Tracking people flow in cities for services optimization

CityFlow group

16è Concurs d'Idees Ambientals i Sostenibles de la UPC

27/03/2015

TABLE OF CONTENTS

1 SUMMARY

Identifying the most crowded places in a city in a given moment can be very useful for deciding how to manage a fleet of taxis, how to plan the cleaning of the streets or how to optimize the frequency of a bus route. These are some examples of situations in which a good knowledge about the mobility of people around a city can lead to a more sustainable environment.

Data is mainly gathered from social networks, like Instagram or Twitter, and combined with information obtained from the different sensors of people flow deployed around the city. The mixture of these two sources of data allows to capture and extract the main patterns of people behavior, and also forecast future events based on historical knowledge. After some processing, the data is stored in a database ready to be delivered to the user by means of a website or an application, whatever channel the customer prefers in order to access the information.

One of the keystones of the project is data. In such data-driven projects, huge amounts of data have to be dealt with in a very short period of time to provide the best service to the user. In this scenario, *Big Data* techniques are used to accomplish the two-fold requirement. Additionally, prediction of future events can also be performed from the historical data to obtain further value from this data.

The project targets two different types of users. Firstly, regular citizens could use the information generated to know, for example, which is the trendiest neighborhood in the city in a particular moment. On the other hand, institutions and companies could benefit from highly processed information in order to optimize their resources or predict future demands.

2 PROJECT DEVELOPMENT

2.1 CONCEPT DEFINITION

Knowing at each moment which are the most crowded places in a city can give value in a range of different ways and for different agents in the management of a city.

Our approach to obtain this information is to acquire data from two main sources: social networks and sensors. Those are two very different kinds of data which have to be aggregated in a clever way in order to give the most accurate results. Finally, other kinds of data sources can be used such as web scraping in event sites and message boards; processing of live camera feeds, number of mobile terminals connected to a base station, etc.

Using different tools and techniques all this data is processed in a short time to extract the best value out of it and then stored, waiting for its consumption by the users of the platform.

Users will then be able to access this information in a number of different ways, depending on its preferences and the way it will be used, as will be further explained in Section 2.5.

This project can be divided in different parts according to different steps that appear in a data processing platform. Those include data acquisition, cleaning, processing, storage, visualization and delivery.

All that means having lots of data to acquire, process, store and deliver in a very small amount of time. The way to deal with that is using Big Data tools for the different steps composing a data platform as the one that this project should be running on, from acquisition to storage without forgetting the processing of the data.

The most used solution for Big Data platforms is Apache Hadoop. It is being used by companies such as Facebook, Amazon or Netflix for different purposes but also by researchers in many different fields, from economics to astronomy or even genomics.

In this document a proof of concept is described for a data platform in the city of Barcelona. It includes all parts of the project but simplified as it only uses geo-localized Instagram posts as a source and an area limited in size so data volumes are small enough to be treated with simpler tools.

The next figure shows a schema of the different parts of the data platform already implemented in this project.

