International Journal for Quality Research 13(4) 769–782
ISSN 1800-6450

José Luís Pereira [1]
Marco Costa

Article info:
Received 05.06.2019
Accepted 05.09.2019

UDC – 004.451.5:005.311.6
DOI – 10.24874/IJQR13.04-01

FROM NOSQL DATABASES TO DECISION SUPPORT SYSTEMS: DEVELOPING A BUSINESS INTELLIGENCE SOLUTION

Abstract: *We are living a time in which the data generated by humans and machines has reached levels never seen before – the so-called era of Big Data. Everyday, vast amounts of data, coming from different sources and with different formats, are created and made available to organizations. First, with the rise of the social networks and, more recently, with the advent of the Internet of Things (IoT), data with enormous potential for organizations is being continuously generated. In order to be more competitive, organizations need to explore all the richness that is present in those data. Indeed, data is only as valuable as the insights organizations gather from it to make better decisions, which is the main goal of Business Intelligence (BI). In this paper, we describe the development of a decision support system in which data obtained from a NoSQL database is used to feed a BI solution.*

Keywords: *Big Data; NoSQL Databases; Decision Support Systems; Data Warehouses; Business Intelligence.*

1. Introduction

In the last decades we have witnessed an increase in the volume of data that is produced by organizations and by people in their daily life activities. In the latter case, as a result of the boom occurred with the social networks, increasing amounts of data are being generated by people, either by themselves and/or as a result of the interaction with other people. These new data have great potential for organizations as a source of insight about people needs, opinions, market tendencies, and so on.

According to IDC, ninety percent of the existing world's data was created over the past two years (Vesset et al., 2012). As of March 2017, there were 1.28 billion active Facebook users exchanging and sharing information in daily basis (source: Facebook). By the same time, 1.3 billion Twitter users were sharing approximately 500 million tweets a day (source: Twitter). Today, Youtube users upload 300 hours of new video every five minutes. In the telecommunications and financial sectors, the volume of transactional data generated daily is becoming overwhelming. In the near future, with the so-called Internet of Things (IoT), in which virtually any electronic device with processing capacity will be integrated in the Internet, generating and consuming data, the amount of data we will have to deal with will increase dramatically. According to the Statista Portal, the global Big Data industry will worth close to 45 billion dollars in 2017 (Statista, 2017).

The new data come in larger amounts, at higher rates, from different sources, and with distinct features. In this context, one might distinguish among three different kinds of data to store and process (Halper & Krishnan, 2013):

[1] Corresponding author: José Luís Pereira
Email: jlmp@dsi.uminho.pt

- **Structured data** – data with a rigid and previously known structure, in which all elements share the same format and size. This is the kind of data, traditionally found in business applications, that has been stored in relational databases;
- **Semi-structured data** – data with a high degree of heterogeneity, which is not easily represented in fixed data structures. Typically, this kind of data has been stored using specific languages/formats such as XML (*Extensible Markup Language*) data, RDF (*Resource Description Framework*) data, JSON (*JavaScript Object Notation*), and so on;
- **Unstructured data** – data without a structure, such as text, video, or multimedia content. This kind of data has grown exponentially during the last decade, with some estimates pointing that, nowadays, 80% to 90% of the generated data is unstructured data. Examples include documents, images, photos, email messages, videos, and so on.

Although, much of the data growth has been in unstructured data, IDC estimates that by 2020, business transactions on the Internet will reach 450 billion per day (Reinsel & Gantz, 2012).

With Big Data organizations understood the enormous potential underlying those vast amounts of available data. They only had to use the right tools to treat those data, in order to better understand their business and their markets. *Business Intelligence* (BI) solutions are what organizations need to access those data and extract the insights needed to make the best decisions and outshine their competition.

By definition, BI is the collection of methods and tools that allow organizations to transform data into valuable information to support decision-making (Kimball & Ross, 2013). Since data may come from different sources and in a multitude of formats, BI tools need to have the capacity to *Extract* data from those sources, to *Transform* those data (selecting, cleaning, joining, calculating, coding/decoding, etc.) according to the purpose of the solution, and to *Load* the data into a repository commonly known as Data Warehouse (DW). In addition to the **ETL** capabilities, BI tools provide the mechanisms to build suitable information delivery front-ends for decision makers, such as reports and dashboards (see Figure 1).

