The Integration of Web-Based Information and the Structured Data in Data Warehousing

Jacek Maślankowski

Department of Business Informatics, University of Gdańsk, Poland
jacek@ug.edu.pl

Abstract. The article presents the concept of the solution for feeding the data warehouse from website forums including opinions about selected products. The key of the solution is to add a new data warehouse dimension called Variable that allows identifying both structured and unstructured data. In suggested solution the results of websites analysis will be stored in the same repository as the data from traditional corporate systems, such as CRM or ERP. The concept was presented regarding Internet shops that offered a selected kind of products.

Keywords: data warehouse, big data, unstructured data, data analysis.

1 Introduction

In recent years it has been observed that the value of unstructured data is increasing. This is the result of the fact that a large amount of valuable information is available on the Internet. For instance consumer preferences can be checked by looking deeper into Internet forums. To search through the websites to seek for consumers' behaviours and habits, Big Data technology can be used. It involves semantic algorithms to search through websites trying to find matching text. Those websites may include Twitter and Facebook as valuable sources for retrieving users' opinions on various topics [Bartram, 2013, pp. 28]. But the data just from the website can not be enough for the company to make the right decision.

The goal is to integrate Big Data approach with the existing databases in the company. Nowadays the typical analytical database used in companies is the data warehouse [Reddy et al., 2010, pp. 1186]. The data warehouse with OLAP tools has been treated as the most significant technology in business data processing since it was developed [Thomas & Datta, 2001, pp. 83].

The thesis of the article is to show that the data can be more valuable when it is loaded from structured and unstructured sources. The aim is to propose a concept of the framework to feed data warehouse directly by using Big Data tools. The way the goal has been formulated indicates that the data warehouse should be a repository to store the analytical data from several sources, including structured and unstructured information. The suggested data warehouse environment should be used in businesses selling different kind of products, such as Internet shops.

S. Wrycza (Ed.): EuroSymposium 2013, LNBIP 161, pp. 66–75, 2013.

The article was divided into six parts. The first part is introduction. In the second part there is a theoretical background of the possibilities of unstructured data analysis. The third part shows proposal for the alternative method to gather the result of analysis based on unstructured data sources in the data warehouse. In the fourth part there is a short comparison between proposed solution and acknowledged methods of unstructured data integration. The last part shows conclusions.

2 Possibility of Unstructured Information Analysis

2.1 Using Big Data to Filter Websites

Decisions made in organizations are usually based on the Data-Driven Decision-Making Process in which the decision process is based on information systems as well as internal and external factors. The result is the course of action [Picciano, 2012, pp. 12]. Those external factors are not only derived as structured data, but also as unstructured information.

Although there is no consensus on the Big Data term, there are several definitions that show the idea of the Big Data technology. Big Data is a generic term that assumes that the information or database system used as the main storage facility is capable of storing large quantities of data longitudinally and down to very specific transactions [Picciano, 2012, pp. 12]. Big Data technology is necessary when the data are too big for traditional systems to handle it [Gobble, 2013, pp. 64]. From the business point of view, Big Data is usually defined as three V's: volume, velocity and variety [Chen et al., 2012, pp. 1182], but sometimes it is defined together with the four V – the fourth one is cited as veracity [Harris, 2013, pp. 29].

But one of the goals of the Big Data is to support an analysis of large amount of unstructured data. More than 80% of all potentially useful business information is unstructured data, including e-mails, social media, videos, images, sensor readings, console logs and others [Das & Kumar, 2013, pp. 153].

Useful information on consumer habits and preferences can be obtained by collecting and analysing the information about product returns, warranties and customer complaints. But this information should be used to potential [Bughin et al., 2011, pp. 104-105]. In literature we can find several concepts of frameworks to analyse social networks, such as blogs. In early stage of analysing the blogosphere, several techniques like data and text mining were used to extract information from blogs by analysing its content [Chau & Xu, 2012, pp. 1190]. Data and text mining used together are well known as the duo mining technology [Maślankowski, 2006, pp. 973].

On the other hand it is important to study the media effects in the electronic media environment, such as Internet sphere with several techniques, using for example social networks analysis [Guo, 2012, pp. 617]. Social network analysis and cluster analysis are wide used to analyse keywords in text documents [Khansa et al., 2012, pp. 20].

The most referred way to make an analysis using Big Data technology is developing the solution on the Apache Hadoop platform [Teplow, 2013, pp. 38], [Harris, 2013, pp. 29]. One of the results of the importance of the Big Data technology is increasing number of Internet users. Based on the Eurostat data, the number of household having access to the Internet in EU27 countries has increased from 41% in 2005 to 76% in 2012 [Eurostat, 2013]. It can be presumed that this increase will correspond to the number of users using Web 2.0 tools.

