International Journal of Information Systems in the Service Sector
Volume 9 • Issue 2 • April-June 2017

Framework for a Hospitality Big Data Warehouse:

The Implementation of an Efficient Hospitality Business Intelligence System

Célia M.Q. Ramos, CEFAGE & ESGHT, University of the Algarve, Faro, Portugal
Daniel Jorge Martins, LARSyS & ISE, University of the Algarve, Faro, Portugal
Francisco Serra, ESGHT, University of the Algarve, Faro, Portugal
Roberto Lam, LARSyS & ISE, University of Algarve, Faro, Portugal
Pedro J.S. Cardoso, LARSyS & ISE, University of the Algarve, Faro, Portugal
Marisol B. Correia, CEG-IST & ESGHT, University of the Algarve, Faro, Portugal
João M.F. Rodrigues, LARSyS & ISE, University of the Algarve, Faro, Portugal

ABSTRACT

In order to increase the hotel's competitiveness, to maximize its revenue, to meliorate its online reputation and improve customer relationship, the information about the hotel's business has to be managed by adequate information systems (IS). Those IS should be capable of returning knowledge from a necessarily large quantity of information, anticipating and influencing the consumer's behaviour. One way to manage the information is to develop a Big Data Warehouse (BDW), which includes information from internal sources (e.g., Data Warehouse) and external sources (e.g., competitive set and customers' opinions). This paper presents a framework for a Hospitality Big Data Warehouse (HBDW). The framework includes a (1) Web crawler that periodically accesses targeted websites to automatically extract information from them, and a (2) data model to organize and consolidate the collected data into a HBDW. Additionally, the usefulness of this HBDW to the development of the business analytical tools is discussed, keeping in mind the implementation of the business intelligence (BI) concepts.

KEYWORDS

Big Data Warehouse, Business Intelligence, Customer Behaviour, Hospitality, Hospitality Big Data Warehouse, NoSQL Databases, Online Reputation, Tourism, Web Crawler

1. INTRODUCTION

To Sheldon (1989, p. 589) "the information is the true blood of the tourism industry." Despite having almost three decades, Sheldon's remark stills as a fact in our days, i.e., travelers, travel agents, hoteliers and stakeholders in the tourism supply chain need broad and trustworthy information. To tourist organizations, an efficient information management will improve the flow of information, the response times to requests from abroad demanders, and the company's development in an increasingly competitive society, which is heading towards very similar technological bases, in order to ensure their survival (Ramos, Rodrigues, & Rodrigues, 2015b).

DOI: 10.4018/IJISSS.2017040102

In today's society, the technological bases of the tourism organizations, in general, and of the hoteliers, in particular, make relevant that marketers and managers have access to data intelligence, and make the best use of it (Peter, 2014). In this sense, those professionals have invested heavily in recent years, organizing strong scientific teams, including statisticians and database (DB) experts, well equipped to build and analyse the contents of their Data Warehouses (DW).

However, the exclusive development and use of the hotel's DW is no longer sufficient to ensure the required competitive advantages (Caldeira, 2012), being necessary to consider the development and use of Big Data Warehouse (BDW) (Di Tria, Lefons, & Tangorra, 2013) architectures, consisting of internal and external data sets (Di Tria et al., 2013; Martins et al., 2015b; Mohanty, Jagadeesh, & Srivatsa, 2013; Ramos, Correia, Rodrigues, Martins, & Serra, 2015a). The concept of BDW refers commonly to the activity of collecting, integrating, and storing large volumes of high velocity data, coming from data sources which may contain both structured and unstructured data (Di Tria et al., 2013). Associated with the concepts of business intelligence (BI), the potentialities related with Big Data are described as technologies that promise to fulfil a fundamental tenet of research in information systems (Schermann et al., 2014), which is to provide the right information, to the right receiver, in the right volume and quality, at the right time (Burke & Hiltbrand, 2011; Ramos et al., 2015a).

Due to the need to survive and to increase competitiveness, hotel marketers are trying to promote their services at travel sites that hold the highest market shares, e.g. Booking or Trivago. This means that the hospitality industry (Chen, Samidjen, Tsai, & Chen, 2013) is using the web as a global vitrine where specialized sites operate, thus providing publicly available information that can be collected into the BDWs with the right technological tools (note, this is not "hacking"). This means generating/retrieving large sets of data, which can be used for hospitality BI purposes, such as, providing a comparison of offers for similar products, or to promote and sell rooms at the best possible price, to the right customers. In essence, hotel managers strive to achieve the best possible revenues but, in order to do it, they need to be in possession of actual and reliable information about their competitive set (e.g., hotels with similar location, facilities, class of service, and number of rooms).

One possibility to get the required information is to buy, from online travel agencies (OTA) of interest, the access to the required businesses data, which is usually made accessible through commercialized Application Programming Interfaces (APIs). Another far more complex but possibly cheaper solution is to extract the information from websites (web scraping), simulating the behaviour of a user (again, this is not the same as "hacking"). In this last alternative, a web robot (bot) or crawler is used to extract the desired data (e.g., prices, room types, or guest reviews), filtering it by country, region, city, time slot, room type, etc. (GuestCentric, 2014; Martins, Lam, Rodrigues, Cardoso, & Serra, 2015a). These extracting technologies are not novel, as several researches have been conducted on data extraction from web information systems (Baumgartner, Gottlob, & Herzog, 2009; Ferrara, De Meo, Fiumara, & Baumgartner, 2014; Zhai & Liu, 2005; Zhao, Meng, Wu, Raghavan, & Yu, 2005), but only a small number of studies have been published on the subject of business to consumer (B2C) (Ghobadi & Rahgozar, 2011; Martins et al., 2015a; Potra, Izvercian, & Miclea, 2016).

This paper presents a framework to develop a Hospitality Big Data Warehouse (HBDW) to manage hotel's activities, including relevant tools to analyse the online reputation, in a way to maximize the revenue and to understand the consumer's behaviour. The main contributions are the architecture of the framework and the description of the process to automatically extract information from travel websites. The automatic extraction comprises a crawler, which periodically accesses targeted web pages and extracts information about a set of hotel features, the data model to organize the collected data, and the process to consolidate the retrieved information. Additionally, the usefulness of this BDW to the development of business analytic tools, to implement the concepts of a BI system applied to the hospitality business, is also discussed.

To enhance the contributions of the present paper, when comparing the proposed framework with others previous studies, such as Lerman, Knoblock and Minton (2001), Rahardjo and Yap (2001) and Martinez-Torres, Rodriguez-Piñero and Toral (2015), the present framework extracts from the web

and stores (on-the-fly) the complete information available on each OTA. For instance, Lerman et al. (2001), only extracts lists and tables from web sources. When comparing with e.g., Crescenzi, Mecca and Merialdo (2001) and with Reis, Golgher, Silva and Laender (2004), the present framework does more than merely finding pattern in different structures, it does an intelligent information structure, using each explicit data extracted from the OTAs for each particular module of the framework. When comparing with Zhai and Liu (2005) and Chiu, Chiu, Sung and Hsieh (2015), the present framework also consolidates the ("same") information extracted from different OTAs. For more details, see also the Contextualization and State of the Art section.

The article is structured as follows: besides the introduction, the second section presents a thorough contextualization and the state of the art of the subject of study. The third section develops the concepts and presents the steps associated with the methodology to build a HBDW, and the fourth section highlights the relevance of BDWs to the hospitality and tourism organizations. The final section presents a discussion, conclusions and suggestions for future work.