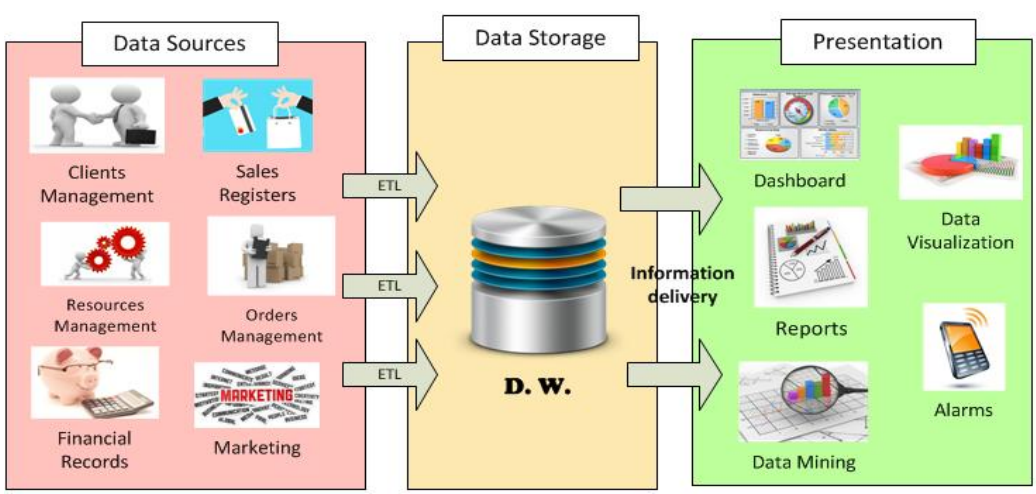


Figure 1. Global Perspective of a Business Intelligence Solution

In the era of Big Data important changes are occurring in the architecture of a BI solution, namely in the "Data Sources" part of the architecture, regarding the technologies used to store the data that feeds the BI solution (Chen, Chiang & Storey, 2012).

What characterizes the era of Big Data are the huge amounts of structured, semi-structured and unstructured data, produced and consumed at increasing higher rates. These new features constitute an enormous challenge to the more traditional relational database technology. To answer to the new challenges created by Big Data, a new family of database technologies has emerged – the NoSQL databases (Atzeni, Bugiotti & Rossi, 2013; Kuznetsov & Poskonin, 2014).

Nowadays, there are four families of NoSQL databases: *Document*, *Column*, *Key/Value* and *Graph* databases (Sadalage & Fowler, 2012). Each one of them with their own characteristics, strengths and weaknesses, but all sharing the same goal: to deal with the new challenges brought by Big Data.

Despite their youth, NoSQL databases are becoming major players in the database market. For instance, the well-known DB-Engines Ranking, which ranks databases according to their popularity, puts three NoSQL databases in the Top 10: MongoDB (a document-based database), Cassandra (a column-based database) and Redis (a Key/Value database) (DB-Engines, 2017).

This paper expands and improves two other papers from the same authors (Costa & Pereira, 2015; Pereira & Costa, 2016). We describe the development of a decision support system, materialized in a BI solution, in which the principal data source consists of a NoSQL database. This system is now being used by a Portuguese firm, which deals with the detection and monitoring of people movements in closed spaces, using a very common and disseminated technology – GSM (*Global System for Mobile Communications*), which we all use, nowadays, in our mobile phones. In this project, client movements inside a shop of a large shopping mall were used as data. The purpose of the BI solution is to provide decision makers with information about the habits of the shoppers, the time spent in shopping, the locals visited, etc., which is relevant for them in order to decide how to better organize the store space.

Concerning the structure of this paper, after a very brief presentation of the main concepts around Big Data and the database technology that promises to solve its major challenges - NoSQL databases, we made a very concise introduction to the area of Business Intelligence, stressing its value to support decision-making in organizations. In the next sections, we describe a development project in which data captured from a NoSQL database is used to feed a specific BI solution. To begin with, we describe the real context in which the BI solution was conceived and then we quickly advance to its development. We divided this project in two parts: the first part deals with the Extraction, Transformation and Loading (ETL) of the NoSQL data into a local database; the second part of the project involves the construction of a set of dashboards to present information in order to support decision-making. Finally, some conclusions about the project are presented and future work is envisaged.

2. The Context of the Project

In order to better manage a shopping store, decision makers would like to know simple facts such as "how many visitors walk by a shop?", "how many visitors enter a shop?", "how many visitors made acquisitions?", "which are the busiest and the quietest hours?", "How much time shoppers spend in the shop?", "which are the zones most visited in the shop?", and many more. To accomplish that, a system for the detection and monitoring of people movements in the store must be in place. Luckily, nowadays almost everyone uses mobile phones so, making use of the GSM technology, in particular using the IMEI (*International Mobile Equipment Identity*) and the IMSI (*International Mobile*

Subscriber Identity), one can easily trace the movements of people in a monitored space. This is a very convenient solution, as those "mobile identifiers" are never switched off (mobile phones only stop emitting a signal if their battery is removed). Therefore, with a convenient distribution of GSM sensors in a given space one can trace the movements of people in that area.

The data used in this project were obtained mostly through sensors installed in a sporting goods store located in a Portuguese shopping mall. Data are collected and stored in a NoSQL database (in this case, a Cassandra database system is used), all day long, every day of the week, non-stop. Using an API (*Application Program Interface*), the Cassandra database provides access to the data in the JSON format (*JavaScript Object Notation*), which is a very simple and convenient format. These data are used to feed the developed BI solution (see Figure 2).

