Challenges in Building a Big Data Warehouse Applied to the Hotel Business Intelligence

DANIEL MARTINS
Instituto Superior de Engenharia
University of the Algarve, LARSyS
Campus da Penha, 8005-139 Faro
PORTUGAL
djmartins@ualg.pt

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

JOÃO M.F. RODRIGUES, PEDRO J.S. CARDOSO, ROBERTO LAM
Instituto Superior de Engenharia, LARSyS and CIAC
University of the Algarve
Campus da Penha, 8005-139 Faro
PORTUGAL
{jrodrig, pcardoso, rlam}@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: - All the tourism stakeholders, mainly the hoteliers, need a state of the art online information system to reply to the customers searches with the necessary and updated information. On the other hand, the hoteliers also need information about their organization competitive set, which implies having access to information about the clients, competitors, and all the stakeholders associated to the hospitality activity. To satisfy the hotels and consumer needs, it is essential to have access to a Business Intelligence (BI) system, which consolidates all the relevant data to be used by the analytical tools. In this sense, BI systems have some challenges related to the definition of an adequate methodology to integrate and store the retrieved data into a hospitality Big Data Warehouse. This paper presents the challenges and some of the necessary steps to overcome the problems associated with the information management and consolidation in a hotel Big Data Warehouse.

1 Introduction

Common travelers plan their vacations or travels using the Internet to search for information about tourism products that they intend to consume, such as accommodation, transportation and entertainment. In addition, they search for information about other traveler's opinions, to know if they had a good experience in the destinations,

they intend to visit. Therefore, it is common to have travelers making their decisions, about what they want to experiment in their holydays, using the information they have access to in the internet. In resume, they will buy according to their preferences and the opinions of others travelers [1].

In this environment, for all the tourism professionals, in particular for hoteliers, marketers and organization managers, it is relevant to have access to analytical information of what are the traveler's commentaries, for instance, which commentaries emerge in the internet, and which were more considered by the consumer in the moment of their decision. These commentaries become part of the online reputation of the tourism destination. For the hotel, the online reputation and the travelers who produce the commentaries are considered as hotel representatives or hotel brand agents [2, 3].

For the hotel decision makers, the access to a technological analytic tool that collects the online information about their business and their competitor (which integrates analyses to support the decision maker) is mandatory. The need to define strategic actions to create new values or increase their competitiveness, can make the difference between success and failure, once the analytic tool can achieve new ideas to implement new actions and create new values to the consumers [4].

However, the creation of tools with these characteristics, which will coexist in a technological architecture, have some challenges to overcome. One of those challenges is the definition of the procedure to store and manage the data collected from different web sources [5]. Maybe even more important, it is the information consolidation in a Big Data Warehouse (BDW), in order to be used by the analytical components in a context of Business Intelligence (BI) and support the different kinds of management in a hotel business [6]. In hotel revenue management activities, which is supported on the manager's experience and available information, must be constantly updated. These challenges assume a great relevance due to their nature.

In addition, in a BDW some tasks must be performed: (i) integration of the information collected from different internet sources. (ii) Update and maintenance of the new information gathered from the internet (once the information about accommodation is constantly changing), e.g., the rooms prices or the promotions offered in according with the customer profile. (iii) Maintenance of the historical data, which will be relevant to the hotel

big data analytics, in order to achieve the business intelligence systems requirements [6].

Based on a primary database [6, 7] founded by data collected by WebCrawlers from different web sources (such as Expedia [www.expedia.com] or Booking [www.booking.com]), this paper presents the challenges and the necessary steps to overcome the problems associated with the definition of an adequate structure and suitable process of information consolidation in a BDW, here designated as a secondary database (see also [6]). The collected data is then to be used by the analytical tools to fulfil the needs of a Hotel Business Intelligence Manager.

The article is structured as follows: after the introduction, the second section presents the contextualization of the subject of the study. The third section highlights the procedure to the consolidation of the extracted data in a BDW. Finally, we will present some conclusions and suggestion for future work.

2 Contextualization

In the area of tourism and hospitality, the volume of information associated with tourism activity, to represent all the different activities related to the tourism sector, is huge. When the tourists plan their travels they search for information about aspects that are related with their preferences, as well as the opinions about the same touristic destinations made by others travelers. From the moment when the destination is chosen and the purchase occur, they are buying a product based on the information mainly available in the web [8]. The relevant touristic information that exists on the internet is updated constantly and has to be managed several times a day to reflect the business needs and the consumer needs.

