Big Data Warehouse Framework for Smart Revenue Management

CÉLIA M.Q. RAMOS, MARISOL B. CORREIA
Escola Superior de Gestão, Hotelaria e Turismo
University of the Algarve
Campus da Penha, 8005-139 Faro, PORTUGAL
{cmramos, mcorreia}@ualg.pt

JOÃO M.F. RODRIGUES, DANIEL MARTINS,
Instituto Superior de Engenharia, LARSyS and CIAC,
University of the Algarve
Campus da Penha, 8005-139 Faro, PORTUGAL
{jrodrig, djmartins}@ualg.pt

FRANCISCO SERRA
Escola Superior de Gestão, Hotelaria e Turismo
University of the Algarve
Campus da Penha, 8005-139 Faro, PORTUGAL
fserra@ualg.pt

Abstract: - Revenue Management's most cited definitions is probably "to sell the right accommodation to the right customer, at the right time and the right price, with optimal satisfaction for customers and hoteliers". Smart Revenue Management (SRM) is a project, which aims the development of smart automatic techniques for an efficient optimization of occupancy and rates of hotel accommodations, commonly referred to, as revenue management. One of the objectives of this project is to demonstrate that the collection of Big Data, followed by an appropriate assembly of functionalities, will make possible to generate a Data Warehouse necessary to produce high quality business intelligence and analytics. This will be achieved through the collection of data extracted from a variety of sources, including from the web. This paper proposes a three stage framework to develop the Big Data Warehouse for the SRM. Namely, the compilation of all available information, in the present case, it was focus only the extraction of information from the web by a web crawler – raw data. The storing of that raw data in a primary NoSQL database, and from that data the conception of a set of functionalities, rules, principles and semantics to select, combine and store in a secondary relational database the meaningful information for the Revenue Management (Big Data Warehouse). The last stage will be the principal focus of the paper. In this context, clues will also be giving how to compile information for Business Intelligence. All these functionalities contribute to a holistic framework that, in the future, will make it possible to anticipate customers and competitor's behavior, fundamental elements to fulfill the Revenue Management.

Key-Words: Revenue Management, Data Warehouse, Big Data, Business Intelligence, Semantic Web, Tourism, Hospitality, Marketing.

1 Introduction

In the area of hospitality, the information to be managed has very specific features since it comes from different activities related to the tourism sector, such as facilities and transportation, among others. It is constantly undergoing changes as, for example, the tariffs offered to potential customers who want to book a room.

To the hotel, it is relevant that marketers and managers have access to intelligence, and make the best use of it [1]. These professionals have invested heavily in recent years, organizing strong scientific teams, including statisticians and database (DB) experts, well equipped to build and analyze the contents of their Data Warehouses.

However, the development and use of internal data sources is no longer sufficient to ensure competitive advantage [2]. This type of Data Warehouse consists of information from the transactions that occur within the organization, while, nowadays, it is necessary to consider the

ISBN: 978-1-61804-280-4

current trend that favors the development and use of Big Data Warehouse architectures consisting of internal and external data sets [3, 4].

The concepts associated with Big Data [5] are describe as technologies that promise to fulfill a fundamental tenet of research in information systems, which are to provide the right information to the right receiver in the right volume and quality at the right time. Following the same path, the concept of Big Data Warehouse refers commonly to the activity of collecting, integrating, and storing large volumes of data coming from data sources, which may contain both structured and unstructured data. Volume alone does not imply Big Data. Other specific issues are related to the velocity in generating data, their variety and complexity [3].

Nowadays, hospitality industry and its partners, hotels, airline companies and travel agents are promoting their services on the web. Consequently, the World Wide Web (WWW) has become a global vitrine where specialized sites, e.g., Global Distribution Systems (GDS) and Online Travel Agents (OTA) operate, thus, providing publicly available information that can be collected, generating large sets of data, that can be used for business intelligence purposes, providing a comparison of offers for similar products.

In the early days of web-based business, data could be freely acquired from specialized websites, because it was in the business company's interest to promote their products [6]. However, nowadays, this panorama is rapidly changing, and information is not free and easy to collect. Nevertheless, hotel marketers need to have access to this kind of information, to define their revenue management policies and to redefine their business tactics and strategies, by using Business Intelligence and analytical techniques to promote and sell their rooms, at the best possible price to the right costumers.

Smart Revenue Management (SRM) is a project in development by the University of the Algarve and VISUALFORMA - Tecnologias de Informação, SA, which aims to develop a set of tools to integrate in a Revenue Management (RM) system. This paper, presents the conceptual and some practical stages in development to construct a Big Data Warehouse (BDW), that will allows the detection of knowledge and the development business intelligence analytics applications.

The article is structured as follows: besides the introduction, the second Section presents a thorough contextualization of the subject of study. The third Section highlights the relevance of Big Data Warehouse, mainly to the hospitality and tourism organizations. The fourth Section, presents the process to develop business analytic tools, based on the BDW, including the analyses of the challenges in hand and the proposed solution to solve it. Finally, we will present some discussion, conclusions and suggestions for future work.

2 Contextualization and State of Art

In the current society, information, creativity and knowledge play an important role in any organizational process and strategy. To cope with globalization, it is essential to use mechanisms that allow the collection and treatment of essential information for the organizations. The optimization of that information in a differentiated way for management tactical and strategic purposes is essential in all organization levels; which aims the reduction of uncertainty in the decision-making processes and track the most sensitive parameters of the organizational performance [7].

