On Data Integration and Data Mining for Developing Business Intelligence

Ping-Tsai Chung, Dept. of Computer Science, Long Island University, Brooklyn, NY, *SeniorMember, IEEE*
Sarah H. Chung, American Express Corporation & St. John's University, NY, *Member IEEE*

***Abstract* —** *Business Intelligence (BI)* allows a corporation's executives to acquire a better understanding of their customers, the market, supply and resources, and competitors in order to make effective strategic decisions. BI technologies provide historical, current and predictive views of business operations such as reporting, online analytical processing, business performance management, competitive intelligence, benchmarking, and predictive analytics. Web Services technologies responded quickly to help such evolution and in many situations the Web Services application is driving businesses and dictating a new way of doing business. Web information usually contains multimedia data with unstructured fashions. Through the effective analysis of company's Web information, we could make effective market analysis, compare customer feedback on similar products, discover the strengths and weaknesses of their competitors, retain highly valuable customers, and make smart business decisions.

In this paper, we discuss two case studies on data integration and data mining. The first case is for the traditional data analytics using relational database techniques such as *Oracle* database and *Cognos* BI tool for integrating and mining a company's web site. The second case is for multimedia data analytics using *Monago* database and *Pentaho* BI tool for integrating and mining multimedia data presented in a company's web site. We compare both cases in aspects of *Data Integration*, *Metadata*, *Query Performance* and *Data Analytics*. Finally, we present experimental results for using the above data mining techniques and tools to better understand features of each customer group and develop customized customer reward programs.

Keywords: Data integration, data mining, web server, database, and business intelligence.

I. INTRODUCTION

In the era of globalization, new businesses led companies to be everywhere in the world to respond for the new needs of gaining new markets and enforce its existence in acquired ones in order to stay in living in extreme competitions that do not recognize countries boundaries. Business Intelligence (*BI*), allows a corporation's management executives to use data on customer purchasing patterns, demographics, and demand trends to make effective strategic decisions to help the company plan its business, lower its inventory levels, and maximize profitability. As companies expand their reach into the global marketplace, the need to analyze how customers use company websites to learn about products and their purchasing preferences, is becoming increasingly critical to survival and ultimate success. It is critical for businesses to acquire a better understanding of the commercial context of their organization, such as their customers, the market, supply and resources, and competitors. Business Intelligence (BI) technologies provide historical, current and predictive views of business operations. Examples include reporting, online analytical processing, business performance management, competitive intelligence, benchmarking, and predictive analytics [1]. Web Services technologies responded quickly to help such evolution and in many situations the Web Services application is driving businesses and dictating a new way of doing business. Web information usually contains multimedia data with unstructured fashions. Through the effective analysis of company's Web information, we could make effective market analysis, compare customer feedback on similar products, discover the strengths and weaknesses of their competitors, retain highly valuable customers, and make smart business decisions [6][7]. Along the changes of life style in the era of globalization, more and more our daily information are getting from internets and web searches. It forces commercial websites paying tremendous efforts to add information into their websites. Using a Food &Wine web site as an example (See Figure 1.), compared to traditional websites, it has added photos, videos and social networking communities , etc … Volume and variety of data are spread rapidly via facebook, Twitter, LinkedIn, Google+,…., etc.

Data Mining is the core of BI. Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories. A popular trend in the IT industry is to perform a preprocessing step: *data cleaning* and *data integration* (where it involves combining data residing in different sources and providing users with a unified view of these data.) before the data mining step, where the resulting data will be kept in a database system. In management circles, people frequently refer to data integration as "Enterprise Information Integration" (EII) [10] [24]. Applications of data mining include web page analysis: from web page classification, clustering to PageRank & HITS algorithms; Collaborative analysis & recommender systems; major dedicated data mining systems/tools (e.g., Oracle Data Mining Tools, Microsoft SQL-Server Analysis Manager, and SAS Business Analytics Software and Services); Mining Web Data (Web content, web structure, and

web usage mining); Pattern Discovery and Inductive databases; Basis of data mining: Discover patterns occurring in the database, such as associations, classification models, sequential patterns, etc. Data mining is the problem of performing inductive logic on databases. The task is to query the data and the theory (i.e., patterns) of the database [4][5][6][7][8].