For the organizations, mainly those related with hospitality, it is important to have access to the information associated with their business, both in the supplier and in terms of demand perspective. This information allows to better understand and analyze the entire environment that surround the organization, including the customer's preferences, online reputation, business trends, among others.

These organization environments (on the internet) are associated with the hotel activity and are characterized by a big volume of data [9, 10]. The processed data comes from different sources, with great diversity, in an unstructured format. In this context, nowadays, the concepts associated with Big Data have to be considered by the organizations

in general [6], and have higher relevance to the ones whose business is situated in the hospitality activity.

For the hospitality, the stakeholders need an information systems, which provide the right information to the right receiver in the right volume and quality, at the right time [11]. Furthermore, to the managers and marketers, to have access to information with these characteristics, can make the difference in their ability to increase the competitiveness and ensure their survival, in a society that all days emerge new competitors and new trends that influence the consumer and offer new values that satisfy their preferences and needs.

It is therefore important and necessary to consider analytical tools that integrate historical analysis from several years, with functions to support decision-making in terms of management and in terms of strategy. Those tools permit the construction and development of Key Performance Indicators (KPI) [12, 13, 14] for the hoteliers and marketers.

For all the touristic organizations, and in particular to the ones related to the hospitality activities, the knowledge management and the BI are the areas that contributes to improve the quality and quantity of information. With that information on hand, it is possible to increase the business and organization performance, supported on decisions that are more accurate and to define expert strategically plans to the organization [15].

The BI systems are supported on a set of phases that includes several kinds of technology and concepts, namely: (i) Data Integration, (ii) Data Warehousing, (iii) Online Analytical Processing Cubes, (iv) Data Mining methods, and (v) Analytic Tools.

The first phase, designated by (i) Data Integration, is constituted by the Extraction, Transformation and Load (ETL) process that integrate the data that exist in different sources in a Data Warehouse [16], with all the information that is pertinent to the organization, from internal and external sources (see Fig. 1, and also [6]). After the identification of the several sources (e.g., search engines), the data is selected and extracted in the extraction process.

The following step is the data transformation, which include tasks such as the cleaning and standardization of data, among others that contributes to the integration of all data that is relevant to consider in the data warehouse to support the analytical tools. The last task, the load, is related

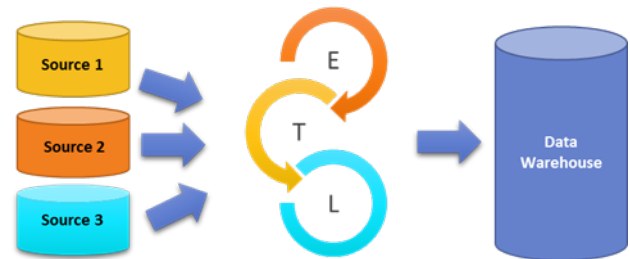with the storing and refreshment of the data in the Data Warehouse.

Fig. 1 – Extraction, Transformation and Load Data in a Business Intelligence System.

The second phase, (ii) Data Warehousing, includes the technology to manage and store the data in a "Data Warehouse". In some cases, business decision makers can also consider the use of Data Marts which are databases constituted by an organization subset of the data, generally related with a department or an activity, which can be independent or dependent of the Data Warehouse [15]. This division allows to fulfil the needs of the organization, taking in consideration the adequate information structure multidimensional model.

The third phase contains the appropriate technology to the (iii) Online Analytical Processing (OLAP) [16, 17], which permit the creation of cubes to explore the information in the Data Warehouse, or in the Data Marts (see Fig. 2). These technologies allow to analyze the information on different business perspectives (dimensions).

The fourth phase, (iv) Data Mining methods, is optional on the Business Intelligence Systems and can be used together with the OLAP cubes. The Data Mining methods consist in the application of artificial intelligence algorithms to discover knowledge in the historical data and, at the same time, to make forecasting to different areas or activities in the organization. The data mining tasks can be divide in two: descriptive and predictive. The descriptive task is considered to identify rules, which characterizes the data, and includes several techniques, such as Clustering and Summarization. The predictive task is pertinent to identify new models that define a variable behavior, which can be used to estimate the future variable values [15].

