Big data in smart cities: a model for Rio bus service

Executive Summary

Columbia University, 2014/2015

16/01/2015
Líderes Cariocas Program
André Peixoto, Francisco Galvão, Gisele Brito, Jessick Trairi, Marcia Marques, Paulo Hirsch.

SUMMARY

1 INTRODUCTION

1.1 SEEDS

The experience of 36 (thirty six) public workers of Rio de Janeiro City Hall (PCRJ), along the Leadership and Management Skills Program, held by Columbia University, should bring tangible results to the cariocas, funders of the investment. More than a course task, join all the information we had access and apply them to develop and improve the city in which we live and work is a mission. A journey that has no end after presentation of the final project, but is going on as a permanent improvement process. The exchange of experiences, learning opportunities and teamwork must be the basis of this and other projects in PCRJ. This spirit involved our group along final project development.

1.2 PURPOSE

The aim of this study is, using the concept of Big Data, to propose a model to support the planning and operation of the public transportation service by bus in the city of Rio de Janeiro. There are plenty of structured data of public transportation by bus in Rio, dispersed across multiple systems. In addition, there are complaints from 1RIO and social networks users. The present study intends to integrate those data under a single model, which enlarge the service demand understanding, user behavior, patterns and factors that impact system performance. The idea is, by providing means to predict scenarios and event analysis, to support a proper management of the system and a fast decision-making in ordinary or emergency situations.

1.3 METHODOLOGY

The first concern of our group after receiving, in New York, the subject of this work was to formulate something implementable and aligned to the demands and strategies of the City Hall. With these premises defined, the first step was to seek technical advice of Pensa Group – a public agency that researches and implements solutions based on Big Data in the city of Rio de Janeiro.

The Pensa Group manager, Pablo Cerdeira, was very receptive to our team. Having in mind that Rio de Janeiro is going through a time of huge transformation, when infrastructural works and mega events occur simultaneously, he suggested the development of an urban mobility model focused on public transport, a major concern of Mayor Eduardo Paes.

After adjusting the focus of the project, the second step was to meet the managers of the Municipal Transportation Agency (SMTR), an important stakeholder. The meeting between our group and deputy assistant Secretary in the Transportation Agency, Helio Faria, aimed to align expectations and design the initial approach for the model. A first draft of mental map was presented and served as the basis for the meeting.

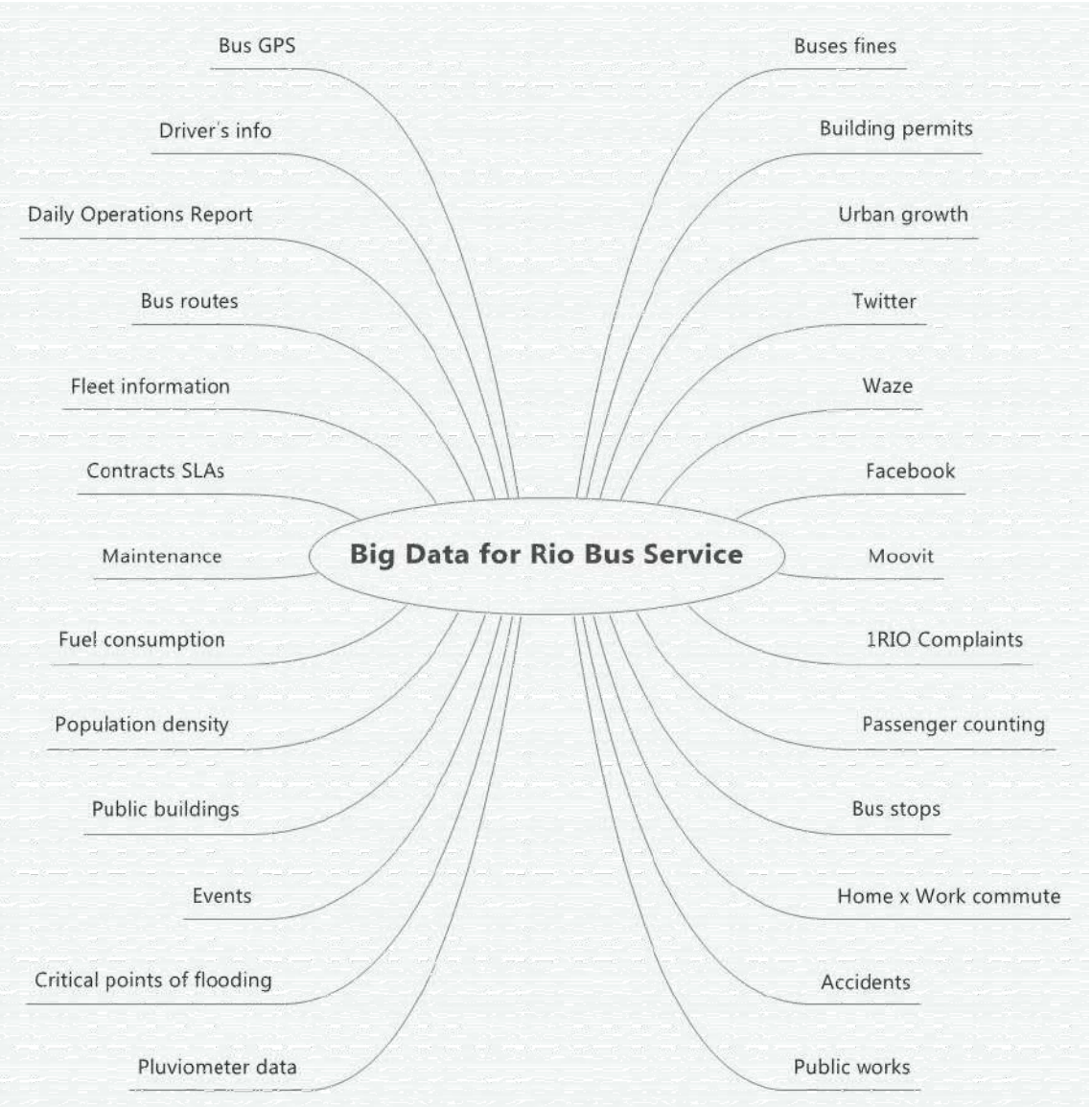


Figure 1 - Mental Map

Throughout the meeting, the group raised SMTR´s needs and questions that should be answered by the model. From mapping the questions, the necessary information sources were identified and analyzed for their existence and information systems already in use. Then, the respective managers, databases locations and attributes were identified.

Another fundamental step was to test the model, to ensure that the architecture is sufficiently comprehensive and flexible to achieve the main goal. For this, our group has partnered with Stone Age, a service provider in the area of Big Data. The Pensa Group allowed access to needed information, in order to design a Proof of Concept - PoC appropriate and feasible to the project deadline. So, this cutout focused in the analysis of city bus intervals, by time band, comparing to the contracted frequency.

1.4 STRUCTURE

The project presents highlights of academic Big Data concept and its application in transportation sector, provides an overview of urban mobility challenge in the city and the world, describes the operation of bus transportation in Rio de Janeiro, proposes a model in Big Data to support the management of this service, glimpses the next steps and the perceptions of the group.

2 WHAT IS BIG DATA?

The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s [1]; as of 2012, every day 2.5 exabytes (2.5×1018) of data were created [2]; as of 2014, every day 2.3 zettabytes (2.3×1021) of data were created [3] [4].