2. CONTEXTUALIZATION AND STATE OF ART

Big Data Warehouses are considered as excellent systems to be used in the marketing research areas, such as to understand: the customer's behaviour (Nadeem, Andreini, Salo, & Laukkanen, 2015; Phillips-Wren, & Hoskisson, 2015), the customer's preferences (Marrese-Taylor, Velasquez & Bravo-Marquez, 2014; Martinez-Torres et al., 2015; Xiang, Schwartz, Gerdes, & Uysal, 2015), or the hotel's online reputation (Chiu et al., 2015; Schuckerta, Liu, & Law, 2015; Xiang & Law, 2013; Zhou, Ye, Pearce, & Wu, 2014). For the hospitality business, a BDW is full of data which can be analysed in several dimensions with the appropriate analytical tools, aiming the knowledge discovery of better targeted social influencer marketing, more accurate business insights, the segmentation of customers, the recognition of sales and market opportunities, etc. (Belfo, 2013; Offut, 2014; Russom, 2011).

For some time now, the potentialities of Big Data in the hospitality business aroused the interest of hotels' managers and several researchers, which highlights the market research importance in the era of digitization and will broaden the horizons of hospitality and tourism research. The acquisition of data is therefore a major issue for those researchers, making extremely relevant to process and collect valuable data from the web by an automated process, such as a web crawler (Chiu et al., 2015; Marrese-Taylor et al., 2014; Martinez-Torres et al., 2015; Schuckert et al., 2015; Xiang, & Law, 2013; Xiang et al., 2015).

Resorting to the potentialities of Big Data analytics for knowledge generation in tourism and hospitality (Fuchs, Höpken, & Lexhagen, 2014), stakeholders can obtain valuable knowledge about their business if the information that is available on the internet, and is relevant to the development of their activity, is included in their BDW. For the hoteliers is important to have access to the information associated with their business, to better analyse and understand the customers' preferences, the hotel's online reputation, the business trends, among others (Martins et al., 2015b). The analyses to be considered are associated with business analytics, where advanced techniques operate on Big Data sets to compare all the different characteristics related to the hoteliers' products information. The Big Data analytics is really about the two things - Big Data and analytics - plus how the two have teamed up to generate one of the most profound emerging trends in BI (Russom, 2011).

The necessities of the hospitality BI systems are very explicit, with well-defined requirements that have to be developed to incorporate historical data into the analytical tools (Martins et al., 2015b). In a technological environment that is characterized by providing the decision makers with timely relevant data and knowledge, which supports their resolutions and creates intelligence (Fuchs et al., 2014; Santos & Ramos, 2009), the relevant data can be collected from several web sources resulting in different formats and structures, as already mentioned. In general, the automatic data extraction process, from those unstructured sources (websites), incorporate a web crawler to collect the (external)

information associated with the business (Martins et al., 2015a, 2015b; Schuckerta et al., 2015; Xiang et al., 2015;), which is then integrated with the hotel's (internal) historical data in a BDW.

Returning to the extraction of data from the web, the process can be seen as a tedious and exhausting effort for humans to carry out, thus justifying the development of tools, web robots or web crawlers, which executes it in an automatic way (Martins et al., 2015a; Papadakis, Skoutas, Raftopoulos, & Varvarigou, 2005). In this sense, several studies about the automatic extraction of information from the web have been conducted. For example, Lerman et al. (2001) presented a fully automatic system that can extract lists' and tables' data from web sources. Rahardjo and Yap (2001) used an algorithm that requires the user to identify the relevant data to extract. A system that compares two HTML pages to find patterns was proposed by Crescenzi et al. (2001). The Tree-Edit Distance algorithm was used by Reis et al. (2004) to find patterns between different structures. A similar implementation, based on pattern analysis and Document Object Model (DOM) trees uses the Edit Distance Algorithm (Qiu & Yang, 2010). The VINTs (Visual Information and Tag structure based wrapper generator) was proposed by Zhao et al. (2005) to extract data from search engines results. Papadakis et al. (2005) exposed a way to figure out the format of the information contained in web pages and discover the associated structure. Zhai and Liu (2005) presented a system that only requires a sample page labelling. They use a method called Sufficient Match to identify the similitude between the objective page and the main sample page. The ViDE (Vision-based Data Extractor), proposed in (Liu, Meng, & Meng, 2010), is a data mining method that relies on the visual aspect of how the data is presented.

In automatic extraction there is a challenge related with the occurrence of dynamic pages which, for instance, use JavaScript to trigger dynamic changes in the HTML. In fact, JavaScript is a programming language that is employed in many e-commerce sites to hide information, making more difficult the automatic data extraction process once these scripts are used to make changes in the client's side structure of the HTML code. Those changes are often triggered by the interaction with the website pages in order to display information that is at first invisible. To overcome this problem, Baumgartner and Ledermiiller (2005) presented a method designated by Lixto Visual Wrapper which integrates the Mozilla browser driver to interact with the web page, in order to present the invisible information from the backend database.

Another problem to be overcome is the fact that a system which aims to extract data from multiple e-commerce sites will find different ways of describing and/or quantifying the same product (Qiu & Yang, 2010; Sambhanthan & Good, 2014). Even the same e-commerce site can change its form of presenting the data, making almost impracticable to create a well-structured database schema for that heterogeneous/dynamically structured set of data (Martins et al., 2015a). In this sense, to store that unstructured information, it might be useful to use schemaless databases making most adjustments to the database become transparent and automatic. One of the most well-known schemaless database is, probably, the MongoDB database (MongoDB, 2015; Redmond & Wilson, 2012). MongoDB is a NoSQL document-oriented database, presents high performance, high reliability, easy scalability (vertically and horizontally through replication and auto-sharding, respectively) and map-reduce support. A MongoDB database is structured as a set of collections, which store documents. These documents are BSON objects (a binary JSON document format (JSON, 2015)), allowed to have a dynamic schema, i.e., documents in the same collection are not forced to have the same structure.

In conclusion, most of the times the HBDW should be prepared to accommodate the historic data (internal), usually stored in structured format such as in relational databases (RDB) (Ramos et al., 2015a), along with data collected from external sources (e.g., websites), potentially stored in non-relational databases (NoSQL databases) (Martins et al., 2015a, 2015b). In the development of the HBDW, taking in consideration the different collections of data coming from internal and external sources (e.g., GDS -Global Distribution Systems, OTA, DW, and PMS - Property Management Systems), it is necessary to integrate the information from the NoSQL databases and the RDB into a final DB capable of being analysed by the correct tools. Presently, the RDBs continue to be the

more prevalent data storage (Offut, 2014), allowing the view of the data in multiple formats and to different stakeholders, even to the ones with activities not related with the DB technologies. In this sense, it is useful to transform the information stored in the NoSQL databases and in the DWs into a RDB, i.e., in a BDW (Ramos et al., 2015a), also called HBDW, as proposed in the following sections.

3. HOSPITALITY BIG DATA WAREHOUSE

In summary, as concluded in the previous section, the HBDW described in this work comprises the process of integrating external data (stored in a NoSQL Database; it will be further explore latter) and internal sources (historical business data - DW) in a final BDW (RDB database), see Figure 1. The methodology considered to develop the HBDW is constituted by several steps, namely: (1) automatically collect information from websites using a web crawler; (2) store the extracted data in a MongoDB Database; (3) define the BDW model; and (4) extract, transform and load (ETL) the data into the BDW.

3.1. Automatic Collection of Information by Web Crawler

For the automatic collection of information from the web, a set of crawlers (Qiu & Yang, 2010) must run periodically in order to find and retrieve pertinent and updated data to the hospitality business. Furthermore, different hospitality business models use distinct sets of extracted data, and from different sites (e.g., Booking, Expedia, TripAdvisor) it is possible to extracted dissimilar information for the same hotel (Martins et al., 2015a). Due to the above reasons, distinct web crawlers were developed, one for each of the analysed sites (e.g., Booking, Expedia, TripAdvisor). Each web crawler has the function of extracting the existing information relative to different periods of stay and simulating different number of guest (e.g., 2 adults, 2 adults and 1 kid), etc.

