Analysing Public Transport data through the use of Big Data tecnhologies for urban mobility

Hugo Antunes
CTS, UNINOVA, Dep.º de Eng.ª
Electrotécnica, Faculdade de Ciências e
Tecnologia, FCT, Universidade Nova
de Lisboa 2829-516 Caparica, Portugal
ha@uninova.pt

Paulo Figueiras
CTS, UNINOVA, Dep.º de Eng.ª
Electrotécnica, Faculdade de Ciências e
Tecnologia, FCT, Universidade Nova
de Lisboa 2829-516 Caparica, Portugal
paf@uninova.pt

Ruben Costa
CTS, UNINOVA, Dep.º de Eng.ª
Electrotécnica, Faculdade de Ciências e
Tecnologia, FCT, Universidade Nova
de Lisboa 2829-516 Caparica, Portugal
rddc@uninova.pt

Joel Teixeira
Link Consulting Tecnologias de
Informação SA
Portugal
joel.teixeira@linkconsulting.com

Ricardo Jardim-Gonçalves
CTS, UNINOVA, Dep.º de Eng.ª
Electrotécnica, Faculdade de Ciências e
Tecnologia, FCT, Universidade Nova
de Lisboa 2829-516 Caparica, Portugal
rg@uninova.pt

***Abstract*— Over the last years, new strategies focusing on mobility have been implemented, especially in big urban areas, in order to solve the mobility problems brought by the urban exodus. The demand for different mobility modes, leads to a complex transportation network that needs to adapt to different mobility requirements. The presented work analyses the current situation of the transportation network of Lisbon's area, where mobility means are provided by different transport operators, private and public, serving a population of around 4M people. The challenge addressed by this work, is to analyze the demand and supply side of the transportation network of Lisbon's metropolitan area, considering ticketing data transactions provided by different transportation operators, which until now, such analyses were essentially obtained through observation methods and questionnaires. This paper explores the ability of Big Data technologies to cope with data collected from transport operators, by inferring automatically and continuously complex mobility patterns in the form of insightful indicators (such as connections, transshipments or pendular movements).**

Keywords*— Big Data Analytics, Machine Learning, Ticketing, Urban Public Transportation, Mobility patterns.

I. Introduction

Nowadays, the world is witnessing the manifestation of a major trend, the urban exodus[1]. This phenomenon is characterized by the movement of a large number of people from rural areas to urban centers. At the same time, coupling this trend and the consequent increase of people in cities [2] the need to improve mobility solutions within the cities increases.

Thus, to solve a set of problems brought by the overcrowding of cities, such as pollution and decrease of quality of life, among others, some governments and companies are investing on new technologies [2]. That way, one of the sectors where cities are making huge investments is the mobility and transportation. The offer provided by transport operators, is crucial for citizens to commute fast, easily and comfortable within and around cities. Nowadays, especially in big urban areas, the expansion of traffic networks' capacities, through the construction of more road infrastructure, is extremely costly as well as environmentally damaging, making it an unfeasible process [3]. A more efficient usage of the existing network is vital in order to sustain the growing travel demand.

Every day, thousands of public transports move through and around cities allowing people go to work, home, or simply to accomplish their routines. According to the last survey of the International Association of Public Transport, carried out in 2015, around 57,3 billion journeys were made during that year, only in Europe (considering 27 European countries in the study), highlighting the importance of this type of transportation in daily lives [4].

In this way, mobility could be regarded as one of the most important areas to maintain the efficient growth of big urban centers [5]. Mobility in cities can be seen from two scopes: individual mobility, in which people use their own means of transportation to get around, or public transportation (PT), in which people commute within cities via transport modes managed by external entities, private or public. This paper will focus on the second, mobility through public transport.

Over the last years, aligned with the demand for new mobility solutions, huge investments have been made in order to enable new mobility solutions, especially in public transportation. Today, all public transports are being "sensorized", meaning that it is possible to know in real-time (with some exceptions), where a certain mode of transport is, or how much time it will take to arrive to the next stop. Such features are crucial for a demanding urban population, that requires better service quality when using public transportation. There's a big effort to provide urban commuters with different multimodality options, according to their needs, involving guarantees for a significant level of interoperability at the data level, from the different transport operators.