The last phase is mainly constituted by the use (v) analytic tools. Those tools allow the analytical investigation about the organization data producing enterprise reporting (also called management reporting) or dashboards, which may take the form of graphics, text and tables. The outcomes of the analytical tools can include results from data mining tasks, interactive queries, key performance

indicators, cubes, balanced scorecards, forecasting methods, among others [15, 18].
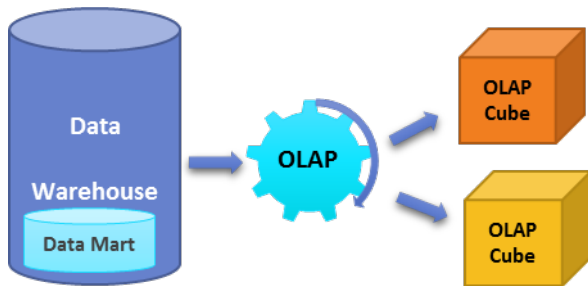

Fig. 2 – Online Analytical Processing in a Business Intelligence System.

However, all of these kinds of technology that coexist in this kind of architecture have a main challenge related to the definition of an adequate methodology to integrate, consolidate and store the data in a BDW [19], which can be relevant to be used by the analytical tools to achieve the requirements of a hospitality business intelligence system.

In the hospitality activity, this challenge assumes a great relevance due to the nature of this activity that is completely supported on information, as presented before. In addition, to achieve the data consolidation, is necessary to find a suitable process to store and manage the data in a secondary database, the BDW. In this case, we are considering that a primary database is constituted, e.g., by the data collected by a WebCrawler from several sites in the web, such as Booking.com. A major problem arises from the fact that different sources have different structure and different meaning to the same hotel features [7]. For example, the "cleanliness" and "room cleanliness" appear in different sources.

In our investigation, the collection of data in the primary database (see Fig. 3) was collected by a web robot or crawler [7, 20] and stored in a NoSQL database, namely MongoDB [21]. The data is then consolidated in a posterior step by integrating it in a secondary database, a BDW [6, 16, 20], as presented above. It is the secondary database that the hotel big data analytics is done.

However, the ETL process in a hospitality BDW takes a higher scale of complexity once the data was collected from different unstructured web channels where, as already mentioned, data with the same meaning can be classified with different designations, which raises the need of consolidation of the extracted data to ensure that the information stored in our BDW is consistent and reliable to the hotel business.
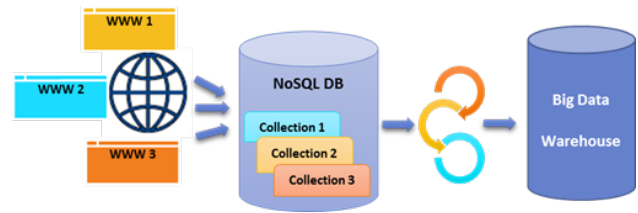

Fig. 3 – Data integration in a Big Data Warehouse.

3 Consolidation of extracted data

In a first phase, the data was extracted from different web channels and stored in a MongoDB database (DB1), the primary database. Four MongoDB collections were used: *AboutHotel*, *Rooms*, *Comments,* and *Scores* [7]. The collection designated by *AboutHotel* contains the hotels characterization, which includes information about hotel name, location, hotel features and rooms amenities. The collection *Rooms* has the information about the rooms and prices, namely including data about the room's name, and number of adults and children that are permitted in the room. In *Comments* are 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 reviews that tourists have attributed to hotels which contributes to define the hotel online reputation.

The DB1 serves as an intermediate database between the web sites / web robots and the BDW, also called secondary database (DB2), which is a relational database. To create the DB2, it is necessary to apply a set of rules, namely: define the data forwarding rules, identify possible conversions, in order to make them readable data for a particular application. In the following sections will be addressed the techniques used for this purpose (see Fig. 4).
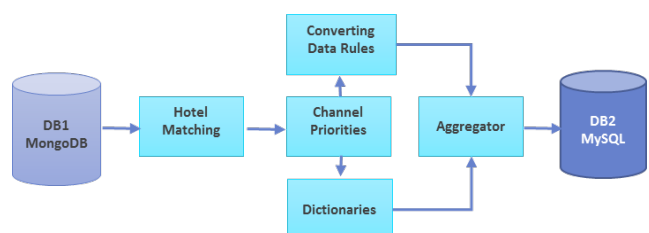

Fig. 4 - Diagram explaining the process for processing data to be consolidated.