The automated extraction of information from the websites must take into account that the information is prepared and displayed for human consumption, i.e., considering what is most appropriate for the website's customer. In general, these sites work in a similar manner: (1) the homepage has a form, which allows searching for a type of hotel, a city and a period of stay. After submitting the form, a (2) list of hotels that match the form criteria will be returned by the webserver. Clicking on one of those listed hotels, (3) a new page will be showed, exposing to the user information about the selected hotel, namely: available rooms, prices, feature, amenities, policies, guest comments, etc.

Important is also the fact that, from time to time, some of those sites change their page layout (design), as well as the attributes values of tags, which contain important information for our purposes.

The algorithm of extraction is summarised in the following phases:

1. Define the URL of the website to be crawled, set the data to fill the website form and other parameterizations;
2. Automatically fill an instance of the webdriver which models the behaviour of the users over the website to be crawled, and do a request to the corresponding (website) server;
3. Store the response to the server request, as a list of links referencing hotels and boarding houses that satisfy the search defined on Step 1;
4. For each link (hotel) in the list built on Step 3, do a second level of request to the server and extract all of the hotel's relevant data;
5. Store the retrieved data in a collection on the MongoDB database.

The process to find the pertinent HTML elements was described in (Martins et al., 2015a). Those HTML elements contain the relevant data to the hotel's business, being the target tags manually provided and placed in a database. Once the webmasters of the e-commerce websites do change the

structure of the website, usually from time to time, there is no other solution than to redo/redefine the target tags as presented by Martins et al. (2015a).

After the HTML tags' recognition process, the intended information, usually located between the begin and end tags of the elements (e.g., the "Hotel x" name in the HTML code Hotel x), is extract. However, some information is not showed as explicit text. For instance, some websites show the number of stars of the hotels as images (see Figure 2). In this case, a relevant attribute value is extracted taking into consideration that the images used in e-commerce website usually have attributes to help their posing in the page and defining the semantic importance of the image, see Figure 2 (bottom).

Furthermore, considering that each website has distinct designs and structures, the choice to store the extracted data was MongoDB. This selection was also supported on the big and diverse variety of data found on the different websites and even under the same one.

3.2. Storing the Extracted Data into a NoSQL Database

The data extracted from the different web sites is stored in a MongoDB database, organized in four collections: *AboutHotel*, *Rooms*, *Comments*, and *Scores* (Martins et al., 2015b). The collection designated by *AboutHotel* contains the hotels' characterizations, which includes information about the hotels' names, locations, features and rooms amenities. The collection *Rooms* has the information about the rooms and prices, namely including data about the rooms' names, and allowed number of guests (adults and children). In *Comments* is the data concerning the reviews of the hotels, which includes information about the customer segment and their country of origin. Finally, in the *Score* collection are the score reviews that tourists have attributed to hotels, which contribute to define the hotel's online reputation.

Figure 3 (top) presents an example of data relative to the scores of a hotel, retrieved from the Expedia website. After being collected by the web crawler, the same data was stored in the Score collection (of the MongoDB database) as a binary-encoded serialization of the JSON presented at the bottom of the same figure.

In the present work, MongoDB is an intermediary database between the websites and the BDW, which implemented over a relational database, as used for instance by Offut (2014). To create the BDW, with the concepts associated with the relational database, it is necessary to define the corresponding relational data model, as explained next.

3.3. Definition of the BDW Model

One of the snags associated with considering a MongoDB database to develop a BI system is the fact that its data (collected from de web) is in general stored in a set of unstructured documents, which is not to the best choice to develop and use analytical tools. In this sense, a relational database is considered as a better choice to develop the functionalities supporting the decision-making process, with the objective of increasing the business insights and the organization performance (Offut, 2014; Ramos et al., 2015a). Therefore, the next steps consist in extracting the collected data from the MongoDB database, transform it and upload the result into the RDB (ETL – Extract, transform and load – process), later the BDW.

In the process of passing the data from the MongoDB database to the RDB, is relevant to define an adequate RDB model, with a structure capable of storing data collected from different websites and making also possible the inclusion of information that exists on the organization's DW. In the definition of the RDB storage design, the business' information system was analysed leading to an entity-relationship model (Chen, 1976). For example, Figure 3 showed part of the information extracted from the Expedia website and stored in the MongoDB's *Scores* collection, which was then transposed into the RDB using the entities presented in Figure 4, an excerpt of the final entity-relationship model.

The architecture of the database (tables and the relations between them) results from the entity-relationship model, being placed in the relational database management system. Figure 5 shows a part

of the full architecture, derived from the previous entity-relationship model excerpt. In more detail, Figure 5 presents the Scores table and its most relevant relationships with others tables, namely: *AboutHotel*, *Channel*, *Segment*, and *ScoreCategory*. A tuple of the *Scores* table identifies an hotel (*idHotel* attribute), the segment of the evaluators (*IdSegment* attribute), a channel (*IdChannel* attribute) which stores the source of the data (e.g., Booking or Expedia), the extraction date (*ExtractionDate* attribute), the category of the evaluation (such as a reference to a "Room Cleanliness" entry in the *ScoreCategory* table, stored in the *IdScoreCategory* attribute), and the score (*Score* attribute) which keeps a numeric evaluation of the score. The next section will address the ETL into the BDW.

3.4. ETL the Data to the BDW

In the global procedure, the next step is to upload the data from the MongoDB database into the RDB, following an ETL procedure. The transformation of the data into the appropriate fields in the RDB is not a linear process (Ramos et al., 2015a). For example, the number of stars of a hotel can be retrieved in different forms/types by the web crawler (e.g., text, image, or image captions), depending on the site that is being processed. Even for a same site, the forms/types can change along the time (for more details, see (Martins et al., 2015a)). Another example of distinct types of values, which are related with the same observed parameter, is the *Review Score*. For example, Booking allows *Review Score* values between 0 and 10, and does a *Score Breakdown* in 7 fields; while Expedia allows *Review Score* from 0 to 5, and does a *Score Breakdown* in 4 fields. Other problems arise from the fact that different sources have different structures and different meanings for the same hotel features (Martins et al., 2015a), e.g., the *cleanliness* and *room cleanliness* score categories appear in different sources with the same meaning.

The above observations raise the need to consolidate the extracted data, in order to ensure that the information stored in BDW is regular and unfailing to the hotel's BI system. In this sense, a set of guidelines must be considered (see Figure 6), namely: define the data forwarding rules and identify possible conversions (Martins et al., 2015b). The data consolidation process is addressed by performing the following steps:

1. Extract the information from MongoDB;
2. Define the data conversion rules;
3. Create / load the data dictionaries;
4. Set the correspondence between the data collected by the web crawlers (different channels) for the same hotels.
5. Set the channels priorities;
6. Define the data flow between the databases; and
7. Upload the information into the BDW.

In more detail, the first step consists in the extraction of the information from the MongoDB by the consolidation program, which is responsible by the adequate processing and conversion of the data to be stored in the BDW. The definition of the data conversion rules is made in the second step. These conversions are needed since the majority of the data stored in the MongoDB database, by web crawler, is in string format, being essential to make the necessary conversions to the correct domains (e.g., numeric, date/time, or GPS coordinates).

The creation and loading of the data dictionaries (third step) is essential, once most of the data stored in the MongoDB database is free text, written by humans. The data dictionaries are lists of words and synonyms that the program will try to find in those texts. For now, these dictionaries are manually defined by the user and updated each time a new synonym for a word is founds. The data dictionaries are stored as documents in a MongoDB collection and are structured in four fields: *Source*, *Type*, *Word* and *Alias*. The *Source* field indicates the channel (e.g., *Booking*) where that particular dictionary should be investigated. The *Type* field indicates the context in which the dictionary should

be analysed. For instance, the name-value "*Type*": "*Amenities*" means that this dictionary should be consulted when looking for the hotel's amenities in a sentence. The *Word* field concerns the word that is being searched (e.g., "swimming pool"). Finally, the *Alias* field is a list of synonyms for the word (for more details, see (Martins et al., 2015b)).