There still exists a nontrivial gap between generic data mining methods and effective and scalable data mining tools for domain-specific applications [4]. For example, for Data Mining for the area of Retail and Telecommunication Industries, there are huge amounts of data on sales, customer shopping history, e-commerce, etc. Applications of retail data mining may involve the following tasks: (1) Identifying customer buying behaviors, (2) Discovering customer purchasing patterns and trends, (3) Improving the Quality of Service (QoS), (4) Achieving better customer retention and satisfaction, (5) Improving goods consumption ratios, (6) Suggest adjustments on the pricing and variety of goods, (7) Design and developing effective goods transportation and distribution policies. For the Telecommunication and many other industries: we could share many similar goals and expectations of retail data mining.

The outlines of this paper are organized as follows. In Section I, an introduction of this paper is provided. In Section II, we discuss two case studies. In the first case, we study on data integration and data mining for a Food & Wine web site, particularly, we focus on the travel related analytics, using relational database techniques such as *Oracle* database and *Cognos* BI tool. For the second case study, we develop a multimedia data analytics for the same web site information using *Mongo* database and *Pentaho* BI tool. In Section III, we compare both cases in aspects of *Data Integration*, *Metadata*, *Query Performance*, and *Data Analytics*. These comparative results are useful for us to better understand the data integration, and mining techniques to feature each customer group and to develop effective customer reward programs. In section IV, we discuss the future trends of data integration and analytics for business intelligence. In Section V, we make a conclusion of this paper and provide future research directions.

II. TWO CASE STUDIES

In this Section, we develop two case studies. For the first case, we study on data integration and data mining of the travel related analytics of a Food & Wine web site shown in Figure 1 by the traditional data analytics using *Oracle* Relational Database Management System (RDBMS) technique [14] and *Cognos* BI tool [15]. In the RDBMS technique, its modeling contains a set of relational database tables, each table has a unique primary key, and the relationships between tables are linked through a set of referential integrity constraints, where each referential integrity constraint enforced by a pair of primary key and foreign key relationship (i.e., A table's foreign key's value if it's not null, must be matched to an existing primary value in another table). In this traditional data processing technique, we usually use Structured Query Language (SQL) to query and analyze the database information through join tables & other relational database operations. The *Cognos* software is a BI and performance management (PM) tool applied to the Oracle or Microsoft SQL-Server databases for integrating and mining a company's web site [15]. For the second case study, we develop a multimedia data analytics for the same web site information using *Mongo* database and *Pentaho* BI tool. MongoDB (from "hu**mongo**us") is an open source document-oriented database system developed and supported by 10gen, the Mongo DB Company. It is the most popular NoSQL (i.e., "Not only SQL") database system. It provides a simple, lightweight mechanism for storage and retrieval of data that provides higher scalability and availability than traditional relational databases. The NoSQL

Figure 1. A sample web site for Food & Wine services.

data stores use looser consistency models to achieve horizontal scaling and higher availability. It does not use a SQL approach for storing data, i.e. in the form of tables [24]. Binaries are available for Windows, Linux, OS X, and Solaris [16]. Pentaho offers a suite of open source Business Intelligence (BI) products called Pentaho Business Analytics providing data integration, online Analytic processing (OLAP) services, reporting, dashboarding, data mining and Extraction, Transform, and Loading (ETL) capabilities [17].

III. COMPARATIVE STUDY OF TWO CASES

Now, we compare both cases in aspects of *Data Integration*, *Metadata*, *Query Performance*, and *Data Analytics*. These comparative results are useful for us to better understand the data integration, and mining techniques to feature each customer group and to develop effective customer reward programs. Our studies were based on a travel related website, of the Foods & Wines web site, which provides world travel guides, articles, vacation ideas, blogs, photo contests, …, etc and more. We understand that how the information collection was changed along with the rapid expansion of social networking community like Facebook (FB), Twitter, Google+, LinkedIn, …etc; and how the organizations overcome the difficulties to reach their new record high of 4 million unique clicks per month in the past January by switching their content management database from Oracle database to MongoDB.

In Table 1, we compare the schema mapping of objects of Oracle database to the corresponding objects in MongoDB. Based on Figure 2, the attributes of a table in relational data model are predefined and maintained by the system catalog of every oracle database. Information was stored in Oracle (in table), row by row and each row was followed same structured defined in system catalog. So, relational data was identified as *Structured Data*; the tabs of *each DOC (i.e., Document),* are *dynamic.* Information was stored in MongoDB, DOC by DOC. Each DOC allows defining different tabs. This is the major reason why DOC was identified as *Unstructured Data*.