Such mechanisms/stages, in the case of the Smart Revenue Management project, include: (a) the automatic collection of information from several sources, including the internal Data Warehouse (DW) of the hotel, but also from the web (using a web crawlers). (b) The storage of the extracted information, and the (c) selection of the most relevant information to the business, taking into consideration the data model suitable for storage, and for the (future) analyses and information treatments.

The analyses to be considered are associated with business analytics, where advanced analytic techniques operate on big data sets. The Big Data analytics is really about two things - big data and analytics - plus how the two have teamed up to generate today one of the most profound emerging trends in Business Intelligence (BI) [8].

In this paper we will not focus in the extraction of data from the internal sources of the hotel (Data Warehouse, Property Management System (PMS), etc.), we will focus only the extraction of information from an external source – the web. For the automatic collection of information from the web (a), a set of crawlers [9, 10, 11] must run periodically in order to produce suitable data [12], nevertheless not all the data that is extracted can be used in all hospitality business models, and from different sites (Booking, Expedia, TripAdvisor, etc.) it is possible to extracted different and coincident information from the same hotel.

In the SRM project, a different crawler was used for each site: Booking, Expedia, TripAdvisor, etc.; for more details see [12]. The crawler extracts

ISBN: 978-1-61804-280-4

periodically all the information existing in each site about each hotel, over different periods of time, and considering different types of users (2 adults, 1 adult and a kid, etc.), which generates an huge amount of "raw" data, that needs to be stored [12].

Related to the storage (b), means getting and store a high volume and data variety at high speed. To store this information it is usually necessary dynamic storage databases, the one chosen was the MongoDB database [12, 13], which is a NoSQL document-oriented database, structured as a set of collections that store documents, it also presents high performance, high reliability, easy scalability and map-reduce, etc.

The last stage (c) consists on the combination and selection of the relevant information from (a) and (b) for the business, in general, is the constructing of the Data Warehouse [14, 15]. Due to the different collections, the integrating, velocity, and the storing large volumes of data coming from the GDS, OTA, internal (DW, PMS), etc. in reality it is a Big Data Warehouse [3, 4]. It is also necessary to consider data models tailored to the needs of the organization, both in terms of features to consider and in terms of information storage structure; as well as semantic concepts [16] to perform a suitable data storage, according to the structure defined. Another important aspect is the information stored in (b), and semantically analyzed to store in the BDW (c) must include the social networks, that define the online reputation (OR) of a product or organization, to develop personalized recommendations and address various customer purchasing behavior [14].

As already mentioned, to access and use the information considered as Big Data, it was contemplated a set of technologies, as for example a NoSQL database. However, the relational database (RDB) continues to be the more prevalent [14] data storage, which allows viewing of data from multiple formats and for different stakeholders, even the ones that their activity is not related to technology. In this sense (not only, as we will see along the text), it is necessary to integrate the information form the MongoDB database in a RDB database, that allows the storage of a collection of data and the access, management and information processing, where the different professionals are able to use and access the data, in a variety of formats.

To do the above transformation, it is necessary to use data models to transform unstructured information in structured, i.e., in relational database models (RDBM). In situations where it is not possible to structure the data present in the NoSQL DB in a RDB is necessary to consider the concepts of semantics for a suitable data processing and conversion, and only later the storage in an appropriate structure.

3. Big Data Warehouse for SRM

The first phase of the generic architecture of the framework is presented in Fig. 1 (for the second phase see Section 4, as well as for the explanation of the "…" appearing in Fig. 1), from (a) the extraction of "all" the information available in the web, by the web crawlers [12], to (b) the storing of all that raw data in a primary database – MongoDB [12]. (c) The creation of semantic models, lexical databases, data models, rules and principles to select and combine the relevant information (in this case for the Revenue Management), and store this information in a secondary database (RDB). The final integration of these three steps (with the ones presented in Section 4, Fig. 9) is the Big Data Warehouse for SRM. Again, we call the attention that in this article, we do not integrate the information from internal sources of each hotel, but in the SRM project, they are being considered (see Fig. 9).

In this paper we will focus on (c) – the last stage in the implementation of the Big Data Warehouse, being already presented stages (a) and (b) in [12], nevertheless, for the better comprehension of the following Sections it is necessary a brief explanation and examples about stage (b).

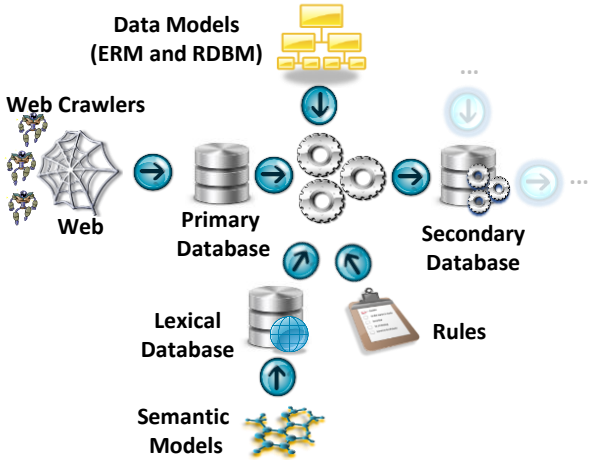

Fig. 1 – Web extraction, selection and conversion of information for the SRM Big Data Warehouse; see text, and Section 4 for the "…" explanation.