The correspondence between the data collected for the same hotel by the web crawlers (different channels) is done on the fourth step, once the web crawlers only collect the data not proceeding to its analyzes. As an example, there is no warranty that a hotel identified as "Hotel X" in one of the sources is the same "Hotel X" on another source. This particular case occurs because the names of the hotels can change slightly according with the channel. In this investigation/product, the correct correspondence is crucial to make an accurate consolidation of the hotel's data. The solution considered to solve the problem was developed as an algorithm, see Figure 7, which matches the extracted data from the different channels sources in three parameters, namely the hotel's name, GPS coordinates, and full address.

The fifth step sets the priorities of the channels, to unravel/decide between dissimilar information (obtained from different sources) before it is stored in BDW. For example, if conflicting information for a hotel exists in four different channels and the channel's priority is defined, in descending order, by "channel 1", "channel 2", "channel 3", and "channel 4", then stored data would be from "channel 1". This conflicting data can occur even in the most improbable parameters, as are the number of stars of the hotels.

After the data transformation by the conversion rules, the application of the data dictionaries, and the establishment of the correspondences between the data collected by the web crawlers, and application of the channel's priorities, it is necessary to upload the information into the right place, according with a data flow defined to comply with a specific data storage (sixth step). In the seventh and final step, the data will be upload into the BDW.

From time to time, the consolidation of information extracted from the web is a task that stills need some human supervision. For instance, when a channel finds new words it is necessary to add them to the data dictionary. Once the data dictionary becomes more complete, thus diminishing the number of new words found, the consolidation system will also become more stable. The data stored in the BDW (i.e., in the HBDW) can now be used by the BI system as explained in the next section.

4. EXTRACTING KNOWLEDGE FROM THE BI SYSTEM

Business Intelligence is a way to identify new opportunities and implement effective strategies according with those opportunities. In this sense, the implementation is supported on insights, or intelligence, that can offer to the business competitive aptitudes that give a market advantage, a long term stability (Ramos et al., 2015a; Rud, 2009), and the access to a collective intelligence (Martínez-Torres et al., 2015).

The traditional methods of analysis are no longer adequate to extract information from a HBDW (Caldeira, 2012; Di Tria et al., 2013; Offut, 2014; Ramos et al., 2015a). Nowadays, given a HBDW, the BI should be done in two steps: extract knowledge from the HBDW and assess the intelligence extracted from that knowledge (Offut, 2014; Santos & Ramos, 2009). The first step, the knowledge extraction, transforms the data into meaningful and useful information for business analysis purposes, presenting it in dashboards or in a management reporting (Moundalexis & Nag, 2013). Intelligence assessment, second step, means the possibility of extracting business value for the tourism organization by several techniques (Caldeira, 2012; Santos & Ramos, 2009).

The knowledge extraction should include: OLAP (Online Analytical Processing) techniques, which allow to analyse the information on different business perspectives/dimensions; data mining methods, that consist in the application of artificial intelligence algorithms to discover knowledge from the data; and methods of forecasting, to identify new models that define a variable behaviour, which can be used to estimate future variable values.

Among others, the techniques associated with the intelligence assessment can be categorized in (Abderrahim & Benslimane, 2015; Offut, 2014): recommendation systems, for example when social networks refers "People you may know likes de hotel X"; social networks analysis, to identify the influence over others; new products analysis, to test new products or ideas and obtain instant feedback; competitors' pricings analysis, to compare products prices; and sentiment analysis, which permit to define the customer sentiment towards products, services, destinations, hotels, etc.

In a HBDW, the integration of the extraction of knowledge with the assessment of the intelligence should be developed in order to analyse and manage the information associated with specific hotel business areas, such as: Revenue Management (RM), Online Reputation Management (ORM) or Customer Relationship Management (CRM), as presented in the Figure 8. As deductible from Figure 8, data mining techniques integrated with the social networks data can discover associations between the information, e.g., a specific segment of guests (e.g., families, couples or business travelers) prefer a room with a better view or the guest from a certain country are very demanding with the room's aspect. Therefore, in face with the correct and necessary information, the hotel manager can improve the relationship with the customer. By another perspective, the integration of the OLAP techniques with the revenue management and the online reputation allows the detection of new trends in the consumer's behaviours that ties the social network profile with the room's price, contributing to predict the result of new pricing policies.

In summary, analytical tools ought to be conjugated with forecasting models applied to historical data, data mining techniques can be used to detect patterns, which can reveal some insights about the business, and online analytical cubes allow to analyse the data according with its most relevant dimensions.

In conclusion, the results of these analytical techniques and tools can be used to understand the opinions of the hotel's guests, and creating many opportunities to understand the positioning of the consumers on marketing campaigns and preferences for products. At the same time, they should allow the detection of new business trends and to implement new strategies to captivate new customers and new markets.

5. DISCUSSION

In the hospitality industry, big data analytic tools are a powerful way to support the decision makers and to control their organization. Supported by a HBDW, those tools can provide the conditions to develop a valuable BI system to the hotel.

On the other hand, with a proper BI system, the hotels' stakeholders can visualize the information of their business in real time, detect urgent situations and automate to attain immediate answers. However, a HBDW is fundamental in this process, once the traditional Data Warehouse are no longer sufficient to satisfy the needs of the hoteliers and marketers. The hospitality BI system will permit the hoteliers to get valuable information to organize data series with predictive algorithms to decide on the best prices and service-mix strategy, in order to obtain higher revenues. Aiming to satisfy this need, the web crawlers must run periodically in order to collect data in a suitable and updated way.

In this paper, a web crawler framework was presented, aiming to demonstrate that the automatic "browse" e-commerce websites is possible. The automatic "browsing" of the referred sited includes the identification of relevant elements and their extraction to a (NoSQL) database. In addition, a small excerpt of the relational data model was presented, as an example of how to implement a HBDW of a hotel. The framework takes into consideration the steps associated with the process of extracting (from the NoSQL database), transformation and consequent upload of the retrieved data into a relational database. In these steps is included the process of information consolidation, which, for now, is a task that steel needs human supervision from time to time. Over time, the consolidation system is expected to become more stable, as the data dictionary grows into a more complete dataset of possible words describing features.

The current framework still presents some limitations, it does not include the treatment (semantic analysis) of the (written) commentaries and guest opinions, which represent a disadvantage to the hotel manager once in our days the management of online reputation is defined not only by the quantitative values in the reviews, but also by the written commentaries. This integration is important for the hoteliers, so they can create strategies to increase the overall online reputation, and at the same time, to control and manage the quality of the services that can appear in the negative review. Another limitation, is from time to time occur "extreme" changes in the structure of the OTAs and there is no other solution (for now) than to redo/redefine the web crawler code, i.e., manually detect the new target tags (from the HTML) and store this new tags in the database, as a way to maintain the system to work completely.

In terms of future work, and to over whelming the limitations, one of the next steps is to complete the development of the application/software with text mining techniques, incorporate gamification concepts into the hotel's websites, and the development of intelligent interfaces with information centred in the user, part of which is already under development. The text mining techniques will permit to analyse the social networks reviews in semantic terms and understand the customer's opinions. The gamification concepts will permit to analyse the consumer's behaviour in order to identify what motivates their decisions. The intelligent interfaces will create a more pleasant navigation by showing information to the customers according with their preferences. These new developments will be incorporated with the data mining techniques conjugated with the OLAP functionalities to produce data insights and intelligence that will be present to the hotelier in dashboards or reports to support the decision-makers and to discover patterns associated to new guest behaviour, in a way to detect new competitors, new markets or new prospective partners.

ACKNOWLEDGMENT

This work was supported by project SRM QREN I&DT, no. 38962 and FCT projects LARSyS (UID/EEA/50009/2013), CIAC (PEstOE/EAT/UI4019/2013), CEFAGE (PEst-C/EGE/UI4007/2013) and CEG-IST - Universidade de Lisboa. The authors also thanks to project leader VisualForma - Tecnologias de Informação S.A.