2.2 Unstructured Data and Data Warehousing

Integration of unstructured and structured data in data warehousing was one of the issues taken into consideration during the data warehouse evolution. Well known proposal of the unstructured data inclusion in the data warehouse was suggested by W.H. Inmon in the fourth edition of the book Building the Data Warehouse. His approach was to match all the information from unstructured documents, such as e-mails, text files, spreadsheets and similar. The matching was mostly based on probabilistic match as well as themed match, to find a relationship between structured and unstructured documents. Two basic approaches were presented – first was to pull data over into the structured environment. The second was to create the two-tiered data warehouse – one tier for unstructured and another for structured data. In the same elaboration he also wrote about Granularity Manager that is used to collect the data from the web logs, mostly clickstream data. In this approach the extraneous data must be created in the single record, then incorrect data edited, after that the data must be converted, summarized and aggregated [Inmon, 2005, pp. 290-291, 305, 311-313, 320]. Therefore the main goal is to provide the mechanism to create a single record that will be processed by Granularity Manager, which is also responsible for passing the refined data into the data warehouse. Unstructured textual data were placed in the data warehouse repository in the approach called Data Warehouse 2.0 [Inmon et al., 2008, pp. 34-35].

However, the solution for filtering web data to feed the data warehouse had been published before the DW 2.0 concept was developed. The first important step into feeding the warehouse from web data was made by defining the web farming Data Warehouse. In this type of the warehouse business-relevant Web content is the input to the data warehouse [Hackathorn, 1997, pp. 43]. This approach was used mostly to collect clickstream logs in the whole user interaction process [Hu & Zhong, 2008, pp. 296-297]. Another approach was named "enhanced Data Warehouse" and the goal of this type of the warehouse was to use the Reader to filter business information from the Web to the data warehouse [Abramowicz et al., 2000, pp. 5, 105, 121-122]. Nowadays we can find the term Hadoop Data Warehouse that is referred to the data warehousing related to Big Data [McKenna, 2013, pp. 9-10].

3 The Concept of the Warehousing Environment

3.1 The Architecture of Data Warehousing

This chapter shows proposal of the schema of the infrastructure used to search through websites and put it into the data warehouse.

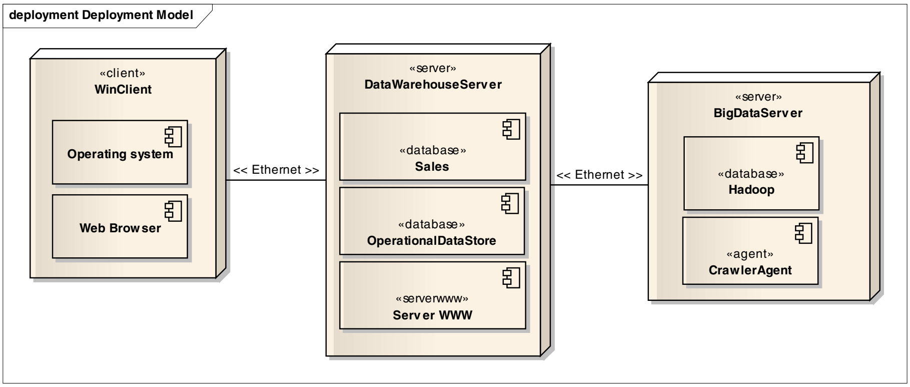

Fig. 1. The data warehouse environment with components to filter unstructured data (Source: own elaboration)

The proposed data warehouse environment is shown in figure 1. The concept of using component and deployment diagrams of UML framework to present the architecture of data warehousing is based on [Lujan-Mora & Trujillo, 2004, pp. 51-53].

The suggested data warehouse environment consists of the following subsystems that were presented as components in figure 1:

- Sales – a data warehouse which stores a data in a star schema as shown in figure 2.
- OperationalDataStore – a staging area in which the unstructured and structured data are gathered.
- Hadoop – tools used to support unstructured data analysis, by parallel processing large amount of data.
- CrawlerAgent – the agent used to gather information from websites.
- Server WWW – server used to host the result of analysis to clients.
- Operating system and Web Browser are typical components used to access the result of analysis hosted by Server WWW.

3.2 The Structure of the Data Warehouse

To store the results of unstructured data processing in the data warehouse, following rules concerning the data warehouse schema must be applied:

- there should be a dimension called Variable that will be able to indicate and describe the keywords that were analysed,
- other dimensions must be the same as in typical data warehouse.