3.1 Reading information from DB1

The first step of data consolidation is the reading of data from collections stored in DB1. To do that, it is necessary to know where the data is stored in DB1 and load it into the memory of the consolidation program. This data will be processed and converted, in an adequate way to be stored in the DB2, without

the preoccupation of having an inappropriate elimination or changes in the original data extracted by web robots.

3.2 Data conversion rules

The vast majority of DB1 data is in string format. Although some fields are straight copies from DB1 to DB2, sometimes it is necessary to make some conversions, for instance to numerical values.

An example are the dates of the reviews for a particular hotel. For instance, on *Booking,com* dates appear as "22 March 2015", and therefore it is necessary convert the date to a valid format to store in DB2, i.e., in the format "22/05/2015". Another conversion that is necessary to do are the GPS coordinates of the hotels. These GPS coordinates were extracted from the web, in format "*latitude*, *longitude*" where *latitude* and *longitude* are decimal numbers. For this reason, the string need to be converted into two separate decimal fields.
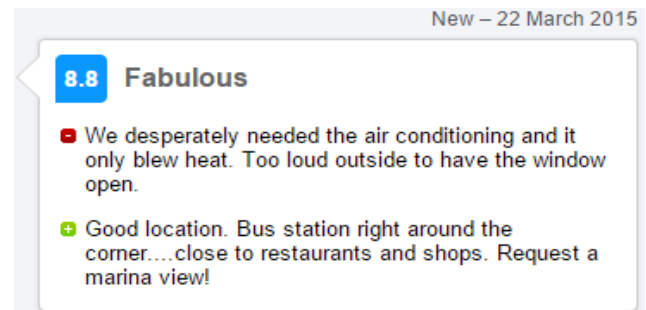
3.3 Data dictionaries

As stated in previous section, the most of the data stored in DB1 is free text, i.e., text written by a human, see [7]. To give meaning to those texts, the consolidation program should use data dictionaries.

These dictionaries are added manually by the user and updated each time the user finds a new synonym for a word. The dictionaries are stored in a MongoDB collection and are structured in four fields: *Source*, *Type*, *Word* and *Alias* (see Fig. 5). The field *Source* indicates the channel (e.g., *Booking.com*) where that particular dictionary should be consulted. The *Type* indicates the context in which particular dictionary should be consulted. For instance "*Type*": "*Amenities*" means that this dictionary should be consulted when looking for the hotel amenities in a sentence. The *Word* field concerns the word that is being searched. Finally, the *Alias* field is a list of synonyms of the word that is necessary to find.

```
{
    "_id" : ObjectId("54f868e13b878b3bc4455493"),
    "Source" : "All",
    "Type" : "Comment",
    "Word" : "PositiveDescription",
    "Alias" : [
        "positive",
        "pros"
    ]
}
```

Fig. 5 - Example of a MongoDB (JSON) document related to "comments".

In the hotel reviews, each channel has different forms to display the customer's opinions, and do it differently from each other. There are channels in which the comments are divided into "positive" and "negative", and there are others where it is just free text. The comments where there is no distinction between positive and negative are stored directly in DB2 without any treatment. Case the channels contains comments separate in positive and negative, it is necessary to refer to data dictionaries in order to distinguish between a positive and a negative review. This is necessary due to the different designations that each channel gives to the positive and negative comments. For example, in *Booking.com* the terms are "positive" and "negative", while in *Expedia* are "pros" and "cons" (see Fig. 6).

Fig. 6 - Example of review comments (*Booking.com* appears above and *Expedia* appears below).

3.4 Correspondence between extracted hotels in different channels

As already mentioned, the web robots are concerned to extract data from web, but not to process/analyse it. The same hotel extracted from different sources is not identified as being the same hotel, i.e., the IDs of the documents of the databases are different.