REFERENCES

Abderrahim, N., & Benslimane, S. M. (2015). STRESS: A social trust-aware system for recommending web services. *International Journal of Information Systems in the Service Sector*, *7*(3), 40–58.

Baumgartner, R., Gottlob, G., & Herzog, M. (2009). Scalable web data extraction for online market intelligence. *Proceedings of the VLDB Endowment*, *2*(2), 1512–1523. doi:10.14778/1687553.1687580

Baumgartner, R., & Ledermiiller, G. (2005). Deepweb navigation in web data extraction.*Proc. Int. Conf. on Intelligent Agents, Web Technologies and Internet Commerce* (Vol. 2, pp. 698-703).

Belfo, F. (2013). A framework to enhance business and information technology alignment through incentive policy. *International Journal of Information Systems in the Service Sector*, *5*(2), 1–16. doi:10.4018/jisss.2013040101

Burke, M., & Hiltbrand, T. (2011). How gamification will change business intelligence. *Business Intelligence Journal*, *16*(2), 8–16.

Caldeira, C. (2012). *Data warehousing*. Lisboa: Edições Sílabo.

Chen, P. P. (1976). The entity-relationship model: Toward a unified view of data. *ACM on Database Systems*, *1*(1), 9–36. doi:10.1145/320434.320440

Chen, W.-T., Samidjen, M., Tsai, C.-W., & Chen, T.-F. (2013). Global hospitality and tourism management technologies (book review). *International Journal of Information Systems in the Service Sector*, *5*(3), 85–87.

Chiu, C., Chiu, N. H., Sung, R. J., & Hsieh, P. Y. (2015). Opinion mining of hotel customer-generated contents in Chinese weblogs. *Current Issues in Tourism*, *18*(5), 477–495. doi:10.1080/13683500.2013.841656

Crescenzi, C., Mecca, G., & Merialdo, P. (2001). Roadrunner: Towards automatic data extraction from large web sites.*Proc. VLDB* (Vol. 1, pp. 109-118).

Di Tria, F., Lefons, E., & Tangorra, F. (2013). Big data warehouse automatic design methodology. In *Big Data Management, Technologies, and Applications* (pp. 115-149).

Ferrara, E., De Meo, P., Fiumara, G., & Baumgartner, R. (2014). Web data extraction, applications and techniques: A survey. *Knowledge-Based Systems*, *70*, 301–323. doi:10.1016/j.knosys.2014.07.007

Fuchs, M., Höpken, W., & Lexhagen, M. (2014). Big data analytics for knowledge generation in tourism destinations - A case from Sweden. *Journal of Destination Marketing & Management*, *3*(4), 198–209. doi:10.1016/j.jdmm.2014.08.002

Ghobadi, A., & Rahgozar, M. (2011). An ontology based semantic extraction approach for B2C eCommerce. *The International Arab Journal of Information Technology*, *8*(2), 163–170.

GuestCentric.com. (2014). *Booking.com: Your worst best friend?* Retrieved from http://www.guestcentric.com/boo kingcom-your-worst-best-friend/

JSON. (2015). *Javascript object notation*. Retrieved from http://www.json.org/

Lerman, K., Knoblock, C., & Minton, S. (2001). Automatic data extraction from lists and tables in web sources. *Proceedings of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining* (p. 98).

Liu, W., Meng, X., & Meng, W. (2010). Vide: A vision-based approach for deep web data extraction. *IEEE Transactions on Knowledge and Data Engineering*, *22*(3), 447–460. doi:10.1109/TKDE.2009.109

Marrese-Taylor, E., Velásquez, J. D., & Bravo-Marquez, F. (2014). A novel deterministic approach for aspect-based opinion mining in tourism products reviews. *Expert Systems with Applications*, *41*(17), 7764–7775. doi:10.1016/j.eswa.2014.05.045

Martínez-Torres, M. D. R., Rodriguez-Piñero, F., & Toral, S. L. (2015). Customer preferences versus managerial decision-making in open innovation communities: The case of Starbucks. *Technology Analysis and Strategic Management*, *27*(10), 1226–1238. doi:10.1080/09537325.2015.1061121

Martins, D., Lam, R., Rodrigues, J. M. F., Cardoso, P. J. S., & Serra, F. (2015a). A web crawler framework for revenue management.*Proc. 14th International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases (AIKED '15)* (pp. 88-97).

Martins, D., Ramos, C. M. Q., Rodrigues, J. M. F., Cardoso, P. J. S., Lam, R., & Serra, F. (2015b). Challenges in building a big data warehouse applied to the hotel business intelligence.*Proc. 6th Int. Conf. on Applied Informatics and Computing Theory (AICT'15), in Recent Research in Applied Informatics* (pp. 110-117).

Mohanty, S., Jagadeesh, M., & Srivatsa, H. (2013). *Big data imperatives: Enterprise 'big data' warehouse. 'BI' Implementations and Analytics*. Apress. doi:10.1007/978-1-4302-4873-6

Mongo DB. (2015). Retrieved from http://www.mongodb.com/

Moundalexis, M. L., & Nag, B. N. (2013). Decision-making, dashboard displays, and human performance in service systems. *International Journal of Information Systems in the Service Sector*, *5*(4), 32–46. doi:10.4018/ijisss.2013100103

Nadeem, W., Andreini, D., Salo, J., & Laukkanen, T. (2015). Engaging consumers online through websites and social media: A gender study of Italian generation Y clothing consumers. *International Journal of Information Management*, *35*(4), 432–442. doi:10.1016/j.ijinfomgt.2015.04.008

Offutt, B. (2014). Big data: Redefining travel business decision making (White Paper). *Sponsored by UNIT4 Business*. Phocuswright.

Papadakis, N., Skoutas, D., Raftopoulos, K., & Varvarigou, T. (2005). Stavies: A system for information extraction from unknown web data sources through automatic web wrapper generation using clustering techniques. *IEEE Transactions on Knowledge and Data Engineering*, *17*(12), 1638–1652. doi:10.1109/TKDE.2005.203

Peter, T. (2014). *Use hotel data to drive growth.* Retrieved from http://www.hotelnewsnow.com/Article/14553/Use-hotel-data-to-drive-growth

Phillips-Wren, G., & Hoskisson, A. (2015). An analytical journey towards big data. *Journal of Decision Systems*, *24*(1), 87–102. doi:10.1080/12460125.2015.994333

Potra, S., Izvercian, M., & Miclea, S. (2016). Changes in CRM approach: Refined functional blocks for customer creative engagement in services. *International Journal of Information Systems in the Service Sector*, *8*(1), 45–57. doi:10.4018/IJISSS.2016010104

Qiu, T., & Yang, T. (2010). Automatic information extraction from e-commerce web sites.*Proc. Int. Conf. on E-Business and E-Government (ICEE)* (pp. 1399-1402). doi:10.1109/ICEE.2010.355

Rahardjo, B., & Yap, R. (2001). Automatic information extraction from web pages.*Proc. Int. ACM SIGIR Conf. on Research and development in information retrieval* (pp. 430-431).

Ramos, C. M. Q., Correia, M. B., Rodrigues, J. M. F., Martins, D., & Serra, F. (2015a). Big data warehouse framework for smart revenue management.*Proc. 3rd NAUN International Conference on Management, Marketing, Tourism, Retail, Finance and Computer Applications (MATREFC '15)* (pp. 13-22).

Ramos, C. M. Q., Rodrigues, P. M. M., & Rodrigues, J. M. F. (2015b). Opportunities, emerging features and trends in electronic distribution in tourism. *International Journal of Information Systems and Social Change*, *6*(4), 17–32. doi:10.4018/IJISSC.2015100102

Redmond, E., & Wilson, J. R. (2012). *Seven databases in seven weeks: a guide to modern databases and the NoSQL movement*. Pragmatic Bookshelf.

