**OCTOBER 2016**

First and foremost, I have to thank my parents for their love and support throughout my life. Secondly, I would like to sincerely thank my supervisor, Prof. Richard Sinnott, for his constant guidance, communication, feedback, encouragement, direction and support throughout this project, and especially for his confidence in me. Finally, to all my friends, thank you for your understanding and encouragement and special thanks to Yikai Gong for his input and expertise on SMASH Architecture.

# Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this report are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This report is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. This report contains approximately 8362 words including appendices, bibliography and figures.

Arjun Chaudhary

October 2016

# Abstract

Air pollution regulation, monitoring and enforcement are inadequate in Australia. According to a estimate about 3000 Australian die premature deaths each year due to urban air pollution [1].

In 2013, a Senate Committee investigation concluded that regardless of improvements in Australian air quality it is still an important problem in many parts of Australia [2]. Some communities are much more influenced than others, depending on their proximity to pollution sources particularly communities living in or near CBDs.

PM pollution is a problem in both large cities and rural towns. The severe health effects from exposure to the diverse sources of air pollution are now well recognized. There is no 'safe' level of exposure for many of these pollutants. Though, pollutant effects varies there are harmful impacts from exposure at levels even lower than the present air quality standards [3].

This reality has directed many doctors desire to better recognize the correlation among air pollution and health problems such as lung cancer, COPD and asthma. A lot of cancer cases can happen with no obvious causality i.e. no history of cancer in the family and no history of smoking. And hence, in order to find whether these cases may result due to $PM_{2.5}$ concentration, leads to need of highly disaggregated data (both temporally and spatially) to be recorded measuring air pollution ($PM_{2.5}$), preferably in real-time.

Heavy city traffic is one of the prime sources to air pollution and health problems particularly near CBD area. Hence, knowing about amount of traffic in a particular region could help to find out impact of pollution in that region. But, both of these dataset are not easily available. Hence, we tried to develop a model based on twitter to correlate amount of traffic, concentration of particulate matter and number of tweets made on roads and to perform data analytics. In order to get correlation we need data and hence we are using 2014 scat and twitter data. For, pollution data, a network of 45 Citizen Sensors will be distributed from University Square to Lincoln Square to Cardigan gardens, finishing at Carlton Gardens. These Citizen sensors will continuously monitor air temperature, relative humidity, noise levels, PM 2.5 concentrations using AirBeam sensors. The data from sensor is sent to Aircasting android App and then in turn to a scalable Ruby server for different analytics and correlations.

As we can see we have huge amount of real time data. At first, we have performed clean air analytics using our traditional system using technologies like MySQL. However at the end of thesis, we did benchmarking and comparison of our traditional system with SMASH, a generic and highly scalable Cloud-based architecture and showed that data we are dealing with is "Big data" which is commonly defined as data that are too large, created too quickly, or structured in such a manner as to be difficult to collect and process using traditional data management systems. Hence, requires more sophisticated system like SMASH which includes a distributed file system for capturing and storing data, a high performance computing engine able to process such large quantities of data, a reliable database system able to optimize the indexing and querying of the data, and geospatial capabilities to visualize the resultant analyzed data.

**Key words:** Twitter, Traffic, PM 2.5, Data Analytics, Big Data, SMASH.

# Table of Contents

## List of Figures

## List of Table

# 1 Introduction

## 1.1 Background and Related Work

Air pollution regulation, monitoring and enforcement are inadequate in Australia. According to an estimate about 3000 Australian die premature deaths each year due to urban air pollution [1].

Particulate matter pollution (also known as PM or particulate pollution) is considered as one of the main and concerning type of pollution, according to new emerging evidence exposure to PM pollution causes lung cancer, cardiovascular disease and stroke [4].

Particulate matter pollution particles are considered as coarse, fine or ultrafine particles ($PM_{10}$, $PM_{2.5}$, or $PM_{0.1}$) depending on the particles size. PM pollution is classified by size rather than as one particular chemical substance [5]. For example $PM_{2.5}$ is 2.5 micrometres in diameter and are airborne particles. So, if we compare the size it is actually less than (around 1/40 the average width of a human hair). $PM_{0.1}$ and $PM_{2.5}$ are more hazardous to human health than $PM_{10}$ because due to their small sizes they is more possibility for them to be easily drained deep into the lungs and are strongly connected to severe health impacts, mainly heart and lung disease.

Causes of PM pollution are extremely diverse. The chief source of PM pollution in Australia is:

- coal combustion for power generation
- mining
- discharges from combustion processes such as petrol and predominantly diesel vehicles
- smoke from wood burning
- bushfires
- dust storms

PM pollution is a problem in both large cities and rural towns. The severe health effects from exposure to the diverse sources of air pollution are now well recognized. There is no 'safe' level of exposure for many of these pollutants. Though, pollutant effects varies there are harmful impacts from exposure at levels even lower than the present air quality standards [3].

This reality has directed many doctors desire to better recognize the correlation among air pollution and health problems such as lung cancer, COPD and asthma. A lot of cancer cases can happen with no obvious causality i.e. no history of cancer in the family and no history of smoking. And hence, in order to find whether these cases may result due to $PM_{2.5}$ concentration, leads to need of highly disaggregated data (both temporally and spatially) to be recorded measuring air pollution ($PM_{2.5}$), preferably in real-time.

As described above, emissions from combustion processes such as petrol and particularly diesel vehicles is major source for PM, hence we can easily relate amount of traffic in a particular region with amount of PM concentration. However, organisation like EPA Australia and VICROADS doesn't make their data publicly available.

Though, we can model traffic based on twitter data [7] [8].Twitter is a social media platform, and with the growth of Internet, people are progressively more likely to share their status, their opinions and the views they experience in real life on Twitter.

With the increased in popularity and usage of Twitter, it has become a trend and now many government organisation have created government-authorized Twitter accounts on many subjects and share subject specific information through Twitter. For example the account of VICROAD is one example of a government authority, who publishes traffic related issues or information on Twitter. With the Twitter function of re-tweet, the published message can be shared quickly. This account can also accept other re-tweet message related to traffic issues. Based on this interaction between twitter users and the government Twitter accounts, traffic issues can be identified and shared quickly.

According to [9], in order to get pollution data, we can use AirBeam which continuously monitor air temperature, relative humidity, noise levels, PM 2.5 concentrations.

## 1.2 Motivation

Owing to the impact of air pollution and its main source i.e. emissions from combustion processes such as petrol and particularly diesel vehicles and difficulty to get data about them. It would be beneficial to find a correlation between Air pollution, Amount of traffic and Tweets made on the road. Once we have implemented this, we could also benchmark the performance of our Big Data technology Architecture with respect to traditional system to find the need of using big data technologies in order to do such large scale processing in real time. For cities like Melbourne which is a liveable city, organisation like City of Melbourne can use such Architecture to perform Clean Air analytics and its impact on environment example trees in real time.

## 1.3 Contribution

There are five main contributions we have explored through our designed system implementation:

1. There is  a strong correlation between increase in amount of traffic and number of tweets made on road in morning during office hours i.e. 8-10 A.M and also when people leaves from their work from 4-6 P.M. On weekdays especially on sat night from 6 P.M to 9 P.M again strong correlation was seen.
2. There is also a strong correlation between increase in $PM_{2.5}$ concentration and number of tweets made on road in morning during office hours i.e. 8-10 A.M and also when people leaves from their work from 4-6 P.M on weekdays. On weekends especially on sat night from 6 P.M to 8 P.M again strong correlation was seen.
3. There is a correlation between Pedestrian Count, $PM_{2.5}$ concentration and Sentiment Analysis of the Tweets for different region in CBD. Correlation shows that PM2.5 concentration and number of negative tweets increases and decreases with increase and decrease of pedestrian count during busy hour. Hence, we can say that due to increase in air pollution people are getting depressed which leads to increase in negative tweets.
   For example, for Grattan St-Swanston St (West) we compared Average Hourly Pedestrian Count over the past 52 weeks with $PM_{2.5}$ concentration. We found strong correlation between both of them at three different intervals at 7-10 A.M (office hour), 11 A.M -1 P.M (lunch time) and 4 P.M - 7 P.M. (students and people leaving for home). All these three timeslot are actually busy hour when we have more cars on road, and we can easily see $PM_{2.5}$ concentration and Pedestrian Count both are high during this interval. Now, for same street intersection and same intervals, i.e. 7-10 A.M (office hour) sentiment changed from positive to neutral showing neutral at peak hour 9 A.M and for 11 A.M -1 P.M and 4 P.M - 7 P.M. we have overall negative tweets/sentiment. So, we can easily see increase in air pollution causing increase in negative tweets. However, other than Air pollution other factors like crowd, noise pollution, work stress and tension (for 11 A.M to 1 P.M slot) could also lead to more negative tweets.
4. There is also a correlation between Numbers of deaths due to diseases related to air pollution VS Amount of traffic (trucks). We saw a strong correlation between them as the area where amount of traffic was large, number of deaths were also more specially in areas like Melbourne (C) Remainder, Brimback (C) Sunshine, Wyndham (C) - North, Gr. Dandenong (C) - Dandenong.
5. We also find out that our SMASH Architecture performs way better than traditional system. For example the most computational task of finding number of tweets made on road which require a Cartesian product between tweets geo-coordinate and roads line geometry is 8 times faster with Spark using (1 node, 4 core, 16 GB RAM) for CBD which is about 10 million rows as compared MySQL cross join.

## 1.4 Limitation

1. Difficulty to find global correlation between increase in amount of traffic and number of tweets made on road, as this correlation was based on 2014 scats data due to non-availability of 2015 or 2016 scats data, we required 2014 geo-located tweets made from road which was very hard to get in large amount. Hence, it was very difficult to see overall relation between both dataset for 24 x 7.
2. Difficulty to find global correlation across Melbourne between increase in $PM_{2.5}$ concentration and number of tweets made on road. As, we didn't have enough pollution data for each street or region of Melbourne and also as a lot our citizen scientist are students and professors who are at their home in night. Hence, we don't have much data to find any correlation after 8 P.M on weekdays.
3. We established a correlation between $PM_{2.5}$ concentration and Negative sentiment of people by connecting them through pedestrian count at a particular region in CBD. As, in order to establish this fact we required very limited set of tweets which were made in a particular region of CBD at particular time interval, we didn't had a huge set of tweet data to come up with a correlation for a particular week or day. Hence, we have to considered tweets for past 52 weeks for coming up with a correlation.
4. For correlation between Numbers of deaths due to diseases related to air pollution VS Amount of traffic (trucks), we observed a strong correlation for much location from plot as specified above. Despite that our adjusted $R^2$ value came between 40 - 60 % for different cases which is not bad. However, it could be better but we had limitation like data about amount of Volume of traffic was for roads unlike death data which was based on SLA11 regions, so the problem was that a particular road was coming inside more than one SLA11 regions, however we tried a workaround for this problem but we could have made better prediction if we could have known exact estimation about amount of traffic for different SLA11 regions. We also had lot of missing values for COPD and Respiratory diseases for most of SLA11 region.

In the following part, we discuss contents as follows. The system architecture is discussed in Section 2; the system methodology in Section 3, and the system implementation and result analysis in Section 4, and the problem finding and future research direction discussion in Section 5, with conclusions in Section 6.

# 2 System Architecture

In this section, we describe how we build the system, including the architecture of system, what kind of data we have used in our system, and what kind of result outputs the user will witness through our system. We will also illustrate the kinds of functionalities our system can provide.