The way the UML class diagram was used to present the star schema in the figure 2 is based on the Data Warehouse Conceptual Schema (DWCS) presented in [Lujan-Mora et al., 2004, pp. 193-194].

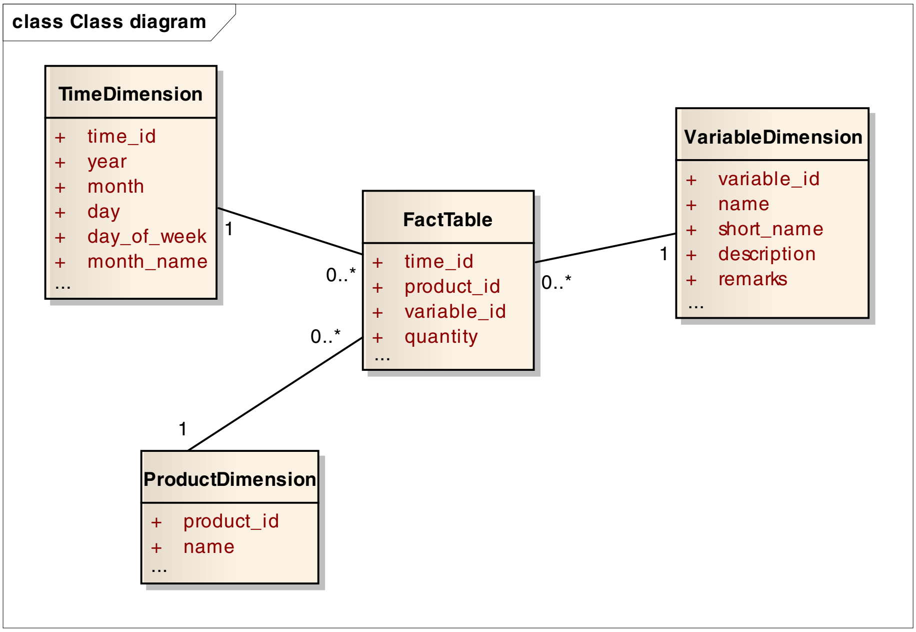

Fig. 2. The data warehouse structure (Source: own elaboration)

As mentioned above, the key of the data warehouse structure is the dimension called "Variable". This dimension is used to describe both structured and unstructured attributes that are used to define the data stored in the warehouse. Using the warehouse terminology, the "Variable" dimension can be treated as business metadata.

The dimension has the following structure, as shown in table 1.

Examples of the data stored in the Variable dimension are presented in table 2.

Table 1. The structure of the "Variable" dimension

Name of the attribute	Data type	Remarks
Variable_ID	Number(5)	ID number – created by using sequences
Name	Varchar2(100)	Used to name the variables
Short_Name	Varchar2(25)	Short names are used in BI tools to be included in header rows or columns
Description	Varchar2(500)	More detailed information about the variable
Remarks	Varchar2(500)	Used to explain a methodology of data gathering

Table 2. The data stored in the "Variable" dimension

Variable_ID	Name	Short_Name	Description	Remarks
1	Number of product sold	Sales	The total number of the particular product sold in the selected period of time.	Structured data source
2	Positive opinion on the product	Positive opinion	The variable is based on the opinion from selected websites forums.	The opinion is based on keywords matching.
3	Negative opinion on the product	Negative opinion	The variable is based on the opinion from selected websites forums.	The opinion is based on keywords matching.
…				

As shown in the table 2, the Variable dimension is used as the metadata to describe the data stored in the fact table of the data warehouse. This concerns both structured and unstructured data.

3.3 Selecting Sources of Unstructured Data

Implementing this solution is just the half of a success, as it is necessary to perform several steps that will ensure company that unstructured data retrieved from selected websites is reliable and can be used as the source system.

The key issue that must be regarded is a proper selection of unstructured data sources. In this article the focus is on assessing the products offered by a particular Internet shops. In this example there are several potential sources to feed the warehouse, which can be mostly adapted to almost every type of products offered by Internet shops:

- Internet shops forums that sell the same products as offered by the shop for which the data warehouse is implemented,
- subject-oriented Internet forums concerning the group of products offered,
- forums on thematic websites, such as eopinions.com,
- website of the shop selling the products,
- e-mails sent to the shop.

It is obvious that some Internet forums on shops websites are filtered and sometimes the opinion published on the website is limited only to the good one. Therefore the suggestion is to create a rank of the websites that can be included in implemented system. The list was prepared to select the right source, as shown in the table 3.