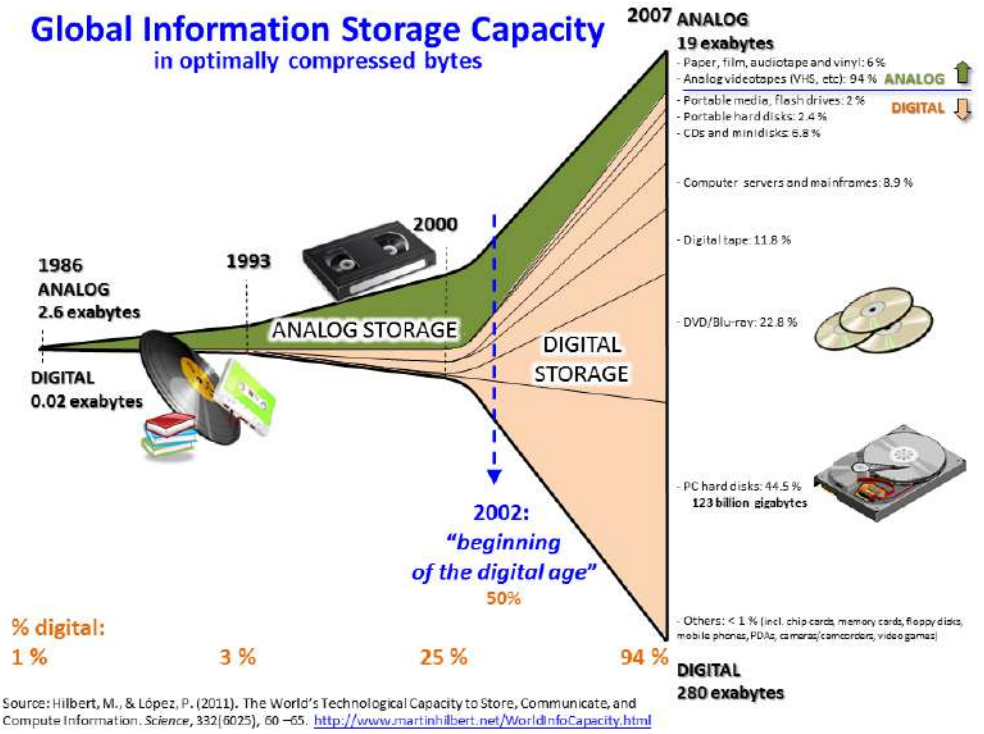


Figure 2: Growth of and Digitization of Global Information Storage Capacity [1]

In this context, Big Data is an evolving term that describes any voluminous amount of data from traditional and digital sources inside and outside your company that has the potential to be mined for information [5].

It can be characterized by 3Vs: the extreme volume of data, the wide variety of types of data and the velocity at which the data must be must processed. Although Big Data doesn't refer to any specific quantity, the term is often used when speaking about petabytes and exabytes of data, much of which cannot be integrated easily.

What is considered "Big Data" varies depending on the capabilities of the organization managing the data set, and on the capabilities of the applications that are traditionally used to process and analyze the set in its domain. Big Data is a moving target; what is considered to be "Big" today will not be so years ahead. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration." [6]

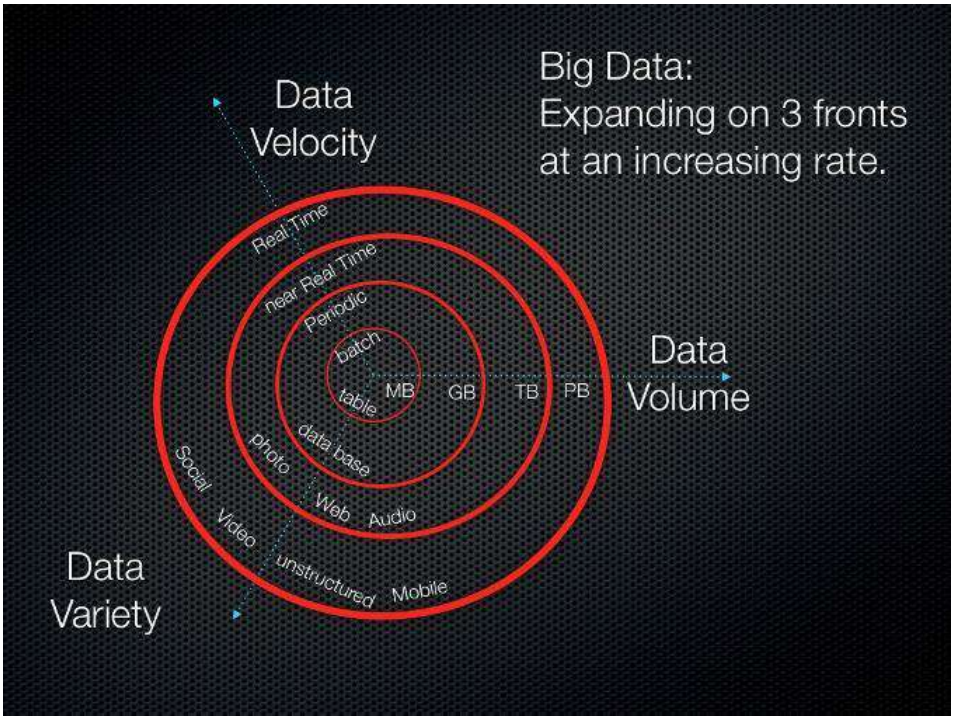


Figure 3 Big Data: Expanding on 3 fronts at an increasing rate [5].

It's important to understand the mix of **unstructured** and **multi-structured** data that comprises the volume of information [7].

Unstructured data comes from information that is not organized or easily interpreted by traditional databases or data models, and typically, it's text-heavy. Metadata, Twitter tweets, and other social media posts are good examples of unstructured data.

Multi-structured data refers to a variety of data formats and types and can be derived from interactions between people and machines, such as web applications or social networks. A great example is web log data, which includes a combination of text and visual images along with structured data like form or transactional information. As digital disruption transforms communication and interaction channels—and as marketers enhance the customer experience across devices, web properties, face-to-face interactions and social platforms- multi-structured data will continue to evolve.

2.1 BIG DATA X BUSINESS INTELLIGENCE (BI)

Regarding data and their use [8], the difference between the approaches are:

- Business Intelligence uses descriptive statistics with data with high information density to measure things, detect trends etc.;
- Big data uses inductive statistics and concepts from nonlinear system identification [9] to infer laws (regressions, nonlinear relationships, and causal effects) from large sets of data with low information density [10] to reveal relationships, dependencies and perform predictions of outcomes and behaviors [8] [11].