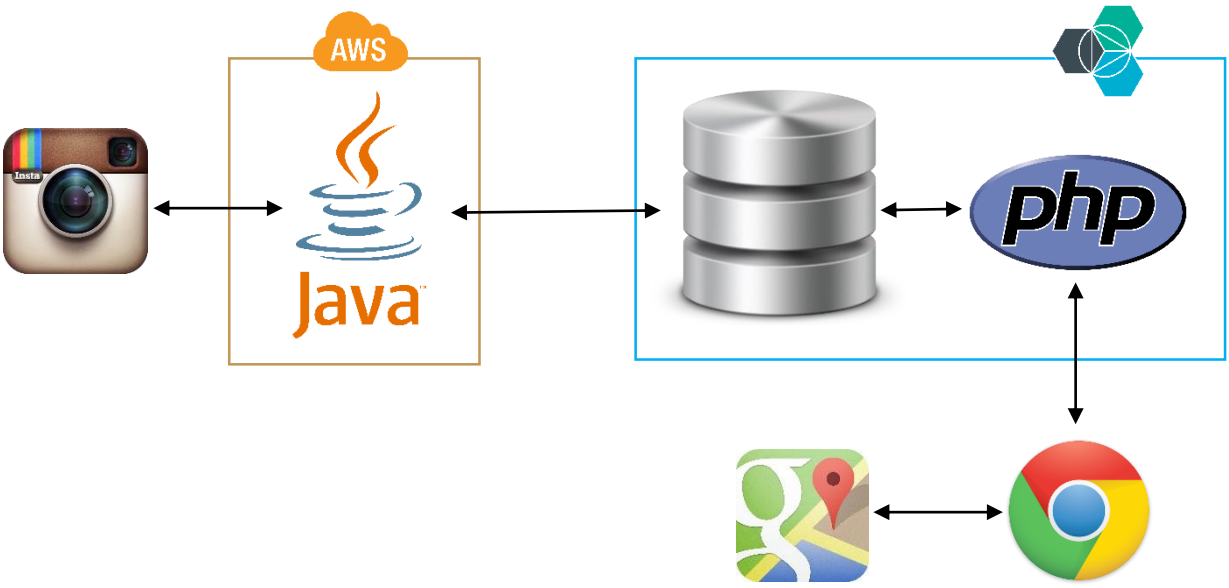

Figure 1. *Schematic of the current proof of concept implementation.*

2.2 DATA ACQUISITION

In our proof of concept the data is acquired exclusively from Instagram. Instagram API is queried to retrieve the last posts in the city of Barcelona periodically. Those posts are then processed and saved to the database.

Adding more data sources is needed for increasing the reliability of the information obtained. However, different sources mean different kinds of data with different natures, structures, etc. A good management of the data flows is compulsory when working with different data sources.

In addition, this aggregation results in a huge data set containing useful data for our purposes but also information that provides no value at all. Thus, cleaning the data set is needed to ease subsequent steps in our way to the useful information for the consumers.

2.3 PROCESSING

By processing vast amounts of data great insights can be obtained from it and it is this information what creates value. In our proof of concept the processing is very simple as we only classify the Instagram posts by neighborhoods and districts depending on their location.

More complex processing including machine learning techniques can allow discovering more information hidden in the data, but it requires using bigger data sets as well. Moreover,

this data has to be somehow mixed to obtain the final results that will be useful for our clients. This is part of the processing. As mentioned, this means heavy computation of vast amounts of differently structured data and tools like Hadoop need to be used. Processing in Hadoop is normally done via MapReduce or Spark programs. Mahout is the Hadoop java library for machine learning. Prediction features in a data platform like this could be a key resource for it as enabling the user of this information to act proactively can really make a difference.

2.4 STORAGE

This is one of the crucial parts when we are handling with huge amounts of data. In the current proof of concept design, it has been implemented using a SQL database taking advantage of its ease of use and the previous experience of several team members on this technology. However, once the project scales up, other approaches have to be used.

In the Hadoop ecosystem different solutions exist for implementing high-availability, high-performance databases in a distributed manner allowing big amounts of data. One solution that uses the distributed file system of Hadoop (HDFS) is HBase, which provides a database on top of it. Having a distributed database in front of the classical database that is in a certain server in a specific physical location has many advantages, being it the most clear one the possibility of storing vast amounts of data using clusters formed by commodity hardware nodes or even virtualized machines in the cloud, thus reducing the cost of storage critically. Also, this new distributed approach can solve the problem of congestion when one single database could be overwhelmed due to the huge amounts of data that it has to process, so it's a very good solution for Big Data projects like the one considered here.

2.5 DELIVERY

The distribution of data is one of the most essential activities the project has as it allows to reach the customers. The delivery of information will be specifically designed for the two main user targets we have identified

The first group of users is composed by the regular citizens. These are driven by curiosity mainly. One of the cornerstone aspects is the way data is presented to the user. Therefore, a very graphical website and a simple mobile application are utilized to reach them. An initial proof-of-concept of the website is already created and can be accessed at http://cityflow.mybluemix.net.