## 2.1 Overall Architecture

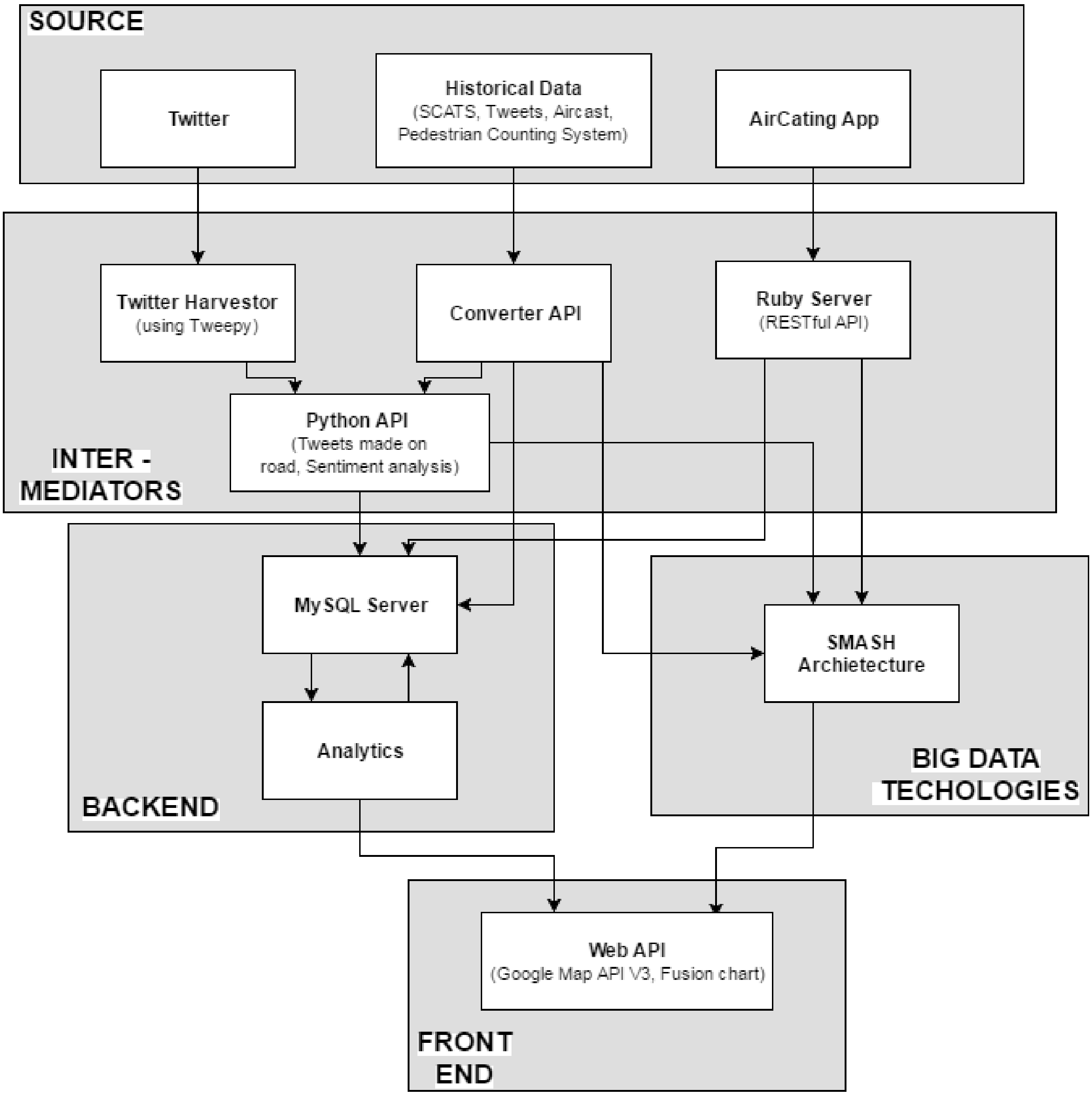


Figure 1 Overall Architecture for the Project

## 2.2 Source

In order to perform Clean Air Analytics, we used data from different sources which is described in Table 1. The data collected from different sources is send to intermediators in order to format it such that it is suitable for storage in different type of database technologies.

| Data Set Name | Data Source | Data Production Methodology |
|---|---|---|
| Twitter | Twitter API | We used different instances on Nectar for running Twitter Harvester using Streaming API and Rest API in order to collect more data. And as our final goal was to get tweets made on road hence we also used another harvester to collect data using Twitter Search API for tweets mentioning keywords like #trafficjam, #traffic and also getting tweets from some twitter official accounts like VicRoads (@VicRoads) because most of these tweets are made by people while driving or waiting for traffic to clear. |
| Aircasting_Data | Aicasting APP | The data recorded by AirBeam devices is sent to Aircasting Mobile Application through Bluetooth and this real time data is sent as streaming data every 15 minute to our Ruby Server which in real time route this real time data for storage to different database technologies for real time analytics. |
| SCATS | Previous SMASH traffic research | We used the 2014 SCATS data collected in previous SMASH traffic research. |
| PSMA | AURIN | **We collected to line geometry of roads from AURIN using PSMA Street Network (August 2016)** dataset. PSMA Street Network data provide national coverage of street network at all levels. Roads data covers everything from major highways to walking paths. |
| Historical_Twitter_Data | Unimelb Twitter Harvestor | To collect more tweets made on the road in 2014 for correlation with SCATS 2014 data. I used 2014 tweet data from University Of Melbourne Twitter Harvester using following cURL command<br>curl -XGET "http://130.56.252.54:8092/melbourne/_design/geo/_view/geoTweets?inclusive_end=true&stale=false&connection_timeout=60000&limit=1&skip=0&descending=false"<br>. |
| Historical_Aircasting_Data | eResearch clean air database | In order to gather more data on pollution I also used previously recorded data using AirBeam devices in clean air database by Citizen scientist. |
| Pedestrian_Count | City Of Melbourne : Pedestrian Counting System | We gathered publicly available Average hourly pedestrian count over the past 52 weeks by City Of Melbourne for different location in CBD from Pedestrian Counting System. |
| Melb_Traffic_Volume | Victorian Road Traffic Volumes | We downloaded the data about 24 Hour Median Midweek Non Holiday Truck Volumes from AURIN in shape file format. |
| Air_Related_Deaths | SLA11 Premature Mortality | We downloaded the data on Deaths from Chronic Obstructive Pulmonary Disease (COPD) - 0 to 74 Years - Count , Deaths from Ischaemic Heart Disease - 0 to 74 Years - Count , Deaths from Lung Cancer - 0 to 74 Years - Count and Deaths from Respiratory System Diseases - 0 to 74 Years - Count from this dataset of AURIN. |

Table 1 Different Data Set used for analytics

| Data Set Name | Data Amount (Rows in database) |
|---|---|
| Twitter + Historical_Twitter_Data | 1952046 (Out of these only: 254526 tweets were made on road in 2014, 2015, 2016 combined) |
| Aircasting_Data + Historical_Aircasting_Data | 5824578 |
| SCATS | 2171412 |
| PSMA | 1961844 |
| Pedestrian_Count (past 52 weeks) | 44 |
| Melb_Traffic_Volume | 9174 |
| Air_Related_Deaths | 79 |
| Total | 11919177 (10 GB of Data) |

Table 2 Sizes of different Data Set

In Table 2, we have defined the amount of data distribution among different dataset. We only collected tweets containing geo-located coordinates; it may look like we have enough number of tweets. However, out of all these geo-located tweets only 264536 were made on road.

For some correlations correlation for example for finding correlation with 2014 scats data we needed tweets made on road in 2014 and  number of tweets made on road in year 2014 is just 54526. We are also finding correlation like correlation between $PM_{2.5}$ concentration and number of tweets for different streets/roads in different region. So, if we look for a particular region like CBD, we only have 95900 tweets which were made on road and only 24245 tweets which were made on road in 2014.

## 2.3 Intermediator

### 2.3.1 Twitter Harvester

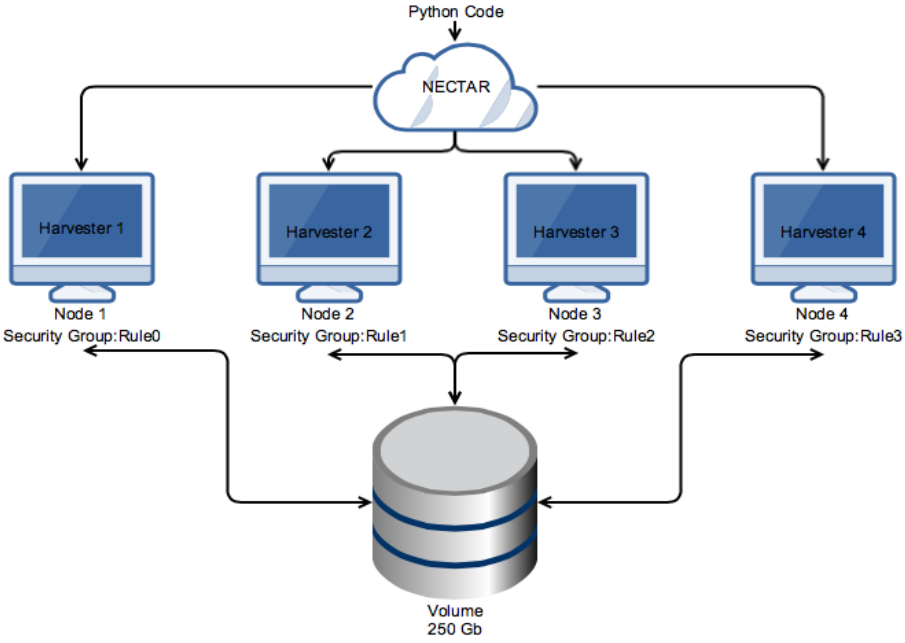


Figure 2 Twitter Harvester Architecture

We used different instances on Nectar for running Twitter Harvester using Streaming API and Rest API in order to collect more data. And as our final goal was to get tweets made on road hence we also used another harvester to collect data using Twitter Search API for tweets mentioning keywords like #trafficjam, #traffic and also getting tweets from some twitter official accounts like VicRoads (@VicRoads) because most of these tweets are made by people while driving or waiting for traffic to clear.

### 2.3.2 Converter API

Converter API includes sets of python script for converting source data format to required data format needed by python API, MySQL and Accumulo for SMASH Architecture. For example AURIN JSON format to proper JSON format suitable for Python API, JSON to CSV for insertion into ACCUMULO

### 2.3.3 Python API

Python API includes python script to read JSON file for historical tweets or doing on the fly real time processing on twitter data collected by twitter harvester. Processing tasks includes sentiment analysis and identifying the tweets that originate on the road network (PSMA) within a proximate distance of +/- 8 metres treating the centre of the road as the point of interest using Cartesian product between road line geometry and tweets geo-location coordinates and Haversine Formula.

### 2.3.4 Ruby Server (RESTful API)

We built a RESTful Ruby server in order to receive real time data from our Aircasting Application and ingest it into MySQL and Accumulo database in real time. In Figure 3, we explained our Ruby Server Architecture and then in subsequent section explained how a request and response is handled by our server. We also have included Github link containing source code and automation steps to clone, install and run our server and Youtube link explaining it's working and types of message sent and received in Appendix.

