Abstract—Big cities show a wide public transportation network, allowing people to travel within cities. However, with the overpopulation of urban areas, the demand for new mobility strategies has increasing. Every day, citizens need to commute fast, easily and comfortably, which is not always easy due to the complexity of the transport network. The main objective of this paper, is to explore the ability of Big Data technologies to cope with data collected from public transportation, by inferring automatically and continuously, complex mobility patterns about human mobility, in the form of insightful indicators (such as connections, transshipments or pendular movements), creating a new perspective in public transports data analytics. The presented work, focus on the Lisbon metropolitan area, aiming to analyze the demand and supply side of transportation network of Lisbon, considering ticketing data transactions provided by the different transportation operators.

I. INTRODUCTION

Over last years, the demand for new strategies to improve quality of life has become one of the main focuses of governments and companies. The business paradigm as well as people’s technology interest, have changed allowing the emergence and exploration of new products and services. An good example, is the emergence of techniques able to process large volumes of data, which is currently a trend and is increasingly exploited by companies [1].

Currently, due to the growth of the Internet of Things (IoT) devices, large amounts of data have been generated daily through different devices, such as mobile phones, GPS devices, electronic cards, among others, being produced 2.5 quintillion bytes of data per day [2]. In the recent years, the exploitation of such data has gained new interest due to the possibility to analyze large volumes of data in a timely manner. Large data sets, when analyzed, may contain valuable information about people’s mobility patterns and their surrounding environments, information that can be useful to create new products and services, aiming to improve people’s quality of life. Aligned with the growth in data generation, and with the boost of technologies [3]–[5], new Big Data processing techniques started to emerge, allowing the creation of new approaches and new data analysis in different sectors.

One of the sectors that have benefited from the development of these new technologies, in particularly technologies to process large amounts of data, is the transportation and mobility [6]. The big data on human mobility have been used to study human behaviors[7], allowing the discovery of new complex mobility patterns in the form of insightful indicators, indicators that can be useful to manage and improve the mobility in cities.

Nowadays, the number of people living near urban centers has increased, due to the urban exodus [8], characterized by the movement of people from rural areas to big centers, overcrowding the urban centers, as well as the services available in them, in particularly the public transportation. Despite all the efforts of governments and private entities, the need to create innovative mobility solutions within cities arise. However, due to overcrowding of people in cities, cities also suffer from overcrowding of infrastructure, which makes it impossible to expand traffic networks. That way, a more efficient usage and exploration of the current network is essential to sustain the mobility demand.

In a metropolitan city like Lisbon, Portugal, millions of people use urban public transport on a daily basis, generating millions of data records regarding ticket validations per day. At that the same time, transport operators struggle to find a way to easily process large volumes of ticketing data, in a timely manner. Thus, most of all data-driven procedures become inefficient and obsolete, mainly because they are based on classical approaches, using traditional database management systems, which don’t scale. The problem becomes more complicated, if consider that in Lisbon area, the transportation network is managed independently and competitively by different operators, contributing to more heterogeneity at the data level.

However, there is a shared infrastructure with the main objective to consolidate the commercial ticketing data, but data sharing between entities is a challenge, raising up legal issues about data privacy. Hence, in the case of the city of Lisbon there is no easy way to process commercial electronic ticketing data coming from different transport operators, and to perform data analysis to get indicators which are difficult to obtain through classical data analysis processes, and nowadays are gathered through human observation methods on the roads and questionnaires to users.

To respond to the previously problem, the presented work explains 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, the kind of information that cannot be inferred automatically with classical processes used by public transport entities, but which is 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 and supported by machine learning algorithms, able to process ticketing data, so that they can later be added in parallel with the current procedures performed by data manager entity.