The second group of users is formed by professionals, either public institutions such as city councils or private companies. Data can be delivered in two flavors, with or without processing. Unprocessed data will give the ability of data mining the raw information to the user and extract the best value out of it. Conversely, access to processed data will free the user from implementing their processing methods, and thus the user will only have to deal with the interpretation of this data. Likewise with regular citizens, we offer the possibility of building the software to visualize the data.

Finally, regardless of the customer segment, we can provide access to real-time data and also to forecasts about future events. This will satisfy a wide range of requirements that the user may have.

2.6 VISUALIZATION

In this proof of concept, only the real-time data is shown. We do not implement any historical statistic, or make any prediction. However, in the real application, depending on the customer segment that we want to reach (municipalities, final users, etc.) the visualization may vary depending on their requirements.

The visualization in the current application is based on Google Maps v3 API. This is, we take a raw map from Google database and then we customize it. When the user enters to the webpage, a Heat Map of Barcelona is shown. With this information the user can see the distribution of uploaded posts on Instagram in the last 30 minutes, in a very visual way.

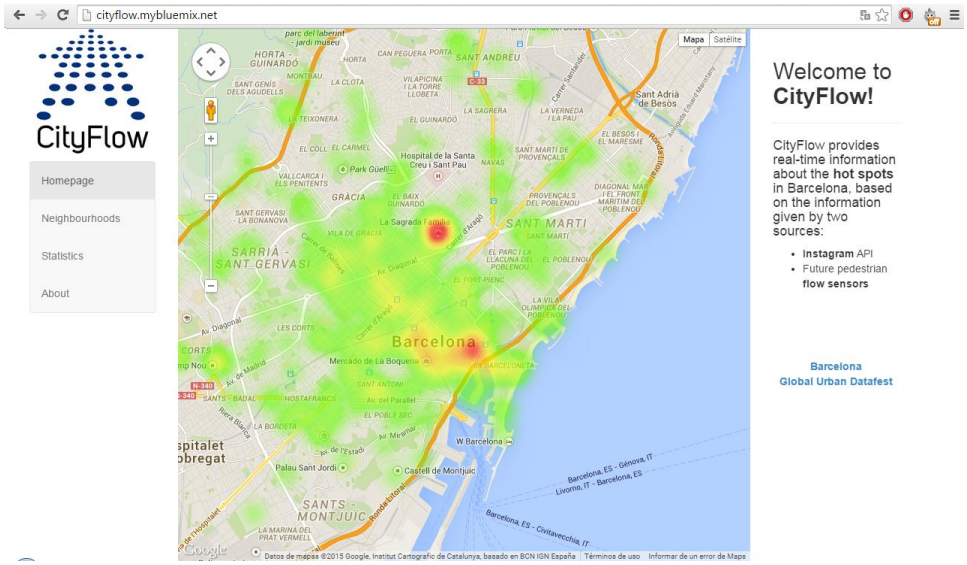

Figure 2. *Heat map view.*

Moreover, we provide a distribution of posts by neighbourhoods, to see whether a particular district or neighbourhood has a heavy Instagram activity. In addition, the last three pictures uploaded at each neighbourhood are shown, with their tags. This last information is more end-user oriented than company-oriented.