3.1 Data Models

As already mentioned, the extracted raw data from the different web crawlers was stored in a MongoDB. Figure 2 presents one example of information retrieved form one site (Expedia), belonging to the collection *Room* (see the remaining

collections, and details in [12], extracted from a specific hotel at a particular date. Different sites (Booking, Expedia, etc.) presents similar and different information extracted about the same topic [12]. In addition, with the structure presented in Fig. 2 it is not possible to make analyses of relationship with other data, for example the price, nor reading the information is intuitive.


```
{
   "_id" : ObjectId("5423f668703c3b04260f0585"),
   "_idHotel" : ObjectId("5423f659a563ee1338ba3484"),
   "Source" : "expedia.ie",
   "ExtractionDate" : ISODate("2014-09-25T11:02:59.005Z"),
   "Search" : {
      "Location" : "      , Portugal",
      "NumberOfAdults" : 2,
      "NumberOfChildren" : 0,
      "NumberOfRooms" : 1,
      "CheckinDelayNights" : 0,
      "DifferenceBetweenCheckinCheckout" : 1,
      "CheckinDate" : ISODate("2014-09-25T11:02:59.005Z"),
      "CheckoutDate" : ISODate("2014-09-26T11:02:59.005Z")
   },
   "RoomName" : "Apartment, 2 Bedrooms",
   "Description" : [
      {
         "Title" : "paragraph-hack",
         "Content" : "1 queen and 2 single\r\nThis room opens to a
furnished balcony. The Select Comfort bed and pillow … This room
is Non-Smoking."
      }
   ],
   "TariffList" : [
      {
         "Conditions" : [
            " FREE Valet Parking",
            "FREE Cancellation before Mon, 13 Oct"
         ],
         "Tax" : [],
         "MaxOccupancy" : [
            {
               "Title" : "max-occupancy",
               "Content" : "Max Occupancy: 4 guests (up to 3
children, 2 infants)"
            }
         ],
         "OldPrice" : {
            "_t" : "TitleValue",
            "Title" : "€",
            "Value" : 16111
         },
         "NewPrice" : {
            "_t" : "TitleValue",
            "Title" : "€",
            "Value" : 14500
         }
      }
   ]
}
```


Fig. 2 – Example of the information about the collection *Rooms* extracted from Expedia, stored in the MongoDB.

To overcome this problem it was considered a model entity-relationship [17], which allows describing reality in terms of a collection of objects and the interaction between them. Taking into account the information presented in Fig. 2 and the concepts of entity-relationship model (ERM) was conducted the analysis of the information system, and has been defined the respective data model, (whose result is presented in a small part in Fig. 3).

Figure 3 shows the association between *Rooms* and *Hotel*, where the "..." represents generically other related entities with the hotel and for which is also being collected information. The entity *Rooms* have some attributes represented in the figure. Namely, RoomName, NewPrice (price with discount), OldPrice (price without discount), NumberOfAdults (number of adults that can be considered to book the room), NumberOfChildren (number of children that can be considered to book the room), and "…" which represent the others characteristics that are also relevant, but aren't represented in the figure.

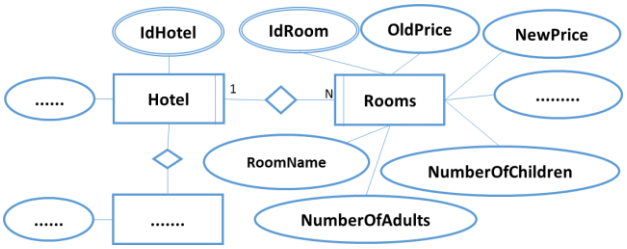

Fig. 3- Excerpt of the ERM.

The next step is to transform the ERM in a structure that it is possible to implement in a RDB. After the analysis, we considered the design of the system, and transform the ERM in a RDBM [18], considering the concepts associated with this data model, where an elementary object will be a table and the association between them will be transformed by specific rules.

The result that ending the conception of an information system, is designated by the specification of the systems and is concretized by the data model to implement in the database system considered, as presented in the Fig. 4.

In the end of the information system conceptualization, the data model includes the tables to create and the relationship to consider between them. In Fig. 4, the table *Rooms* represents the entity *Rooms* in Fig. 3 and the fields that belong to the table, in Fig. 4, are corresponding to the attributes of the *Rooms* entity in the Fig. 3.

The next step is the development of the RDB, or also called the secondary database. In this database is where we will deposit the data collected from the MongoDB, the primary database, according to the logic structure defined by the data model, presented in the Fig. 4. In the data transformation from a NoSQL database to a RDB there are some challenges that the application developer has to face,

such as the insertion of the adequate data in the appropriate fields.