The presented work is structured as follows: Section 2 describes an overview about important topics and related work. Section 3 presents the Big Data architecture. Section 4 describes the data sets used in the work. Section 5 describes the data flow, and algorithms used 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 the human mobility, in particularly in urban centers, the use of Big Data has the potential to achieve significant efficiencies as well as new innovative products and services. Despite the great potential, the Big Data exploitation in transport have been a slower start compared to other sectors such as healthcare and retail [9]. However, in recent years, the potential of Big Data approaches has been applied to the study of human mobility by exploiting machine learning techniques[10], [11], forecasting models [12] or through the application of complex event processing tools [13], [14].

The study of urban mobility has become very important in understanding how people move within cities in order to properly manage transportation services. To acquire knowledge about people’s mobility, some studies address the human mobility through the use of mobile data. An example of this scenario was explored in [15], where was proposed an exploitation of human mobility through the analysis of twitter data. In this study, the author using forecasting techniques to predict the movement of people and thus create mobility patterns throughout the days.

The development of new IoT devices was another important step in the study of human mobility. With the emerge of new IoT devices, such as the smart card, the data acquisition methods have changed, allowing the use of this data in order to create new mobility metrics. Data from smart cards were explored in London [16] (through Oyster card data), where these data were used to analyze the station flow volume in order to infer travel patterns and create different patterns of mobility.

Another way to study human mobility is through the analysis of road traffic data [13]. Through the implementation of complex event processing tools is possible to analyze the road data and automatically identify abnormal traffic patterns in order to quickly predict anomalies in the normal road traffic.

III. ARCHITECTURE

To achieve all the requirements proposed in the present work and to create a platform capable to automatically and continuously process ticketing data, a technical architecture was developed. Thus, to achieve all the requirements, and considering the volume of ticketing data generated every day, all the architecture was based on Big Data technologies, since data collection to data visualization. The architecture is displayed in Figure 1, when is possible to see all the different layers, as well as the different tools used in each layer.

The architecture was divided in four layers, each one with specific tasks. The first layer, denominated data collection and ingestion layer, is responsible to ingest the ticketing data on platform. In the presented work, the ticketing data comes from multiple sources (different public transport operators), being necessary that this data was uniformly ingested in the architecture. In this layer the first procedures of data harmonization and data analysis are performed. Data collection and ingestion is performed using Apache Spark, which can be used as a Big Data collection proxy with excellent performance results, as was proved in [17].

The second layer is the data storage layer. This layer is
responsible to store all data generated, from data ingestion to the data processing. In this layer, the data presents different schemes, in which, although represent same data type, different schemas are need (e.g. validations, station location, ticketing). Due to this requirement, the database chosen to integrate the architecture was MongoDB[18]. This database is a NoSQL schemaless database, optimized to work with different non-schema data.

After first two stage, data ingestion and data storage, the data need to be processed. The data processing occurs in third layer of architecture. This layer is the most important layer and has special focus on section V. The data processing layer is responsible for heavy data processing task, in order to harmonize the data and to optimize them for future analysis. The toll behind this layer is, on more time, the Apache Spark. Apache Spark is a data analysis engine, which uses MapReduce algorithms, to process large amounts of data.[19]

When data processing is complete, the data goes to the data querying and analytics layer. This layer is responsible for creating small data sample, which will be used in data visualization layer. Finally, with the data already processed and the indicator already extrapolated, the data visualization layers expose the processed insights through visualization tools, such as Tableau[20].

To deploy, configure and manage all the Big Data technologies in the architecture across any hardware infrastructure, independently of the number of physical nodes, an approach based on Docker Swarm environment was used.

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 DESCRIPTION

In this section, the datasets used to test and validate the algorithms and architecture are described. This work used two datasets: the first dataset contains the ticketing data of seven different public transport entities operating in Lisbon city; the second dataset represents the Generic Transit Feed Specification (GTFS) data from Lisbon, Portugal. Both datasets cover a temporal dispersion of a month, May 2018.