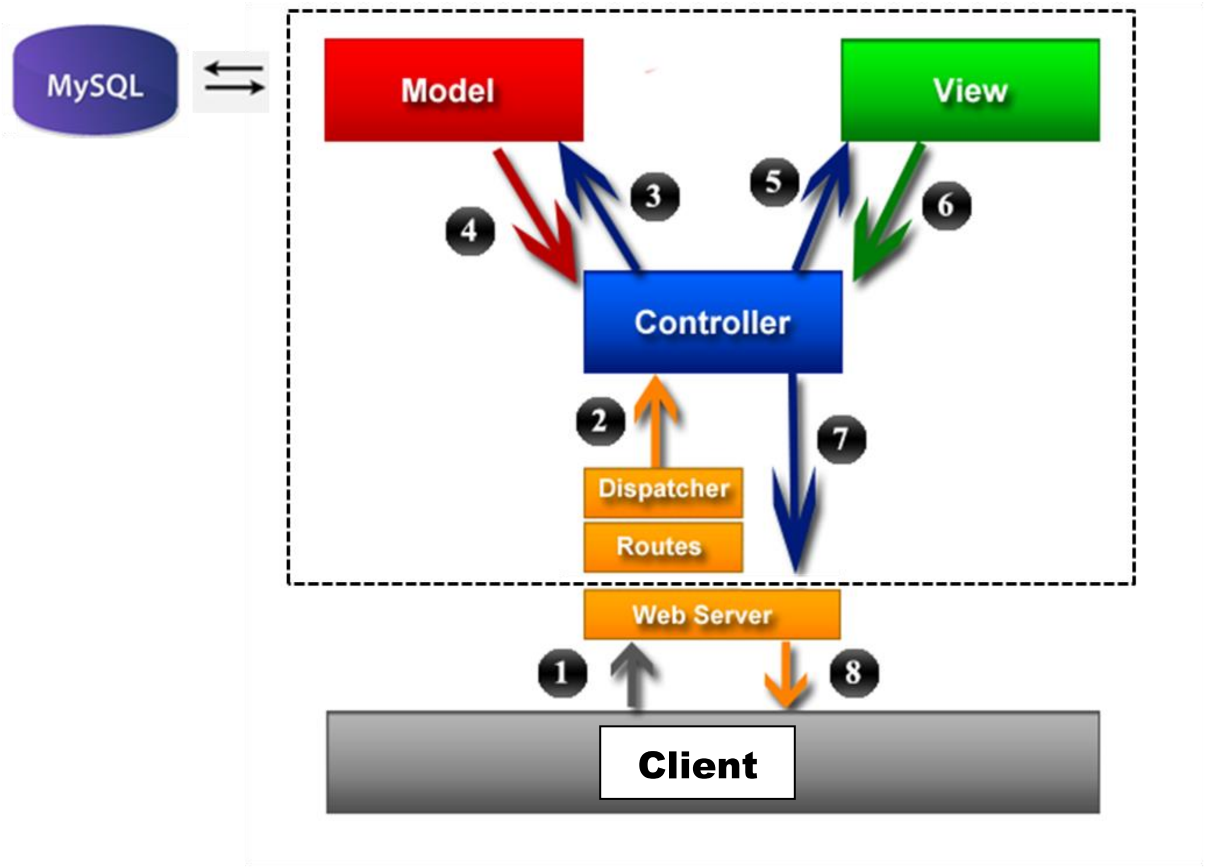


Figure 3 Ruby Server Architecture

**Following steps explains working of our Ruby server:**

1. The client (Aircasting APP and FrontEnd) sends HTTP Requests to the web server. A HTTP Request is in JSON format and it can be an authentication request, request containing measurement data or a request for retrieving particular data.

2. The web server processes the request, determines which route it belongs to and dispatches that request to the corresponding controller method.

3. The controller then asks the model layer for all the necessary information in order to be able to complete the request. Model interacts with MySQL database and if it is an authentication request check whether user exists or not and if exists and then match the supplied password otherwise if it is a request

containing data to be inserted then check whether data format is correct or not or if it is a request for data from server then retrieve required data from server and return.

4. The model layer collects all the information and returns it to the controller.
5. The controller gives the appropriate information to the view, and asks it to render
6. The view renders itself and gives the rendered JSON HTTP Requests (this can be an 200 OK message showing request was successful or a reply containing requested data) to the controller.
7. The controller assembles the reply and gives it to the web server.
8. The web server returns the JSON reply to the client.

## 2.4 BACKEND

### 2.4.1 Traditional System

I am using MySQL database to store the data sent by Python API, Converter API and Ruby Server. MySQL schema is shown in the Figure 4. FrontEnd issues different analytical query to retrieve the result from this database.

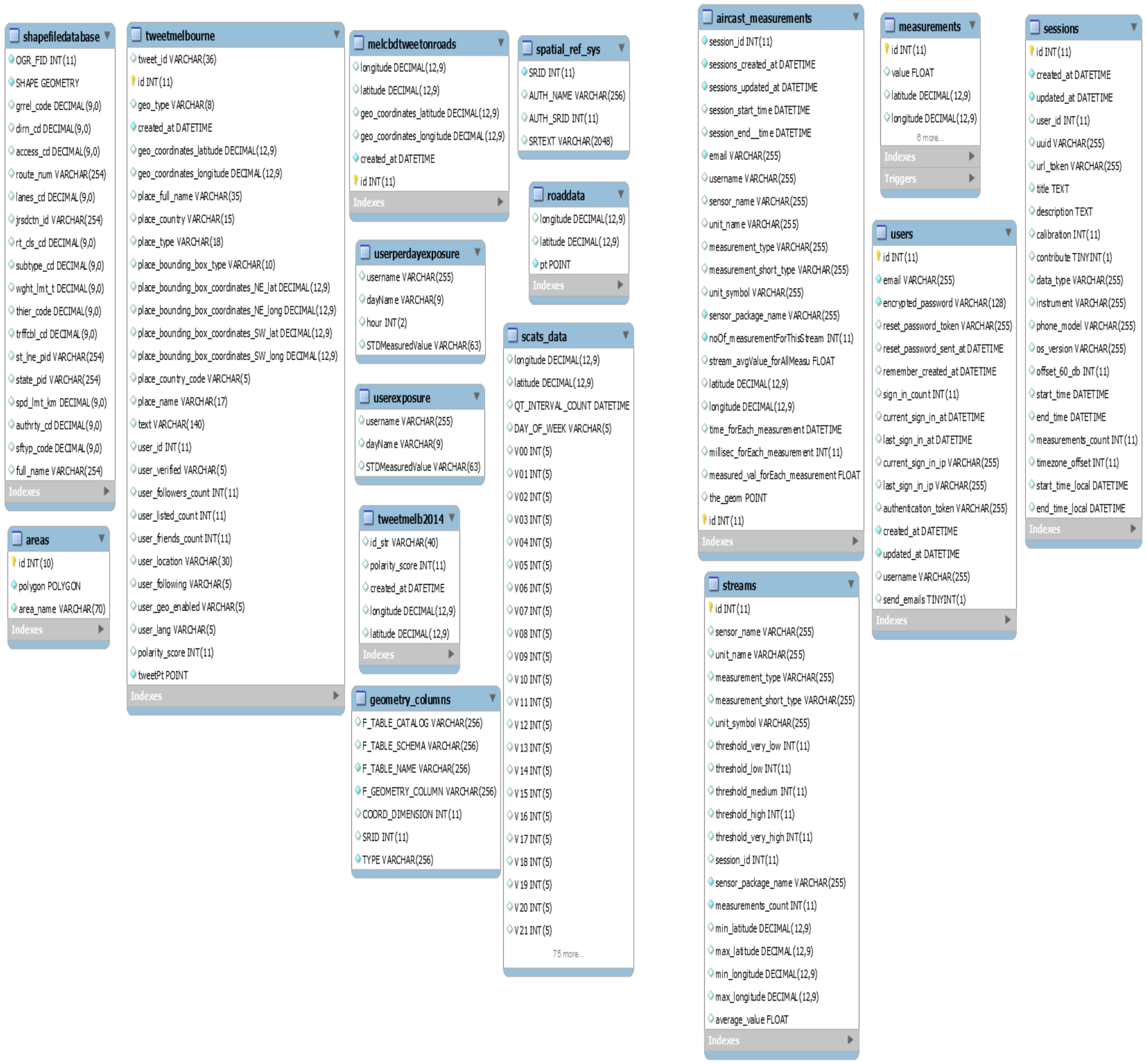


Figure 4 MySQL Schema for Traditional System

## 2.5 FrontEnd (Web API + Analytics)

We have built a Web Application to show different analytics and correlation performed on data stored in MySQL. Each analytics and correlation is explained below.

### 2.5.1 LANDING PAGE

Figure 5, shows the landing page of our Application. Its show trails of every user in the system or $PM_{2.5}$ data collected by all Citizen Scientists during their journey or region they visited.

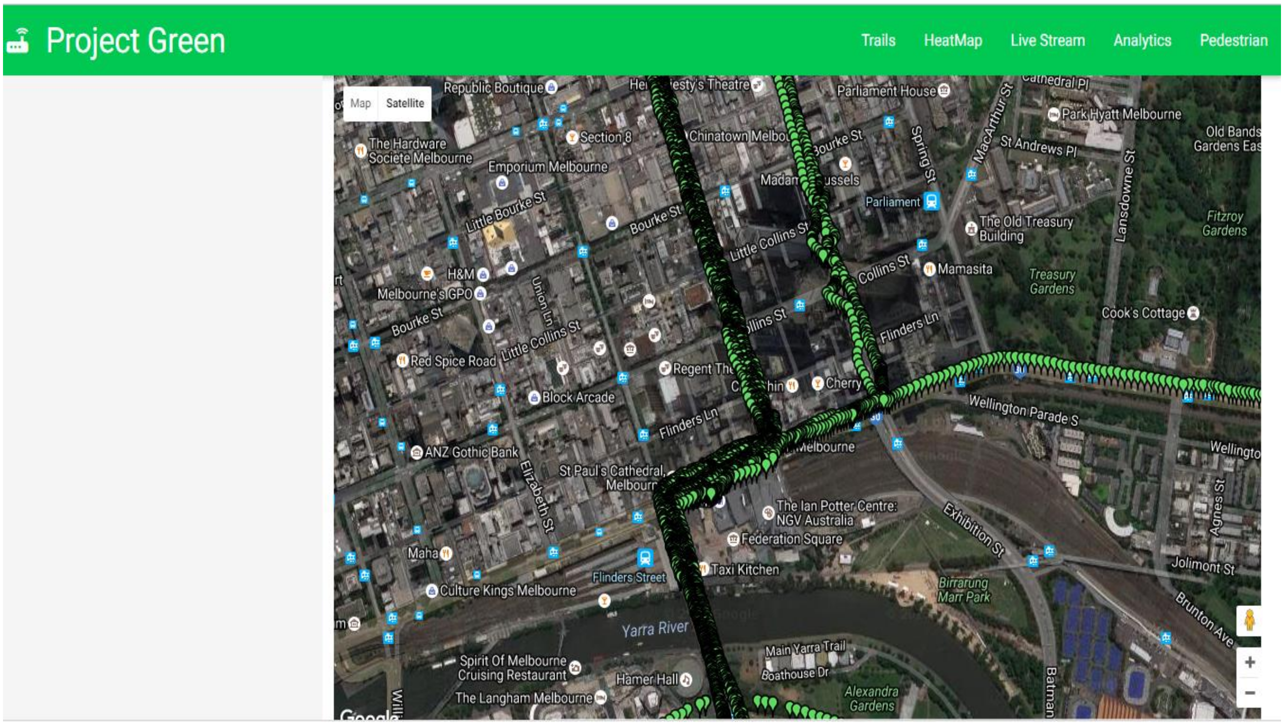


Figure 5 Landing Page for Web Application

### 2.5.2 LIVE STREAM

We can see the real time data collected by a particular citizen scientist in Live Stream mode. If person has switched on the Streaming mode on his Aircasting Application then real time data is sent to system every 15 min and we can show this data on the map in real time for any citizen scientist in the system.