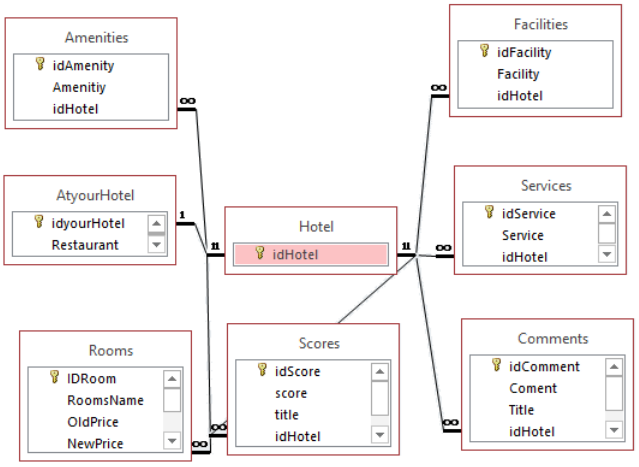

Fig. 4- Excerpt of the RBDM.

For example, and returning again to Fig. 2, considering the information that is formatted in bold, it is possible to see that the content to be inserted in the field *NumberOfAdults* is 2, the number of children that is permitted in the room is zero, the price with discounts is 145.00 euros (in Fig. 2 shown as an integer), which is the content of the field *NewPrice*.

But there are other cases, that aren't so trivial, for example, consider that we have a table *At your hotel* which store the information about the hotel features (property features), for example the number of restaurants, the number of swimming pools, among others considered relevant to the business; see an example in Fig. 5, formatted in bold. To insert these attributes as indicated before, the application developer have to consider the concepts of semantic to find the right information, in the content of MongoDB, so they can be insert in the appropriate field of the RBDM.

Again, we call the attention that even the same information, e.g., the number of stars of a hotel was retrieved from different forms (text, image, image captions) by the web crawler in each site, and even in the same site, it changes along the time, see details in [12]. Nevertheless, in the field *Stars* in the collection *AboutHotel* a number will be available.

For some type of data, during the transformation of the data stored in the MongoDB to the RDB, it will be necessary to map the fields of the first DB into the fields of the second DB. These mappings will not be direct and straight because there is no normalization in the notation used by the different producers of information for the web. For example, at the date this article was written, Booking uses Review Score from 0-10, and Score Breakdown in 7 fields, the Expedia shows Review Score from 0-5, and Score Breakdown in 4 fields.

{
"_id" : ObjectId("5423f65b703c3b04260f0584"),
"_idHotel" : ObjectId("5423f659a563ee1338ba3484"),
"Source" : "expedia.ie",
……..
{
"Title" : "At your hotel",
"Content" : " features a full-service spa, **3 outdoor swimming pools**, an outdoor tennis court, ... room(s)\r\nMeeting rooms\r\nMultilingual staff\r\nFree valet parking\r\nPoolside bar\r\nBeach bar\r\nPorter/bellhop\r\nLuggage storage\r\nArea shuttle (surcharge)\r\n**Number of restaurants - 2**\r\nConference center\r\nOutdoor tennis court\r\nSauna\r\nSpa services on site\r\nSteam room\r\nWedding services\r\nHair salon\r\nBeach/pool umbrellas\r\nTours/ticket assistance\r\nWireless Internet access - surcharge\r\nWired (high-speed) Internet access - surcharge\r\nNumber of outdoor pools - 3\r\nChildren's club\r\nRoom service … tub\r\nSupervised childcare/activities\r\nChildren's pool\r\nIndoor pool\r\nHide"
}
…..
}

Fig. 5 - Example of the information about the hotel property stored in the MongoDB, extracted from Expedia.

On the other hand, as it is known, live languages consist of phrases and words with multiple meanings of difficult understanding for computational systems. Also, the utilization of plurals instead of singulars can worsen this problem. For the interpretation of the meaning of a sentence by a computational application, we need more than a dictionary because language is polysemic, i.e., the same word or phrase can acquire various meanings according to the context in which it operates.

3.2 Lexical database, semantics and ontology

A lexical database as the WordNet, developed by Princeton University (WN.Pr) [19, 20] as a Natural Language Processing (NLP) application, can help to interpret the meaning of the sentences (see also Section 4.1). Lexical information is not organized in word forms, but in word meanings, which is consistent with the human representations of meaning, and their processing in the brain.

As mentioned, there is no normalization of the information used and displayed in the websites, it can happen that two websites as e.g., Booking and Tripadvisor, use different words to designate the same facilities or amenities. For example, they can use different words to designate the same type of room. This can be solved using the mentioned lexical database, which will be responsible for the mappings between the MongoDB and the RDB, and taking in consideration the semantic web concepts; see the structure in Fig. 1.

ISBN: 978-1-61804-280-4

A different support can come from the semantic web [16]. In the semantic web, the organization of the pages structure is different from the current web, as shown in Fig. 6. In the semantic web, the structure consists of software, documents, libraries, images, concepts and people, in the case of current web, each document provides hyperlinks to other documents, which may, or may not be linked between them. The semantic web seeks to understand the meaning more than the content present on the page [21], in order to identify the existing knowledge on the web through in a way that is understandable to all (canonical form, that is, in its simplest form).

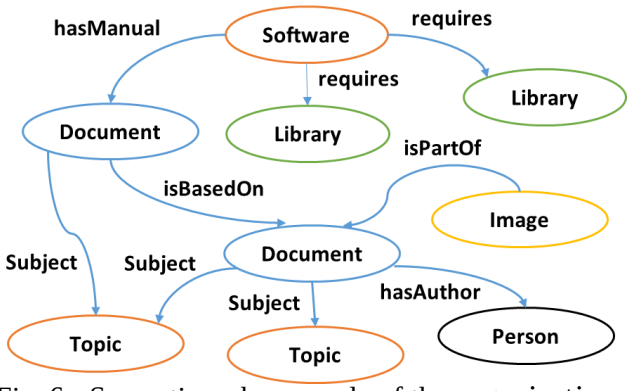

Fig. 6 - Semantic web, example of the organization of the page structure.