In a metropolitan city as Lisbon, Portugal, millions of people use urban public transport on a daily basis, generating millions of data records regarding ticket validations. At that the same time, transport operators struggle to find a way to easily process large volumes of such ticketing data, in a timely manner. Thus, classical data-driven procedures

978-1-5386-9282-0/19/$31.00 ©2019 IEEE

become inefficient and obsolete, mainly because they are based on approaches that use traditional database management and warehousing systems, where scalability is an issue. The problem becomes more complicated, if we consider that in Lisbon area, the transportation network is managed independently and competitively by different transport operators, posing more heterogeneity at the data level. However, there is a shared infrastructure aiming to consolidate commercial ticketing data, but data sharing between entities is a challenge, raising up legal and data privacy issues, which are outside of the scope presented here. Hence, in the case of the city of Lisbon, there is no straight-forward way to process commercial electronic ticketing data coming from different transport operators and drive useful insights about public transportation offering and demand. Achieving reliable key performance indicators, is difficult through classical data analysis processes, where data is gathered through human observation methods on the roads and questionnaires to users. There is a need to automatize and make this process more reliable.

To answer the previous problem, the presented work focuses on the development of a Big Data architecture to automatically and continuously process transportation data, and extract complex information in the form of insightful indicators, such as connections within the same mode of transport, transshipments between different modes of transport, pendulum movements within the network (using the same routes from and to the origin and destination points at different times of day) or abnormal network usage, such as demand peaks or problems in the network. Hence, these insights represent the kind of information that cannot be inferred automatically with classical processes used by public transport entities, but which are possible to calculate and obtain using novel Big Data processing techniques. Thus, the main goal of this work is to implement and develop an architecture, supported by Big Data technologies and machine learning algorithms, able to process ticketing data transactions, so that they can later be added in parallel with the current procedures performed by data management entities.

This document is structured as follows: Section 2 describes an overview about important topics and related work. Section 3 presents the overall Big Data architecture. Section 4 describes the data sets used throughout the work. Section 5 describes the workflow of an algorithm to extract complex information about ticketing data. Section 6 are showed the result of developed work. Finally, section 7 concludes the paper and points out future achievements.

II. Related Work

In recent years, with the explosion in technological development, a term that has gained strength among researchers, developers and industries is Big Data [6]. Data is growing [7], not only in terms of volume, variety and velocity [8], but also in value [9]. By itself, data has not any value, but when data is processed in meaningful ways, it can represent a great value for companies. All the procedures related with larges datasets such as, for example, data collection, data processing, or data analytics can be considered Big Data procedures [10].

The mobility area is an example, where large volumes of data are generated daily, requiring an approach capable of handling data management procedures, such as data extraction, data cleaning and modelling. Examples of the use of Big Data tools to improve mobility are described by [11] and [12], which propose a solution to improve road management processes by applying Big Data tools to forecasting driver behavior. Another approach that uses Big Data processing tools to improve mobility and transport is presented in [13], in which the authors aim at creating an application capable of performing real-time monitoring of road traffic and generating traffic indicators through stream mining for traffic operators, providing information generally not accessible in real time. An exploitation of Big Data techniques applied to process public transportation data is proposed by [14], which proposes a study of passengers' travel behavior, through the creation of mobility patterns. This study explores the behavior of bus users at different periods of the year, proving that the different periods manifest different user behaviors (number of users, places they go).