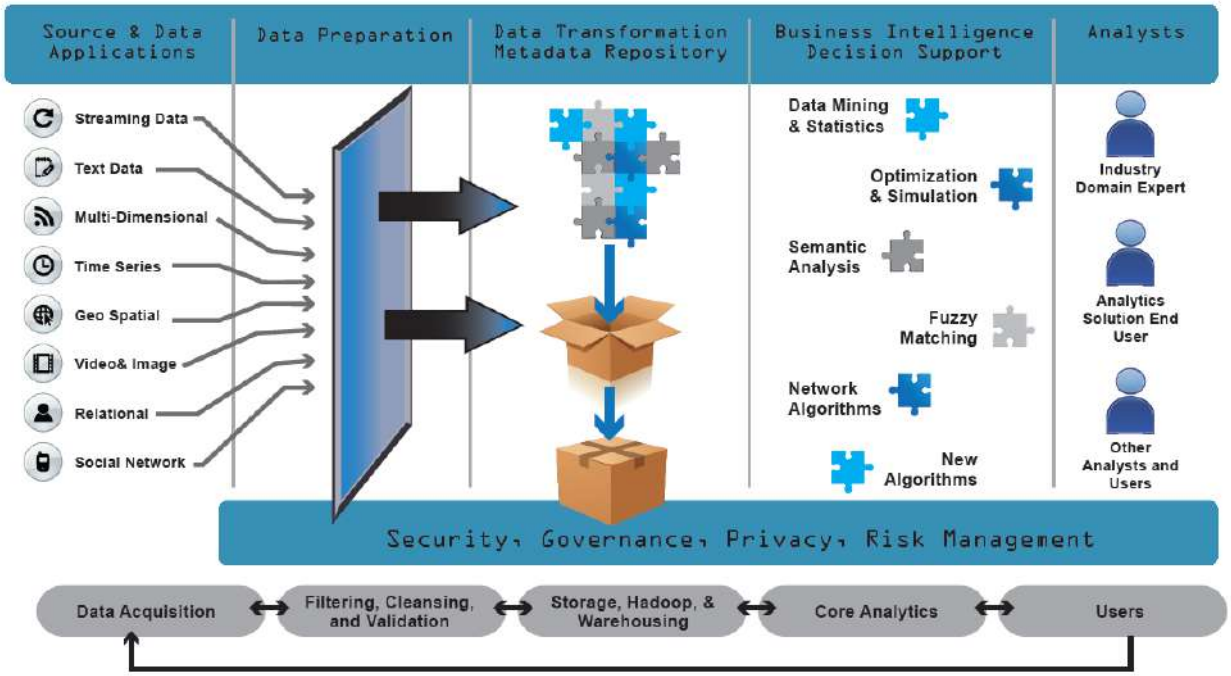


Figure 4: Notional Information Flow – The information Supply Chain [12]

Most existing relational database management systems and desktop statistics and visualization packages are not suitable to Big Data, which requires, instead, "massively parallel software running on tens, hundreds, or even thousands of servers" [13], often associated with cloud computing[1] [12]. The analysis of large data sets in real-time requires hardware platforms that can store large data sets across distributed clusters, and software capable of coordinate, combine and process data from multiple sources. New approaches to storing and analyzing data have emerged that rely less on data schema and data quality[2]. Instead, raw data with extended metadata is aggregated in a data lake[3] where machine learning and artificial intelligence (AI) programs use complex algorithms to look for repeatable patterns.[4]

The large size of data sets impact the work in many areas, including meteorology, genomics [14], complex physics simulations [15], biological and environmental research [16], Internet search, finance and business informatics. Data sets grow in

[1] Cloud computing is computing in which large groups of remote servers are networked to allow centralized data storage and online access to computer services or resources. Clouds can be classified as public, private or hybrid [24].

[2] Models and processes traditionally adopted in transactional and analytical data processing.

[3] A data lake is a storage repository that holds a vast amount of raw data in its native format until it is needed. While a hierarchical data warehouse stores data in files or folders, a data lake uses a flat architecture to store data. Each data element in a lake is assigned a unique identifier and tagged with a set of extended metadata tags. When a business question arises, the data lake can be queried for relevant data, and that smaller set of data can then be analyzed to help answer the question [25].

[4]

size in part because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers, and wireless sensor networks [17] [18] [19].

2.2 BIG DATA IN TRANSPORTATION

Big Data proponents argue that new insights to dealing with many transportation challenges will emerge from exploiting the vast datasets. Others point out that Big Data requires big judgment: Big Data is only useful if policy questions are framed correctly and if datasets are relevant to those [20].

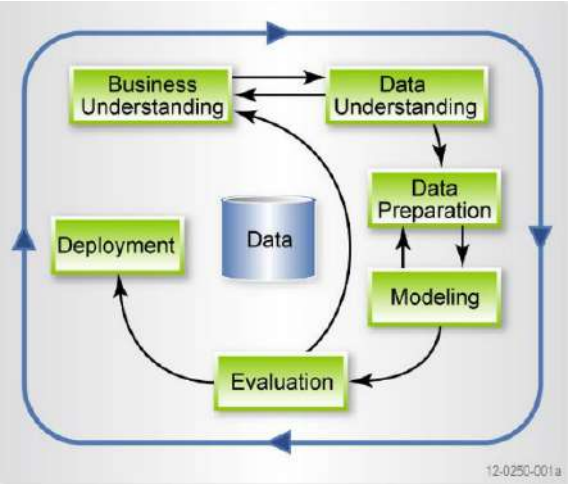


Figure 5: CRISP-DM Process [21]

Yet raw data has no or little value. The amount and speed of data collection is often greater than the capabilities of organizations to manage it. One challenge is to find the resources and skills and the storage capacity to deal with it. The main problem today is not collecting data but processing it to take full advantage of the information it contains. New technologies and analytical tools have enabled real-time data analysis, allowing immediate solutions for transportation challenges. Data is valuable when it creates societal value through, for example, development of policies to reduce congestion and improve infrastructure performance.

The gains from using big data are significant. It can help governments, businesses and individuals to make more informed decisions. Better data can help transportation authorities to understand commuters' behavior, provide targeted information and identify policy interventions. In fact, the biggest gains from using big data may come from changing user behavior. Singapore, for instance, uses data on local traffic conditions in real-time to determine prices for road tolls. This provides an incentive to drivers to avoid driving during the most congested periods and optimizes the use of the road network [22].

The world is becoming increasingly interconnected and intelligent. Around 80% of vehicles in Europe and North America will be two-way connected by 2018. But the full benefits of connectivity can be achieved only if vehicles are also connected to

infrastructure and other service providers. For example, intermodal car navigation proposes not only alternative routes but also alternative modes based on real-time information [23].

Big data provides new ways of gathering new information about transportation infrastructure from passenger and vehicle movements and allows for a shift from passive approaches to active crowd-sourcing with innovative transportation solutions. For example, some GPS systems enable users to inform others about incidents on the roads. This information is transferred to network operators in real-time, allowing for rapid responses to disruptions [23]. In Sweden, GPS data, radar sensor, weather and visibility data, along with other sources, provide information to the intelligent identification of current traffic conditions, estimating how long it would take to travel from point to point in target cities, offering advices on various travel alternatives, such as routes, and eventually helping traffic improvement in a metropolitan area [12].

3 URBAN MOBILITY

3.1 A GLOBAL ISSUE

Between 2010 and 2050, the number of people living in urban areas around the world is expected to grow by 80% - from 3.5 billion to 6.3 billion. This growth will create problems for urban mobility by increasing congestion, raising greenhouse gas (GHG) emissions, accelerating the deterioration of transportation infrastructure, and decreasing citizens' quality of life.

The challenges of urban mobility in big cities have two major causes: inefficient public transportation and massive use of private cars, perhaps because of the first cause (FGV, 2014). In Brazil, the basis of social pyramid has less access to public transportation, resulting in difficulties to go to work or use services, that, of course, impacts the economy. The time spent daily in transportation in nine biggest urban Brazilian areas, according to Akatu Institute (2014), is about 82 minutes that, converted in working hours, can represent R$ 300 billion/year or 7.3% of GDP.

Based on TomTom Company research, using data collected from GPS devices and mobile apps from vehicles of the 120 biggest cities around the world, Rio de Janeiro is ranked in the 3rd position on traffic density.

3.2 THE PROBLEM IN RIO

Rio de Janeiro has around 6 million people living in the city. The main mode of public transportation is the bus that carries daily over 3.3 million passengers [24] (ANNEX B). Bus transportation is responsible for over 62% of regular public transportation in the city, followed by train and subway, respectively by 6.62% and 6.59% of passenger transportation. The graphic below presents the public transportation market share, and it shows a little upgrade in the use of bus. Although it is projected to 2016 a reduction in the use of this modal, the bus still represents a strategic object of study for public policies.