According to W3C [22], the goal of the semantic web is to create a universal medium for data exchange. In the semantic web environment requires the ability to represent and manage the content on the web in the form semantics, i.e., allow an agent to learn the meaning of a term by appointment of a formalization of terms based on metadata, ontologies, or other concepts considered to generate knowledge. Moreover, the extraction of information from a collection of documents has to be done in order to satisfy the needs of the user. This extraction is made in documents written in natural language, stored, represented and organized in different types of systems [23], as presented in the Fig 7 (see also [24]).

An ontology consists of a set of classes, relations, instances and axioms, where the classes represent concepts that belong to a domain which describes the ontology relationships and represent the association between the elements of the ontology, the instances are used to represent a particular element of the class and finally, the axioms are assumed to be true statements [23].

The layer of ontologies is one of the most important because it is responsible for providing the necessary expressiveness to the representation of ontologies. An example of a multilingual ontology for the hospitality sector can be seen in [25]. Figure 8 presents a small extract of one ontology related with tourism. In the figure, the term "is a" means that the name "is part of", for example, "Cultural" is part of the "Tourism", and so on.

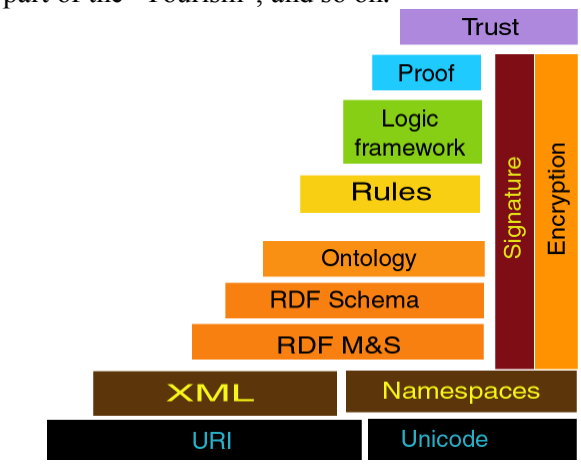

Fig. 7 - Semantic web layers [24].

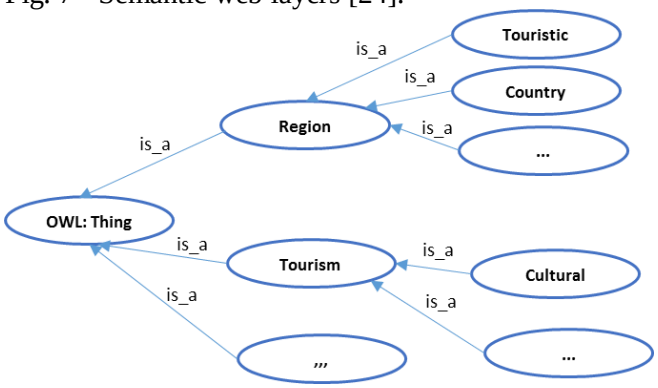

Fig. 8 – Tourism ontology example.

The linguistic ontologies are referenced by their application in natural language processing systems. To work with ontologies are used designated languages for Web Ontology Language (OWL). There are different types of languages, recommended by the World Wide Web Consortium (W3C) [22] to work with different levels of semantic expressivity, as for example the OWL Lite, OWL DL, and OWL Full. In addition, the semantic web will make it possible to find information in the extracted data from the web.

However, when the goal is to extract information from the collected data, stored in the Big Data Warehouse, it is necessary to consider other features.

4. Extracting information for BI

In the above Sections (see also Fig. 1), we show the path of the information from the source, in this case

the web, to the secondary database - Big Data Warehouse. In this Section, we concentrate on complementing the information in the secondary database and the extracting of information from the BDW for the BI. To extract information for the Big Data Warehouse, the use of traditional methods of analysis are no longer adequate [14], making it is necessary to consider new tools, some also described in Section 3.2.

Figure 9 shows the second phase of the generic architecture of the framework, i.e., represents the process to extract knowledge from the data, that is relevant to control and manage the organization and to support the decision maker in the context of Business Intelligence. In the secondary DB (or the SRM Big Data Warehouse), as already mention, and we reinforce the subject, it is necessary to include the Data Warehouse of each hotel (were the SRM will be applied), and use these information system, with the forecasting and RM models most adequate to the business model of each hotel.

With all these data, and with analytic models, and sentiment analysis, an analytic background is created, one, which will make possible for the decision maker to have access to the analytical tools that will facilitate the creation of the business intelligence environment, such as reporting, forecasting and cubing for data analysis.

The analytical tools can be considered for the development of thematic or segmented subsets of the Big Data Warehouse, called Data Marts (DM) to analyze and manage specific areas, such as RM, or Online Reputation or the Customer Relationship Management (CRM) as represented in the Fig. 9.