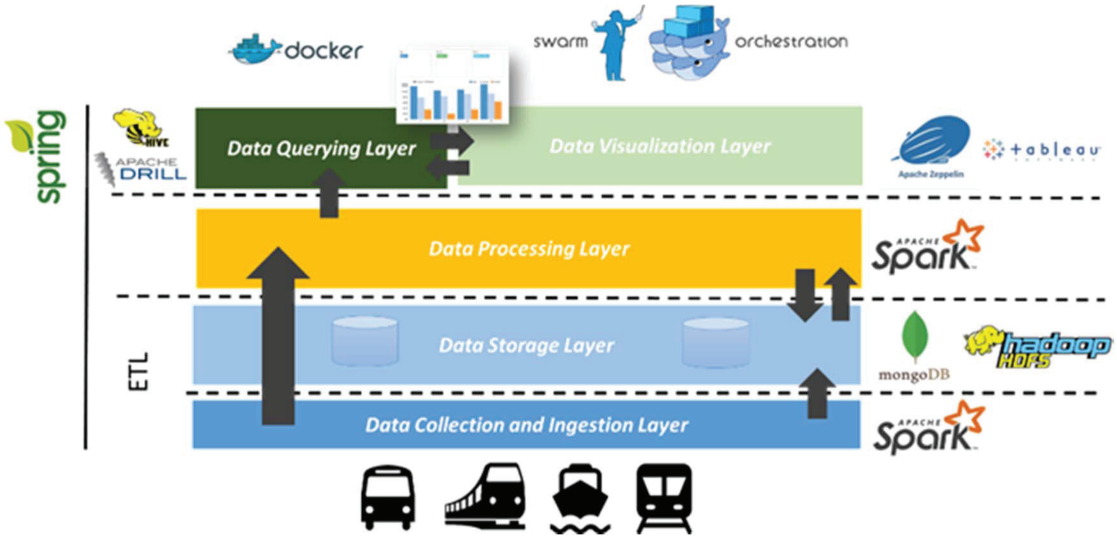

Figure 1 - Big Data architecture overview

Another study that is deemed relevant, and cross Big Data and mobility, was written by the authors of [15]. In this study, an analysis of human mobility data in an urban area was conducted using data from available bicycles in the stations of the community bike program *Bicing* in Barcelona.

III. Work Architecture

To implement the presented approach and to achieve all the requirements, a technical architecture was developed. The proposed architecture comprises the entire Big Data procedures, from data collection to data visualization. The architecture, as well as the tools, technologies and APIs used in each module, are displayed in Figure 1.

The first layer, data collection and ingestion, performs the ingestion of data from different sources. In this case this layer ingests data from validations of different urban public transport vehicles. Data collection is performed using Apache Spark, which can be used as a Big Data collection proxy with good performance results.

After collecting, cleaning and harmonizing, the data is stored on MongoDB, a NoSQL schemaless database. The reason why MongoDB was chosen is that the different datasets used had different schemas, but corresponded to the same data type (e.g. validations, transactions). Hence, a flexible, schemaless format was chosen to encompass validation and transaction data coming from different entities with different formats. The data processing layer, based on Apache Spark, is responsible for heavy, demanding and intensive data processing tasks. Finally, the data querying and visualization layers expose the processed insights through query engines, such as Apache Drill and Hive or visualization tools, such as Tableau or Apache Zeppelin. To easily deploy, manage and scale, the entire Big Data architecture was built on Docker Swarm environment. Docker Swarm is a docker environment manager, easily scalable to clusters.

IV. Datasets

The presented work used two datasets: the first dataset represents the ticketing data for seven different public transport operators; the second dataset was the Generic Transit Feed Specification (GTFS) data for Lisbon, Portugal.

The ticketing dataset represents data acquired by transport operators on a daily basis, through the acquisition of data from electronic ticketing and commercial transactions. Such datasets represent entries and exit registries of the transportation network, through ticket validation, as well as all operations of loading tickets. The present dataset contains more than 55 million of records, representing entry and exit validations in Lisbon's PT network, which were gathered from more than 4500 different stop stations, combined with 1500 different types of tickets. This is just a sample of data, which represents only one month of validations for the seven PT entities operating in Lisbon.

The GTFS dataset [16], firstly introduced by Google to handle Google Maps' PT information, defines a common format for PT schedules and associated geographic information. The GTFS dataset contains information of urban PT schedules, stops and routes, and is used to geographically pinpoint the validations and associate the validations to existing PT routes. This second dataset was acquired from an open source platform available by Lisbon city council through of Transporlis [16].

V. Workflow

This section covers all procedures performed on the previously presented datasets, and will be divided into two parts, data collection and harmonization and data processing. The entire workflow is structured in Figure 2.