Now along unstructured data has been increasing rapidly, organizations are facing a challenge of collecting unstructured data into a relational data store for analysis. This challenge includes either data loss during information collection process due to incompatible data format or requiring the engagement of huge resources for data transformation.

Oracle	MongoDB
Database	Database
Table	Collection
Attribute	Tag
Row	DOC

Table 1. Schema Mapping between relational database and documentary database.

(1) Data Integration –

The challenge of collection DOC (i.e., web page) data into a relational database: Since website has integrated with social network community, such as FB, Twitter, Google+, LinkedIn, …, etc, organizations also want to gather more details about their potential customers by looking into their favorite pages or their "Like" history. However, due to the limitation of relational data model, it requires concurrent efforts of modifying database schema design in order to collect information of FB or twitter's conversations and actions from page viewer. Figure 2 shows a Data Transformation Portal for Relational Data Model. On the other hand, if website using MongoDB as content management database, FB and Twitter's conversations and actions can directly upload into database. Documentary database didn't require schema verification.

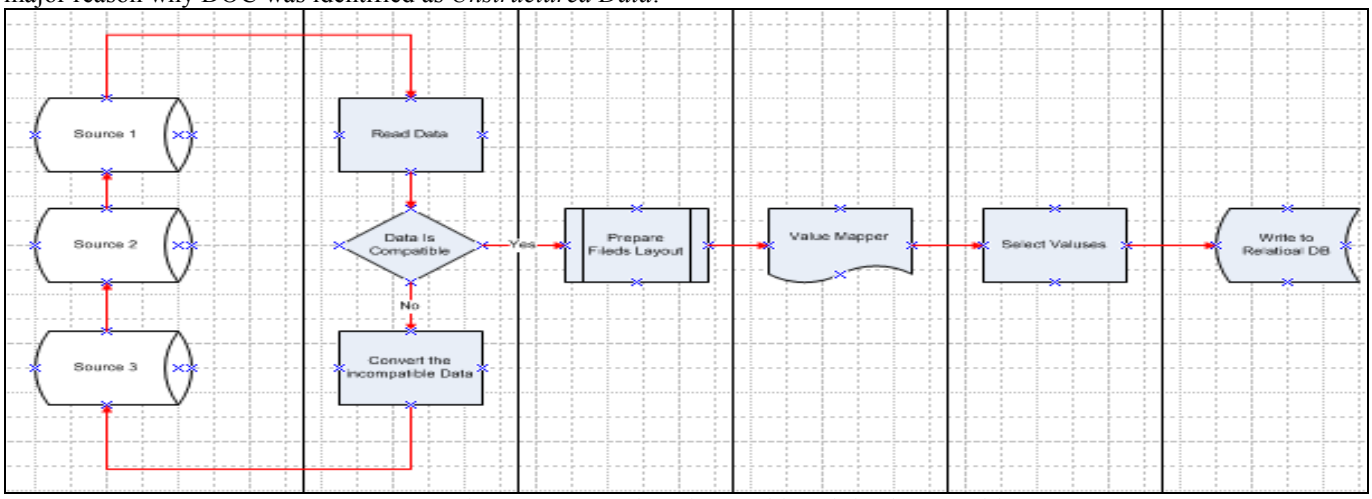

Figure 2: Data Transformation Portal for Relational Data Model.