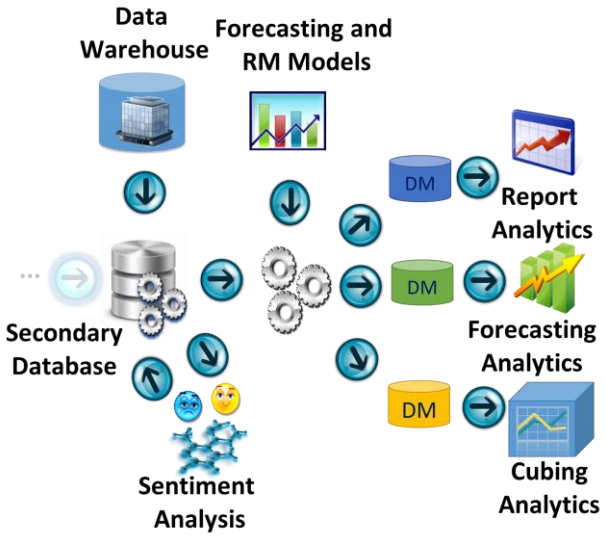

Fig. 9 – Extracting information from SRM Big Data Warehouse for BI (second phase); see text, and Section 3, Fig. 1 for the "…" explanation.

Business Intelligence is a way to identify new opportunities and implementing an effective strategy based on insights, or intelligence, it can offer to the business a competitive intelligence that give a market advantage and a long term stability [26]. The competitive intelligence is developed by considering a set of techniques and tools for the transformation of data into meaningful and useful information for business analysis purposes [26].

In the Business Intelligence process, it is necessary to take in consideration two steps: (a) the extraction of knowledge and the (b) assessment of the intelligence extracted from that knowledge.

Knowledge extraction, a huge information collected is relevant, but it is necessary to use adequate tools to produce knowledge about the organization. The reports and fixed dashboards [14] produced by the Data Warehouse are limited solutions compared with the results from the big data analytics, which can include *ad hoc* queries and discoveries of meaningful relationships between the data. Furthermore, to extract knowledge from the information stored it is necessary to include the data from the Data Warehouse, that have information about the transactional operations of organization, as presented in the Fig. 9, and to include predictive models to help to define the future behavior of the consumers.

Intelligence Assessment, the possibility to extract business value for the organization have captivated several researchers and stakeholders [1, 14, 27, 28, 29]. There are several techniques that are actually considered in the big data environment [14]: (a) recommendation systems, for example in social networks when refers "People you may know likes de hotel X". (b) Analysis of social networks, to identify the influence over others. (c) Analysis of new products, to test new products or ideas and obtain instant feedback. (d) Analyses competitors' pricings, to compare with their prices. (e) Sentiment analysis, which permit to define the costumer sentiment towards products, services, destinations, hotels; among others.

4.1 Sentiment analysis

The sentiment analysis or opinion mining techniques is probably the biggest challenge in the second phase of the SRM Big Data Warehouse, it comprises an area of NLP, computational linguistics and text mining [28], and refers to a set of techniques that deals with data about opinions and tries to obtain valuable information from them [30]. It is constituted by a group of computational techniques used to extract, sort, understand and

ISBN: 978-1-61804-280-4

evaluate the opinions expressed by users about products, services, destinations, cruise companies, hotels, among others; from textual sources. It can be used, for example, to understand the opinions of the hotel clients or product consumers. The emergence of the semantics has created many opportunities to understand the views of the consumers on marketing campaigns and preference for products. Some of the concepts already presented in the Section 3.2 can be used to extract the characteristics and to identify the opinion associated with those characteristics, which may be positive or negative [28, 30].

The semantic is the key, not only one, to find information in the content of textual field in primary DB (MongoDB), but also to consider the customers opinion data (feedbacks) and apply a sentiment analysis to produce intelligence associated to the organization, which is a challenging subject of investigation and with difficult implementation. One of reasons is because of the nature of the associated tourism information, whether referring to data of the hotels, transportation, or entertainment. Another reason, and most relevant, is identify the adequate methodology to apply the ontologies to the tourism and hospitality information. This difficulty is present in the storage of the information from the MongoDB to the secondary database, as in the analyses referred before, with more impact in the sentiment analyses.

Nevertheless, there is at least one solution to solve this challenging problem; it is to consider a lexical database that can help to interpret the different meanings and to find the synonyms of words. WordNet, [19, 20, 31, 32] (see also Section 3.2) can be seen as a "dictionary of meaning," integrating the functions of a dictionary and a thesaurus. As the data extracted by the web crawler can be in different languages: English, Portuguese or Spanish, the adoption of the various adjustments to the WordNet to other languages can be one of the approaches. The other one could be to translate all the data extracted by the web crawler, independently of the language in which it is, to the same language, for instance, to English. By default, the web crawler searches the information in English, nevertheless users comments can appear in several languages.

If the decision is not to translate, in Global Wordnet Association Website [32], there is information about the languages, the name of the resources and the type of license on the various adaptations available worldwide. For instance, adaptations to the Portuguese are three, two for the Portuguese of Portugal (Onto.PT and WordNet.PT) and one for Portuguese of Brazil (OpenWN-EN) [33, 34].

In WordNet, nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms, called *synsets*, each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. Nouns can be connected through *hypernyms*/*hyponyms* and *meronyms*/*holonyms* relationships; verbs are organized by *troponym*/*hyperrnym* and entailment relations; while adjectives are linked to their antonyms, and relational adjectives point to their related nouns. Finally, adverbs mostly derived from adjectives and are linked to them via a *pertainym* relation.