A. Data Collection and Harmonization

The process of data collection and harmonization starts with the ingestion of ticketing data in a database, which will save all the raw data, sent directly by PT entities. After that, the process focused on data cleaning. The first data cleaning process was performed by deleting duplicated data. The original dataset had some duplicate data and some validation registries which had errors (e.g. consecutive entries in the same station with a few seconds of interval). The second data cleaning procedure corresponded to the identification and subsequent elimination of incorrect data. In a first analysis over the ticketing data, some validation without stop locations were detected and deleted. Furthermore, only ticketing data with unique card serial number and validation

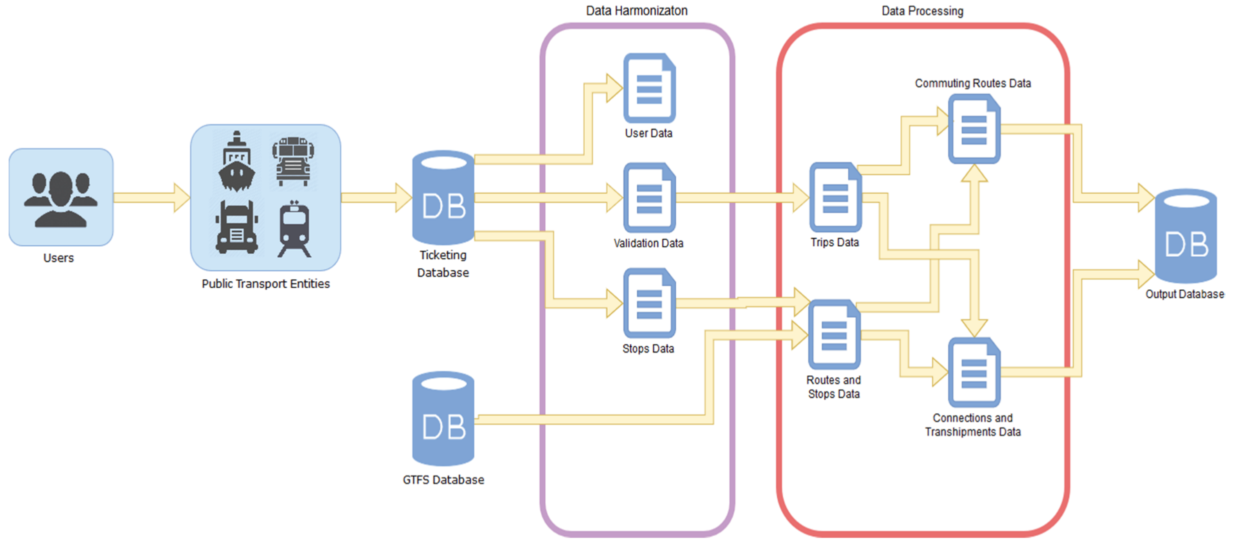

Figure 2- Workflow overview

time were maintained, and entry or exit validations on the same station or stop had to have at least 5 minutes between them.

After the data cleaning process, the next step was to split and store the data. For instance, each ticketing data registry contains card serial number, age group, gender, and other unnecessary information. Before data storage, the data schema had to be reduced, thus ticketing data was divided into three distinct data schemas: user data, station data and validation data.

User data – this new dataset contains all user information that has at least one record in the validation dataset. In this dataset can be found a set of information, such as age, gender, postal code about users.

Station data – this dataset represents a mapping of all information about stations, which had at least one record in validation dataset. To create this dataset, the validation dataset and the GTFS entities information dataset were crossed. Information such as geographic location, station name, station identifier or entity operating in station, are stored in this dataset.

Validation data – this new dataset differs from the raw one because the number of fields in the schema has been reduced. From this point on, the validations dataset is already reduced. In this dataset it is possible to find information's such as validation data, operation identifier, product type, validation type, card serial number that performed the operation, etc.

B. *Data Processing*

This step encompassed the analysis and exploration of data. As shown in Figure 2, the first step in data processing deals with grouping different validations per trip. A trip is a set of one or more validations performed by the same user. The trip criterion was set to 1 hour between validations because after a previous analysis on different times interval (e.g. 00:30h, 00:45h and 01:30h), it was concluded that 01:00h was the best time interval to gather validations corresponding to the same trip. The creation of the trip dataset was the first step to create and analyse automatically insightful indicators such as connections or transhipments. All the trips which had one single validation were excluded for the next steps because such trips didn't contain any connection or transshipment.

The second step occurs in parallel with the first. This step comprised the complex task of linking stops and stations data with GTFS data. In this step all the stops and stations were mapped with one or more routes, by crossing the location in GTFS data and stops and stations data. As the locations of both datasets do not match perfectly, a proximity algorithm was used to allow the matching between both datasets.