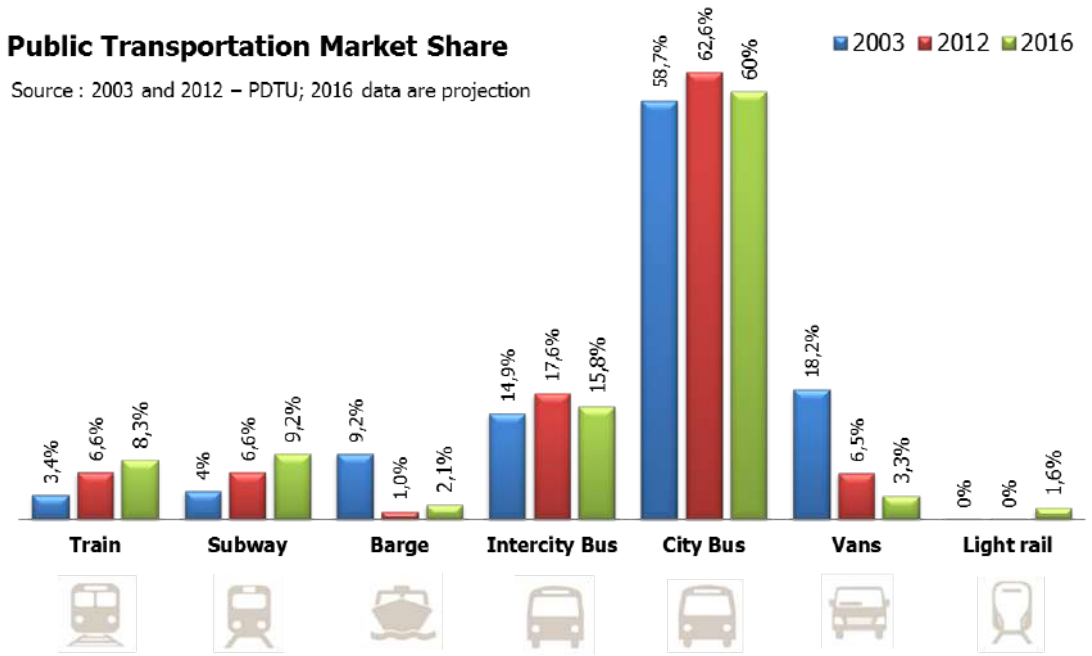


Figure 6 - Public Transportation Market Share

Huge numbers, huge challenges. Provide high quality transportation services, with punctuality, based on demand and optimized according to the city areas and their vocations is essential to improve urban mobility, which affects life quality and national economic performance.

Because of the quality and availability of the service, below the needs and expectations of carioca citizens, public transportation has been passed over in relation to the individual motorized transportation in most urban trips. The use of private cars, as IBGE research in 2012, increased about 120% from 2001 to 2012, with an annual growth of 11.8%. This propensity, in addition to causing significant impacts on infrastructure and the environment, contribute to congestion in big cities. The number of carioca citizens that spend more than two hours to go to work has increased 179% in the last 10 years, and more than 27% take more than one hour (Folha de São Paulo, 2013). Recent study by Federation of Industries of the State of Rio de Janeiro - FIRJAN - signals that traffic jams in Rio de Janeiro cost R$ 29 billion a year.

Further on the economic issue, the impact on quality of life of carioca citizens is disastrous and immeasurable.

Can urban planners and transportation policy decision makers curb some of these trends by better planning? If so, what information is available for them to enhance the relevance and accuracy of their projections? The focus of discussions on urban mobility is rapidly moving from roads to data analysis. (World Bank, 2014)

4 BUS TRANSPORTATION SERVICE IN RIO DE JANEIRO

The Public Bus Passenger Transportation Service - STCO-RJ - of Rio de Janeiro has adopted, since 2010, the concession regime. In the previous regime, 45 companies were authorized to run the transportation public service. After the concession, 4 consortia were given the right to operate, through the next 20 years, renewable for other 20 years, the bus lines in the 5 Regional Transportation Networks - RTR - described in the concession notice.

Figure 7 - Regional Transportation Networks

The consortia Transcarioca, Intersul, Internorte and Santa Cruz share an annual average fleet of 8.718 vehicles, among 697 lines that run almost 16 million trips traveling more than 730 million kilometers [24]. They generate a revenue of about 1,7 billion Brazilian reais for an operational cost of 1,6 billion Brazilian reais, employing directly over 40.000 people.

Playing the role of regulator sector, with the function of auditing and supervising the operation and planning the service, the Transportation Department (SMTR), in behalf of the Executive Power, has classified the transportation regular bus routes according to the regional passenger movements; the functional and hierarchical parameters; the geographical area; and the demand and vehicle characteristics.

SMTR has defined, in the concession notice, the parameterization basic concepts of the bus routes operation listed above:

- declared (or payed), repressed, transported and projected demand
- calculation and temporal variation of the demand (seasonality)
- peak period
- nominal vehicle capacity

- section occupation
- critical trip occupation
- occupation index by capacity
- average flow of passengers
- lines sizing
- trip time
- passengers renovation index
- maximum headway

The concession notice has established, among other, rules of extension, shortening, modification e canceling of regular lines; characteristics, accessibility and life cycle of the vehicles, and the type of fuel used; installation and maintenance conditions of garages, parking lots and bus stops; growth or reduction of the bus fleet that is closely related to the demand and section and peak periods maximum headway.

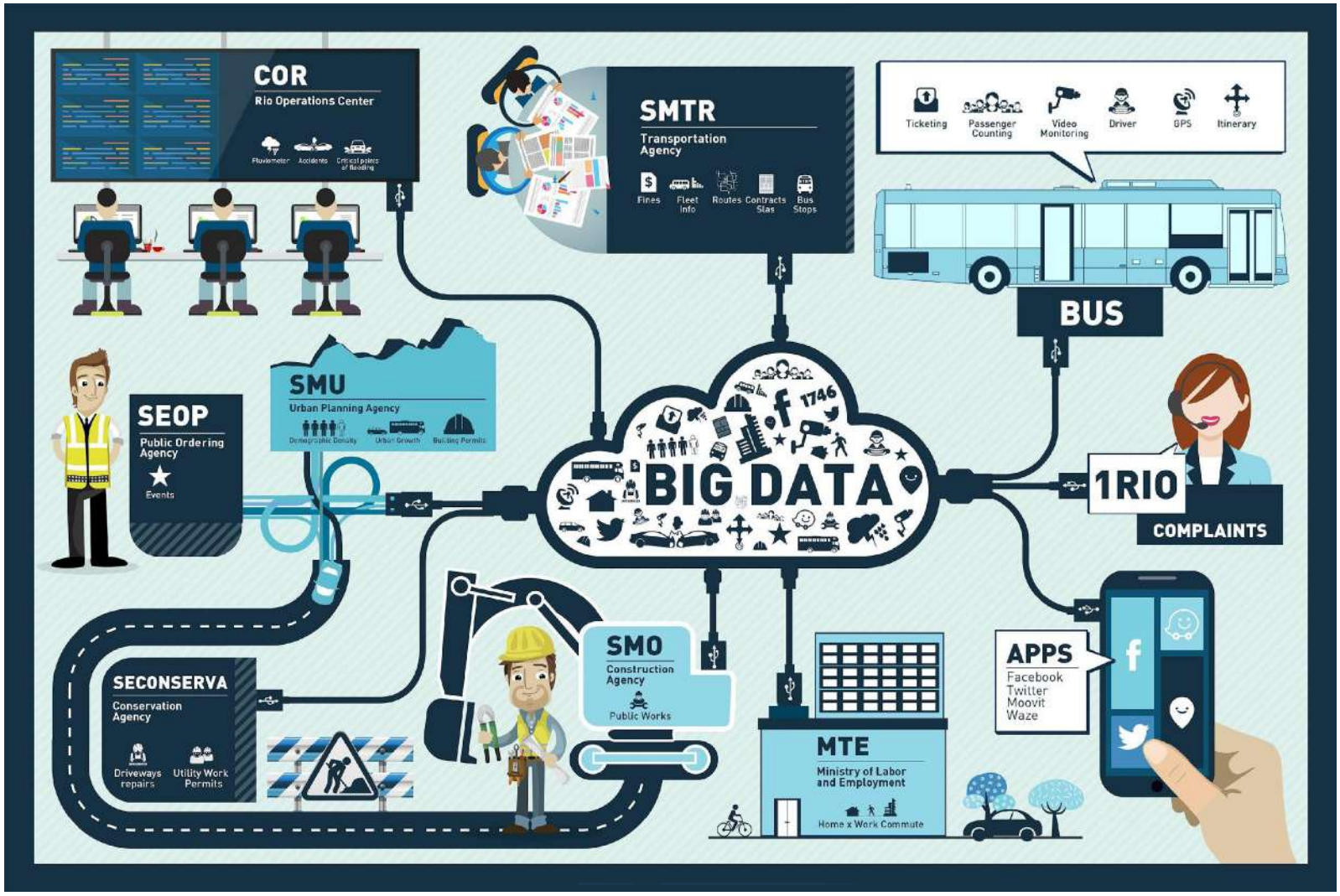
Furthermore, the consortia were obligated to install passenger safety equipment (video camera) and vehicle path monitoring equipment (electronic tachograph and GPS) which are valuable data sources for service delivery supervision and to support our big data model.