### 2.5.3 HEAT MAP

We can generate Heat Map for a particular citizen scientist, showing average amount of $PM_{2.5}$ exposure for him. We grouped all the $PM_{2.5}$ reading on hourly basis over subsequent trips in a particular region and then took standard deviation for each hour and then took standard deviation for resultant standard deviation value for each hour to get 24 hour $PM_{2.5}$ concentration and based on this value we followed the table sown in Figure 7 to choose different colour showing Health Category for different regions on Heat Map. Figure 6, shows the heat map for a username rsinnott, as we can see from the figure, the amount of $PM_{2.5}$ exposure is high in CBD, Federation Square and St Kilda area as respect to other region, which is quite evident as CBD has more traffic and hence more smoke which ultimately increases the $PM_{2.5}$ concentration in that region.

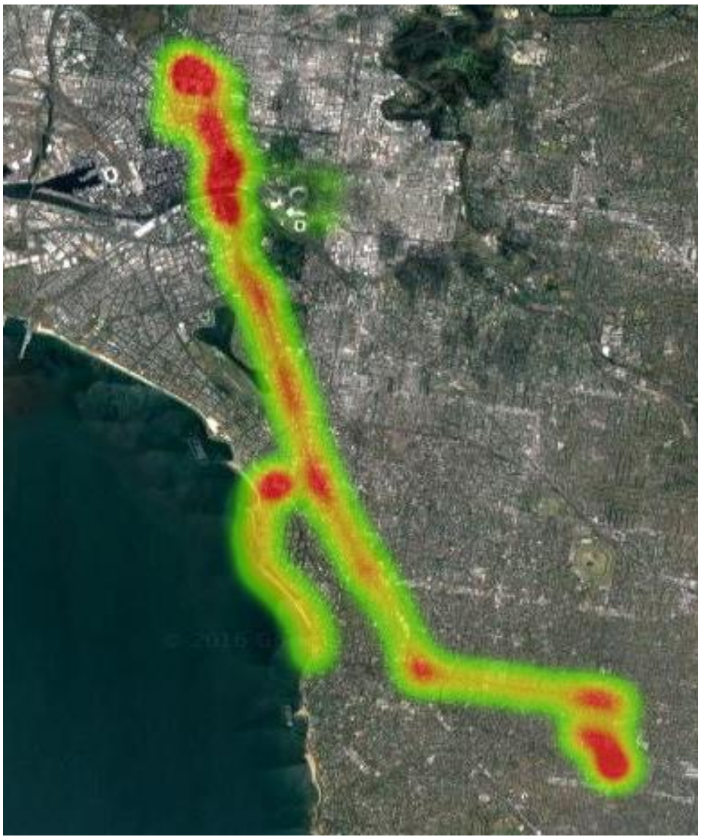
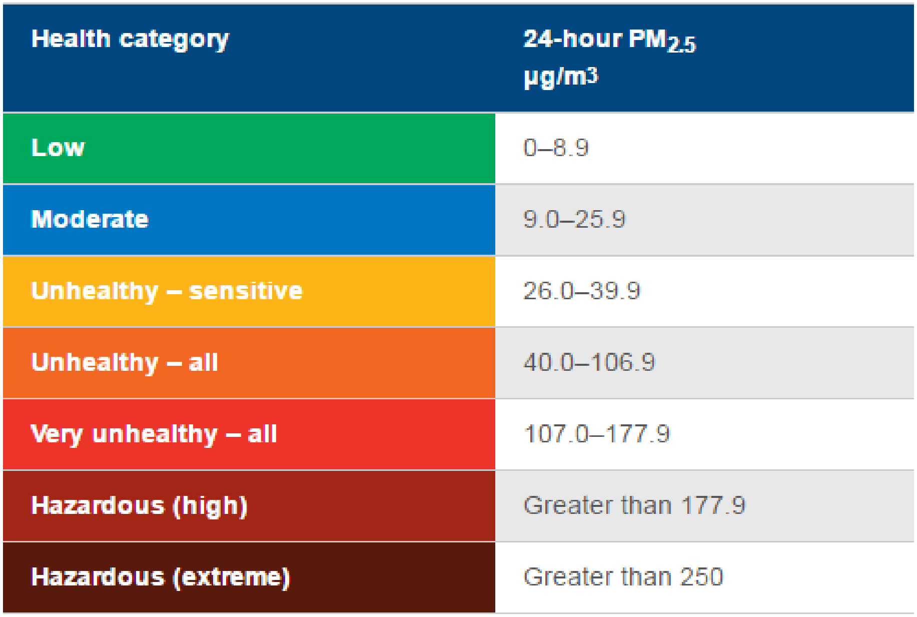


Figure 6 Heat Map showing $PM_{2.5}$ concentration in different region during regular journey of a Citizen Scientist.



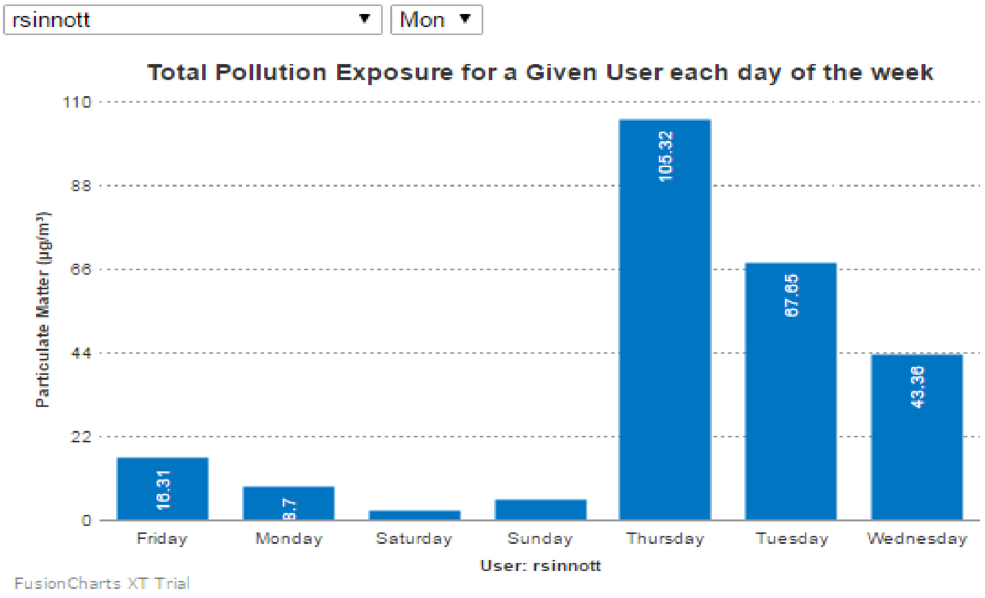
| Health category | 24-hour $PM_{2.5}$ µg/m3 |
|---|---|
| Low | 0–8.9 |
| Moderate | 9.0–25.9 |
| Unhealthy – sensitive | 26.0–39.9 |
| Unhealthy – all | 40.0–106.9 |
| Very unhealthy – all | 107.0–177.9 |
| Hazardous (high) | Greater than 177.9 |
| Hazardous (extreme) | Greater than 250 |

Figure 7 Colour Coded Chart for different Health Category for $PM_{2.5}$ concentration

### 2.5.4 ANALYTICS

#### *2.5.4.1 User Daily and Hourly Exposure*

We can generate graphs showing Daily and Hourly Exposure of $PM_{2.5}$ concentration for any user in our database. Figure 8 shows Total Pollution Exposure for username rsinnott for each day of the week, this value is generated as standard deviation of all the $PM_{2.5}$ value for each day of the week over subsequent trip for selected username. We can also generate Total Hourly Pollution Exposure for a particular day of the week and again, value is generated as standard deviation of all the $PM_{2.5}$ value for each hour for a particular day of the week over subsequent trip for selected username.



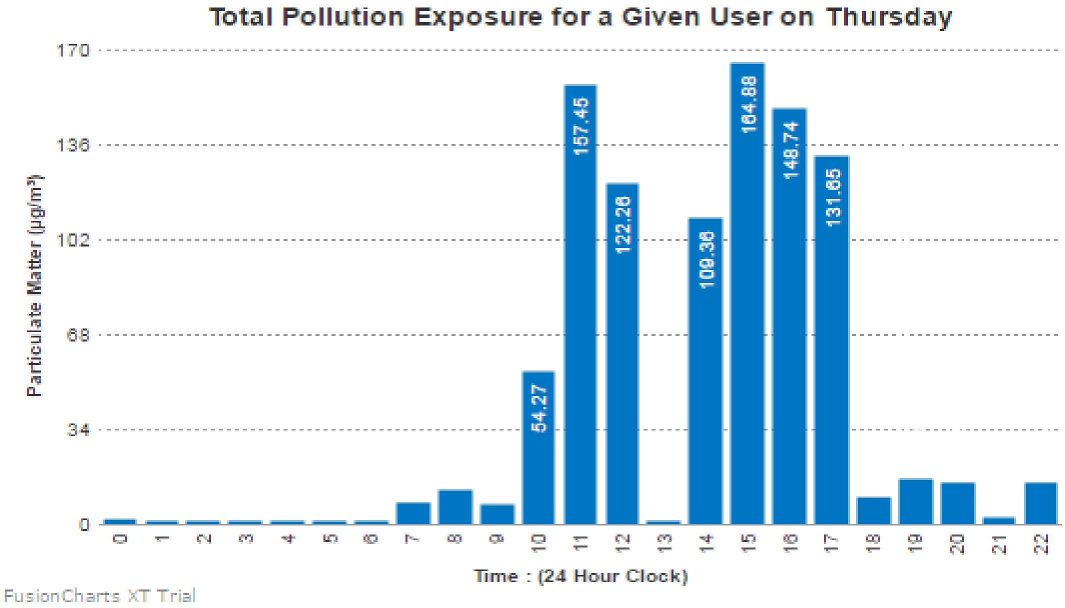


Figure 8 Weekly and Hourly Total Pollution Exposure for a given user.

### 2.5.5 Correlation

#### 2.5.5.1 *Amount of traffic vs. Number of Tweets*

In order to use twitter as a model for traffic, we have to look for correlation between amount of traffic and number of tweets made on road such that we can say that for an increase in X% of tweets we can see an increase in Y% amount of traffic. In order to get that we used the traffic information from 2014 scats data, the red dots in Figure 9, shows the location of sensors around Melbourne measuring number of vehicles passing through them every 15 minute.