Reis, D., Golgher, P., Silva, A. S. D., & Laender, A. F. (2004). Automatic web news extraction using tree edit distance.*Proc. of the 13th Int. Conf. on World Wide Web* (pp. 502-511). doi:10.1145/988672.988740

Rud, O. (2009). *Business intelligence success factors: Tools for aligning your business in the global economy*. Hoboken, N.J: Wiley & Sons.

Russom, P. (2011). *Big data analytics*. TDWI Best Practices Report, Fourth Quarter.

Sambhanthan, A., & Good, A. (2014). Strategic advantage in web tourism promotion: An e-commerce strategy for developing countries. *International Journal of Information Systems in the Service Sector*, *6*(3), 1–21. doi:10.4018/ijisss.2014070101

Santos, M., & Ramos, I. (2009). Business intelligence (2 ed.). FCA Editora: Lisboa.

Schermann, M., Krcmar, H., Hemsen, H., Markl, V., Buchmüller, C., Bitter, T., & Hoeren, T. (2014). Big data - An interdisciplinary opportunity for information systems research. *Business & Information Systems Engineering*, *6*(5), 261–266. doi:10.1007/s12599-014-0345-1

Schuckerta, M., Liu, X., & Law, R. (2015). A segmentation of online reviews by language groups: How English and non-English speakers rate hotels differently. *International Journal of Hospitality Management*, *48*, 143–149. doi:10.1016/j.ijhm.2014.12.007

Sheldon, P. J. (1989). Travel industry information systems. In S. Witt & L. Moutinho (Eds.), *Tourism Marketing and Management Handbook* (pp. 589–592). London: Prentice Hall.

Xiang, Z., & Law, R. (2013). Online competitive information space for hotels: An information search perspective. *Journal of Hospitality Marketing & Management*, *22*(5), 530–546. doi:10.1080/19368623.2012.671563

Xiang, Z., Schwartz, Z., Gerdes, J. H. Jr, & Uysal, M. (2015). What can big data and text analytics tell us about hotel guest experience and satisfaction? *International Journal of Hospitality Management*, *44*, 120–130. doi:10.1016/j.ijhm.2014.10.013

Zhai, Y., & Liu, B. (2005). Web data extraction based on partial tree alignment.*Proceedings of the 14th international conference on World Wide Web* (pp. 76-85). doi:10.1145/1060745.1060761

Zhao, H., Meng, W., Wu, Z., Raghavan, V., & Yu, C. (2005). Fully automatic wrapper generation for search engines. *Proceedings of the 14th international conference on World Wide Web* (pp. 66-75). doi:10.1145/1060745.1060760

Zhou, L., Ye, S., Pearce, P. L., & Wu, M. Y. (2014). Refreshing hotel satisfaction studies by reconfiguring customer review data. *International Journal of Hospitality Management*, *38*, 1–10. doi:10.1016/j.ijhm.2013.12.004

APPENDIX

Figure 1. Data integration process in a Hospitality Big Data Warehouse

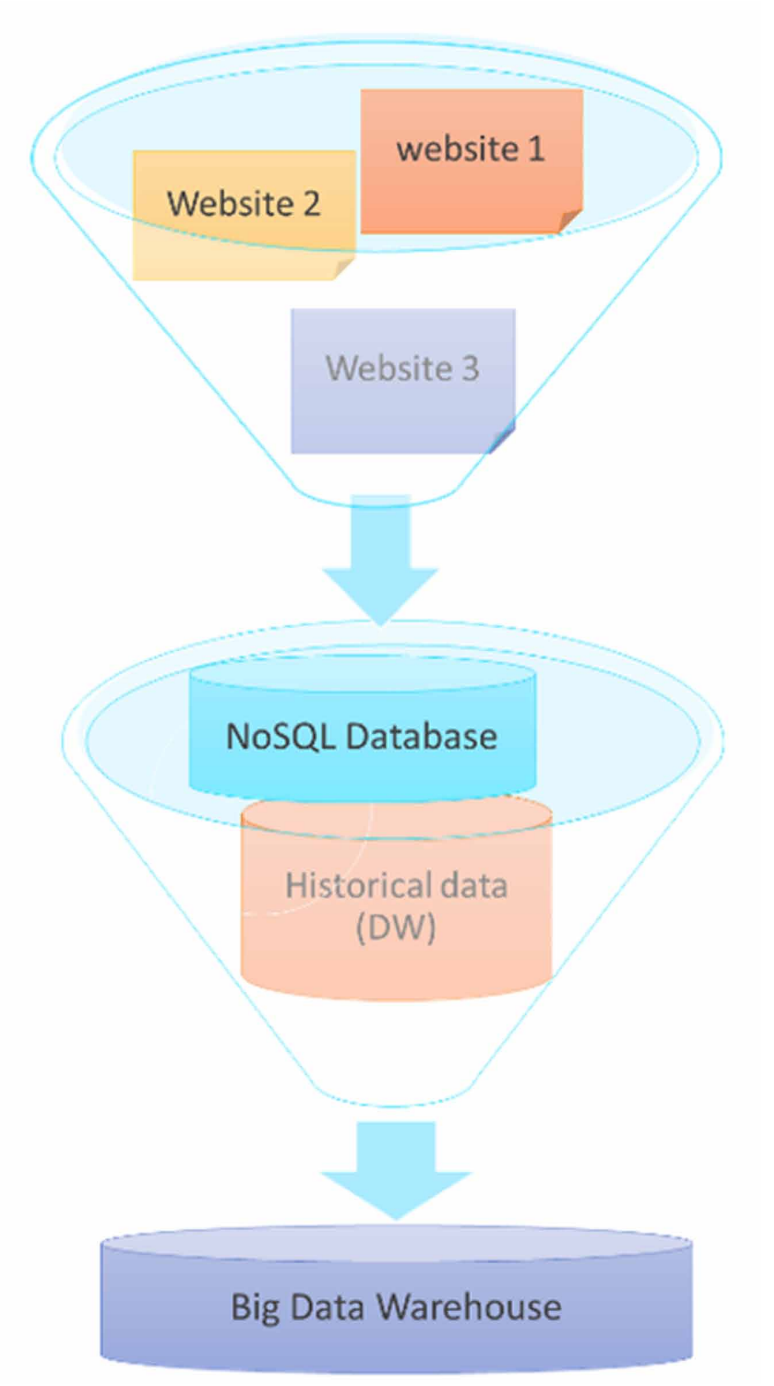


Figure 2. At the top, the rendered view showing the image that indicates the number of stars and, in the bottom, the HTML code that shows the tag without explicit information. The number of stars is marked in the top and bottom by a blue rectangle

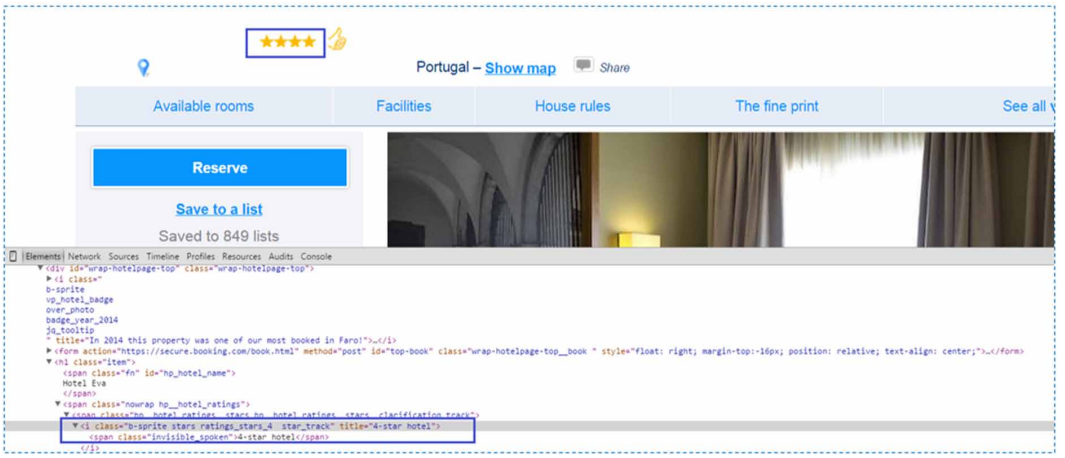
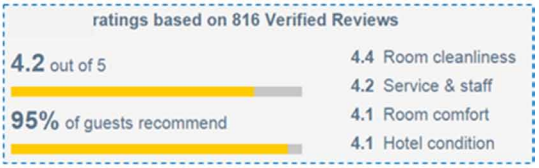