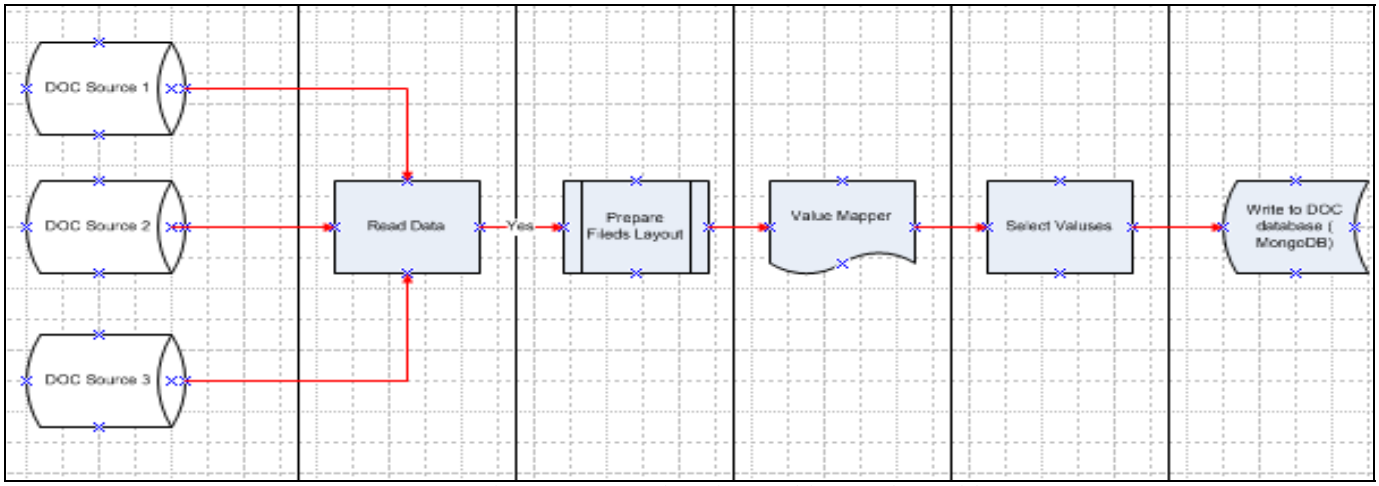

Figure 3: Data Transformation Portal for DOC data model.

Please refer to Figure 3, in which we could see that the benefits of using documentary database (MongoDB) as content management database can avoid data loss due to unmatched schema and can improve the lead time during data integration.

(2) Metadata –

NoSQL database systems are often highly optimized for retrieval and appending operations and often offer little functionality beyond record storage (e.g. key–value stores). The reduced run-time flexibility compared to full SQL systems (e.g. the Oracle or Microsoft SQL-Server databases) is compensated by marked gains in scalability and performance for certain data models [24]. There are 246 tables, 463 indexes and hundreds of constraints were implemented when used Oracle as content management database; However, It implements by 5 collections and 5 indexes when used MongoDB as content management database. In general, more objects require more resources for data maintenance and create complexity in SQL query.

(3) Query Performance –

We compare the following two queries for showing the complexities of data processing from Oracle and for MongoDB databases. Along with the increasing of the unstructured data in our daily life, the traditional RDBMS are really facing a lots of difficulties for a company used as content management database. Suppose we are looking for attractions of New York City, the following are the query statements from MongoDB and Oracle databases.

MongoDB Scripts:

```
db.weekend_getaways_cities.find({geo_id:436}).pretty()
```

Oracle Scripts:

```
SELECT DISTINCT [geos].name GeoName, geo_id, venues.id VenuesID, venues.name VenuesName, online_desc Description, 'New York City',is_out_of_business,show_on_tl
FROM [geos] INNER JOIN geo_relationships ON geos.id = geo_relationships.child_id join venues on venues.geo_id=geos.id
WHERE geos.id in ( SELECT DISTINCT geos.id
FROM [geos] INNER JOIN geo_relationships ON geos.id = geo_relationships.child_id
WHERE ([geo_relationships].parent_id =436)
UNION
SELECT DISTINCT geos.id
FROM [geos] INNER JOIN geo_relationships ON geos.id = geo_relationships.child_id
WHERE geos.id=436)
UNION
SELECT DISTINCT [geos].name GeoName, geo_id, venues.id VenuesID,venues.name VenuesName, online_desc Description, 'New York City',is_out_of_business,show_on_tl
FROM [geos] INNER JOIN geo_relationships ON geos.id = geo_relationships.child_id join venues ON venues.geo_id=geos.id
WHERE geos.id in ( SELECT DISTINCT geos.id
FROM [geos] INNER JOIN geo_relationships ON geos.id = geo_relationships.child_id
WHERE ([geo_relationships].parent_id =436)
UNION
SELECT DISTINCT geos.id
FROM [geos] INNER JOIN geo_relationships ON geos.id = geo_relationships.child_id
WHERE geos.id=43)
```

(4) Data Analysis –

For the following, we compare the Frequent Data Mining Objects in counting the number of Unique Entries (UEs) for a case study. Suppose that we have the following information:

(1) The Unique Entries of the travel related analytics of Food & Wine web site: **1,000**.
(2) Facebook Friends: **2,000**.
(3) The average of each "Friends of Friend": **50**.
(4) The average of each Friend has **20** Favorite Pages.
(5) Google+: **1, 500**.

Therefore, there are 1000+2000+1500 = **4,500** unique entries for the Oracle database system. On the other hand, there are 1000 + 2000 + 2000x50 + 2000x20 + 1500 = **144,500** unique entries for the MongoDB database system (See Table 2).