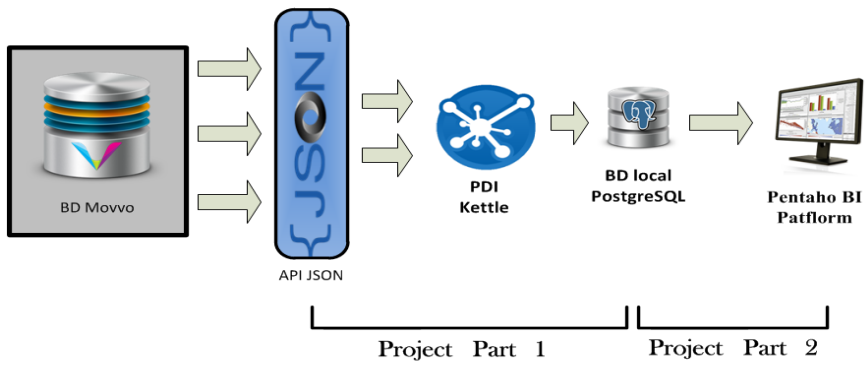


Figure 2. Architecture of the BI Solution

The development of the BI solution involved two parts. In the first part, data are retrieved from a NoSQL database – a Cassandra system - using the provided API and, after some processing tasks, are stored in a local database. In the second part a suitable set of dashboards is developed, according to the elicited requirements of decision makers (Costa & Pereira, 2015; Pereira & Costa, 2016).

Regarding the technologies used in the development of the BI solution, in addition to the PostgreSQL used to manage the local data repository, we selected the Pentaho family of products. In particular:

- **Pentaho Data Integration (Kettle)** – the solution offered by Pentaho for ETL. In this project Kettle was used to extract data from the provided Cassandra API, do the necessary treatments, and store the resulting data in the PostgreSQL database (PDI, 2015);
- **Pentaho BI Platform** – A platform that allows us to take the data from the repository and turn it into useful information for decision makers, by providing tools for creating reports and information panels, or dashboards (PBIP, 2015).

In the next section, the first part of the BI solution is described.

3. The BI Solution - Data Extraction, Transformation and Loading

In order to develop a BI solution adequate to support the needs of decision makers, regarding the management of a sporting goods store located in a shopping mall, we need to access data from which we may extract some metrics. The following table (Table **1**) summarizes the metrics that were considered relevant by the store managers.

Table 1. Metrics needed for Decision-Making

Metric	Description
space-tickets	Number of registered sales in store
space-walk-bys	Number of detected people passing in front of the store
space-visitors	Number of detected people inside the store
space-visiting-time	Average duration of visits to the store
zone-visitors	Number of detected people in each zone of the store
weather	About the weather and temperature

To get those data using the provided API, an HTTP request such as the following has to be issued, in which the metric "space-visitors" is requested:

```
https://(...)/days/2015-08-
15T00:00:00Z/2015-08-
16T23:00:00Z?metrics=space-
visitors:hour:series
```

In this example we issued a request to search for data about the number of visits to that store from 15 of August to 16 of August of 2015. An extract of the resulting response follows (in JSON format):

```
{
  ▾ "space-visitors:hour:series": [
      ▸ {… },
      ▾ {
            date: "2015-08-15T12:00:00+01:00",
            value: 346
        },
      ▾ {
            date: "2015-08-15T13:00:00+01:00",
            value: 322
        },
      ▾ {
            date: "2015-08-15T14:00:00+01:00",
            value: 428
        },
      ▸ {… },
```

As can be seen, 346 visits were detected in the store at Noon on 15 of August of 2015. In fact, this represents a total of visits to the store between 11:00 AM and Noon; 322 visits between Noon and 01:00 PM; and 428 visits between 01:00 PM and 02:00 PM. The result set continues through the remaining hours of the day, in the two days requested.

The API provided allows us to group the various metrics in a single request, simply by writing the HTTP request as follows:

```
https://(...)/days/2015-08-
15T00:00:00Z/2015-08-
25T23:00:00Z?metrics=space-
tickets:hour:series,space-walk-
bys:hour:series,space-visiting-
time:hour:series,space-
visitors:hour:series,zone-
visitors:day:series
```

To process the data obtained from the Cassandra database system and to store them into a local database (in this case, a PostgreSQL database system) a set of ETL steps were developed in Kettle. This specific ETL was developed in order to be autonomous, that is, it does not require the user to enter the dates in the requests to the API provided. In **Error! Reference source not found.** we can see the developed ETL steps used to Extract data from the provided API, do the necessary Transformations, and finally Load the data into the local database (tables Zones, Dates, Hour_records and Meteo).

The experience using the graphical interface of Kettle (known as Spoon) to develop ETL has been quite interesting and rewarding. Spoon has a wide range of available steps, such as Data Input and Output, Statistics, Validation, Mapping, Utilities, and so on, which may be added to the workspace in a drag-and-drop fashion.