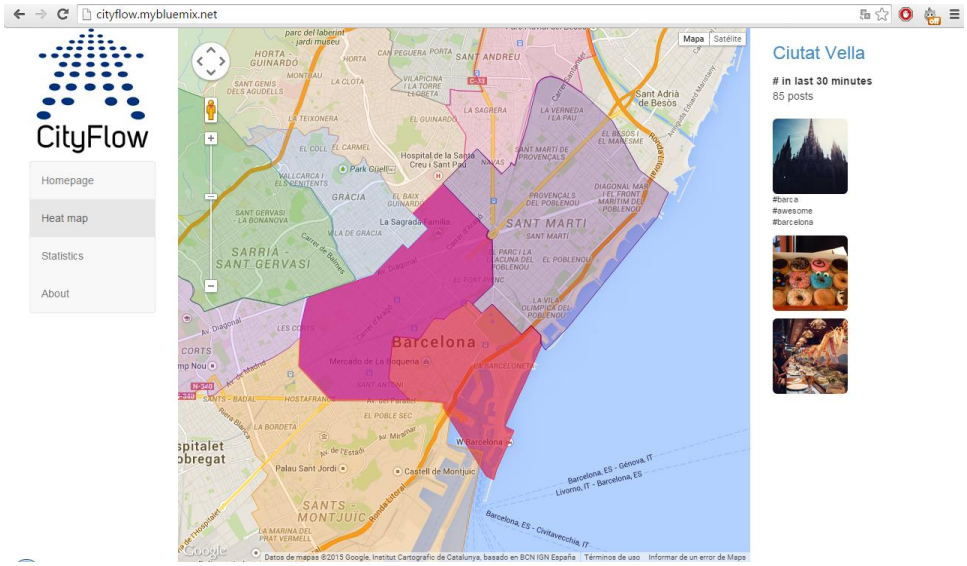

Figure 3. *Neighborhood view.*

Finally, we have presented a first approach to the statistics section. In this case, only the distribution on the last 30 minutes is shown, but in a near future, we will be able to show the distribution of posts at a certain time a day of the week. So, for example, if the user wants to show the historical distribution of people on Thursday at 3 PM we will provide this info.

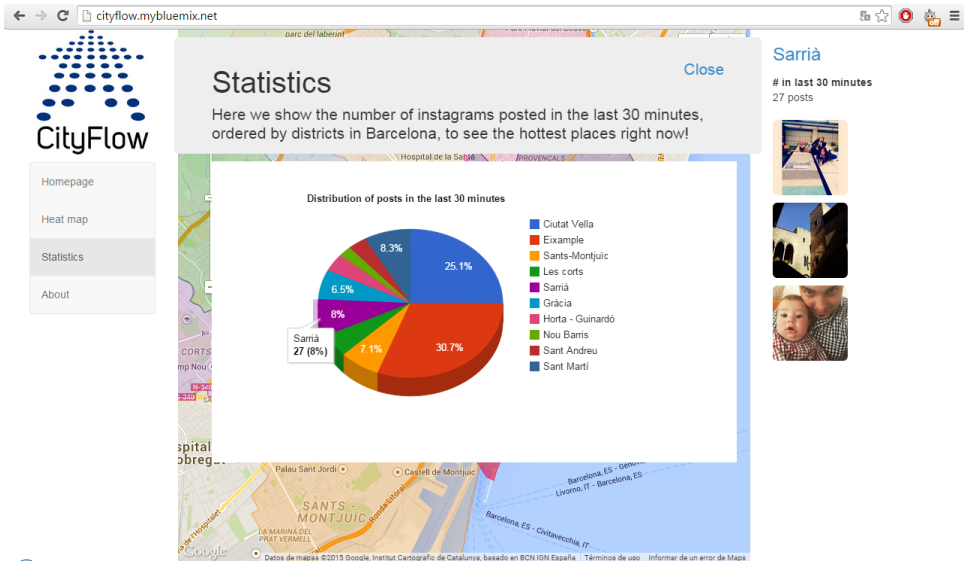

Figure 4*. Statistics view.*

3 ENVIRONMENTAL IMPACT

As we have stated above, we can have several applications for the same concept. Identifying the distribution of people at a given moment, and having this data stored, is a very potential tool also for reducing the pollution in a city.

For instance, we can provide the data to bike renting companies (e.g. Bicing in Barcelona) in order to redistribute their bike fleet depending on the distribution of people in the city. So, by having more bikes in the most crowded areas, we will promote the use of a clean transport means, which will decrease the pollution of the city. Moreover, using bikes will also reduce acoustic contamination and will improve the image of the city for both citizens and tourists.

In a similar way, we can provide our data to entities in charge of the distribution of buses across the city (e.g. TMB in Barcelona). With this info, the frequency of the bus routes can be changed, and placing more buses in the most crowded hours along a day will make a bus route more efficient and therefore the amount of CO_2 per person will decrease.

Figure 5. *Redistribution of bike fleet and reorganization of bus route frequency*

Another example would be cleaning companies. If the distribution of people is given to the cleaning squad, they can go to the most crowded places in order to have more cleaning effort at places where there are more people, since these places would be dirtier. Furthermore, the municipality could also promote the recycling in these areas, making then the city more sustainable.