	Oracle	**MongoDB**
Unique Entry (UE)	4,500	144,500

Table 2. A comparison of number of unique entries of Oracle and MongoDB databases.

Note that if the back-end database is a relational database, we only can collect the unique name in FB or Google+; If the back-end database is an MognoDB, we can collect information by viewing pages (DOC), That is, we can find out not only pages (DOC), but also "Friends of Friend" in FB, favorite pages, likes, …, etc, which will help organizations to have better understanding of their potential customers. Using Oracle as back-end database for viewing pages, if the Next travel destination is Paris, we have the following the Average Page View Per Entry:

- Promotion,
- Paris City Pass,
- Eiffel Tower,
- Palace,
- Subway.

If we use the MongoDB, the possible viewed pages including friends bookmark like "favorite restaurants", "Shopping tips in Paris", "shows" ,"London", "Promotion for Paris and London" and "Buckingham Palace", …,etc. Therefore, the following table is a summary of the Average Page View (APV) of Oracle and MongoDB databases for our case studies.

	Oracle	**MongoDB**
Average Page View (APV)	5	11

Table 3. A comparison of Average Page View (APV) Per Entry of Oracle and MongoDB databases.

For the Data Analytics for Business Intelligence, we could identify Foods & Wine web page viewers' next vacation destinations, and project the consumer's trend by cross referencing websites' page viewers between restaurants and vacation destinations in order to design and promote shopping, lodging and eating package as revenues generator.

IV. FUTURE TRENDS OF DATA INTEGRATION AND ANALYTICS FOR BUSINESS INTELLIGENCE

In the era of *Big Data*, every day, we create 2.5 quintillion bytes of data. There is 90% of the data in the world today has been created in the last two years alone. This Big Data comes from everywhere: sensors used to gather different kinds of information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals, …, etc. In [21], it states that Big Data spans four *V*'s (i.e., four dimensions): *Volume*, *Velocity*, *Variety*, and *Veracity*.

(1) Volume: Enterprises are awash with ever-growing data of all types, easily amassing terabytes—even petabytes—of information.
(2) Velocity: Sometimes 2 minutes is too late. For time-sensitive processes such as catching fraud, big data must be used as it *streams* into your enterprise in order to maximize its value. The *Computing* models close to human are the weather forecast and stock exchange which have similar computing requirements, *huge amount of information* are continuously flowing; it is *time sensitive* and require *scalabilities* for the system to grow. To advance the concept of computing in such an environment, *Streams Computing* is developed as a new platform for computing.

(3) Variety: Big Data is any type of data - *Structured Data* such as relational database data and *Unstructured Data* such as text, sensor data, audio, video, click streams, log files and more. New insights are found when analyzing these data types together. For example, we monitor hundreds of live video feeds from surveillance cameras to target points of interest. Another important discovery would be that we exploit the 80% data growth in images, video and documents to improve customer satisfaction.

(4) Veracity: One in three business leaders don't trust the information they use to make decisions. How can you act upon information if you don't trust it? Establishing trust in Big Data presents a huge challenge as the variety and number of sources grows.

Big Data is more than simply a matter of size; it is an opportunity to find insights in new and emerging types of data and content, to make your business more agile, and to answer questions that were previously considered beyond your reach. Until now, there was no practical way to harvest this opportunity [21]. For the following, we list some *Trends of Data Mining*, especially for *Privacy-Preserving*

(privacy-enhanced or privacy-sensitive) mining [3][4][5][6] [7][8][10] below.

(1) *Integrating structured, semi-structured (tagged), unstructured data and Analytics into Business Intelligence (BI)*; Mining multimedia, text and web data; Integration of data mining with Web search engines, database systems, data warehouse systems and cloud computing systems; *Mining social and information networks*; Extract information from data in other modalities (speech, image, video), visual and audio data mining.
(2) *Privacy Protection and Information Security* in data mining.
(3) *Continuous Data Mining*, mining spatiotemporal, moving objects and cyber-physical systems; Distributed data mining and *Real-time Data Stream Mining*; Scalable and interactive data mining methods; Data mining with software engineering and system engineering.
(4) *Application exploration: dealing with application-specific problems.* For example, mining biological and biomedical data.