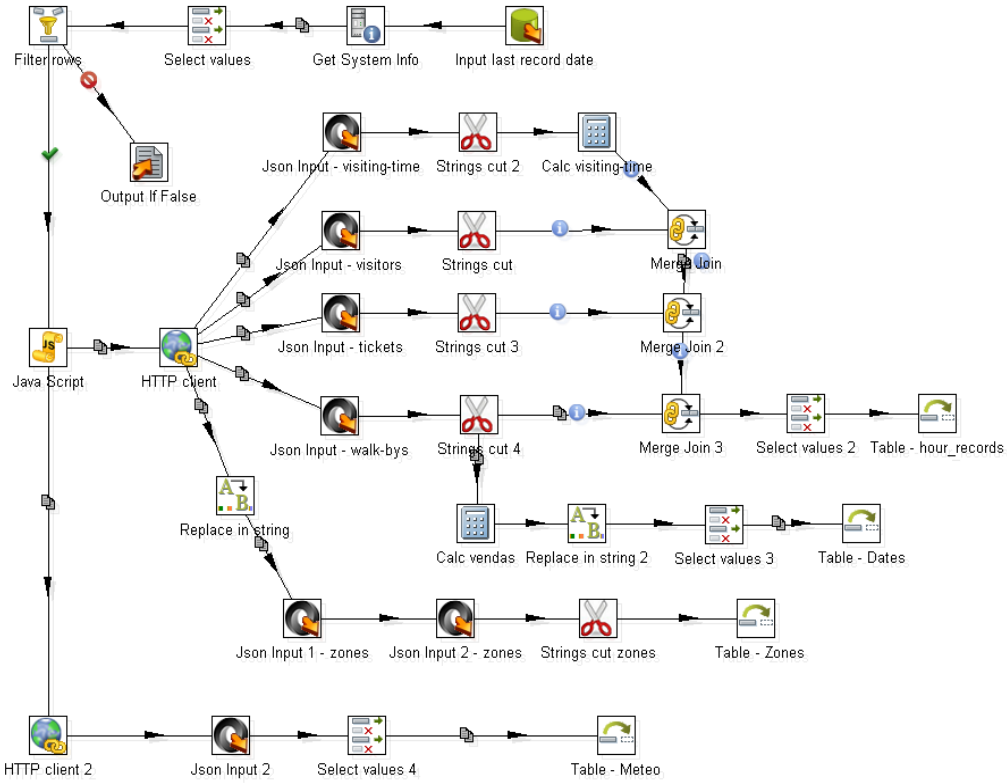


Figure 3. The ETL used to load and refresh the local PostgreSQL database tables

4. The BI Solution - Dashboards

A dashboard is a suitable visual representation of relevant information, which allows decision makers to quickly recognize the most pertinent aspects of their business (Malik, 2005). Dashboards may help organizations to consolidate the information about the business in a visually attractive format that is concise and easy to interpret.

In this project the development of the BI solution involved the exploitation of the Pentaho BI Platform tool to visually display the data through a set of dashboards. This set of dashboards is composed of several visual components, such as graphics, tables and even a map of the store. They were developed using a tool named CDE (*Community Dashboard Editor*), an open source tool designed to simplify the creation and editing of dashboards.

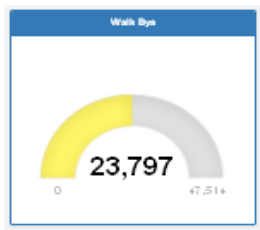
The first thing to do in the development of a dashboard is to identify "what" information we would like to have and "how" it should be displayed. In this project, after a careful requirements analysis involving the relevant decision makers, it was decided that the set of dashboards should include the following eight elements:

- **Dashboard 1** - The value of the metrics 'space-walk-bys', 'space-visitors' and 'space-tickets' for a certain month, compared to the maximum ever recorded in a previous month;
- **Dashboard 2** - Proportion of the metrics 'space-walk-bys', 'space-visitors' and 'space-tickets' over a given month;
- **Dashboard 3** - Overview of the metrics 'space-walk-bys', 'space-visitors' and 'space-tickets' over a given year;

- **Dashboard 4** - The relation between the metrics 'space-visitors' and 'space-tickets' in the form of an area chart;
- **Dashboard 5** - Overview of the metric ‘space-visiting-time’ over a given day;
- **Dashboard 6** - A 'Pie Chart' graphic and table with the data about the store zones and correspondent 'zone-visitors';
- **Dashboard 7** - Map of the store indicating the most visited zones;
- **Dashboard 8** - List of the 10 weeks in which there were more sales and its comparison with the number of visits.

Regarding the **Dashboard 1**, one can use a ‘Gauge Component’ to visualize each one of the three metrics required. With this component, the metric values are displayed within a range, thus giving the user a better sense of the magnitude of its value. The range is set between 0 and the maximum value ever recorded in a previous month. This component was built to be feed with eight values: the title, the value of the metric, the minimum value of the scale, the maximum value of the scale, the main color of the component, the color for the minimum value, the color for the medium value, and the color for the maximum value. In Figure 3 we show the aspect of the component used to visualize the metric 'space-walk-bys', and the SQL code used to feed it.

a)



b)

```
Select
    CONCAT(' ') as title,
    aux2.walk-bys,
    CONCAT('0') as min,
    MAX(aux.walk-bys),
    CONCAT('#E0E0E0'),
    CONCAT('#FFF567'),
    CONCAT('#FFF567'),
    CONCAT('#FFF567')
from (
    select hour_records.date, sum(hour_records.space-walk-bys) as walk-bys
    from hour_records,dates
    where dates.date=hour_records.date and month in
                (select month from dates where date=${pdate})
        group by hour_records.date
        ) as aux,
        (select sum(outdoor) as walk-bys
    from hour_records
    where date=${pdate}
    group by hour_records.date) as aux2
group by aux2.walk-bys
```