As a summary, if we are able to make our city more sustainable, we will reach our basic goal when we started this project, which was to provide tools for improving the lifestyle of both citizens and tourists, not only by reaching them by an end user app but also by providing tools to the companies to enhance the everyday services.

4 SOCIAL BENEFITS

The social benefits of our application are in conjunction with the Environmental Impact explained in Section 3. There is a bunch of applications that could use our data to improve or extend their functionalities. As we have done before, we will state some examples that will make the final user benefit from our project.

Regarding the bike companies distribution, a user (whether tourist or citizen) will feel much happier if he wants to rent a bike and there are bikes available. Unfortunately, in many places, nowadays, when a user wants to rent a bike in a very crowded area there are no bikes available, and then he has to take a car, a bus, or whatever. The promotion of the bike use will benefit the name of the city as well, since, for example, Barcelona would have a cleaner name and then the citizens would be happier.

We have also talked about taxi companies. If the taxi fleet is distributed in advance, the waiting times would decrease drastically and therefore their use would be more massive.

These two toy examples would have a similar impact in the final user: reduction of stress and improvement of their everyday lifestyle. Actually, the waiting times are one of the most important factors in the user's perception and satisfaction. As you can see from this example made in California by *H. Iseki et al.* [1], CR2 (transport on time) is in the rightmost part of the figure.

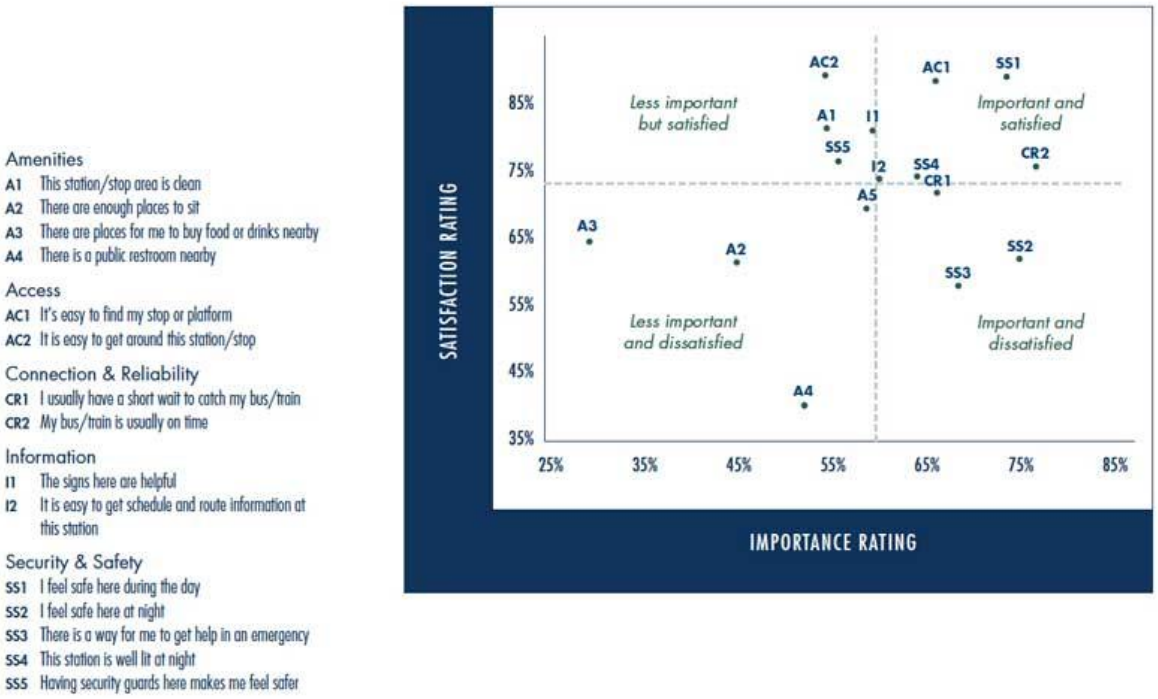

Figure 6. *Attributes users find most important and satisfying at transit stops and stations.*

Finally, visitors of the city could also benefit from this data as they could know which are the trendiest neighborhoods at each moment and choose the place to go out.