After these two procedures, the aggregation of validation by trips and the identification of routes for each stop, it was possible to analyze these two new data sets and to identify complex indicators, such as connections, transhipments, and pendular movements.

Hence, the two last steps in data processing were similar. The first step consisted in the categorization of trips into connections and transshipments. To identify connections all trips that change route but keep the same PT mode and entity were considered. To identify transshipments all trips that combined more than one PT mode and entity were considered. The second step consisted in the identification of pendular movements. This identification was possible through the creation of an origin-destination matrix for each trip, based on the routes, and whenever a user made two trips with origin-destination pairs opposite to each other, these trips were identified as being a pendular movement.

The developed algorithms allowed to map individually all user routes, and to extract connections, transshipments and pendulum movements, as well as to identify the stops and stations' usage throughout the day.

VI. Data Visualization – Results

This section addresses all the results obtained from the previously explained procedures. The main goal with this study was to calculate insightful indicators, such as connections, transshipments or pendulum movements in order to understand the complex movement of people in the city of Lisbon.

To validate the model of Figure 2 a dataset with more than 55 million records of ticketing data was used, which generated the records exhibited in table 1.

Table 1 - Number of records in each dataset

Dataset	Number of records
Ticketing data (original data)	55,110,230
Validations	52,730,245
User	2,346,247
Stop location	4,459
Trips (with more than one validation record)	2,016,320
Connections and Transshipments	15,516,791
Connections	9,304,132
Transshipments	6,212,659
Pendular movements	2,644,569

A. *Connections and Transshipments*

To understand the potential of connections and transshipments indicators three visualization charts are presented for the purpose of exemplification.

Figure 3 represents the geography projection of destiny stations when connections were made (can also be made for transshipments), and when the departure station is Entrecampos. In the Figure 3, the size of the circles indicates the quantity of people using these stations as destiny. The data used to produce this figure corresponds to May 1, 2018. This day was randomly chosen as an example to visualize these indicators.

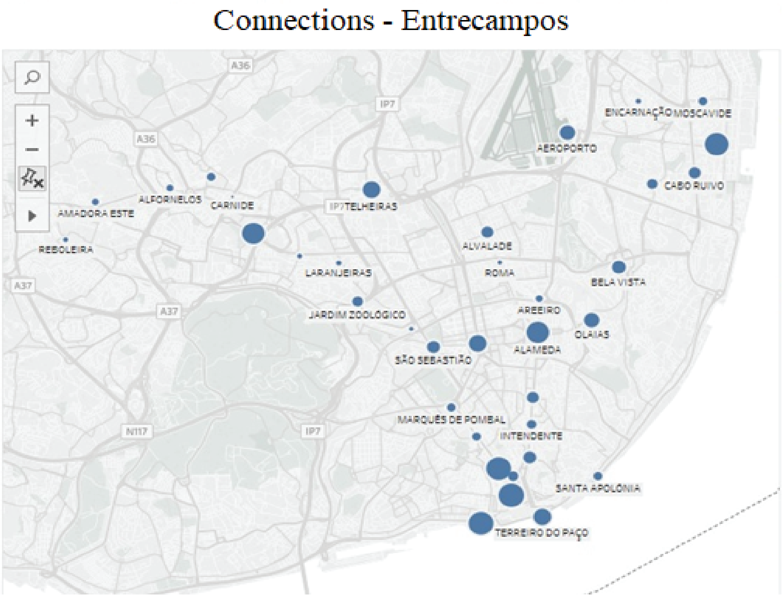

Figure 3 - Destination stations when a connection was made, and the departure station was Entrecampos.

Another interesting indicator can be seen in Figure 4. This figure shows the transshipments' percentage for each destination entity, when the origin entity is the entity with the identification 67. This indicator is very important to understand the correlation between entities, enabling the creation of new combined tickets through the analysis of such information, for instance. To create this figure, data from the first two weeks of May 2018 was used.