Figure 3. a) A component of Dashboard 1; b) The SQL needed to feed it

The **Dashboard 1** is very relevant to decision makers as it allows them to have a notion of “how many people have passed in front of the store in a given month, and the maximum value ever recorded”; “how many visits the store has received in a given month, and how does it compares to the maximum recorded previously”; and “how many sales were made

in a given month, compared to the maximum number of sales ever occurred in a month". For instance, in **Error! Reference source not found.**, we can see that, at a given month, 33.622 people have passed in front of the store (from a maximum of 47.514), from which 3.316 have entered the store (from a maximum of 7.426), resulting in 727 sales (from a maximum of 1.403 sales).

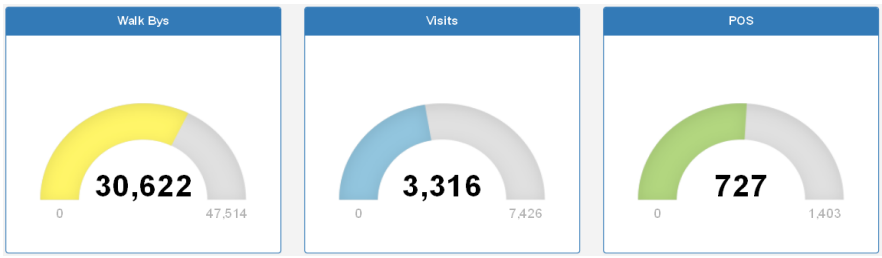


Figure 5. Information about the metrics 'space-walk-bys', 'space-visitors' and 'space-tickets'

Dashboard 2 may be considered complementary to the Dashboard 1, as it gives users of the BI solution a more comprehensive analysis of the metrics 'space-walk-bys', 'space-visitors' and 'space-tickets', in a given month, showing the proportion between these three metrics (Figure 4). This visual element allows decision makers to have a better notion of "which is the proportion between the people who has passed in the front of the store, the people who had entered the store, and the people who had actually shopped something".

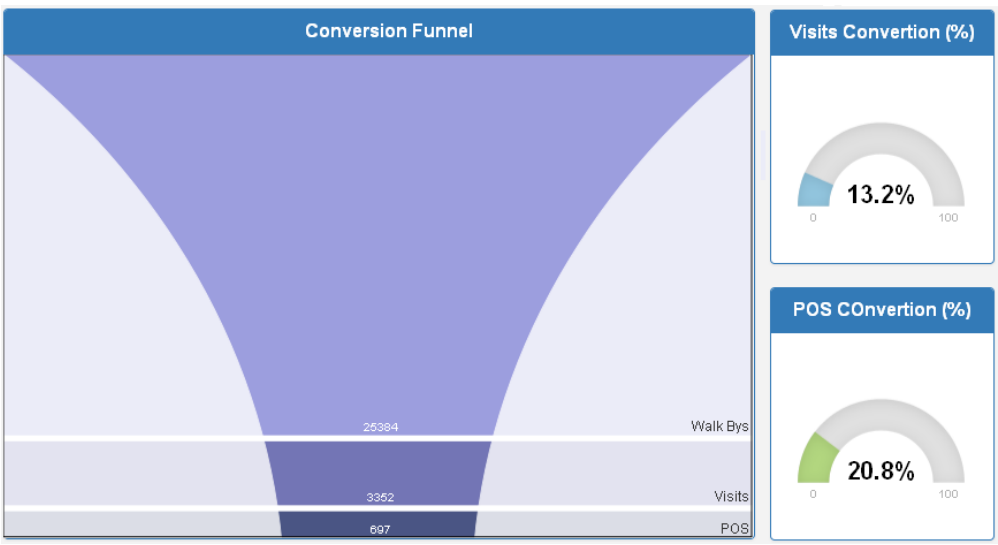


Figure 4. Proportion of the 'space-walk-bys', 'space-visitors' and 'space-tickets' in a given month

Dashboard 3 allows users of the BI solution a widespread analysis of the metrics 'space-walk-bys', 'space-visitors' and 'space-tickets' throughout a year. To do that the dashboard includes a line chart with the data regarding those three metrics, obtained during a year (Figure 5). For a more detailed view this component allows us to display each metric,

individually, by simply hiding the others.