Figure 3. Example of data retrieved by the web crawler from the Expedia Website, and the corresponding JSON, which was saved in the MongoDB's Score collection




```
{
  "_id" : ObjectId("5423f6b4703c3b04260f0694"),
  "_idHotel" : ObjectId("5423f659a563ee1338ba3484"),
  "Source" : "expedia.ie",
  "ExtractionDate" : ISODate("2015-07-24T11:04:13.928Z"),
  "ReviewerType" : "Everyone",
  "ScoresList" : [
    {
      "Title" : " Room cleanliness",
      "Value" : 440
    },
    {
      "Title" : " Service & staff",
      "Value" : 420
    },
    {
      "Title" : " Room comfort",
      "Value" : 410
    },
    {
      "Title" : " Hotel condition",
      "Value" : 410
    }
  ],
  "ReviewersCount" : 816
}
```

Figure 4. Excerpt of the entity-relationship model used to store the hotels' Scores

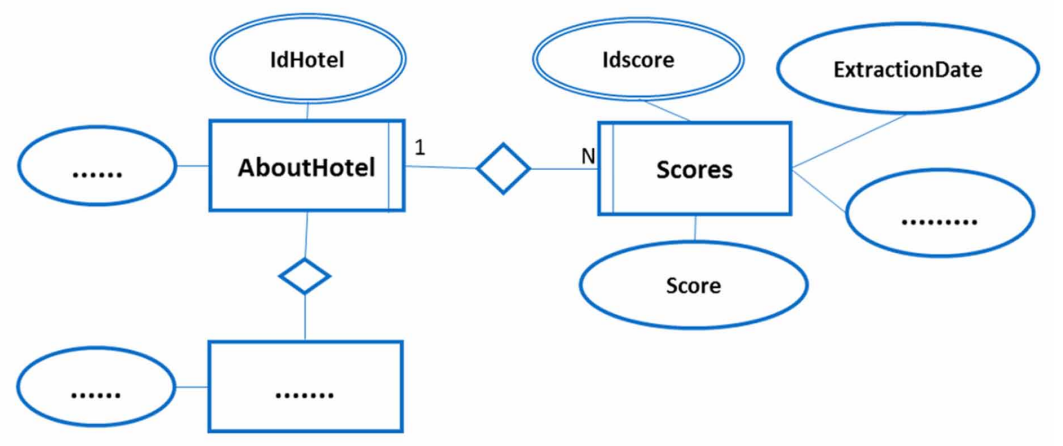


Figure 5. Excerpt of the relational database model presenting the Scores table and the most relevant relationships with others tables

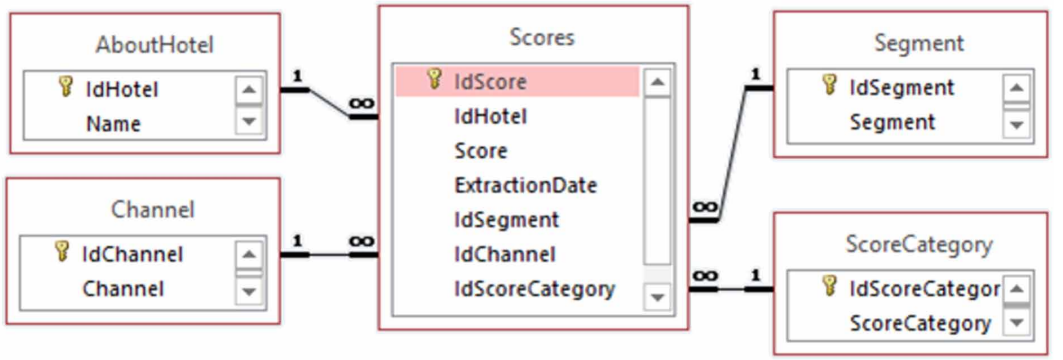


Figure 6. Diagram showing the flow of the data consolidation process

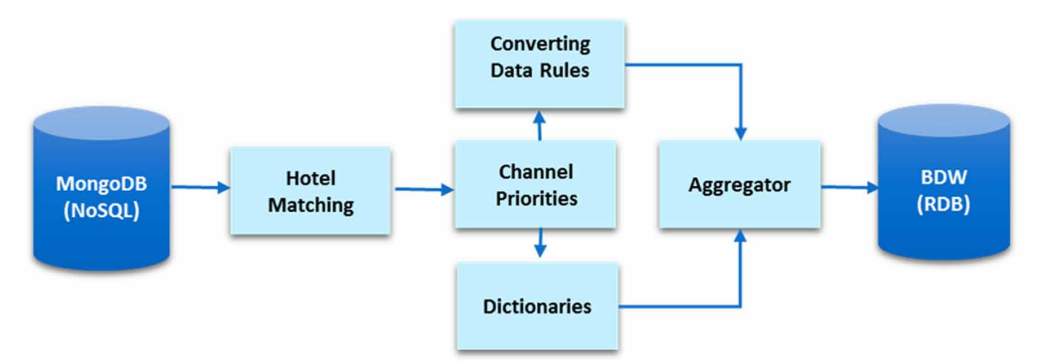


Figure 7. Flowchart of the algorithm to match the data retrieved from different channels for "Hotel X"