5 ECONOMIC STUDY OF THE IMPLEMENTATION

Related with all the information given before, this section shows an economic analysis of the resources the platform needs to be operative and the potential income that will be achieved.

5.1 COSTS

Table 1 summarizes the main costs that the project would incur in a first stage of the development.

Staff cost	**Annual**
Senior Programmer x2	25.000,00 € each
Intern x2	5.000,00 € each
Sales	20.000,00 €
Total Personal Cost	**80.000,00 €**

Table 1. *Personal cost table.*

CityFlow is going to hire two Senior Programmers to implement the back-end and front-end of the application. Two interns will support the work of the seniors. Finally, a salesperson will look for new customers.

Other cost	**Annual**
Material cost	5.000,00 €
Communication cost	2.000,00 €
Total cost	**7.000,00 €**

Table 2. *Other cost table*

We can anticipate an initial cost of material and a cost of communication, to advertise the platform in the relevant media.

5.2 POTENTIAL INCOME

In Table 3, the income for the first year is detailed.

Potential customers	**Annual**
Municipalities	40.000,00 €
Taxi companies	15.000,00 €
Cleaning municipal services	10.000,00 €
Total income	**65.000,00 €**

Table 3. *Income during the first year.*

As we can see, the first year, the income is not enough to cover the costs, so the company will use different methods to finance the project.

Financial terms	**Annual**
Crowdfunding	15.000,00 €
Grants	2.500,00 €
Loan	5.000,00 €
Total amount financed	**22.500,00 €**

Table 4. *Financial terms table.*

Initially, the project will be upload to a crowdfunding platform to reach 15.000€ which will be the first investment. As is an innovative project and related with environmental matters, we will apply to some grants, that in the worst of the cases will be 2.500€ [4].

Finally, to cover the last expenses and to have some operative margin, we will require a bank loan.

During the first year, the project will experiment lots of changes due to customer interaction and new needs that they can require. We will use customers to improve the platform, adapting it to their needs.

Another potential use with the available data after the first year could be to discover new possible use cases in order to find different groups of interest.

6 APPLICATION FEASIBILITY

In order to make a proper economical study, we need to analyze the different items that make the project feasible.

Social networks public users are the main source of information in our platform, because they are going to provide us with information and generating a database with all this information. As we can see in the graph below, provided by statista.com, the number of social networks' users worldwide is growing dramatically.