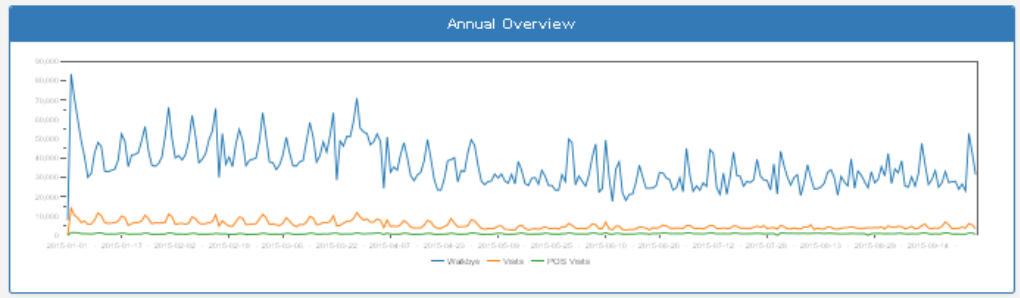


Figure 5. Evolution over a year of the metrics 'space-walk-bys', 'space-visitors' and 'space-tickets'

Dashboard 4 consists of an area chart which allows users to relate the metrics 'space-tickets' and 'space-visitors' in a given day, along with three text components, highlighting the hours and the maximum and minimum values recorded for the metric' Space-visitors' and the average value of the selected day. As can be seen in Figure 6, in a given day, around 3:00 PM (15:00) the store visits reaches the peak with 337 visitors, while in the same day at 6:00 PM (18:00), the store is in its quietest moment, with only 148 visitors. Another relevant information, which might be observed in this dashboard, is that after 10:00 PM (22:00) there are no more sales in the store, with people merely walking by.

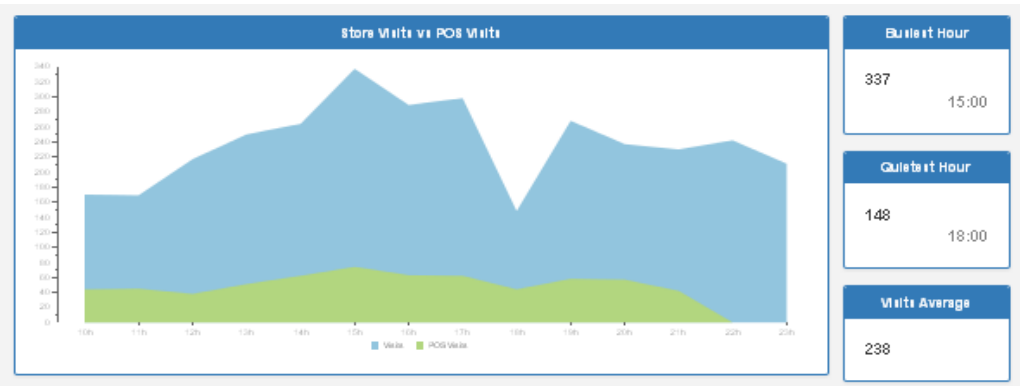


Figure 6. The relation between the metrics 'space-visitors' and 'space-tickets'

Dashboard 5 uses the metric ‘space-visiting-time’ to give decision-makers a notion of how long visitors stay inside the store. This visual element allows decision-makers to know “how long costumers remain inside the shop during the day”. As can be seen in Figure 7, depending on the time of the day, there are some differences in the visits duration of the shoppers, with the longest visits (around 20 minutes) at 05:00 PM (17:00), 07:00 PM (19:00) and 10:00 PM (22:00), and the shortest visits (around 7 minutes) at 04:00 PM (16:00). This is very relevant information for decision makers as they can use it to better manage the store.

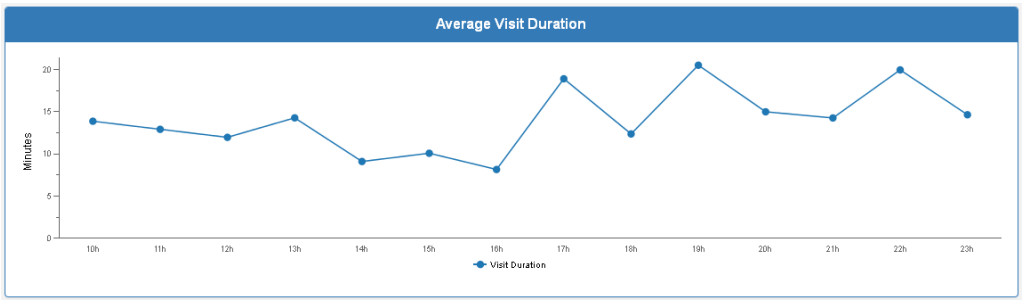


Figure 7. Overview of the metric 'space-visiting-time' over a given day

In **Dashboard 6** we can see a Pie Chart indicating the distinct zones of the store and the number of visitors in those zones during a given day. The table on the right list the zones in descending order of the number of visitors. By observing Figure 8 we can conclude that there are substantial differences in the number of visitors per zone in the store, with the "Fitness Equipment" zone (with 1577 visitors) largely outperforming the "Water Sports" zone (which had only 36 visitors) in that given day.