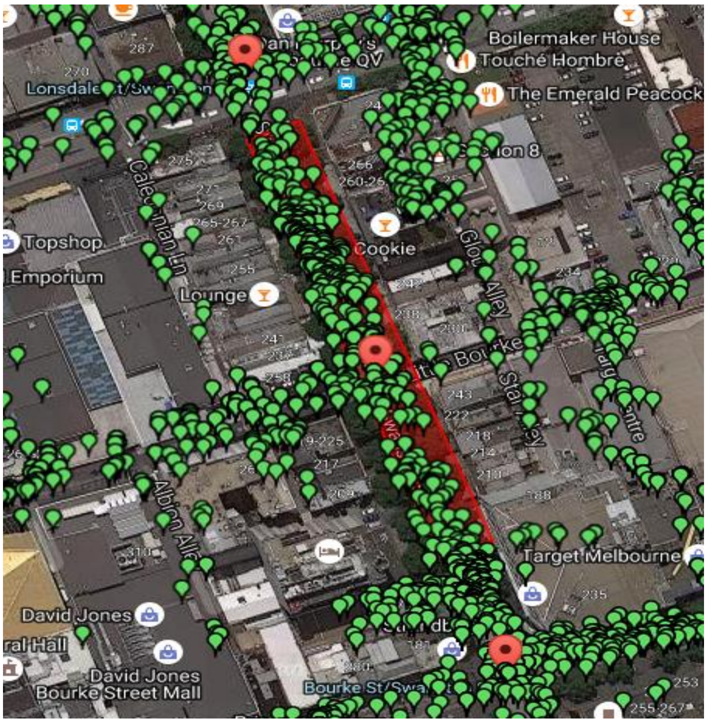


Figure 9 Street selection through a polygon for finding correlation for that particular part of street.

The green dots are the total tweets made on road from 2014 to 2016. We can select any road or street or part of road/ part of street by drawing a polygon around it. In the given Figure 9, the red polygon is drawn to select the Swanston Street. Now, in order to draw correlation between amount of traffic and number of tweets made on road, we are just considering 2014 tweets from database because we have 2014 scats data.

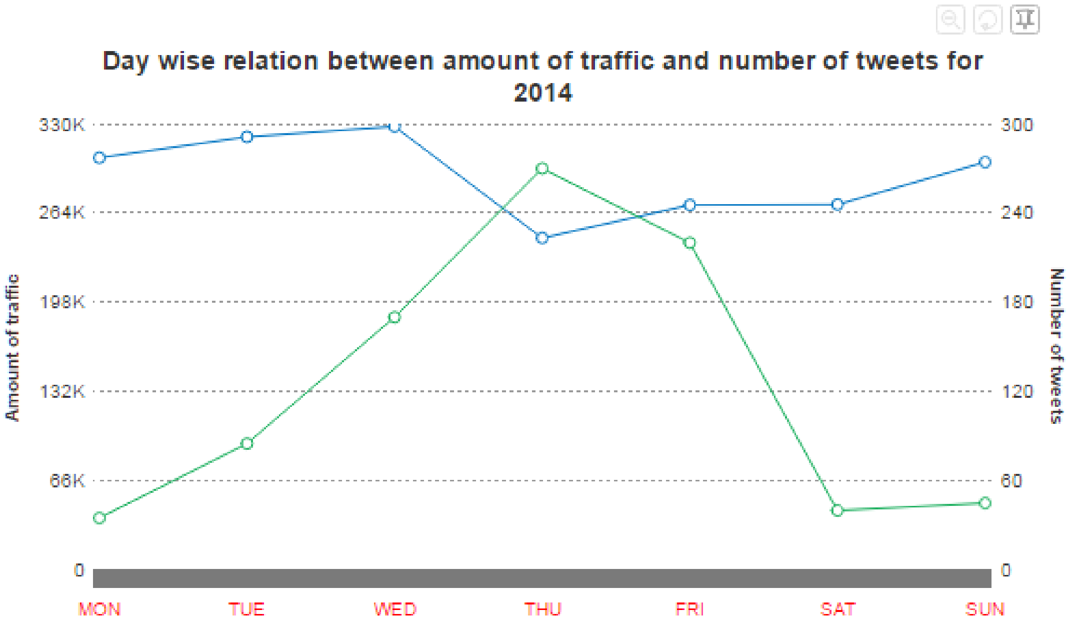


Figure 10 Day wise relation between amount of traffic and number of tweets for 2014.

As soon as you select the required street you want a correlation for, a graph is automatically generated showing Day wise relation between amount of traffic and number of tweets. Figure10 represents the graph generated for Swanston Street. We can't see any strong correlation as such; however there is an increase in traffic with number of tweets for weekdays except Thursday.

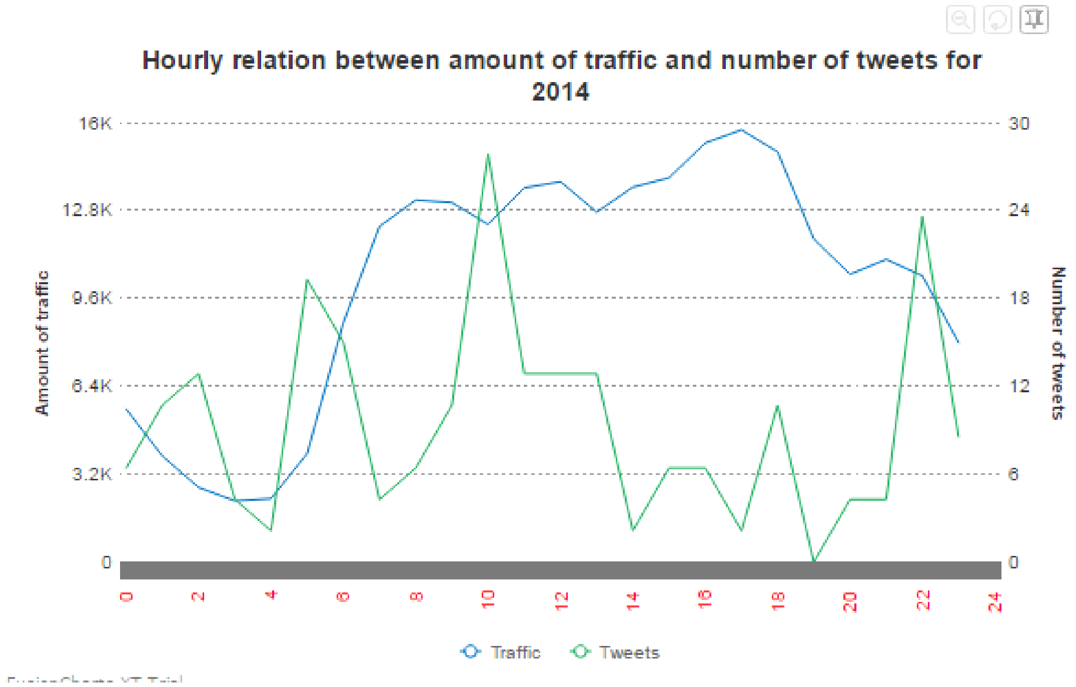


Figure 11 Hourly relation between amount of traffic and number of tweets for 2014.

However, we can also generate hourly correlation between amount of traffic and number of tweets for any day of the week. Figure 11, shows hourly relation for Swanston street on Thursday, we can see a strong correlation between both of them during peak hour or office hours i.e. 9 to 11 A.M and also when office leaving hour i.e. 4 to 6 P.M.

### 2.5.5.2 *Pollution Exposure vs. Number of tweets*

Once, we have set a correlation between amount of traffic and number of tweets. Now, we want to look for correlation between Pollution Exposure and Number of tweets made on roads. As soon as you select the required street as shown in Figure9, a graph is automatically generated showing Day wise relation between pollution exposure and number of tweets.

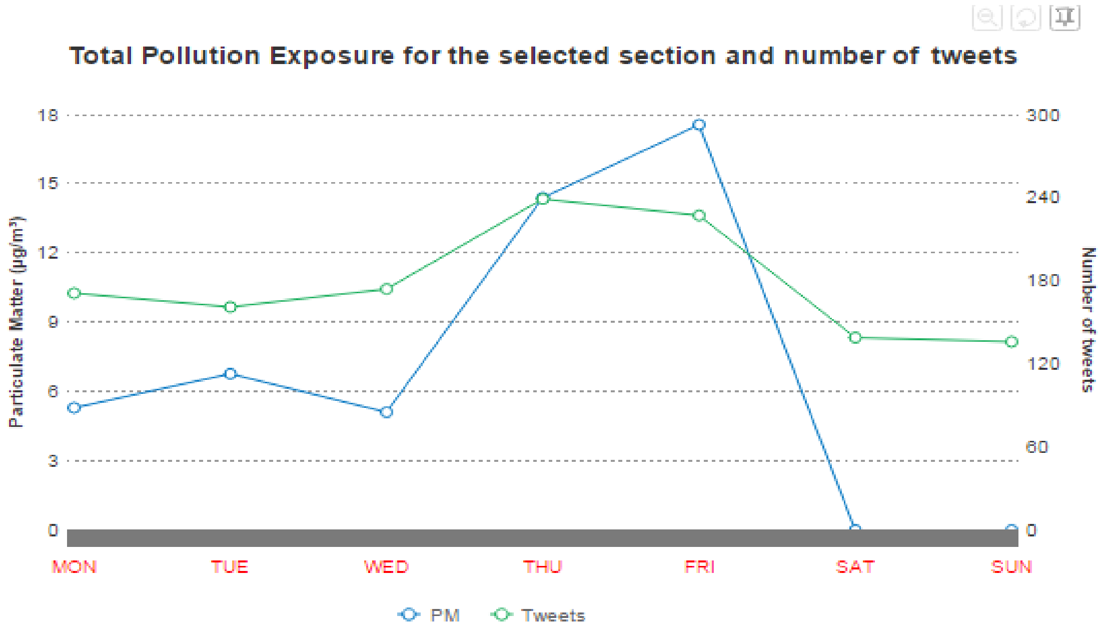


Figure 12 Weekly Total Pollution Exposure for selected section of street or road and number of tweets

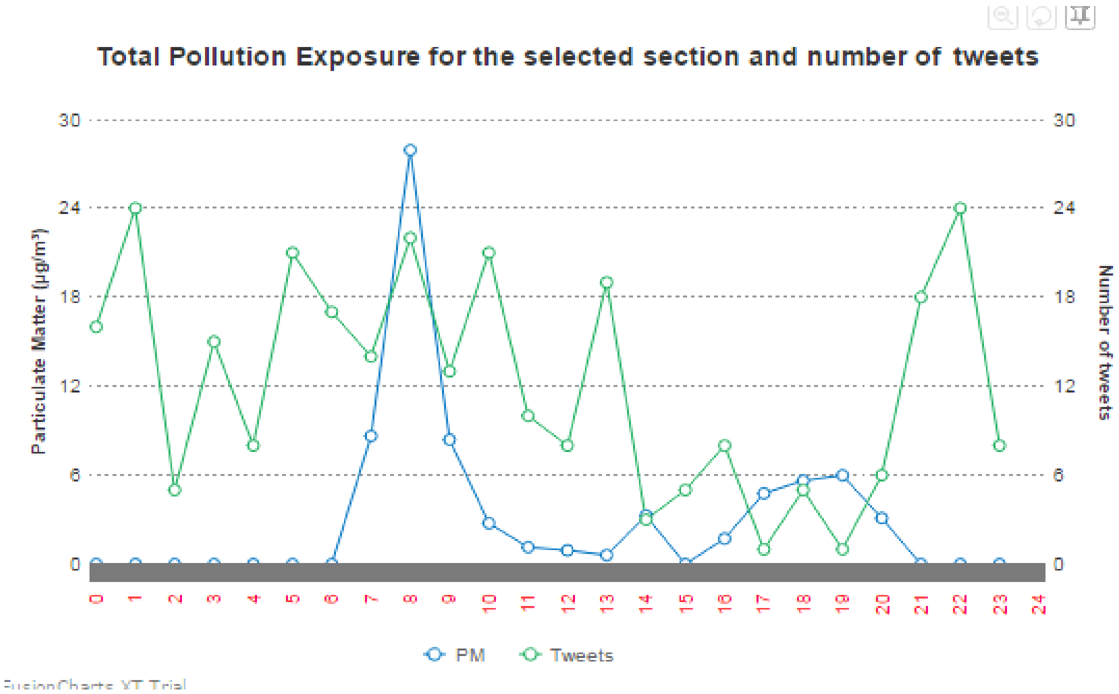


Figure 13 Hourly Total pollution Exposure for selected section of street or road and number of tweets.