There are several API for WordNet, as for Java, C# and Python [35, 36, 37]. One of them that is particularly well known is the Java API for WordNet Searching (JAWS) [38], which is an API that provides Java applications with the ability to retrieve data from the WordNet database. Futhermore, for each *synset*, WordNet shows the several relationships. *Hypernyms* are the *synsets* that are more general and the *hyponyms* are the *synsets* that are more specific. The *synsets* and the relationship *hypernyms*/*hyponyms* are the principal relationships that will be expected to be used for the SRM project.

However, there are others types of relationships. The *holonyms*/*meronyms* relationship is one of them, where the *holonyms* are used to denote a whole and the *meronyms* are used to denote things that are a part of something. For example, given: "floor," "wall" and "room light" were some of the *meronyms* found and "building" and "edifice" were the *holonyms* reached.

The WordNet 3.0 contains 155.287 distinct words, distributed in 117.659 *synsets*, resulting in 206.941 pairs of word/meaning [31, 32].

In resume, WordNet and the other adaptations of this lexical database to other languages are popular NLP applications, which allow disambiguating senses of words, measuring their relatedness to others and defining and describing their meaning. In this research, these lexical databases are used in the normalization of the data during the transformation of the primary DB into the secondary DB. However, WordNet cannot be used for a complete semantic analysis of a text or corpus, which may require detecting and processing sentiments. To do this, sentiment analysis or opinion mining can solve this problem.

5 Discussion

The Big Data collected about the environment that surrounds an organization, mainly in the hospitality

industry, and apply to them big data analytic tools is a powerful way to support the decision maker and to control the organization.

Big Data are mainly velocity, volume and variety [14], the huge amounts of data, collected from different sources and high velocity will, in conjunction with the data of the organization itself, constitute the Big Data Warehouse. It is a necessary information system once the travel business is in constant change, and the stakeholders need to visualize the information business in real time, to detect urgent situations and automate with immediate answers [15]. For example, with new policies applied to the rooms, dynamic pricing, taking in consideration the competitive set of the hotel.

With the concepts associated with the development of an information system it is possible to develop and implement a Big Data Warehouse. However, the traditional Data Warehouse as to complemented with external sources, for that it is necessary to take in considerations some technologies such as: web crawlers and NoSQL databases. By other side and in the context of the present work, the concepts and techniques of semantics have to be included in the system to overcome the problems that are founded in the creation of a system with this dimension, this variety, and this quick analytics tools. The WordNet was considered as a semantic tool to solve some limitations; however, this solution is not fully adequate to a sentiment analysis or opinion mining.

This work is an asset to the hotel managers and marketers and tourism stakeholders, as it features a set of stages and intermediate phases and steps necessary for creating a Big Data Warehouse and the development of big data analytic tools whose capabilities for managers and for business intelligence are obvious and go with the current trend in the society.

In terms of future work, consist in the completing the development of the application/software, once part is already under development, and presenting promising results (see [12]), and solve the limitations regarding the sentiment analysis, that will be addressed and developed through the use of concepts and semantic techniques.

Acknowledgements: This work was supported by project SRM QREN I&DT, n.º 38962 and FCT and FEDER/COMPETE projects LARSyS (PEst-OE/EEI/LA0009/2013), CIAC (PEstOE/EAT/UI4019/2013), CEFAGE (PEst-C/EGE/UI4007/2013) and CEG-IST - Universidade de Lisboa. We also thanks to project leader VisualForma - Tecnologias de Informação S.A.

References:

[1] Tim Peter, *Use hotel data to drive growth*. http://www.hotelnewsnow.com/Article/14553/ Use-hotel-data-to-drive-growth, accessed 26/10/2014.

[2] Carlos Caldeira, *Data Warehousing*, Edições Sílabo, 2012.

[3] Di Tria, Francesco, Ezio Lefons, and Filippo Tangorra, *Big Data Warehouse Automatic Design Methodology,* Big Data Management, Technologies, and Applications, 2014, pp. 115-149.

[4] Mohanty, Soumendra, Madhu Jagadeesh, and Harsha Srivatsa, *Big Data Imperatives: Enterprise Big Data Warehouse*, 'BI'Implementations and Analytics, Apress, 2013.

[5] Michael Schermann, Helmut Krcmar, Holmer Hemsen, Volker Markl, Christoph Buchmüller, Till Bitter, Thomas Hoeren, Big Data - An Interdisciplinary Opportunity for Information Systems Research, *Business & Information Systems Engineering*: Vol. 6: Iss. 5, 2014, pp. 261-266.

[6] Caryl Helsel, Kathleen Cullen, *Dynamic Packaging – 2005 White Paper series*, Hotel Electronic Distribution Network Association (HEDNA), the SolutionZ Group, VA, 2005.

[7] Kenneth C. Laudon, Jane P. Laudon., *Management information systems: managing the digital firm*, New Jersey, 13, 2013.

[8] Philip Russom, *Big data analytics*, TDWI Best Practices Report, Fourth Quarter, 2011.

[9] Taofen Qiu, Tianqi Yang, Automatic information extraction from e-commerce web sites, In Proc. *Int. Conf. on E-Business and E-Government* (ICEE), IEEE, 2010, pp. 1399–1402.