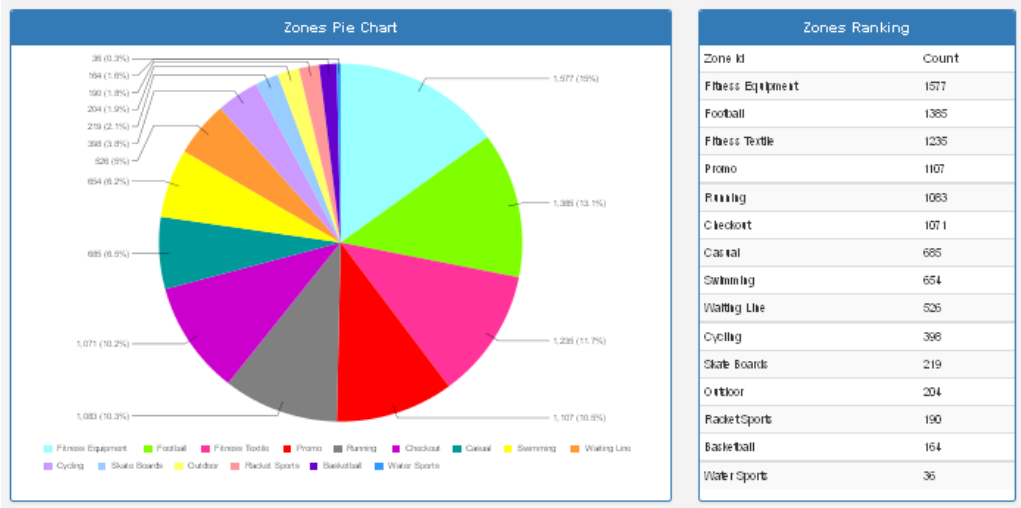


Figure 8. The zones of the store and their visitors in a given day

Dashboard 7 is a very insightful one as it illustrates the areas of the store that are most visited by shoppers (represented in a darker blue color), allowing decision-makers to know what are the zones of the store which deserve more attention and careful arrangement of the space. Thus, by just looking at the dashboard 7 (Figure 9), which depicts the map of the store with the several different zones in which it is organized, this visual element allows decision makers to answer the question "Which store areas are receiving more visits by people?".

Finally, **Dashboard 8**, as was asked by decision makers, proposes to represent the ten weeks of a given year in which there were more sales in the store (Weekly POS Visits), making a comparison with the number of visitors in those periods (Weekly Store Visits). As might be observed in Figure 10, a higher number of visitors to the store in a certain week does not necessarily means a higher number of sales in that same week.