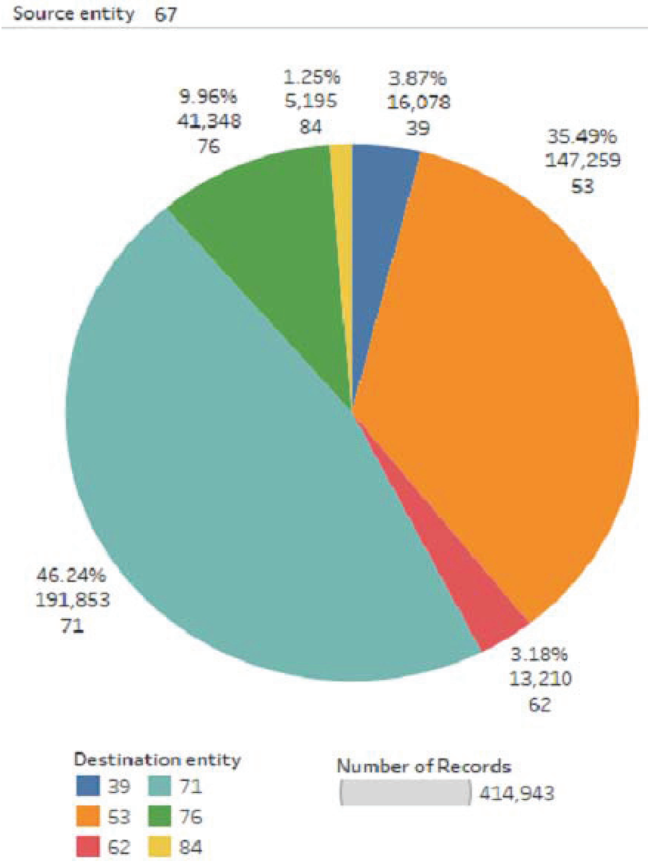

Figure 4 - Transhipments' percentage for each entity when the origin entity is the entity with ID 67

The exploration and analysis of these indicators (connections and transshipments) did not finish here. Other types of visual analysis have been explored, such as the times of the day that more connections and transshipments are made. These two analyzes shown in Figure 3 and Figure 4 were chosen because they were of great interest for the entities that integrated the study.

B. *Pendular movements*

The analysis of pendular movements represents another interesting indicator. The pendulum movements can be seen as the combination of user movements that are going from point A to point B, and then from point B to point A. The analysis of pendular movements is shown in Figure 5. This figure represents the pendular movements occurred in the first two weeks of May 2018, for all users that entered and exited in the Fogueteiro station.

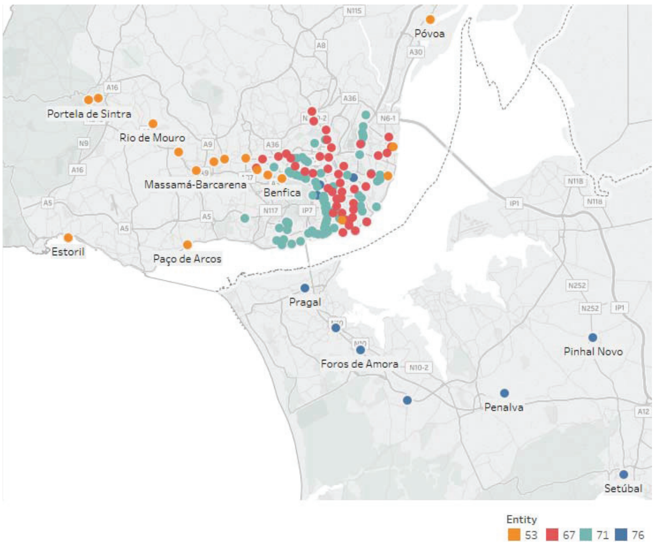

Figure 5 - Represents the pendulum movements that occurred when the users had as first entry point, the Fogueteiro station.

VII. Conclusions

The work presented here enables a novel approach combining a Big Data architecture and machine learning techniques to process PT ticketing data. Until now, the analysis performed in Lisbon's PT network were based only on surveys, and through this work, a new and sophisticated way of gathering knowledge from ticketing data was created.