[10] Nikolaos K. Papadakis, Dimitrios Skoutas, Konstantinos Raftopoulos, Theodora A .Varvarigou, *Stavies: A system for information extraction from unknown web data sources through automatic web wrapper generation using clustering techniques*. IEEE Tr. on Knowledge and Data Engineering, Vl. 17, Nº 12, 2005, pp. 638–1652.

[11] Robert Baumgartner, G. Ledermiiller, Deepweb navigation in web data extraction, In *Proc. Int. Conf. on Intelligent Agents, Web Technologies and Internet Commerce*, IEEE, Vol. 2, 2005, pp. 698–703.

[12] Daniel Martins, Roberto Lam, João M.F. Rodrigues, Pedro J.S. Cardoso, Francisco Serra, A Web Crawler Framework for Revenue Management, submitted to *14th International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases* (AIKED '15), Tenerife, Canary Islands, Spain, 2015.

[13] Eric Redmond, Jim R Wilson, *Seven databases in seven weeks: a guide to modern databases and the NoSQL movement*, Pragmatic Bookshelf, 2012.

[14] Bob Offutt, *Big Data: Redefining Travel Business Decision Making*, A White Paper Sponsored by UNIT4 Business, Phocuswright, 2014.

[15] Mike Gualtieri, Rowan Curran, Holger Kisker, Martha Bennett, Boris Evelson, David Murphy, *The Forrester Wave™: Big Data Streaming Analytics Platforms*, Q3, 2014.

[16] Tim Berners-Lee, James Hendler, Ora Lassila, The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities, *Scientific American*, Vol. 284, Nº 5, 2001, pp. 1-5.

[17] Peter Pin-Shan Chen, The Entity-Relationship Model: Toward a Unified View of Data, *ACM on Database Systems*, Vol. 1, No. 1, March, 1976.

[18] Edgar Frank Codd, Relational Model of Data for Large Shared Data Banks, *Communications of the ACM*, Vol. 13, Nº 6, 1970, pp. 377-387.

[19] Christiane Fellbaum, *WordNet: An Electronic Lexical Database*, Cambridge, MA: MIT Press, 1998.

[20] George A. Miller, WordNet: A Lexical Database for English, *Communications of the ACM*, Vol. 38, No. 11, 1995, pp. 39-41.

[21] Y. Y. Yao, N. Zhong, J. Liu, S. Ohsuga, Web Intelligence (WI) research challenges and trends in the new information age, In *Web intelligence: Research and development*, Springer Berlin Heidelberg, 2001, pp. 1-17.

[22] W3C, *World Wide Web Consortium*, http://www.w3.org/, accessed in 25/11/2014.

[23] Claudia Deco, Bender Cristina, Adrián Ponce. *Proposal of an ontology based web search engine*, XIV Congreso Argentino de Ciencias de la Computación, 2008.

[24] Tim Berners-Lee, *Enabling Standards & Technologies - Layer Cake*, http://www.w3.org/2002/Talks/04-sweb/slide12-0.html, acessed in 20/11/2014.

[25] Marcirio Chaves, Cássia Trojahn, Towards a multilingual ontology for ontology-driven content mining in social web sites, *Proceedings of the ISWC 2010 Workshops*, Vol. I, 1st International Workshop on Cross-Cultural and Cross-Lingual Aspects of the Semantic Web, 2010.

[26] Olivia Rud, *Business Intelligence Success Factors: Tools for Aligning Your Business in the Global Economy*, Hoboken, N.J: Wiley & Sons, 2009.

[27] Kevin May, *Crawling is the new API – a legal and technical rough guide for the travel industry*, http://www.tnooz.com/article/API-new-crawling-legal-technical-guide/, accessed in 21/11/2014.

[28] Edison Marrese-Taylor, Juan D. Velásquez, Felipe Bravo-Marquez, A novel deterministic approach for aspect-based opinion mining in tourism products reviews, *Expert Systems with Applications*, Vol. 41, Nº 17, 2014, pp. 7764-7775.

[29] Albert Weichselbraun, Stefan Gindl, Arno Scharl, Enriching semantic knowledge bases for opinion mining in big data applications, *Knowledge-Based Systems*, Vol. 69, 2014, pp. 78–85.

[30] Bing Liu, Web data mining, Springer-Verlag Berlin Heidelberg, 2007.

[31] WebONTO, http://projects.kmi.open.ac.uk/webonto/, accessed in 25/10/2014.

[32] *Wordnet Princeton*, http://wordnet.princeton.edu/, accessed in 25/10/2014.

[33] OntoLingua, http://www.ksl.stanford.edu/software/ontolingua/, accessed in 25/10/2014.

[34] Jorge Morato, Miguel A. Marzal, Juan Lloréns, José Moreiro, WordNet applications. In *Global Wordnet Conference*, Vol. 2, 2004, pp. 270-278.

[35] JWNL (Java WordNet Library), http://sourceforge.net/projects/jwordnet/, accessed in 21/11/2014.

[36] WordNet C#, https://wordnet.codeplex.com/, accessed in 21/11/2014.

[37] PyWordNet, http://osteele.com/projects/pywordnet/, accessed in 21/11/2014.

[38] Java API for WordNet Searching (JAWS), http://lyle.smu.edu/~tspell/jaws/, accessed in 21/11/2014.