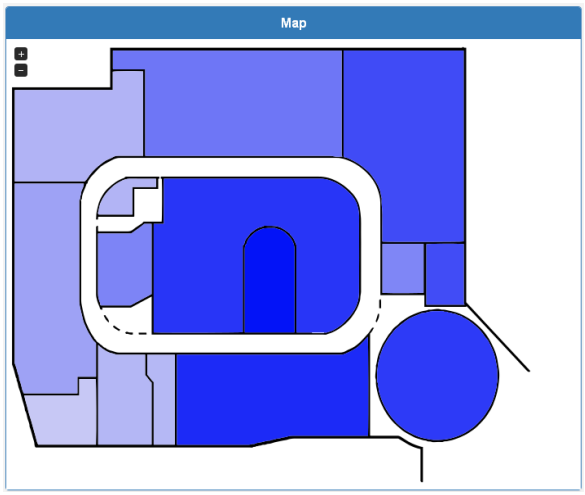


Figure 9. The most visited areas of the store

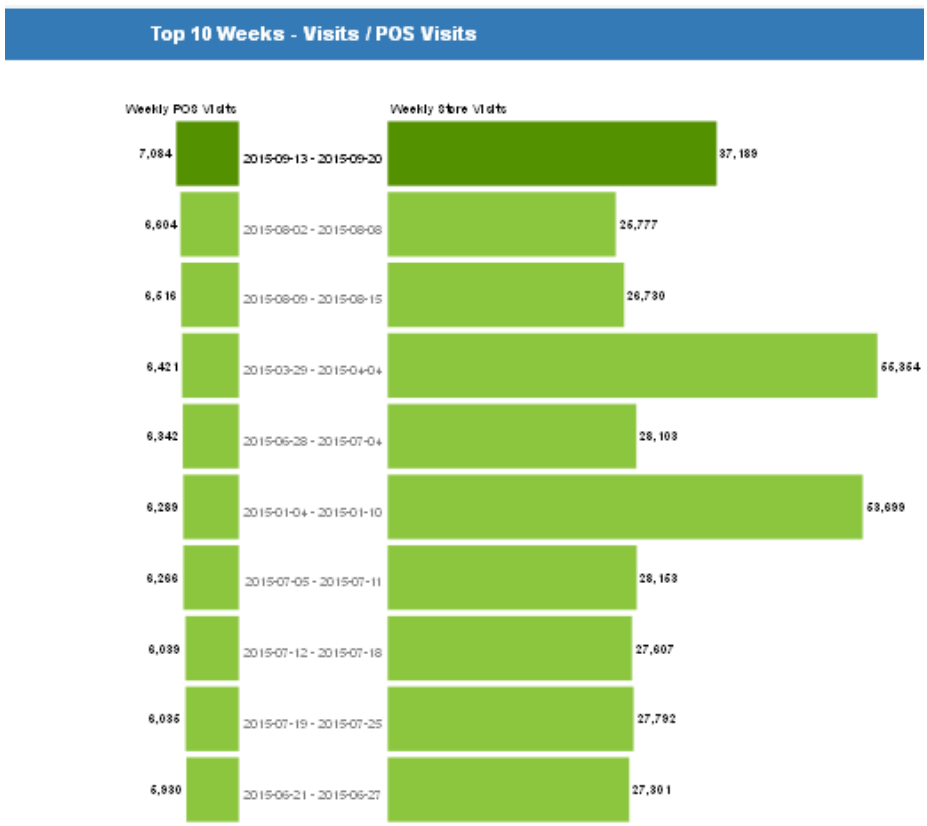


Figure 10. The Top Ten sales weeks and their correspondent number of visitors

As mentioned before, these dashboards were developed as a response to the initial requirements of the desired decision support system. Nowadays, due to the relevance and success of the BI solution, new visual information components are being considered

for the next version of the system. Indeed, the feedback received from the local decision-makers is very positive, as the system allows them to better understand what happens in the store, thus making them more able to make decisions and manage their business.

5. Conclusions

This paper completes the work described in two other papers, from the same authors, in which an innovative BI development project, that gets its data from a NoSQL source, is described (Costa & Pereira, 2015; Pereira & Costa, 2016). The business environment in which the project was developed involves a sporting goods store located in a large shopping mall. The movement of visitors, around and inside the store, as well as the time spent in each zone of the store, are continuously monitored and the corresponding data registered in a Cassandra database system. Those data, complemented with data about sales, allowed the definition of several metrics, used to better understand the behavior of visitors.

The result of the project was a functional decision-support system, which remains in use today, constructed around a BI solution whose data source is a non-traditional one – a NoSQL database. Using the capabilities of ETL tools the non-traditional data source was straightforwardly integrated in the final BI solution.

BI solutions allow decision-makers to easily recognize what is going on with their businesses, in order to make the best decisions. In the era of Big Data, with the enormous amounts of data which are available to organizations, BI solutions are even more relevant to their success. This project confirms that, using suitable BI tools, one can develop solid BI solutions, very rapidly and with a small amount of resources.

In the case of the BI tools used in this project (the *Pentaho Business Intelligence Platform*), despite the wide range of components ready to use and available to the Pentaho community, not all of the desired functionality existed. This has not proved to be a substantial obstacle since, as usually happens with other open source systems, it was possible to edit existing components in order to create the new desired functionalities.

Acknowledgments: This work has been supported by FCT - Fundação para a Ciência e Tecnologia, within the Project Scope: UID/CEC/00319/2019.

References:

Atzeni, P., Bugiotti, F., & Rossi, L. (2013). Uniform access to NoSQL systems. *Information Systems*, *43*, 117-133.

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly: Management Information Systems*, *36*(4), 1165-1188.

Costa, M., & Pereira. J. L. (2015). From a NoSQL Data Source to a Business Intelligence Solution: An Experiment. *Proceedings of the 4th International Conference on Virtual and Networked Organizations, Emergent Technologies and Tools. ViNOrg 2015.* Póvoa de Varzim - Portugal.

DB-Engines (2017). Knowledge Base of Relational and NoSQL Database Management Systems. Retrieved from http://db-engines.com/en/ranking (accessed in July, 2017)

Halper, F., & Krishnan, K. (2013). *TDWI big data maturity model guide: Interpreting your assessment score.* The Data Warehousing Institute.

Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: The definitive guide to dimensional modeling* (3rd Ed). John Wiley & Sons.

Kuznetsov, S. D., & Poskonin, A. V. (2014). NoSQL data management systems. *Programming and Computer Software*, *40*(6), 323-332.

Malik, S. (2005). *Enterprise dashboards design and best practices*. New Jersey. Wiley.

PBIP (2017). Pentaho Business Intelligence Platform. Retrieved from http://www.pentaho.com/product/business-visualization-analytics (accessed in July, 2017).

PDI (2017). Pentaho Data Integration - Kettle ETL tool. Retrieved from http://etl-tools.info/en/pentaho/kettle-etl.htm (accessed in July, 2017).

Pereira. J. L., & Costa, M., (2016). Decision Support in Big Data Contexts: A Business Intelligence Solution. *Proceedings of the 4th World Conference on Information Systems and Technologies. WorldCist'16.* Recife - Brazil.

Reinsel, D., & Gantz, J. (2012). The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the Far East, *IDC*, December, 2012.

Sadalage, P., & Fowler, M. (2012). *NoSQL distilled: A brief guide to the emerging world of polyglot persistence*. Upper Saddle River. Addison-Wesley.

Statista (2017). The Statistics Portal. Retrieved from http://www.statista.com (accessed in July, 2017)

Vesset, D., Woo, B., Morris, H. D., & Villars, R. L. (2012). *Worldwide big data technology and services 2012-2015 Forecast.* IDC, March, 2012.

José Luís Pereira
Information Systems Department &
ALGORITMI Centre
Minho University, Campus de Azurém, Guimarães, Portugal
jlmp@dsi.uminho.pt

Marco Costa
Information Systems Department &
ALGORITMI Centre
Minho University, Campus de Azurém, Guimarães, Portugal
44037@alunos.uminho.pt