In other words, when a web robot extracts the hotels, there is no warranty that the Hotel *X*, e.g., in *Booking.com* is the same as the Hotel *X* on *Expedia*. This happens because the names of the hotels can change slightly according to the channel. What happens is that the same hotel will get two different ID, depending on the channel from where its data was extracted. This is a serious problem, because it is essential to make a correct consolidation of the

hotel data, for the other existing tables that depend on the table of the hotels.

To solve this problem was developed an algorithm that uses the name, address and hotel's GPS coordinates to match between hotels extracted from different channels. A function (*MatchStrings*) was developed to help the algorithm to verify the degree of similarity between two strings. This function detects the number of words in the same strings and, also the ratio (range 0 to 1) versus the number of equal words in each string. The ratio equation is given by $R = 0.5 \times EQW \times (L1 + L2) / (L1 \times L2)$, where *EQW* is the number of equal words between the strings, and *L1* and *L2* are the number of words in each string, *string1* and *string2* respectively.

The hotel matching algorithm has the purpose of finding the correspondence between hotels in different channels and works as follows (see Fig. 7):

(i) Search for hotels that have the same name and address of Hotel *X*. If true ($R > 0.99$), the hotel ID is found, go to vi). If the search returns empty then go to step ii).

(ii) Search hotels that have the same GPS coordinates of Hotel *X* (equal to the second decimal place). Only checks to the second decimal place between channels because the coordinates range slightly changes from the third decimal place, both latitude and longitude.

(iii) Use the *MatchStrings* function to check the degree of similarity between the Hotel *X*, name and the names of each of the hotels resulting from the previous step, returning only hotels with a degree of similarity from a given threshold (e.g., $R > 0.5$).

(iv) The same as step (iii) but now with the addresses of the hotels.

(v) If the above steps do not result a match, then it is considered that the hotel was not yet acquired from any other channel. If there are hotels that match the search, then all of those hotels receive the same ID. Exceptions can still occurs, for instance two hotels in the same street one side with the other and with a very similar name. In those cases, both hotels ID will be the same. In those exceptions, it is impossible to know with absolute sure that the hotel is the same, or if they are two different hotels.

(vi) Hotel ID defined.

3.5 Channels priority

Once the data is extracted from different channels, it happens that there are sometimes redundant data. For example, after the application of the

MatchStrings algorithm (explained in the previous section), it is possible to obtain several distinct names for the same hotel.
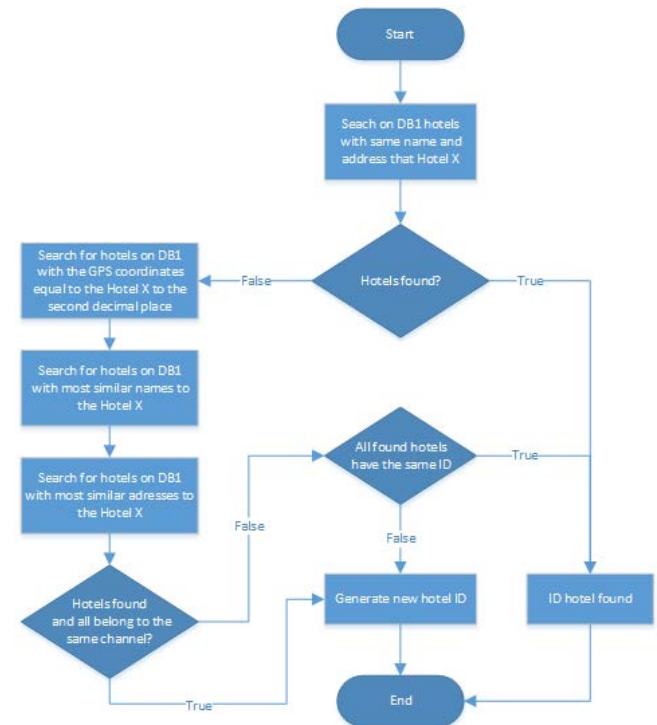

Fig. 7 - Flowchart of the algorithm to find hotel match in different channels. The "Hotel X" is the unknown hotel that we wish to find a correspondence.