Table 3. The list of potential data sources

No.	Name of the source	Remarks	Reliability
1	eopinions.com	A lot of opinions about miscellaneous products.	high
2	amazon.com	Mostly about products offered by Amazon.	high
3	consumersearch.com	Lots of professional reviews.	high
…			

Based on the list of potential data sources, the proper as well as the most reliable data sources must be selected to be included in the data warehouse.

3.4 Matching the Keywords

The matching keyword algorithm used to filter Internet forums to find useful opinions about offered products is usually limited to one language. The effect of the globalization is that we can search through the Internet and find some useful opinions about products in several thematic websites from different countries. It means that the algorithms of matching keywords should be extended to be able to compare the keywords from different languages in one variable, such as pattern:

is good – ist gut – es bueno – est bon – é bom – är bra – on hyvä

must be equivalent to each other and should increase the value of the Variable "positive opinion". On the other hand the list of the following keywords means the same:

not so expensive – not expensive – no expensive – cheap – rather cheap

But in some cases there are some patterns that use the same words but are different:

so expensive – not so expensive

The last key issue that should be regarded before making algorithms to filter the websites is that the sentences published on the Internet forums include lots of mistakes and the following patterns should be regarded as the same:

expensive – expensve – expensiev (intentionally written with mistakes)

On the other hand there are some mistakes such as misspellings that can be ambiguously interpreted:

on expensive – no expensive

Therefore the solution proposed in this article will never be fully reliable. But looking deeper into psychological issues, Internet forums are full of subjective opinions from users, sometimes just because they are not satisfied from their shopping, not the product they bought.

In this article there is no recommendation on semantic algorithm used to matching the keywords included in different opinions, as it is not the goal of this paper.

4 The Comparison between Proposed Solution and Existing Ones

The goal of the article was to show a concept solution to include information from Internet forums in the data warehouse. The biggest advantage of the suggested solution is that it doesn't affect the warehouse environment that is used in most of the companies.

In suggested solution, a traditional data warehouse used in a company can be enhanced by adding new dimension called Variable. This dimension will make the data warehouse more flexible to apply to changes in the environment, such as describing new phenomenon that will occur in the future. This phenomenon could be a new voting mechanism for products or any other social forum about products.

The proposed solution uses one repository for structured and unstructured data, and unstructured data are only available as a result of analysis included in the structured star schema. Looking into other proposals, such as DW 2.0 proposed by W.H. Inmon, there are two different repositories, separately for structured and unstructured data (compare the figure 2.2 in [Inmon et al., 2008, pp. 28]). However the DW 2.0 is much more sophisticated than proposed architecture in this article because it assumes including almost any type of unstructured document, while the solution described in this article focused on websites only.

Other proposals of integrating web feed opinions with the data warehouse assumes to use several fact tables and dimensions as well [Moya et al, 2011, pp. 22-23], which differentiate from the solution proposed in this paper that includes only one fact table and several dimension tables, depending on the need of including them in the warehouse schema.

The suggestion of implementing the solution is to develop it with Oracle database, as it offers Big Data connectors to Apache Hadoop, which can be used for parallel processing the websites. The data warehouse implemented with guidelines suggested in this article leads to the conclusion that the following information, presenting in table 4, can be received from the warehouse.

Table 4. Example results of the analysis

Product_name	Month	Year	Positive opinion	Negative opinion	Sales
A	January	2013	23	2	110
A	February	2013	42	3	159
A	March	2013	73	4	169
B	January	2013	34	5	210
B	February	2013	55	6	204
B	March	2013	62	4	190

Based on the table above which presents a snapshot of the data report from the suggested warehouse repository, we can easily compare the sale amount and the number of positive and negative opinions written on Internet forums. This is the value added to the data warehouse. Efficient use of the Big Data technology and unstructured data filtering will allow company to build their market strategy supporting by the list presented above.

5 Conclusions

The goal of the article was to show an alternative way to design and implement a solution to integrate both structured and unstructured data. The main advantage of suggested solution to integrate structured and unstructured data together is that it is rather simple and easy to implement.

As it was written in this article, unstructured data are one of the most relevant sources of information in business today. Rapid increase in the number of users on the Internet has a big impact on the way how company is perceived by society. In that sense, Internet is one of the sources for lots of people trying to get an opinion about particular products. This lead to the conclusion that companies today cannot lose their chances to rapidly react to changes in the market, based on opinions of different Internet forums.

To have a strong and fast reaction to changes on the market, it is necessary to implement a solution that will be able to gather and process users' opinions about different products. These opinions are freely available on various websites. The goal is to ensure that the data warehouse will provide rapidly and on time information to make changes in the rules of products offered. It means that the information must be processed continuously and automatically, what is one of the aims of the Big Data technology.