Figure 12, represents the graph generated for Swanston Street. We can't see any strong global correlation for Day wise relation as such. However, we can also generate hourly correlation between pollution exposure and number of tweets for any day of the week and also we don't have pollution data for weekend hence it's shown as zero in graph it doesn't mean that pollution on weekend was negligible. Figure 13, shows hourly relation for Swanston street on Thursday, we can see a strong correlation between both of them during peak hour or office hours i.e. 6 to 9 A.M and also when office leaving hour i.e. 4 to 7 P.M. We don't have any data after 9 P.M to see any correlation at night time.

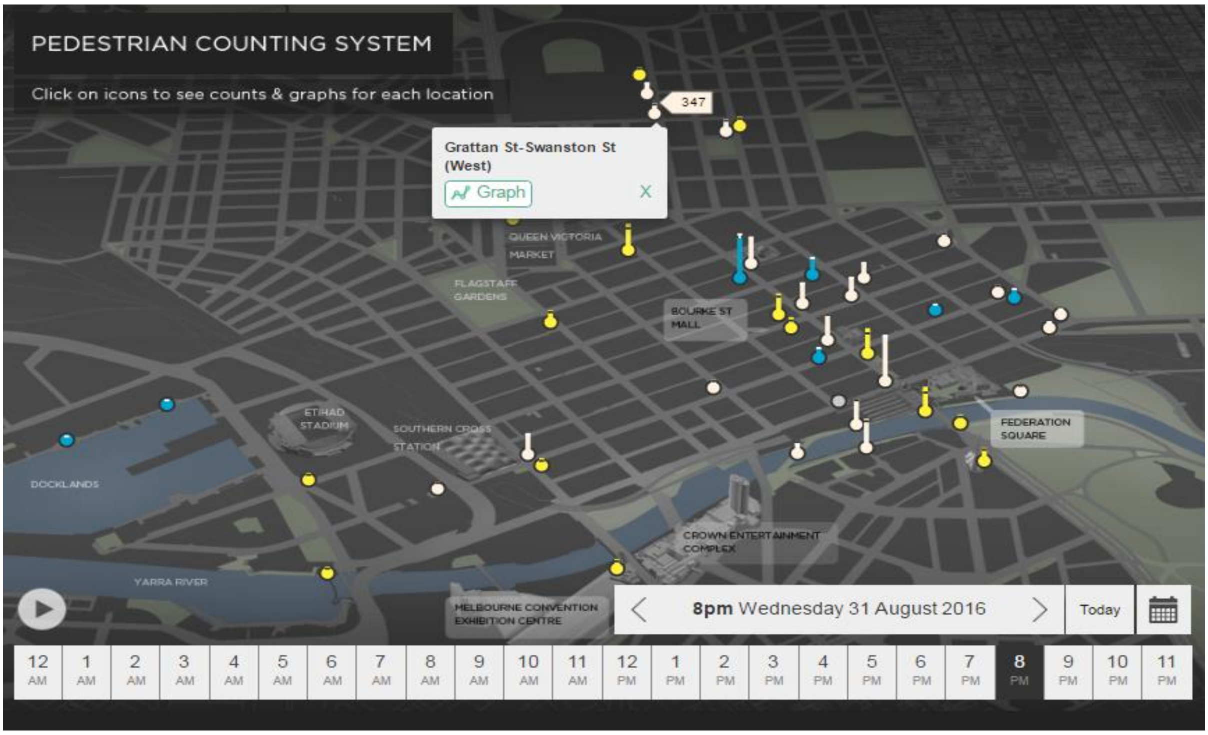


Figure 14 Pedestrian Counting system Web Application provided by City Of Melbourne.

### *2.5.5.3 Number of Pedestrian vs. PM 2.5 concentration vs. Sentiment Analysis of Tweets*

We found a correlation between pedestrian count and $PM_{2.5}$ concentration in the CBD. Correlation shows that PM2.5 concentration the increases and decreases with increase and decrease of pedestrian count. The reason behind that can be the relation between pedestrian count and amount of traffic. So, high pedestrian count signifies busy hour like office hour and hence traffic (number of cars on road) must also be high at that time.

After that we also found correlation between Pedestrian Count and Sentiment Analysis of the Tweets for different region in CBD, which shows that as Pedestrian Count increases number of negative tweets increases during busy hour.

Hence, we can say that due to increase in air pollution people are getting depressed which leads to increase in negative tweets.

For example:

For Grattan St-Swanston St(West) as shown in Figure 14, if we compare Average hourly pedestrian count over the past 52 weeks with $PM_{2.5}$ concentration. If you observe closely in figure 15, both the curves are almost same and most important thing to look is strong correlation between both of them at three different peaks in graph at 7-10 A.M (office hour), 11 A.M -1 P.M (lunch time) and 4 P.M - 7 P.M. (students and people leaving for home). All these three timeslot are actually busy hour when we have more cars on road, and we can easily see $PM_{2.5}$ concentration and Pedestrian Count both are high during this interval.

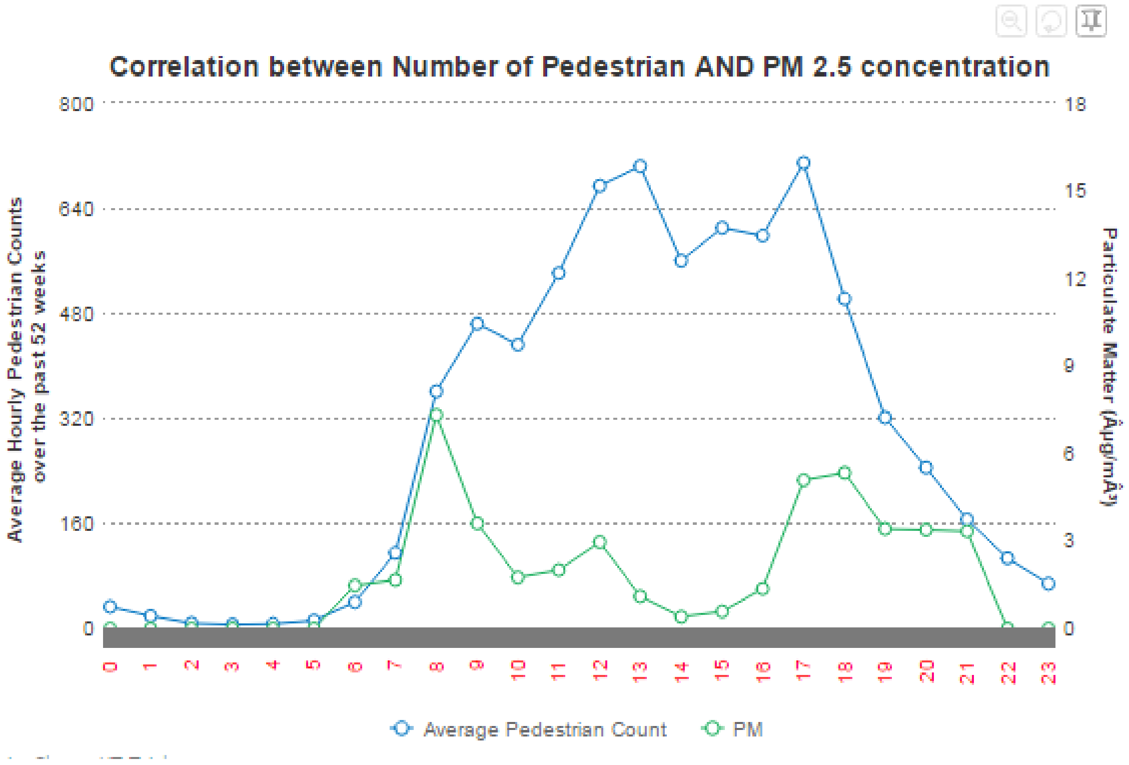


Figure 15 Correlation between number of pedestrian and $PM_{2.5}$ concentration.

Now, for same street intersection and same intervals, we can see from Figure 16, that for 7-10 A.M (office hour) we have changed from positive to neutral showing neutral at peak hour 9 A.M and for 11 A.M -1 P.M and 4 P.M - 7 P.M. we have overall negative tweets.

So, we can easily see increase in air pollution people leading to increase in negative tweets.

However, other than Air pollution other factors like crowd, noise pollution, work stress and tension (for 11 A.M to 1 P.M slot) could also lead to more negative tweets.

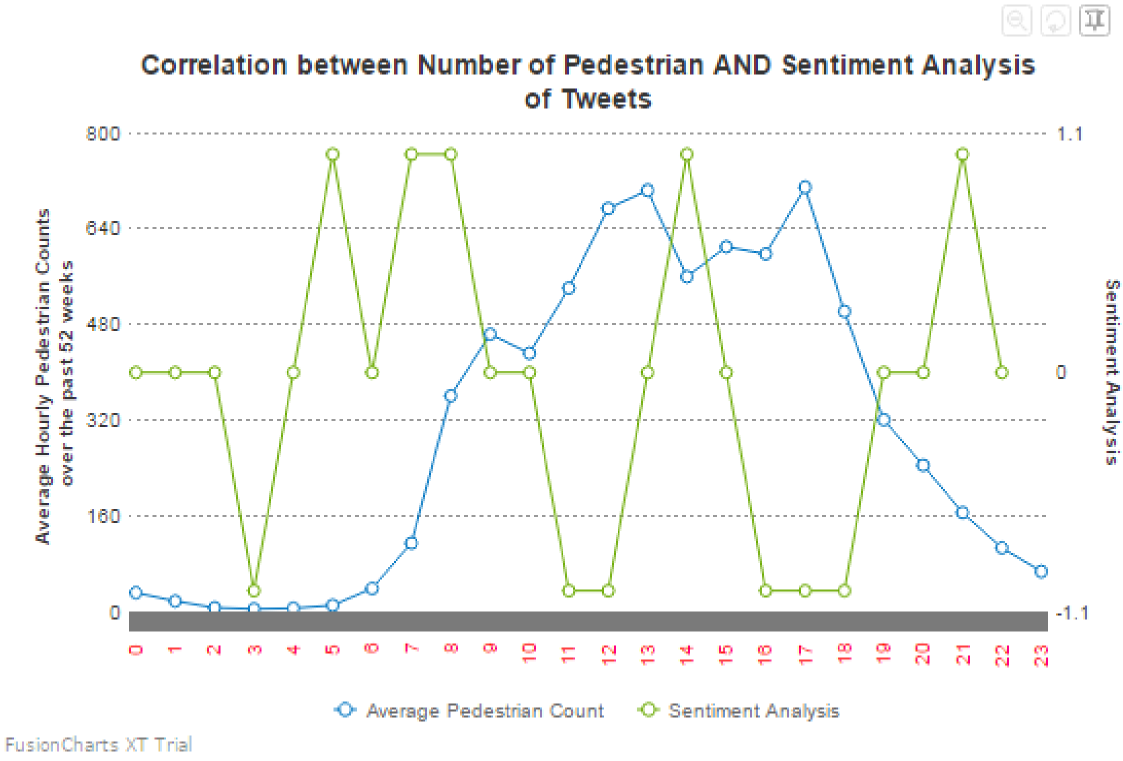


Figure 16 Correlation between number of pedestrian and sentiment analysis of tweets.