In this case, the administrator will decide the name to select and stored in DB2. For this reason, it is necessary to define priorities to the channels, to withdraw and to store the information in DB2. In other words, if a hotel exists in four different channels and the priority to the channel is set, in descending order, as *channel 1 → channel 2 → channel 3 → channel 4*, then the priority channel should be *channel 1* and so the data stored will be from that channel. Another example in which these priorities must be defined is the number of stars of the hotels. Usually there is no consensus on the number of hotel's stars from channel to channel, i.e., an hotel that has four stars on a channel can have three stars in another channel. Then, once again, the rule of priority to decide which channel is more reliable is applied.

This mechanism will only be applied in cases where there are evidence of ambiguities. In the case of the comments, for example, this is not necessary since every review is a different opinion and all comments must be stored, even if there are two identical comments, what matters is that they are two different people.

3.6 Routing rules / data flow

After the extracted data from DB1 were transformed by all conversion rules, dictionaries, correspondence and channel priorities it is necessary route them to the right place. Once there are some tables in DB2, which have relations of many-to-many and for these cases, it is necessary to comply with a data storage order.

For instance, consider the three tables: *Hotel*, *Amenities* and *HotelAmenities*, where *Hotel* and *Amenities* are tables, which are related to each other by a many-to-many relation, supported on the *HotelAmenities* table. To store the data in these tables, first, the data are stored in *Hotel* and *Amenities* to create the corresponding IDs, and after that, will be stored the data in *HotelAmenities* table (see Fig. 8).
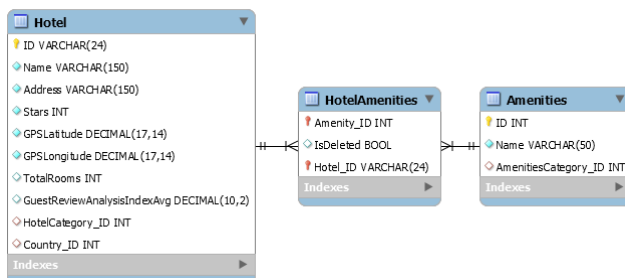

Fig. 8 - Schemes of the relationships between *Hotel, HotelAmenities* and *Amenities* tables.

3.7 DB2 Information storage

Finally, the data will be stored in DB2. When storing the data in DB2, the program searches in DB1 the origin of such data and updates the consolidation date. Thus, it is possible to know later the date and time when it was performed the consolidation for the data. Once stored in DB2, the data can be used directly for the final application in the BI system.

4 Discussion

The requirements of the hospitality BI system are very specific, with well-defined needs, and have to be developed to integrate analytical tools applied to historical data. The data includes the internal and external information that is pertinent to the hotel. The objective is to provide the decision makers with timely relevant data and a shared vision of the future and knowledge that encompasses the decision makers' resolution and create intelligence, providing a BI system to the organization [15].

In a hospitality BI system, which includes data from several web sources, with different formats and structure, it is essential to consider the development of a BDW. In this kind of Data Warehouse, the data integration phase starts when the web crawler collects information that is presented in relevant websites, related with the hotel business, and store it in collections of different kind of information (in a NoSQL database).

The NoSQL database, also called in our case as primary database, is constituted by collections of data that are stored in an unstructured format and aren't consolidate, which represent a problem to implement a BDW. Therefore, it is necessary to clean and transform the data, and after that to upload the consolidated extracted data into a database, also called a secondary database. This secondary database permits the development and implementation of analytical tools, which includes OLAP and Data Mining, to elaborate enterprise reporting which supports the hotel decision maker activities.

The process of data transform and consolidate the data in an adequate structure and format in the secondary database is constituted by several tasks: i) reading information from DB1, ii) data conversion rules; iii) data dictionaries, iv) correspondence between extracted hotels in different channels, v) channels priority, vi) routing rules/data flow, and vii) storage information on DB2.

The consolidation of information extracted from the web is a task that needs supervision from time to time. Each time a channel, add a new word that does not appear in a data dictionary it is necessary to add it to the data dictionary. It is expected that over time the data dictionary start becoming increasingly complete and thus the number of new words that may arise will decrease. Consequently, the consolidation system will become more stable over time.

References:
[1] K-H. Lee and S. S. Hyun, A model of behavioral intentions to follow online travel advice based on social and emotional loneliness scales in the context of online travel communities: The moderating role of emotional expressivity, *Tourism Management*, Vol. 48, 2015,pp. 426-438.