In conclusion, the objective of this work, the development of a Big Data architecture to automatically and continuously process transportation data and extract complex information in the form of insightful indicators, was achieved. It is important to highlight the performance of the architecture. The architecture was tested in a single machine, with the following specifications: an AMD Ryzen 5 1600 - 12CPU's, with 32GB RAM (Corsair Vengeance LPX) and a SSD 120GB and a 1TB HDD, achieving much better performance, 4 hours (Reading/writing to MongoDB on each stage, with no indexes), comparing with traditional DW processes, which performed the same actions in days. Thus, with these new performance values achieved by the architecture, operators have at their disposal a platform capable of processing their ticketing data in real time, something that with the current system was impossible to perform.

Another important achievement is the acquisition of new knowledge about Lisbon's PT network, through insightful indicators. These indicators allow to create a new overview and new approaches for public transportation networks, enabling an improvement of such networks, such as better route planning, better route management, better transport management, among other improvements. With these indicators, other studies of public transport networks can be performed, and new knowledge can be created.

As future work, it is planned to test the architecture in a cluster of computers, in order to validate and analyze the performance of the architecture when the number of nodes in the architecture increase. In the near future, the exploitation of new frameworks, such as CEP (complex event processing) or streaming frameworks, in order to automatically identify fraud patterns in purchased tickets, is planned to be integrated in the developed architecture. Other feature for future work encompasses real-time ticketing data streaming, allowing for an improvement in current methods of mobility analysis. Finally, the combination of abnormal event prediction and the real-time mobility analytics is also considered for future development.

References

[1] H. Roser and M. Ritchie, "Urbanization," *Published online at OurWorldInData.org*, 2018. [Online]. Available: https://ourworldindata.org/urbanization.

[2] P. Neirotti, A. De Marco, A. C. Cagliano, G. Mangano, and F. Scorrano, "Current trends in Smart City initiatives: Some stylised facts," *Cities*, vol. 38, pp. 25–36, 2014.

[3] E. Rodríguez-Núñez and J. C. García-Palomares, "Measuring the vulnerability of public transport networks," *J. Transp. Geogr.*, vol. 35, pp. 50–63, 2014.

[4] UITP, "URBAN PUBLIC TRANSPORT IN THE 21ST CENTURY." Brussels, Belgium, 2017.

[5] P. C. Cheshire and D. G. Hay, *Urban Problems in Western Europe: an economic analysis*. Routledge, 2017.

[6] C. L. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Inf. Sci. (Ny).*, vol. 275, pp. 314–347, 2014.

[7] A. Ben Ayed, M. Ben Halima, and A. M. Alimi, "Big data analytics for logistics and transportation," in *Advanced Logistics and Transport (ICALT), 2015 4th International Conference on*, 2015, pp. 311–316.

[8] H. Chen, R. H. L. Chiang, and V. C. Storey, "Business intelligence and analytics: from big data to big impact," *MIS Q.*, pp. 1165–1188, 2012.

[9] O. Kwon, N. Lee, and B. Shin, "Data quality management, data usage experience and acquisition intention of big data analytics," *Int. J. Inf. Manage.*, vol. 34, no. 3, pp. 387–394, 2014.

[10] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 137–144, 2015.

[11] R. Costa, P. Figueiras, P. Oliveira, and R. Jardim-Goncalves, "Understanding Personal Mobility Patterns for Proactive Recommendations," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, 2015, pp. 127–136.

[12] B. Kažič, J. Rupnik, P. Škraba, L. Bradeško, and D. Mladenič, "Predicting Users' Mobility Using Monte Carlo Simulations," *IEEE Access*, vol. 5, pp. 27400–27420, 2017.

[13] P. Figueiras, Z. Herga, G. Guerreiro, A. Rosa, R. Costa, and R. Jardim-Gonçalves, "Real-Time Monitoring of Road Traffic Using Data Stream Mining," in *2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, 2018, pp. 1–8.

[14] S. Tao, J. Corcoran, I. Mateo-Babiano, and D. Rohde, "Exploring Bus Rapid Transit passenger travel behaviour using big data," *Appl. Geogr.*, vol. 53, pp. 90–104, 2014.

[15] A. Kaltenbrunner, R. Meza, J. Grivolla, J. Codina, and R. Banchs, "Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system," *Pervasive Mob. Comput.*, vol. 6, no. 4, pp. 455–466, 2010.

[16] "Open Data." [Online]. Available: https://www.transporlis.pt/Default.aspx?tabid=36.