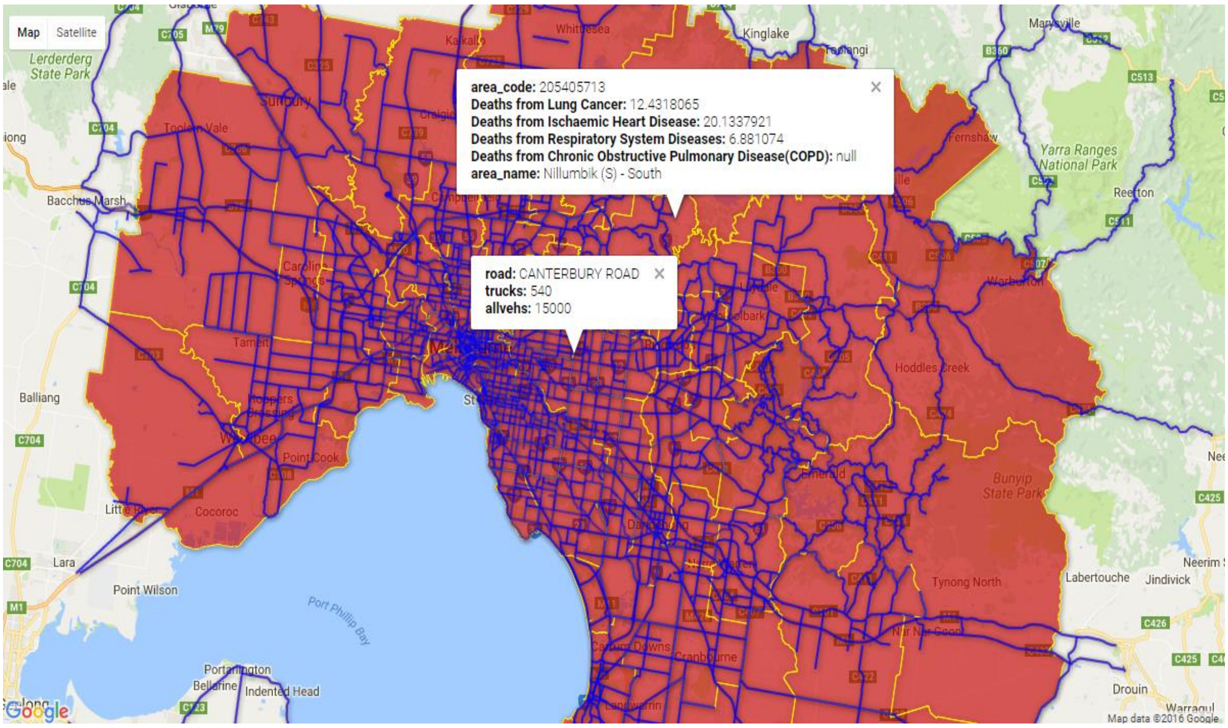
**NOTE:**

1. We took Average of all the sentiment score of the tweets made on Grattan St-Swanston St(West) road within 8-10 meter of every point line-string geographical coordinate for road for past 52 weeks and considered the value range -0.5 to + 0.5 as neutral i.e. 0, -1 to -0.5 as negative i.e. -1 and 0.5 to 1 as positive i.e. 1.

2. To determine the sentiments of all the tweets, we used Word2vec (CBOW) model. Google published Word2vec, in 2013. Word2vec is a neural network implementation that learns distributed representations for words. Training for Word2vec is relatively fast as compared to other models and Word2Vec does not require labels in order to produce meaningful representations. This is practical, because the majority data in the real world is unlabeled. We used the outstanding implementation of word2vec from the gensim package in Python. The process, method and code for sentiment analysis is published and explained on Github and included as link in Appendix.

### *2.5.5.4 Number of deaths due to diseases related to air pollution VS Amount of traffic (trucks)*

In order to find correlation between data from AURIN about deaths due to diseases related to air pollution like Chronic Obstructive Pulmonary Disease (COPD), lung cancer, respiratory diseases and Ischaemic Heart Disease and Amount of traffic, we wanted to relate them through a common field. Because, for VicRoads data we had multi-line geometry for roads to plot but death data only had SLA codes and no geometry data. In order to make spatial query and plot on map, we needed geometry for death data. For that we have downloaded, Statistical Local Area Digital Boundaries (ASGC 2006) data in ESRI Shapefile Format from http://www.abs.gov.au/ausstats/abs@.nsf/DetailsPage/1259.0.30.0022006?OpenDocument, which contains geometry of SLA and their codes, hence we used the SLA codes as common field to link death dataset and SLA digital boundary dataset.



area_code: 205405713
Deaths from Lung Cancer: 12.4318065
Deaths from Ischaemic Heart Disease: 20.1337921
Deaths from Respiratory System Diseases: 6.881074
Deaths from Chronic Obstructive Pulmonary Disease(COPD): null
area_name: Nillumbik (S) - South

road: CANTERBURY ROAD
trucks: 540
allvehs: 15000

Figure 17 Plotting dataset of number of deaths related to air pollution in different SLA11 region, SLA11 region geometry and amount of traffic volume of trucks on road lying or intersecting different SLA11 regions.

Then after this I have plotted both the dataset on a MAP as shown in Figure 17 in which:

- Red Blocks: These are SLA code or SLA areas. And if you click on it, will tell you total patients in that area having (COPD, Heart Disease and Lung cancer).
- Yellow Lines: Boundary between different SLA regions.
- Blue Lines: Showing geometry of roads around Melbourne. When you click any line inside the red box. It will tell you road name and 24 Hour Median Midweek Non Holiday Truck Volumes on that road.

We used three different approaches to find the correlation between number of deaths and amount of traffic:

- **First Approach**: We considered all the road geometry intersecting or within polygons, we are assuming that if there are X amount of Volume of trucks on a road then all those trucks will be crossing through every region which contains even a small part of road. And from Figure 18 you can see a strong correlation between them as the area where amount of traffic is large, number of deaths were also more specially in areas like Melbourne (C) Remainder, Brimback (C) Sunshine, Wyndham (C) - North, Gr. Dandenong (C) - Dandenong.

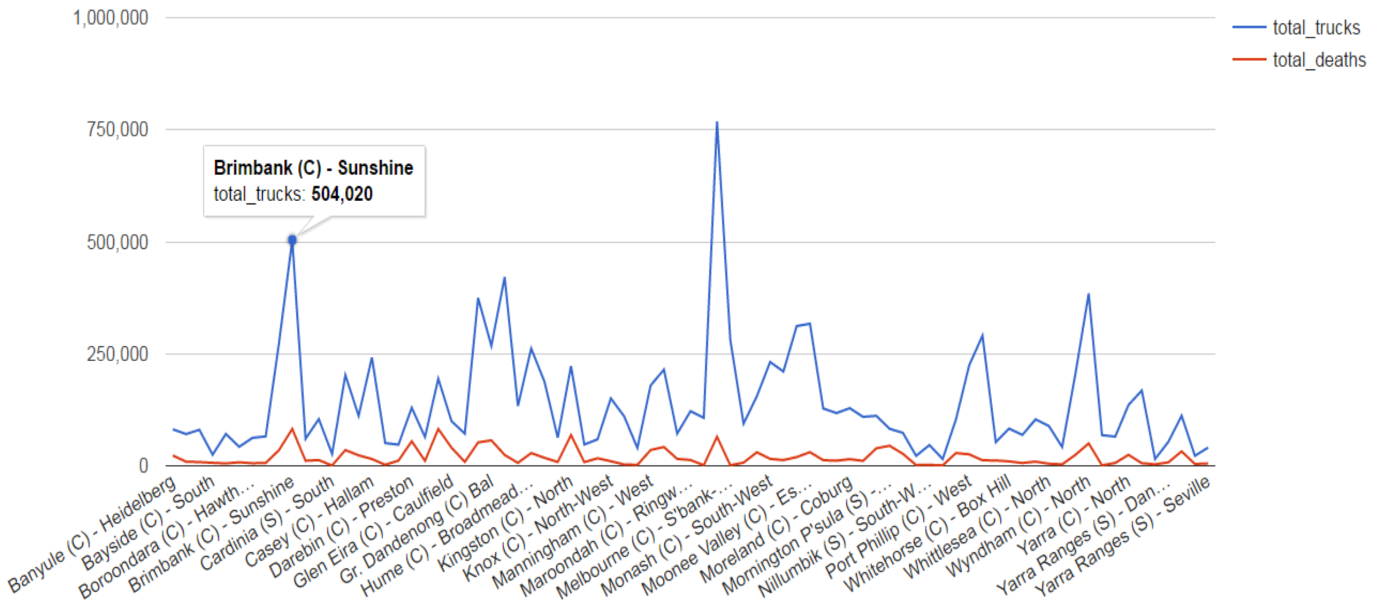


Figure 18 Correlation between total amount of volume of trucks and total number of deaths due to diseases related to air pollution for different SLA11 regions.

- **Second Approach:** We considered all the Road Geometry intersecting or within polygons but this time we are considering weight age of road lying inside a polygon if the road exist in more than one polygon, which was done based on the assumption that if a road has 350 trucks and it completely lies in an SLA then it is easy. But, if that road crosses two SLAs and one has 90% of the road and the other only 10% of the road, then we are assuming SLA with 90% road has 315 trucks and the other SLA has 35 trucks. We did this in order to make our research more close to real word scenario as in previous approach we were assuming that all the trucks on a road will cross all SLA region containing even a slight section of Road but it's not actually true in real world. Basically truck transport is from one region to another i.e. they have source and destination point, so they are not on move always. For, example DONNYBROOK ROAD which lies in both Hume (C) - Craigieburn and Whittlesea (C) - North has 510 trucks. But, may be only 210 trucks actually has source in one of the SLA region and destination in other one and hence crosses the regions. Otherwise, these both regions are so big that other 290 trucks may only use the section of road lying within this region and hence never crossed the region.

In order to find the dependence between Number of deaths due to diseases related to air pollution VS Amount of traffic (trucks) for all the above two approaches we did Regression Analysis which helps understand how the dependent variable changes when any of the independent variables is varied, while any other independent variables are held fixed. And Table 2 shows adjusted $R^2$ values which give you an idea of how many data points fall within the line of the regression equation  for the two approaches.So, we can see that our first approach is better in estimating the relationships between Number of deaths due to diseases related to air pollution and Amount of traffic (trucks). However, we observed a strong correlation for much location from plot in Figure 18. Despite that our adjusted $R^2$ value came between 40 - 60 % for different approaches which is not bad. However, it could be better but we had problem that data about amount of Volume of traffic was for roads unlike death data which was based on SLA11 regions, hence a particular road was coming inside more than one SLA11 regions. We tried a workaround for this problem in Approach 2 but it actually reduced our Adjusted $R^2$ by about 20% as compare to Approach 1, which shows that our assumption of considering truck value by percentage of road section in different SLA region is not correct. The reason behind that is if you closely observe the plotting on Map of both dataset in Figure 17 as we go deep near CBD in Melbourne size of roads and SLA region has decreased drastically and hence our assumption of Approach 1 is becoming more real, as portion of road lying inside other region is becoming very small, in fact number of roads contained inside a SLA region is far more than intersection roads.