[2] M. Šerić, I. Gil-Saura and A. Mollá-Descals, Loyalty in HighQuality Hotels of Croatia: From Marketing Initiatives to Customer Brand Loyalty Creation, *Journal of Relationship Marketing*, Vol. 12, No. 2, 2013, pp. 114-140.

[3] M. Sigala, Exploiting Web 2.0 for New Service Development: Findings and Implications from the Greek Tourism Industry, *International Journal of Tourism Research*, Vol. 14, 2012, pp. 551-566.

[4] C. K. Anderson, S. Kimes and B. Carroll, Teaching Revenue Management at the Cornell University School of Hotel Administration, *INFORMS Transactions on Education*, Vol. 9, No. 3, 2009, pp. 109-116.

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

[6] C.M.Q. Ramos, M.B. Correia, J.M.F. Rodrigues, D. Martins, F. Serra, Big Data Warehouse Framework for Smart Revenue Management, *3rd NAUN International Conference on Management, Marketing, Tourism, Retail, Finance and Computer Applications* (MATREFC '15), Tenerife, Canary Islands, Spain, 10-12 Jan, 2015, pp. 13-22.

[7] D. Martins, R. Lam, J.M.F. Rodrigues, P.J.S. Cardoso, F. Serra, A Web Crawler Framework for Revenue Management, *14th International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases* (AIKED '15), Tenerife, Canary Islands, Spain, 2015, pp. 88-97.

[8] H. Oppewal, T. Huybers and G. I. Crouch, Tourist destination and experience choice: A choice experimental analysis of decision sequence effects, *Tourism Management*, Vol. 48, 2015, pp. 467-476.

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

[10] M. Mohanty, M. Jagadeesh, and H. Srivatsa, Big Data Imperatives: Enterprise Big Data Warehouse, *'BI'Implementations and Analytics,* Apress, 2013.

[11] M. Schermann, H. Krcmar, H. Hemsen, V. Markl, C. Buchmüller, T. Bitter, T. Hoeren, Big Data - An Interdisciplinary Opportunity for Information Systems Research, *Business & Information Systems Engineering*, Vol. 6, No. 5, 2014, pp. 261-266.

[12] D. Kala and S.C. Bagri, Balanced Scorecard Usage and Performance of Hotels: A Study from the Tourist State of Uttarakhand, India, *Asia-Pacific Journal of Innovation in Hospitality and Tourism* (APJIHT), Vol. 3, No. 2, 2014, pp. 153-173.

[13] L. Cvelbar and L. Dwyer, An importance–performance analysis of sustainability factors for long-term strategy planning in Slovenian hotels, *Journal of Sustainable Tourism*, Vol. 21, No. 3, 2013, pp. 487-504.

[14] D. Parmenter, *Key performance indicators (KPI): developing, implementing, and using winning KPIs*. John Wiley & Sons, 2010.

[15] M. Santos and I. Ramos, *Business Intelligence*, 2ed. FCA Editora: Lisboa, 2009.

[16] M. Fuchs, W. Höpken and M. Lexhagen, Big data analytics for knowledge generation in tourism destinations – A case from Sweden, *Journal of Destination Marketing & Management*, Vol. 3, 2014, pp. 198-209.

[17] W. Höpken, M. Fuchs, G. Höll, D. Keil and M. Lexhagen, Multi-dimensional data modeling for a tourism destination data warehouse. In: L. Cantoni, & Ph. Xiang (Eds.), *Information and communication technologies in tourism*, New York: Springer, 2013, pp. 157-169.

[18] G. Hill, *A guide to enterprise reporting*, http://ghill.customer.netspace.net.au/reporting/ definition.html, acessed 31/03/2015.

[19] M. Varga, K. and Curko, K., Some aspects of information systems integration, MIPRO, *35th International Convention*, 2012, pp.1583-1588.

[20] Z. Xiang, Z Schwartz, J.H. Gerdes Jr., M. Uysal, What can big data and text analytics tell us about hotel guest experience and satisfaction?, *International Journal of Hospitality Management*, Vol. 44, 2015, pp. 120-130.

[21] Inc. MongoDB. *Mongodb*. http://www.mongodb.com/, 2014, accessed 28/10/2014.