V. CONCLUSIONS AND FUTURE WORKS

In this paper, two case studies on data integration and data mining were presented. The first case is for the traditional data analytics using relational database techniques such as *Oracle* database and *Cognos* BI tool for integrating and mining a company's web site. The second case is for multimedia data analytics using *Monago* database and *Pentaho* BI tool for integrating and mining multimedia data presented for the travel related analytics of Food & Wine web site. We compared both cases in aspects of *Data Integration*, *Metadata*, *Query Performance* and *Data Analytics*. In these studies reveal that NoSQL database management systems are very useful when working with a huge quantity of data when the data's nature does not require a relational model. The data can be structured, but NoSQL is used when what really matters is the ability to store and retrieve great quantities of data, not the relationships between the elements. Usage examples might be to store millions of key–value pairs in one or a few associative arrays or to store millions of data records. By using this kind of organization is particularly useful for statistical or real-time analyses of growing lists of elements (such as Twitter posts or the Internet server logs from a large group of users), we could effectively develop data mining strategies and methods to better understand features of each customer group and develop customized customer reward programs.

In the era of Big data, it refers to the organizational utilization of data that exceeds the volume (amount), velocity (speed of flows), and variety of data typically stored using traditional structured database technologies. However, for the Veracity issue**,** there is about one in three business leaders don't trust the information they use to make decisions. Therefore, establishing trust in Big Data is a grand challenge as the variety and number of sources grows. For the future works, we will develop studies on Integrating structured, semi-structured (tagged), unstructured data and Analytics into Business Intelligence (BI) and research into the Privacy-Preserving and Information Security in our data mining projects.

REFERENCES

[1] P. Chung, S. Chung, and C. Hui, *"A Web Server Design Using Search Engine Optimization Techniques for Web Intelligence for Small Organizations",* IEEE LISAT 2012: Long Island Systems, Applications and Technology Conference, May 2012.
[2] G. Marakas, Modern Data Warehousing, Mining. And Visualization – Core Concepts, Prentice Hall, 2003.
[3] S. Suh, Practical Applications of Data Mining, Jones & Bartlett Learning, 2012.
[4] J. Han, M. Kamber, and J. Pei. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed. 2011.
[5] B. Liu, Web Data Mining, Springer 2006.
[6] R. Akerkar and P. Lingras, Building an Intelligent Web: Theory and Practice. Jones and Bartlett Publishers, 2008.
[7] Z. Markov and D. T. Larose, Data Miniing the Web: Uncovering Patterns in Web Content, Structure, and Usage. 2007.
[8] S. Chakrabarti. Mining the Web: Statistical Analysis of Hypertex and Semi-Structured Data. Morgan Kaufmann, 2002.
[9] D. Easley and J. Kleinberg. Networks, Crowds, and Markets: Reasoning about a Highly Connected World. Cambridge University Press, 2010.
[10] A. Doan, A. Halevy, and Z. Ives, Principles of Data Integration, Morgan Kaufmann, 2012.
[11] K. Banker, MongoDB in Action, Manning, 2012.
[12] W. Inmon, A. Nesavich, Tapping Into Unstructured Data – Integrating Unstructured Data and Textual Analytics Into Business Intelligence, Prentice Hall, 2008.
[13] T. Davenport, Enterprise Analytics – Optimize Performance, Process, and Decisions Through Big Data, Pearson Education Inc., FT Press, 2013.
[14] Oracle Database - http://www.oracle.com/.
[15] Cognos BI Tool - http://www-01.ibm.com/software/analytics/cognos/.
[16] Mongo Database - http://www.mongodb.org/.
[17] Pentaho BI Tool - http://www.pentaho.com/.
[18] "*Oracle Information Architecture: An Architect's Guide to Big Data"*, An Oracle White Paper in Enterprise Architecture, Oracle Corporation, August 2012.
[19] "*Oracle: Big Data for Enterprise"*, An Oracle White Paper, Oracle Corporation, January 2012.
[20] "*Big Data: The Next Frontier for Innovation, Competition, and Productivity"*, McKinsey Global Institute, June 2011.
[21] J. McKendrick, "*Big Data, Big Challenges, Big Opportunities: 2012*", IOUG Big Data Strategies Survey, Unisphere Research, a Division of Information Today, Inc., Sept. 2012.
[22] F. Halper, "*Four Vendor Views on Big Data and Big Data Analytics: IBM*", Hurwitz & Associates, January 2012.
[23] D. Stuttard and M. Pinto, The Web Application Hacker's Handbook – Finging and Exploiting security Flaws, Second Edition, Wiley, 2011.
[24] Wikipedia, the free encyclopedia. http://www.wikipedia.org/.