We could have achieved better Adjusted R-Squared value, if we could have known exact estimation about amount of volume of traffic for different SLA11 region roads instead for roads. Another reason for overall low Adjusted R-Squared value for all SLA11 regions can be missing values for COPD and Respiratory diseases for most of SLA11 region.

| Approach | Adjusted R-Squared |
|---|---|
| All the Road Geometry intersecting or within polygons | .6131 |
| All the Road Geometry intersecting or within polygons considering weight age of road lying inside polygon. | .4061 |

**Table 3** Adjusted R-Squared values as a result of regression test for our different approach.

# 2.6 BACKEND

## 2.6.1 Big Data Technologies

### *2.6.1.1 SMASH Architecture*

We have to examine large amounts of data in order to uncover all the above hidden patterns, correlations and other insights in real time. The traditional system based on MySQL we have been using is slower and less efficient. On the other hand with big data technology, it's possible to analyse our data and get answers from it almost immediately.

For example, most expensive operation for our analytics was to find tweets made on the road from a collection of tweets as it involved a Cartesian product of line geometry of roads and geo-coordinate of tweets. It was so computationally expensive that it was impossible to do it for whole Melbourne with our traditional system even after one day process was not able to complete. Hence, we have to divide Melbourne into regions for example CBD and then do the Cartesian product separately for every region. We used MySQL cross join for Cartesian product and it took 16 hours for CBD which was about 10 million rows to find all the tweets made from road for CBD region. However, when we did same operation with SPARK it took 2 hours. So, according to our benchmarking SPARK was 8 times faster using (1 node, 4 core, 16 GB RAM) for CBD, with total 10 million rows in database.

This rapid increase in efficiency motivated us to move to big data platform. We created our own Architecture known as SMASH Architecture. This Architecture is more efficient, faster and scalable as compared to our traditional system. SMASH Architecture uses GeoMesa which is distributed, spatio-temporal database built on distributed cloud data storage systems Accumulo and HBase which makes storing of large spatio-temporal datasets highly efficient. And as our data is spatio-temporal as we are dealing with real time data which changes with movement, Geomesa/Accumulo is far more beneficial with added advantage of spatial querying then MySQL as MySQL has limited spatial support. SPARK is used for providing faster computation and we are using Zookeeper a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services for maintaining our distributed applications. Visualisation of data is handled by GeoServer which provide capability for real-time aggregation and area selection.

As we have, already seen that SPARK was 8 times faster by just using (1 node, 4 core, 16 GB RAM). But with SMASH architecture we can increase capability of SPARK by using more NECTAR nodes and hence make this operation way faster. By, this simple operation you can imagine the power of Big Data technology.

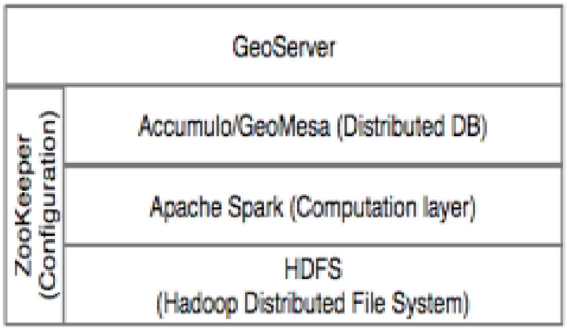

GeoServer
ZooKeeper (Configuration)
Accumulo/GeoMesa (Distributed DB)
Apache Spark (Computation layer)
HDFS
(Hadoop Distributed File System)

Figure 19 SMASH Architecture

## 3 Future Research Direction

In this report, we did analytics based on traditional system; we explored different correlation between number of tweets, amount of traffic and pollution data. We also give suggestions for the future research direction on this topic:

1. We think that using current scats data for 2016 for correlation could increase the efficiency as we could easily gather large amount of data from twitter for recent year. As, we are looking specifically for tweets made on road in 2014, which include two filter i.e. tweets must be geo-located and must lie on road due to which our data subspace got decreased substantially.
2. Collecting more pollution data around Melbourne. At, present we have data for main region or main streets and in between certain time range however if we want to see global correlation, we have to collect data for whole Melbourne.
3. Till now we were only able to benchmark only some part of the traditional system w.r.t to SMASH Architecture and it was found really fast efficient. However, doing Benchmarking by doing full analytics would make our statement for moving to Big Data Platform more convincing.

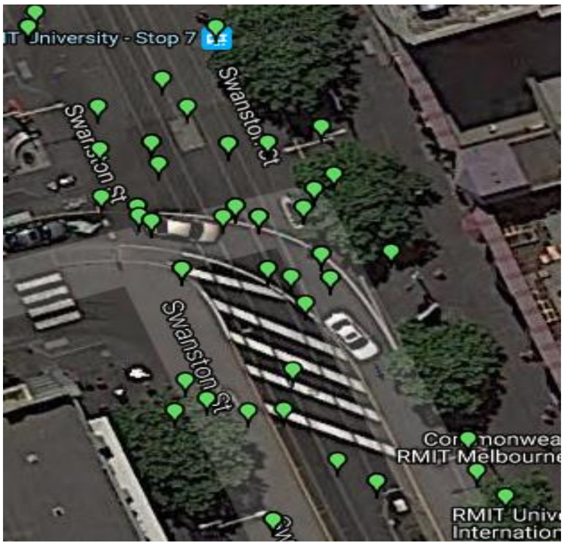


Figure 20 High resolution visualization of tweets made on road.

4. Another important future direction would be to find out the direction of car while the tweet was made and also whether the tweet made from a tram, which is very simple using my approach of plotting the tweet, as tweets made on left side of the road will be opposite in direction to the tweet made on right side of the road. However, it's impossible to do it with traditional system as it is very computational expensive but it is very much.

# 4 Conclusion

In this report we designed a traditional system to analyse to what extent we can use Twitter data to model traffic and then correlate number of tweets with our pollution data. And then based on these analytics we showed that amount of computation and processing power need to analyse this large amount of data in real time requires a Big Data Architecture SMASH or system for fast, scalable and more efficient analysis. Some conclusions and findings can be seen as follows:

1. We showed the regions with high $PM_{2.5}$ concentration through or Heat Map.
2. We found a strong correlation between increase in amount of traffic and number of tweets made on road in morning during office hours i.e. 8-10 A.M and also when people leaves from their work from 4-6 P.M. On weekdays especially on sat night from 6 P.M to 9 P.M again strong correlation was seen.
3. We found a strong correlation between increase in $PM_{2.5}$ concentration and number of tweets made on road in morning during office hours i.e. 8-10 A.M and also when people leaves from their work from 4-6 P.M on weekdays. On weekends especially on sat night from 6 P.M to 8 P.M again strong correlation was seen.
4. We found a correlation between Pedestrian Count, $PM_{2.5}$ concentration and Sentiment Analysis of the Tweets for different region in CBD. Correlation shows that PM2.5 concentration and number of negative tweets increases and decreases with increase and decrease of pedestrian count during busy hour. Hence, we can say that due to increase in air pollution people are getting depressed which leads to increase in negative tweets.

   For example, for Grattan St-Swanston St (West) we compared Average Hourly Pedestrian Count over the past 52 weeks with $PM_{2.5}$ concentration. We found strong correlation between both of them at three different intervals at 7-10 A.M (office hour), 11 A.M -1 P.M (lunch time) and 4 P.M - 7 P.M. (students and people leaving for home). All these three timeslot are actually busy hour when we have more cars on road, and we can easily see $PM_{2.5}$ concentration and Pedestrian Count both are high during this interval. Now, for same street intersection and same intervals, i.e. 7-10 A.M (office hour) sentiment changed from positive to neutral showing neutral at peak hour 9 A.M and for 11 A.M -1 P.M and 4 P.M - 7 P.M. we have overall negative tweets/sentiment. So, we can easily see increase in air pollution causing increase in negative tweets. However, other than Air pollution other factors like crowd, noise pollution, work stress and tension (for 11 A.M to 1 P.M slot) could also lead to more negative tweets.
5. We also found a correlation between Numbers of deaths due to diseases related to air pollution VS Amount of traffic (trucks). We saw a strong correlation between them as the area where amount of traffic was large, number of deaths were also more specially in areas like Melbourne (C) Remainder, Brimback (C) Sunshine, Wyndham (C) - North, Gr. Dandenong (C) - Dandenong.
6. We also find out that our SMASH Architecture performs way better than traditional system. For example the most computational task of finding number of tweets made on road which require a Cartesian product between tweets geo-coordinate and roads line geometry is 8 times faster with Spark using (1 node, 4 core, 16 GB RAM) for CBD which is about 10 million rows as compared MySQL cross join.
7. Finally we also provide some suggestions for the future direction for improvement of our current analysis and correlation and also ways to find the direction of cars or tweets for more efficient correlation.

# References

[1] Begg, Vos, Barker, Stevenson, Stanley & Lopez, The burden of disease and injury in Australia 2003, Australian Institute of Health and Welfare, Cat. no. PHE 82, Canberra (2007), p234, available at http://www.aihw.gov.au/publication-detail/?id=6442467990.

[2] Senate Community Affairs References Committee, Parliament of Australia, Impacts on Health of Air Quality in Australia, 2013, p3.

[3] Doctors for the Environment Australia, Submission no 4 to Senate Community Affairs References Committee, Parliament of Australia, Impacts on Health of Air Quality in Australia, 2013, pp5,8; World Health Organization, Health Aspects of Air Pollution with Particulate Matter, Ozone and Nitrogen Dioxide, Report on WHO Working Group (2003) pp5-6.

[4] World Health Organization, Global Health Observatory: Ambient Air Pollution available at http://www.who.int/gho/phe/outdoor_air_pollution/en/.

[5] National Pollution Inventory, Substance Fact Sheet: Particulate Matter, available at http://www.npi.gov.au/resource/particulate-matter-pm10-and-pm25.

[6] National Pollution Inventory http://www.npi.gov.au/resource/particulate-matter-pm10-and-pm25; CSIRO Submission no 48 to Senate Community Affairs References Committee, Parliament of Australia, Impacts on Health of Air Quality in Australia, p8.

[7] https://pursuit.unimelb.edu.au/articles/taking-a-city-s-pulse-touch-ons-transactions-and-tweets.

[8] Sinnott, Richard O., and Shuangchao Yin. "Accident Black Spot Identification and Verification through Social Media." *2015 IEEE International Conference on Data Science and Data Intensive Systems*. IEEE, 2015.

[9] Malika, Madelin, and Duché Sarah. "Low cost air pollution sensors: New perspectives for the measurement of individual exposure?."

# Appendix A. Source Code and Videos

## Code

1. Website
   - https://github.com/Arjun-Chaudhary/projectGreen
2. Ruby Server
   - https://github.com/Arjun-Chaudhary/rubyServer

## Videos

1. Website
   - https://youtu.be/EUTx5sWYBUc
2. Ruby Server
   - https://youtu.be/s2h5AMmG-TE