The ticketing data is the biggest dataset, containing more than 55 million of records, representing entry and exit validation in Lisbon’s public transportation network. These records are acquired by different transport operator daily, through the acquisition of data from smart cards. All the data records were gathered from more than 4500 different stop stations, combined with 1500 different types of tickets.

The second dataset, the GTFS dataset [21] defines a common format for public transportation schedules and associated geographic information. This dataset was firstly introduced by Google to handle Google Maps’ public transportation information and contains information of urban public transportation schedules, stops and routes, and is used to geographically pinpoint the validations and associate the validations to existing routes. All the records of this dataset are open source and was acquired from an platform available by Lisbon city council through of Transportlis [21].

V. DATA FLOW AND INDICATORS IDENTIFICATION

The section 5 addresses all procedures performed in the data in order to achieve one of the main objectives of the present work, automatic and continuous inference of complex mobility patterns in the form of insightful indicators (such as connections, transshipments or pendular movements). The present section is divided into two parts, data cleaning and harmonization and data processing. The entire data flow, as well as all transformation in data, are shown in Figure 2.

A. Data cleaning and harmonization

The data flow starts with the ingestion of ticketing data and GTFS data in databases. In both cases, all databases save the raw data, sent directly by public transportation entities or by Transpolis. However, in the case of ticketing data, before these data are sent to ticketing database undergo a cleaning and validation process. When data ticketing is collected from public transportation, occasionally it shows some errors, such as duplicated data, data with errors (e.g. consecutive entries in the same station with a few seconds of interval, or validation data without stop locations), that need to be fixed. The elimination of duplicate data and data with errors is ensured through this data cleaning process Thus, only ticketing data with unique card serial number and validation time and entry or exit validations on the same station or stop had to have at least 5 minutes between them were maintained. All the data that do not fill the requirements were eliminated by data cleaning algorithm and were not used to infer the complex indicators.

After data cleaning process, the next steps focus on data harmonization. For instance, each ticketing data record contains a set of information’s, such as card serial number, age group, gender, and other unnecessary information that need to be arranged. Before data processing, the data schema had to be reduced, to create low processing times. Thus, ticketing data was divided into three distinct data schemas: user data, station data and validation data.

User dataset – The user dataset is responsible to store all user information, such as age, postal code, gender, address, among others. A record is only stored in this dataset if the user has at least one record in the validation dataset.

Station dataset – this dataset contains a mapping of all information about stations, which had at least one record in validation dataset. Information such as geographic location, station name, station identifier or entity operating in station can be found on this dataset. To create the dataset, the GTFS entities information data and the validation data were crossed.
Validation dataset – the validation dataset is very similar with the original ticketing dataset. In this case, the number of fields in the schema has been reduced comparing with the original ticketing dataset. 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, among others.

B. Data Processing

After the data cleaning and harmonization, in which the data were selected and restructured. The next steps go through the analysis, processing and exploration of data. The first step in data processing, as shown in Figure 2, deals with grouping different validations per trip. A trip is a set of one or more validation records performed by the same user during a defined time frame. To define the time frame between validation was analysis five different times interval, 00:30h, 00:45h, 01:00h, 01h15 and 01:30h, and it was concluded that 01:00h was the best time interval to gather validations corresponding to the same trip. The definition and organization of validation data into trips was the first step to create and analyse automatically complex mobility patterns in ticketing data, where later they will be extrapolated insightful indicators (e.g. connection or transshipment). At this stage all the trips which had single validation were excluded because such trips didn’t contain any insightful indicators, such as connection or transshipment.

The second step is performed at the same time as the first. This step encompassed the complex task of linking stops and stations data with GTFS data. At this stage, the stations need to be mapped with public transportation routes, in order to associate the station where the user validates with certain routes. Thus, the location in GTFS data and stops and stations data are crossed, connecting the stations and stops with one or more routes. To create this match, a proximity algorithm was used to allow the matching between both datasets.