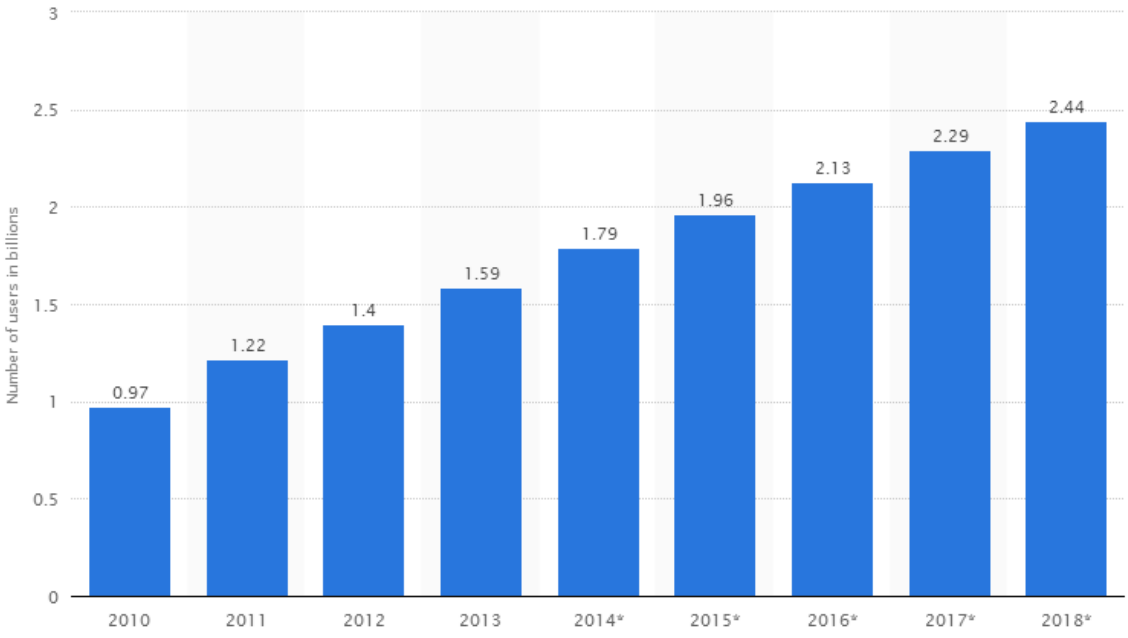

Figure 7. *Growth of the users in social networks. [2]*

This information, shows us that our database will increase over the time, and therefore the platform will release more detailed information which will have a positive impact in the final data provided to the users.

Moreover, some iCity flow sensors are allocated through the city. These sensors' network will complement Social Networks Database to provide more accurate information to the final user.

In prior sections, we explained that we are targeting regular citizens as well as institutions and companies. Here, like in the Delivery section 2.4, we will distinguish between them as our approach towards them will be different.

The first group will mainly contribute to spread the word of our services. It will allow to make publicity of the services that we are offering and to get the attention of possible companies and institutions that could make business with us.

The second group will be our potential customers as they will be our main source of income. Several examples are: municipalities, taxi companies and cleaning municipal services.

Municipalities will be capable of organizing the public transport services depending on different historical statistics of the distribution of people in the city along the different hours of the day. Moreover, the platform can also capture real time data thus providing the possibility of fast reaction to unexpected events. As a result, public transport services will be able to adapt its frequency to provide a better service to the citizens, and optimize the resources.

In the case of Barcelona, the public transportation is managed by TMB. With our data, they could obtain great insights about the distribution of people at any time so they could plan the frequency of the routes proactively and run more efficiently. Such a service, would report them important savings in terms of fuel, maintenance, etc. and also improve the quality of the service they provide enhancing their image. Taking this into account they would probably be willing to pay an important sum for this data.

Taxi companies would know which the most crowded areas are in order to organize their fleets. Using the platform, taxi drivers could avoid driving through the city, and go directly to the zone that there is more people. In this way, they will have more probability to pick up a potential customer.

Knowing the total amount of taxi companies that there are in Barcelona, we could know the taxi licenses, and so, the potential users of the platform. All this information is provided by Institut Metropolità del Taxi, where shows that nowadays are 12 broadcasting taxi stations in Barcelona, with 3.760 vehicles [3]. This will mean a reduction in fuel because taxi drivers could go directly to an area instead of driving along the city during low-occupancy hours.

Cleaning municipal services will be capable to do a more accurate prediction about the areas that should be prioritized because they have had more people movement. Therefore, they could concentrate their fleet capacity in a specific area. This will mean a more efficient service and thus, a more environmentally friendly and clean city.

The most important feature of our project is the data generated and the information that we can extract from it. Thus, other types of companies could also be interested in using the data in more sporadic manner, for example, to know where to place a certain type of advertising hoarding.

7 REFERENCES

[1] Iseka, H., Smart, M., Taylor B. and Yoh, A. (2012) *Thinking Outside the Bus,* Access, No 40, Spring 2012. Available online: http://www.accessmagazine.org/articles/spring-2012/thinking-outside-bus/

[2] Statista.com, *Number of social network users worldwide from 2010 to 2018*. Accessed online on March 2015. Available online: http://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

[3] Àrea metropolitana de Barcelona (AMB), *Taxi Barcelona*. Accessed online on March 2015. Available online: http://www.taxibarcelona.cat

[4] ACCIÓ Gencat.cat, *Ajuts a la Innovació per a empreses*. Accessed online on March 2015. Available online: http://accio.gencat.cat/cat/ajuts-financament/ajuts2014/innovacio/empresa.jsp