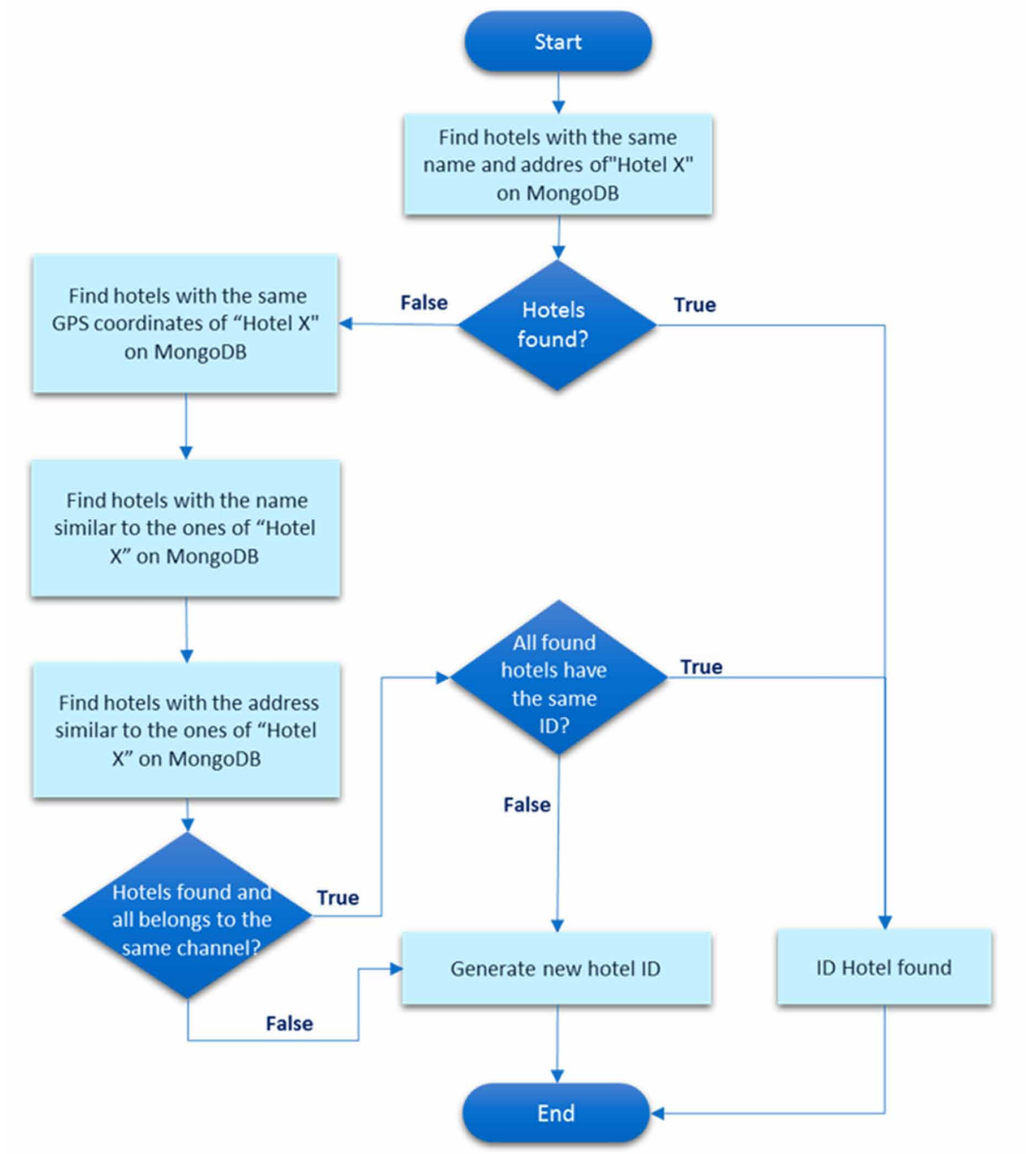
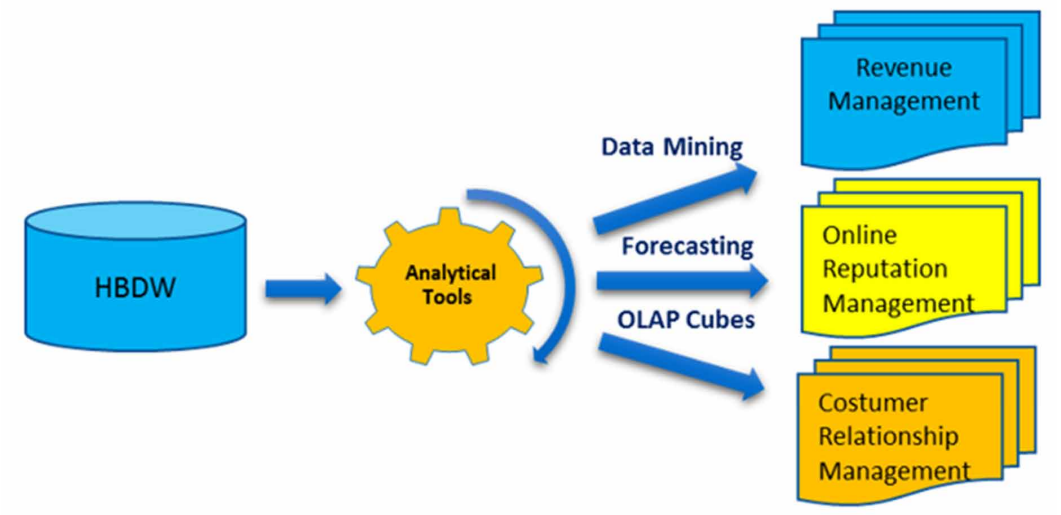


Figure 8. Examples of analytical tools that can exist in a BI system and its applicability to hotel business areas



Célia M. Q. Ramos graduated in Computer Engineering from the University of Coimbra, obtained her Master in Electrical and Computers Engineering from the Higher Technical Institute, Lisbon University, and the PhD in Econometrics in the University of the Algarve (UALG), Faculty of Economics, Portugal. She is Adjunct Professor at School for Management, Hospitality and Tourism, also in the UALG, where she lectures computer science. Areas of research and special interests include conception and development of information systems, tourism information systems, big data, etourism, econometric modeling and panel-data models. Célia Ramos has published in the fields of information systems and tourism, namely, she has authored a book, three book chapters, conference papers and journal articles. At the level of applied research, she has participated in several funded projects.

Daniel Jorge Martins graduated in Electrical and Electronics Engineering in the Instituto Superior de Engenharia - University of Algarve in 2013. He is now finishing his MSc degree in Electrical and Electronics Engineering, also at the University of the Algarve. He participated in 1 financed scientific project, and he is co-author of 5 scientific publications. His major research interests are in Web Crawlers and NoSQL databases.

Francisco Serra received his PhD in Economics and Management Sciences from the University of Huelva, Spain, in 2003. He was hired as an Assistant Professor in 1992 by the School of Management, Hospitality and Tourism of the University of the Algarve, in Portugal. He is the Director of the School of Management, Hospitality and Tourism of the University of the Algarve. He also participates in some research and scientific management networks at national, European and global levels. He has conducted and supervised research in the fields of Hospitality Management, Tourism Development, Regional Economics and Systems Dynamics. From 2006 until 2012 he served at the Administration of a large public hospital in Faro, Portugal and Between 1983 and 2001 he held various positions as assistant and general manager in hotels in various countries, as well as teaching at the Portuguese Institute for Tourism Training and co-owned and administered 3 companies in the Real Estate, Hotel Management Consulting and Catering business areas.

Roberto Lam, graduated in Computer Science in 1995. In 2001 he obtained an MSc degree at the University of the Algarve in Faro, where he lectures computer science courses at the Instituto Superior de Engenharia. Presently he is pursuing a PhD degree in the Vision Laboratory (UAlg). He is member of the LARSyS (Lisbon) and the Portuguese Chapter of Eurographics. His major research interest is tridimensional modelling: 3D object representation, recognition and retrieval.

Pedro J.S. Cardoso holds a PhD in the field of Operational Research from the University of Seville (Spain), a Master in Computational Mathematics from the University of Minho (Portugal) and a Degree in Mathematics - Computer Science from the University of Coimbra (Portugal). He teaches Computer Science and Mathematics at the Instituto Superior de Engenharia of the Universidade do Algarve (UAlg) and is member of LARSyS. He has high knowledge in the fields of databases, algorithms and data structures, and Operational Research. Over the past few years has been involved in 7 national and international scientific and development projects and is the co-author of about 40 scientific publications.

Marisol B. Correia is a professor in the Information Technologies and Systems scientific area of the School of Management, Hospitality and Tourism of the University of Algarve and she is a collaborator of the Centre for Management Studies of IST of the University of Lisbon. She holds a PhD in Electronics and Computer Engineering from the University of the Algarve, a masters in Electronics and Computer Engineering from the University of Lisbon and 5-year undergraduate degree in Informatics Engineering from the University of Coimbra. She is a scientific reviewer for some international journals and conferences and she participated in I&DT projects. She has presented papers at international conferences and has published scientific papers in several journals. Her current research interests include Information and Communication Technologies applied to Hospitality, Management, Tourism and Marketing, Websites Evaluation, Web Semantic, Business Intelligence and Evolutionary Computation.

João M.F. Rodrigues graduated in Electrical Engineering in 1993. He got his MSc in Computer Systems Engineering in 1998 and PhD Electronics and Computer Engineering in 2008 from University of the Algarve, Portugal. He is Adjunct Professor at Instituto Superior de Engenharia, also in the University of the Algarve, where he lectures Computer Science and Computer Vision since 1994. He is member of associative laboratory LARSyS (ISR-Lisbon), CIAC and the Associations APRP, IAPR and ARTECH. He participated in 14 financed scientific projects, and he is co-author of more than 120 scientific publications. His major research interests lie on computer and human vision, assistive technologies and human-computer interaction.