5 A BIG DATA MODEL TO SUPPORT RIO BUS SERVICE

5.1 MODEL VIEW

For a proper development of the model, it is very important acting in a theoretical and investigative way, not only by examining the existing data, but also figuring out all the data needed, considering the municipal plans for transportation.

A graphic schema was designed to ease the comprehension of the model. It shows involved stakeholders and correspondent mapped data sources.


COR
Rio Operations Center
Fluviometer
Accidents
Critical points of flooding
SMTR
Transportation Agency
Fines
Fleet Info
Routes
Contracts Slas
Bus Stops
Ticketing
Passenger Counting
Video Monitoring
Driver
GPS
Itinerary
BUS
SMU
Urban Planning Agency
Demographic Density
Urban Growth
Building Permits
SEOP
Public Ordering Agency
Events
1746
BIG DATA
1RIO
COMPLAINTS
SECONSERVA
Conservation Agency
Driveways repairs
Utility Work Permits
SMO
Construction Agency
Public Works
MTE
Ministry of Labor and Employment
Home x Work Commute
APPS
Facebook
Twitter
Moovit
Waze

5.2 WHAT THE MODEL INTENDS TO ANSWER

The model must be able to answer the following questions:

Where, when and how many buses are necessary to provide a good service?

Are the contracted Service Level Agreements - SLAs being accomplished?

Are the contracted SLAs meeting the city needs?

Are the existing bus routes suitable for the urban demographic growth?

How long does the passengers wait for a given bus route at the bus stop?

What are the quality levels of the fleet?

Are the fares balanced with operational costs?

Are the routes being performed?

How to rearrange the routes, in case of accidents or other events?

What are the citizen's feelings about the city bus service? Which routes have the most of complains?

What is the bus average speed comparing to the road average speed?

Which bus route do the citizens use to go to work and in what periods of the day?

How does an event impact the buses transportation patterns?

Which is the fare payment profile?

How is the driver performance?

What should be the appropriate fleet distribution by slot?

Are the existing bus routes suitable to reach the public buildings?

Real time information about route map and bus schedule.

Real time information to rearrange bus routes in case of rainstorms.

5.3 SOURCES' MAPPING

Going further in the concept of mapping, the sources were identified, their availability was verified, the existence of the IT system that manages them, the managers' identification, the location of these bases (below) and the necessary attributes of each source (Annex A).

Sources	Description	Information Provider	Database Location	IT System	Information owner		Update frequency	
					Agency	Name	Today	Desired
Bus GPS	All city buses have GPS equipments installed. These data are created by Rioonibus, that transmits in real time to a database located in City Hall Datacenter.	RIOONIBUS - Private company, the operates buses in Rio	City Hall Datacenter	GPS Base	Transportation Agency - SMTR	Alberto Nygaard	Real Time	Real Time
Daily Operations Report	This information is provided by Rioonibus, 40 days later, and it is summarized by day.	RIOONIBUS - Private company, the operates buses in Rio	City Hall Datacenter	Transportation BI	Transportation Agency - SMTR	Alberto Nygaard	Monthly	Real Time
Bus routes	Information on bus routes and their georeferenced itineraries	Transportation Agency - SMTR	City Hall Datacenter	SPPO	Transportation Agency - SMTR	Marcelo Estillac	Real Time	Real Time
Fleet information	Fleet information to analyze passenger comfort: air conditioner, wheelchair adapted, accessibility, bikes allowed etc.	Transportation Agency - SMTR	City Hall Datacenter	STU	Transportation Agency - SMTR	Lauro Silvestre	Real Time	Real Time
Contracts SLAs	The contracts are elaborated by Transportation Agency and are avaiable in 2010. These contracts define all service level agreements that must be accomplished by the buses companies.	Transportation Agency - SMTR	There is no IT System					When a new SLA is stablished
Fleet Maintenance	These informations are provided monthly by Rioonibus. They are used, among other informations, to calculate an accurate fare.	Transportation Agency - SMTR	There is no IT System, the information is sent using eletronic worksheets			Alberto Nygaard	Monthly	Monthly
Fuel Consumption	These informations are provided monthly by Rioonibus. They are used, among other informations, to calculate an accurate fare.	Transportation Agency - SMTR	There is no IT System, the information is sent using eletronic worksheets			Alberto Nygaard	Monthly	Monthly
Demographic Density	Information about the Rio population such as territorial distribution, age structure, education, social and economic profile of the families etc.	Urban Planning Institute - IPP - using data from IBGE (Brazilian Institute of Geography and Statistics)	Urban Planning Institute - IPP	Data Storage - Armazém de Dados	Urban Planning Institute - IPP	Luiz Arueira	When a new research occurs	
Public Buildings	Public Buildings georeferenced such as schools, hospitals, sport equipments, police stations, administration buildings etc.	Urban Planning Institute - IPP	City Hall Datacenter	Arcgis Software	Urban Planning Institute - IPP	Luiz Arueira	When a new public equipment is created	
Events	Information about events organized in Rio that could impact public transportation system.	Public Ordering Agency - SEOP	There is no IT System. Operation Center has this information in its databases				When a new event is planned	
Utility Work Permits	Information about utility work permits (scheduled or emergencial) that could impact the public transportation system.	Conservation Agency - SECONSERVA	City Hall Datacenter	SISMAC - Sistema de Monitoramento de ações de conservação	Conservation Agency	Marco Aurélio Regalo	Real Time	Real Time
Public Works	Information about infrastructural public works that could impact the public transportation system.	Construction Agency - SMO	There is no IT System. Operation Center has this information in its databases				When a new event is planned	
Accidents	Traffic accidents that could impact public transportation system.	Waze, COR, Twitter, Facebook	City Hall Datacenter	PENSA database	PENSA - Big Data Group	Pablo Cerdeira	Real Time	Real Time
Bus fines	Traffic fines received by bus drivers, which can be used to evaluate the quality of service.	Transportation Agency - SMTR	City Hall Datacenter	CITRAN	Transportation Agency - SMTR	Fernanda Ojeda	7 days	7 days