Although there is variety of information systems that allows filtering data from websites, the availability of Big Data technology, such as Apache Hadoop, will make it easier to implement and use real-time web filtering system.

Abbreviations

CRM – Customer Relationship Management
DW – Data Warehouse
EU – European Union
ERP – Enterprise Resource Planning
UML – Unified Modelling Language

References

1. Abramowicz, W., Kalczyński, P., Węcel, K.: Filtering the Web to Feed Data Warehouses, pp. 5, 105, 121–122. Springer (2002)
2. Bartram, P.: The value of data. In: Financial Management, p. 28 (March 2013)
3. Bughin, J., Livingston, J., Marwaha, S.: Seizing the potential of 'big data'. McKinsey Quarterly (4), 104–105 (2011)
4. Chau, M., Xu, J.: Business Intelligence in Blogs: Understanding Consumer Interactions and Communities. MIS Quarterly 36(4), 1190 (2012)
5. Chen, H., Chaing, R.H.L., Storey, V.C.: Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly 36(4), 1182 (2012)
6. Das, T., Kumar, P.M.: BIG Data Analytics: A Framework for Unstructured Data Analysis. International Journal of Engineering Science & Technology 5(1), 153 (2013)
7. Eurostat database, http://epp.eurostat.ec.europa.eu/

8. Gobble, M.M.: Big Data: The Next Big Thing in Innovation. Research Technology Management 56(1), 64 (2013)
9. Guo, L.: The Application of Social Network Analysis in Agenda Setting Research: A Methodological Exploration. Journal of Broadcasting & Electronic Media 56(4), 617 (2012)
10. Harris, C.: Dividing into Big Data. Canadian Underwriter 80(2), 29 (2013)
11. Hackathorn, R.: Farming the Web. Byte.com 22(10), 43 (1997)
12. Hu, J., Zhong, N.: Web farming with clickstream. International Journal of Information Technology & Decision Making 7(2), 296–297 (2008)
13. Inmon, W.H., Strauss, D., Neushloss, G.: DW 2.0. The Architecture for the Next Generation of Data Warehousing, pp. 28, 34–35. Elsevier Inc. (2008)
14. Inmon, W.H.: Building the Data Warehouse, 4th edn., pp. 290–291, 305, 311–313, 320. Wiley Publishing, Inc. (2005)
15. Khansa, L., Zobel, C., Goicochea, G.: Creating a Taxonomy for Mobile Commerce Innovations Using Social Network and Cluster Analyses. International Journal of Electronic Commerce 16(4), 20 (Summer 2012)
16. McKenna, B.: King.com gaming site unlocks big data to switch to Hadoop database. Computer Weekly, 9–10 (May 3, 2013)
17. Maślankowski, J.: Integration of Text- and Data-Mining Technologies for Use in Banking Applications. In: Nillson, et al. (eds.) Advances in Information Systems Development, p. 973. Springer Science, New York (2006)
18. Moya, L.G., Kudama, S., Aramburu Cabo, M.J., Berlanga Llavori, R.: Integrating web feed opinions into a corporate data warehouse. In: Proceedings of the 2nd International Workshop on Business Intelligence and the WEB. ACM, New York (2011)
19. Lujan-Mora, S., Trujillo, J.: Physical Modeling of Data Warehouses using UML. In: Proceedings of the 7th ACM International Workshop on Data Warehousing and OLAP, pp. 51–53 (2004)
20. Luján-Mora, S., Vassiliadis, P., Trujillo, J.: Data Mapping Diagrams for Data Warehouse Design with UML. In: Atzeni, P., Chu, W., Lu, H., Zhou, S., Ling, T.-W. (eds.) ER 2004. LNCS, vol. 3288, pp. 191–204. Springer, Heidelberg (2004)
21. Picciano, A.G.: The Evolution of Big Data and Learning Analytics in American Higher Education. Journal of Asynchronous Learning Networks 16(3), 12 (2012)
22. Reddy, G., Srinivasu, R., Rao, M., Rikkula, S.: Data Warehousing, Data Mining, OLAP and OLTP Technologies are Essential Elements to Support Decision-Making Process in Industries. International Journal on Computer Science & Engineering 2(9), 2866 (2010)
23. Teplow, D.: The emperor has no clothes. Business Intelligence Journal 18(1), 38 (2013)
24. Thomas, H., Datta, A.: A Conceptual Model and Algebra for On-Line Analytical Processing in Decision Support Databases. Information Systems Research 12(1), 83 (2001)