After these two steps, the creation of trips and the identification of routes for each stop, it is now possible to combine these two new datasets and analyze them in order to identify complex indicators such as connections, transshipments and pendular movements.

The next step considers the categorization of trips into connections and transshipments (indicators). To be performed this process, an algorithm which analysis individually each trip was developed. This algorithm crosses route and station information with validation records on each trip, and identifies whether the user made a connection, a transshipment, or neither. To identify connections all trips that change route but keep the same public transportation entity were considered. To identify transshipments all trips that combined more than public transportation entity were considered.

Finally, the last step consisted in the identification of pendular movements. The 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.

VI. RESULTS – DATA VISUALIZATION

The section 5 addresses all the results achieved from the data processing process, previously explained. To express the results achieved and to validate the model of Figure 2, a sample dataset with more than 55 million of ticketing data records was used. The results after ingestion this sample in the architecture and after all the data processing procedures are exhibited in table 1.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ticketing data (original data)</td>
<td>55,110,230</td>
</tr>
<tr>
<td>Validations</td>
<td>52,730,245</td>
</tr>
<tr>
<td>User</td>
<td>2,346,247</td>
</tr>
<tr>
<td>Stop location</td>
<td>4,459</td>
</tr>
<tr>
<td>Trips (with more than one validation record)</td>
<td>2,016,320</td>
</tr>
<tr>
<td>Connections and Transshipments</td>
<td>15,516,791</td>
</tr>
</tbody>
</table>
One of the greatest interests to the public transport entities with the present work is the exploration of insightful indicators that they cannot reach with their technologies. To understand the potential these indicators some visualization charts are presented for the purpose of exemplification. Note: in the following graphs, all public transport entities are anonymized.

<table>
<thead>
<tr>
<th>Connections</th>
<th>9,304,132</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transshipments</td>
<td>6,212,659</td>
</tr>
<tr>
<td>Pendular moves</td>
<td>2,644,569</td>
</tr>
</tbody>
</table>

Figure 5 - Second entry station when a transshipment was made, and the departure station was Cais do Sodré.

Figure 3 represents the geography projection of destiny stations when transshipments were made (can also be made for connections), and when the departure station is Cais do Sodré. In this figure, the size of the circles indicates the quantity of people using these stations as destiny. With this type of approach is possible for public transport entities to optimize route planning according to the behavior of their users.

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.

Other types of visual analysis can be seen on Figure 5. This figure displays the number of transshipments made per hour in each public transport.

The analysis of pendular movements represents another interesting indicator. This kind of indicators is very important because it can be used to acquire a better understanding of human mobility in cities. The study of pendular movements is shown in Figure 6. In this case, the figure show pendular movements occurred when the entry and exit station is Portela de Sintra.

Figure 6 - Represents the pendulum movements that occurred when the users had as first entry point, the Cais do Sodré station.
The work presented here enables a novel approach to the study of human mobility in cities. Until now, the exploration of ticketing data analysis, in Lisbon’s public transportation network, was based on surveys, demonstrating innumerable limitations in the type of analysis to be performed. However, with the main goal to explore 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, the present work shows a new and innovative approach to acquire and explore knowledge from ticketing data.

In short, the results of this work can be divided into two parts, the performance of proposal architecture and the improvements that indicators can bring to the users and entities of Lisbon’s public transportation network. 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 taking 4 hours to process one month of data (55 million records). However, when this result is compared to traditional DW processes, which performed the same actions in days, the results are very promising.

The achieved indicators show a source of new knowledge about Lisbon’s PT network. With these new indicator’s, novel approaches for public transportation networks can be created, enabling an improvement of such networks, such as better route planning, better route management, better transport management, among other improvements.

As future work, it is planned to deepen the studies on already calculated indicators, in order to create new knowledge about human mobility. It also planned to test the architecture with more data volume and scale the number of nodes in architecture, creating a cluster of computers. 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


VII. CONCLUSIONS