Sources	Description	Information Provider	Database Location	IT System	Information owner		Update frequency	
					Agency	Name	Today	Desired
Building permits	Real estate projects that may cause the need for redesign of urban transport in the region.	Urban Planning Agency - SMU	City Hall Datacenter	SISLIC	Urban Planning Agency -SMU	Maria Cristina Auler	7 days	7 days
Urban Growth	Information about the Rio population, including territorial distribution and expected growth.	Urban Planning Institute - IPP	Urban Planning Institute -IPP	Data Storage - Armazém de Dados	Urban Planning Institute -IPP	Luiz Arueira	Annual	Annual
Twitter	Listening channel on what people say on Twitter about the bus service in Rio.	PENSA - Big Data Group		Not applied	PENSA - Big Data Group	Pablo Cerdeira	Real Time	Real Time
Facebook	Listening channel on what people say on Facebook about the bus service in Rio.	PENSA - Big Data Group		Not applied	PENSA - Big Data Group	Pablo Cerdeira	Real Time	Real Time
Moovit	Evaluation by Moovit application on the quality of bus service in Rio	PENSA - Big Data Group		Not applied	PENSA - Big Data Group	Pablo Cerdeira	By request	Real Time
Waze	Information about road average speed and accidents.	PENSA - Big Data Group		Not applied	PENSA - Big Data Group	Pablo Cerdeira	Real Time	Real Time
1RIO Complaints	Complaints about the bus service made through the 1746 Call Center (1RIO).	Management Agency - CVL	City Hall Datacenter	SGRC	Management Agency - CVL	André Marques	Real Time	Real Time
Passenger counting	The Operations Daily Report provides consolidated information only. It is important to know how much users traveling on the bus per shift. The ratchets are now automated and can provide information about passengers input. We don't have, however, information on passengers leaving the bus.	Transportation Agency - SMTR	There is no information avaiable					Real Time
Bus Stops	Bus Stops georeferenced.	Transportation Agency - SMTR	There is no IT System				7 days	7 days
Critical points of flooding	The mapping of the critical points of flooding, associated with the pluviometer data, anticipates the need for redesign of bus routes	Operational Control Center - COR	City Hall Datacenter	The data is updated directly to the database	Operational Control Center - COR	Alexandre Cardeman	When the Summer Plan is made	When the Summer Plan is made
Pluviometer data	Identify the increase in rainfall associated with the critical points of flooding, can anticipate the need for redesign of the bus lines.	Operational Control Center - COR	City Hall Datacenter	The data is updated directly to the database			5 min	5 min
Home x work commute	Ministry of Labor and Employment information about place of residence, workplace and times of entry and exit of employees.	Ministry of labour and employment - TEM	There is no information avaiable				Monthly	Monthly
Carnival	During Carnival is important to know the dates, times and locations of blocks to redesign the bus routes.	Tourism Agency - Riotur	There is no IT System				When Carnival is planned	When Carnival is planned
Drivers' information	Bus drivers' information	Transportation Agency - SMTR	City Hall Datacenter	STU	Transportation Agency	Lauro Silvestre	Real Time	Real Time

5.4 QUESTIONS AND SOURCES

The sources and attributes mapped will allow answering, among others things, the following questions:

Questions \ Sources	Bus GPS	Driver's information	Daily Operations Report	Bus routes	Fleet information	Contracts SLAs	Maintenance	Fuel consumption	Demographic density	Public buildings	Events	Critical points of flooding	Bus fines	Building permits	Urban growth	Twitter	Facebook	Moovit	IRIO Complaints	Passenger counting	Bus stops	Home X Work Commute	Accidents	Public works	Pluviometer data	Waze
Where, when and how many buses are necessary to provide a good service?				■					■					■	■					■	■	■				
Are the contracted Service Level Agreements - SLAs being accomplished?	■		■		■	■														■						
Are the contracted SLAs meeting the city needs?				■	■	■			■					■	■	■	■		■	■		■				
Are the existing bus routes suitable for the urban demographic growth?				■					■					■	■					■		■				
How long do the passengers wait for a given bus route at the bus stop?	■																		■		■					
What are the quality levels of the fleet?					■											■	■	■	■							
Are the fares balanced with operational costs?			■				■	■												■						
Are the routes being performed?	■			■																						
How to rearrange the routes, in case of accidents or other events?	■			■							■												■	■		■
What are the citizen's feelings about the city bus service? Which routes have the most of complains?				■												■	■	■	■							
What is the bus average speed comparing to the road average speed?	■			■																						■
Which bus route do the citizens use to go to work and in what periods of the day?				■																■		■				
How does an event impact the buses transportation patterns?	■										■	■												■		■
Which is the fare payment profile?			■																	■						
How is the driver performance?		■		■									■			■	■	■	■							
What should be the appropriate fleet distribution by slot?	■			■		■														■		■				
Do the existing bus routes are suitable to reach the public buildings?				■						■																
Real time information about route map and bus schedule	■			■																						
Real time information to rearrange bus routes in case of rainstorms	■			■								■													■	

5.5 SOURCES AVAILABILITY

Already available

All GPS information, Daily Operations Report - RDO, bus routes, fleet information, contracts SLAs, maintenance and fuel consumption, information about public buildings and its location, events, public works, accidents, bus fines, building permits, estimated urban growth, Moovit, Waze, 1RIO complaints, bus stops, pluviometer data, critical points of flooding, data bus drivers and the planning of Carnival blocks are already available in the Rio City Hall databases.

Medium to long-term availability

The incorporation of some new bases, listed below, depends on other actions:

- Twitter and Facebook – it's necessary to establish an agreement.
- Passenger counting – nowadays, although the number of passengers getting in the bus is known, the quantity of passengers traveling on the bus at a given period of time is completely unknown.
- Home x work commute – it's necessary to establish an agreement with Ministry of Labor and Employment.

5.6 RESTRICTIONS

This model does not include all dimensions needed for a complete analysis of the demand. The following scenarios are not included:

- Passengers who were at bus stop and gave up taking the bus for any reason;
- Users of private cars that would use the bus, if they realize any improvement in the quality of the service.

5.7 RISKS

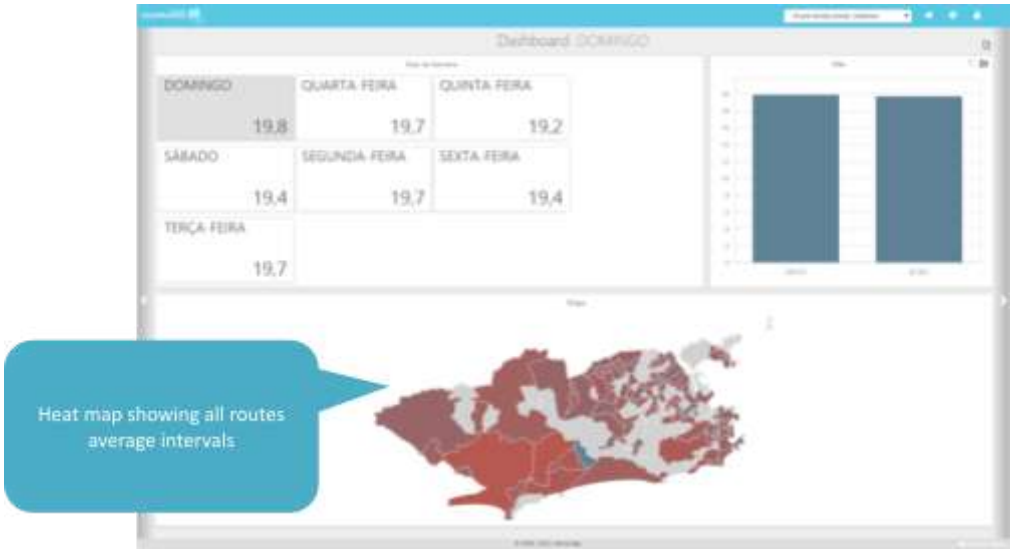
The main identified risks to the Project are shown below:

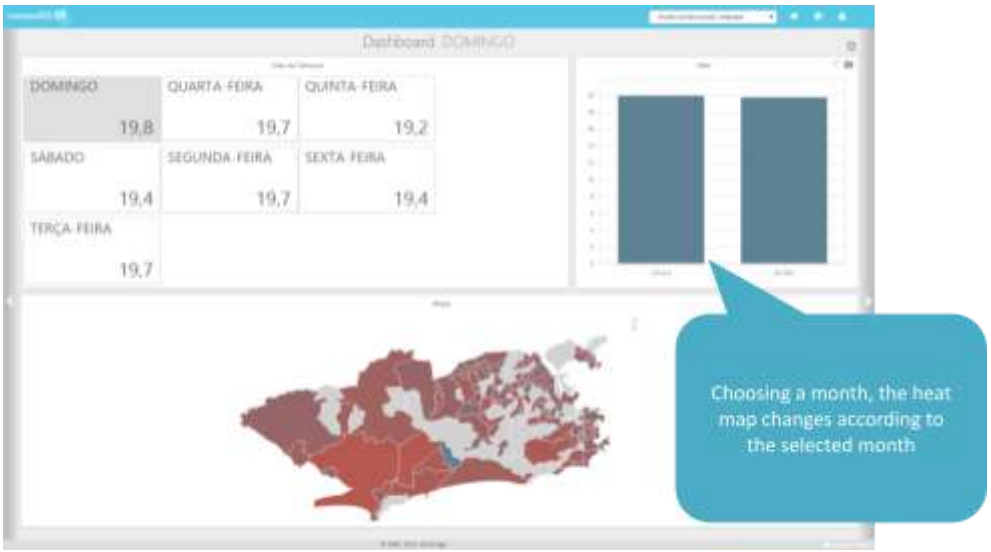
Risk	Probability	Impact
Lack of adherence of Rio transportation managers to the proposed model of Big Data and Public Transport	Low	High
Information managers do not allow access to the databases	Medium	High
Unavailability of hardware/software infrastructure necessary for model implementation	High	High
No implementation of mechanism for the passenger counting on the buses	High	Medium
No establishment of agreement with Ministry of Labor and Employment	Medium	Medium

5.8 PROOF OF CONCEPT

To ensure that the architecture is sufficiently comprehensive and flexible to achieve the main goal, our group has partnered with Stone Age, a service provider in the area of Big Data. The Pensa Group allowed access to needed information, in order to design a Proof of Concept – PoC, appropriate and feasible to the project deadline. So, this cutout focused in the analysis of city bus intervals, by time band, comparing to the contracted frequency.

The question "How long do the passengers wait for a given bus route at the bus stop?" and part of the question "Are the contracts SLAs being accomplished?" were answered by this PoC, using the model data sources: Bus GPS, Contracts SLAs, Bus stops, Bus routes.





6 PROJECT TIMELINE

The milestones for the project are listed below:

Activity	Date
PoC conclusion	Feb 2015
Detailed planning	May 2015
Budgeting	Jun 2015
Mayor approval	Jul 2015
State and federal agreements	Sep 2015

7 GOING FURTHER

Thinking beyond the project and in order to improve the bus service management, following are some suggestions:

- Enrich the model by using data from other transportation modals, such as subway, trains and intercity buses.
- Include pent-up demand data for bus transportation service
- Voice communications are controlled, prioritized and channeled to either one-on-one conversations or to broadcast messages from the dispatcher to groups or all buses.
- Provide text messaging between dispatchers and drivers.
- Videomonitoring with face recognition
- Fleet management dashboards provide system status information in at-a-glance views.
- Emergency alarm trigger aboard the bus to alert management to a situation and enable them to listen to audio and see camera images in real time from within the bus
- Deliver real time information via internet to people who are planning their travel, waiting for an arriving vehicle or already onboard, which allows people to access the predicted arrival time at their stop. This information can also be provided via displays in public places. Accurate and reliable passenger information is critical to increasing mass transportation ridership.
- Deliver real-time information to passengers onboard the vehicle to advise them of upcoming stops using audible announcements and lighted signs inside the bus. Audible announcements outside the bus also inform waiting passengers when the bus arrives at a stop.

8 CONCLUSIONS

The seeds gathered from Columbia program applied to our daily routine and mainly to the project development, brought us the following fruits:

- Establishing internal and external partnerships is a critical factor for a successful project implementation.
- The great part of sources used in our Big Data model is already available and the PoC has proved that it is feasible.
- The model can be used by other transportation agencies that have similar issues.
- The data collected may be freely available to everyone who wishes (open data)
- Joining Big Data analytics and accountability can lead transportation management to a higher step in our city, improving the quality of life in Rio.

9 BIBLIOGRAPHY

[1] M. Hilbert e P. López, “The World's Technological Capacity to Store, Communicate, and Compute Information.,” *Science,* vol. 332, n. 6025, p. 60–65, 2011.

[2] “IBM What is big data? — Bringing big data to the enterprise,” [Online]. Available: http://www.ibm.com/big-data/us/en/. [Acesso em 26 August 2013].

[3] “Big Data at the Speed of Business - What is big data?,” [Online]. Available: http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html. [Acesso em 23 November 2014].

[4] “The FOUR Vs of Big Data,” [Online]. Available: http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg. [Acesso em 23 November 2014].

[5] M. Rouse, “Big Data,” [Online]. Available: http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data. [Acesso em 18 December 2014].

[6] R. Magoulas e B. Lorica, "Introduction to Big Data". Release 2.0, Sebastopol CA: O’Reilly Media, Inc., 2009.

[7] L. Arthur, “What Is Big Data?,” 15 August 2013. [Online]. Available: http://www.forbes.com/sites/lisaarthur/2013/08/15/what-is-big-data/. [Acesso em 18 December 2014].

[8] P. Delort, 2014. [Online]. Available: http://www.bigdataparis.com/presentation/mercredi/PDelort.pdf?PHPSESSID=tv7k70pcr3egpi2r6fi3qbjtj6#page=4.

[9] S. A. Billings, Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains, Wiley, 2013.

[10] P. Delort, “Big Data Paris 2013,” 2013. [Online]. Available: http://www.andsi.fr/tag/dsi-big-data/.

[11] P. Delort, “Big Data car Low-Density Data ? La faible densité en information comme facteur discriminant,” [Online]. Available: http://lecercle.lesechos.fr/entrepreneur/tendances-innovation/221169222/big-data-low-density-data-faible-densite-information-com.

[12] “Demystifying Big Data - A Practical Guide to Transforming The Business of Government,” TechAmerica Foundation, Washington, D.C., 2012.

[13] A. Jacobs, “The Pathologies of Big Data,” *ACMQueue,* 6 july 2009.

[14] “Community cleverness required,” *Nature 455 (7209):1,* 4 9 2008.

[15] “Sandia sees data management challenges spiral.,” *HPC Projects,* 4 August 2009.

[16] O. J. Reichman, M. B. Jones e M. P. Schildhauer, “Challenges and Opportunities of Open Data in Ecology,” *Science,* vol. 331, n. (6018): 703–5., 2011.

[17] C. Surdak, Data Crush: How the Information Tidal Wave is Driving New Business Opportunities, 2014.

[18] J. Hellerstein, “Gigaom Blog,” *Parallel Programming in the Age of Big Data,* 9 November 2008.

[19] T. Segaran e J. Hammerbacher, Beautiful Data: The Stories Behind Elegant Data Solutions., O'Reilly Media, Inc., 2009.

[20] “Big Data in Transport: Applications, Implications, Limitations,” em *International Transport Forum - 2014 Annual Summit*, Leipzig, Germany, 2014.

[21] B. D. Murrow, “What Big Data Means to You?,” 21 March 2013. [Online]. Available: http://murrow.net/Portals/0/Publications/Brian%20Murrow%20-%20What%20Big%20Data%20Means%20To%20You.pdf. [Acesso em 18 December 2014].

[22] T. Y. Lui, “Singapore Urban Mobility,” em *International Transport Forum - 2014 Annual Summit*, Leipzig, Germany, 2014.

[23] E.-M. Huitema, “Smarter Transportation Data… the next natural resource for Smarter Cities,” em *International Transport Forum - 2014 Annual Summit*, Leipzig, Germany, 2014.

[24] Q. Hassan, “Demystifying Cloud Computing,” *The Journal of Defense Software Engineering,* pp. 16-21, Jan-Feb 2011.

[25] M. Rouse, “Data Lake,” [Online]. Available: http://searchaws.techtarget.com/definition/data-lake. [Acesso em 28 December 2014].

[26] J. Pontém, “Single Window - Best Practice and the Way Forward,” 2011.

ANNEX A – Attributes

Sources	Data Attributes
Bus GPS	LatLong
	Speed
	Bus Identification
	Date/time
Daily Operations Report	Bus Route
	Date YYYYMMDD
	Amount of kilometers covered
	Amount of travels
	Amount of passengers
	Amount of gratuities
	Amount of payments using Carioca Transport Card (BUC)
	Amount of payments using BUC integrated with other buses
	Amount of payments using BUC integrated with trains
	Amount of payments using Transportation Ticket
	Amount of payments in cash
Bus routes	Bus Route
	Route name
	Georeferenced itinerary
	Consortium of bus companies
	Bus company
	Maximum occupancy rate at peak hours
	Maximum occupancy rate off-peak hours
	Maximum occupancy rate during the weekend
	Determined fleet
	Working fleet on weekdays
	Working fleet on Saturdays
	Working fleet on Sundays and holidays
	Maximum interval
	Average travel time
	Average flow of passengers
Fleet information	Bus Identification
	Bus License
	Wheelchair adapted (Y/N)
	Air conditioner (Y/N)
	Acquisition date
	Last inspection date
	Vehicle type
	Legal requirements OK (Y/N)
	Documentation OK (Y/N)
	Vehicle features within the standard (Y/N)
	Data collection equipment (Y/N)
	Electronic tachograph (S / N)
	GPS (Y/N)
	Left door (Y/N)
	External layout within the standard (Y/N)
	Allows installation of surveillance and monitoring equipment (Y/N)
	Video camera (Y/N)
	Use fuel within the standard (Y/N)
	Nominal vehicle capacity
	Engine within the standard (Y/N)
Contracts SLAs	Consortium of bus companies
	Start date
	Period
	Operating Region
	Garage within the standard (Y/N)
	Parking area within the standard (Y/N)
	IT Infrastructure within the standard (Y/N)
	Number of garages within the standard (Y/N)
	Number of parking areas within the standard (Y/N)
	Number of towing positions within the pattern (Y/N)
	Accounting information within the standard (Y/N)
	Audit information within the standard (Y/N)
	Penalties Imposed (Y/N)
	Electronic Ticketing System within the standard (Y/N)

Sources	Data Attributes
Maintenance	Consortium of bus companies
	Bus Identification
	Maintenance costs
	Month/Year
Fuel Consumption	Consortium of bus companies
	Bus Identification
	Fuel Consumption costs
	Month/Year
Population Density	District
	Year
	Total area
	Land area
	Number of people
Public Buildings	LatLong
	Public Building Type
	Public Building Name
	Public Building Capacity (number of people)
Events	Event date/time
	Event name
	LatLong
	Place
	Public expected
	Duration
	Event type
Public Works	LatLong
	Scheduled or emergency
	Start date/time
	End date/time
	Responsible agency
Accidents	LatLong
	Accident date/time
	Accident type
Bus fines	Bus license
	Fine date/time
	LatLong
	Fine type
Building permits	LatLong
	Start date
	End date
	Number of housing units by type
Urban Growth	District
	Year
	Estimated growth rate
Twitter	Date/time
	Twitter Id
	Latlong
	Hashtag
	Comment
	Re-tweet from
Facebook	LatLong
	Facebook Id
	Comment
	Date/time
Moovit	LatLong
	Date/time
	Evaluation
	Comment
	Bus Route
Waze	LatLong
	Date/time
	Event type
	Speed
	Comment

Sources	Data Attributes
1RIO Complaint	Date/time
	Bus Route
	Bus Identification
	Bus license
	Type of complaint
	Subtype of complaint
	Description
Passenger counting	Lat Long
	Date/time
	Bus Route
	Bus Identification
	Amount of passengers inside the bus
Buses Stops	Latlong
	Reference point
	Type
	Bus stop within the standard (Y/N)
Critical points of flooding	LatLong
	Place
Pluviometer data	LatLong
	Telemetry (mm)
Home x work commute	Home LatLong
	Workplace LatLong
	Date
	Employee entry time
	Employee exit time
Carnival	LatLong
	Description
	Start date/time
	End date/time
Driver´s info	Name
	Identification number (CPF)
	Driver's license number
	Expiration date
	Issue date
	Birth date
	Sex

ANNEX B – 2013 Monthly Operational Data of Rio de Janeiro

Estado do Rio de Janeiro - Mobilidade Outlook 2012 - © FETRANSPOR 2014
tabela 4
Versão 1 - Último update: 08- Maio -2014

Tabela 4 - Dados operacionais mensais do MUNICÍPIO DO RIO DE JANEIRO ano de 2013

MÊS	Nº LINHAS	FROTA	VIAGENS	QUILOM.	PAX PAGANTES	PAX GRATUIDADES	I.P.K*	P.M.M**	DIESEL***	L / KM	IDADE FROTA	PESSOAL
jan/13	708	8.693	1.375.911	61.959.372	80.775.152	14.064.259	1,30	7.128	23.570.078	0,3804	3,30	40.858
fev/13	710	8.701	1.188.899	53.924.000	71.585.034	13.811.527	1,33	6.197	20.524.693	0,3806	3,35	40.644
mar/13	710	8.690	1.192.348	56.308.286	75.082.500	17.270.592	1,33	6.480	21.985.883	0,3905	3,46	40.763
abr/13	708	8.690	1.305.446	59.668.156	83.323.157	19.584.398	1,40	6.866	22.927.550	0,3843	3,46	41.228
mai/13	706	8.699	1.347.020	61.274.787	85.546.427	20.040.646	1,40	7.044	25.411.722	0,4147	3,54	41.619
jun/13	708	8.714	1.261.280	55.848.277	81.627.348	19.698.094	1,46	6.409	22.833.418	0,4088	3,59	40.482
jul/13	699	8.725	1.328.833	59.265.195	85.281.603	16.255.604	1,44	6.793	26.196.736	0,4420	3,61	41.263
ago/13	693	8.749	1.377.322	74.061.492	87.088.411	19.720.204	1,18	8.465	23.479.776	0,3170	3,65	41.472
set/13	683	8.783	1.307.619	58.118.481	84.811.437	19.279.105	1,46	6.617	24.265.207	0,4175	3,62	41.213
out/13	684	8.734	1.406.385	62.553.378	90.657.291	19.855.569	1,45	7.162	24.325.589	0,3889	3,65	41.150
nov/13	684	8.715	1.506.723	68.908.652	85.232.975	19.466.198	1,24	7.907	21.759.638	0,3158	3,66	42.271
dez/13	687	8.725	1.391.090	61.755.512	82.873.763	15.872.742	1,34	7.078	26.196.663	0,4242	3,74	40.912

Fonte: DATABANK FETRANSPOR / RIOÔNIBUS
Nota: * Índice de passageiro pagante por quilômetro; ** Percurso médio mensal; *** Litros consumidos.

MÉDIA	697	8 718	1.332.488,83		82.823.341,50	17.909.911,50
					Média PAX/dia	3.357.775,10