"HENRI COANDA"
AIR FORCE ACADEMY
ROMANIA

"GENERAL M.R. STEFANIK"
ARMED FORCES ACADEMY
SLOVAK REPUBLIC

INTERNATIONAL CONFERENCE of SCIENTIFIC PAPER
AFASES 2015
Brasov, 28-30 May 2015

USING BIG DATA FOR INTELLIGENT BUSINESSES

Cristian Bucur*

*University of Economic Studies, Bucharest & Petroleum and Gas University, Ploiesti, Romania

Abstract: *Big data changes the way organizations use their data infrastructure and data analytics software platforms. This article presents an overview of the current technologies involved and is presented the use cases of big data systems. Also is presented a general architecture and how data analytics system evolved from traditional data warehouses. Are discussed the main challenges a company is faced when implementing a big data system and how the new technology is perceived by organizations.*

Keywords: *big data, business intelligence, data warehouse, nosql.*
***MSC2010*:** *97P30, 97R10, 97R50.*

1. INTRODUCTION

We face today an extraordinary revolution of communications, which lead to a staggering quantity of data available. Apart from technical evolution that made possible to store and compute these data, an important role is played by improved statistical and computational methods that made possible to analyze and discover knowledge.

Big data is the term used to describe this massive volume of structured or unstructured data collected that is too large to be processed with traditional methods.

According to Gartner (http://www.gartner.com /doc/code/235055 ?ref=ddisp) the definition of big data is: a high volume, velocity and variety of information assets that necessitates innovative forms of processing in order to enhance process optimization or business insights and decision making. To the "three Vs" definition of analyst Doug Laney from Gartner, SAS added two more dimensions: variability and complexity [18]. Data flows can be inconsistent and vary in time, so it can be complicated to deal with peaks and also data has multiple sources, so necessitates be matching and transforming to acquire correlations.

2. USE CASES

A big data system is capable to generate complex processes and deeper business insights than existing data warehouse and business intelligence systems. These systems are not limited to providing support for decision making, could be also used for:

- Marketing and sales growth - One use is developing a recommendation engine for making purchasing suggestions to customers based on their interests compared with comportment of millions of other customers. Another is to optimize sale conversion process by tracking the actions of customer and getting insights on how can be improved. Also big data could help obtain customer segmentation for companies, helping them to make personalized offers and targeted campaigns.
- Risk and compliance management - insurance or credit companies could use a big data system to detect specific associations or

comportments associated with lower risks of default.

- Behavioral analytics - in sociology the scientists can learn about specific people habits and customs.
- Allocation of resources - in public services determining the efficient allocation of people by predicting where is most likely to need them, reduction of costs or optimizing the consumption of resources in a city by monitoring data received from sensors.
- Improvement of performance in operational department - by analyzing companies entire data you can detect the non-functional areas.
- Improving financial performance of enterprise - one example would be the reduction of maintenance costs by determining exactly which equipment is likely to fail.
- Monetization of data - studies made on a specific domain market and customers could be sold to other players in the same domain.
- Fraud detection - in case of financial firms' big data could be used to detect fraud schemes and anomalies in operations by analyzing multiple sources of data in real time.
- Innovation of new services or products - by analyzing the sentiment and expectation of customers regarding companies or competition products.

3. ARCHITECTURE OVERVIEW

Big data analytics systems apply analytic capabilities to large and varied datasets. Such new systems combined to traditional data warehouses or online analytical processing OLAP, enlarged the domain of application for decision support systems. Studies show that big data implementations would not replace data warehouses but due to new specifications in terms of volume, speed and variety of data, the traditional model would be modified to logical data warehouses LDW, which integrates multiple structures and types of data sources interconnected [6].

A typical enterprise data warehouse architecture [11] showed in the figure below is made of several sources of data that provides information for a staging area from where data is transformed and stored in EDW. The presented architecture assumes that data marts are part of EDW and the application know which database to query for data. Is also introduced a sandbox optional area for storing uncertified data.

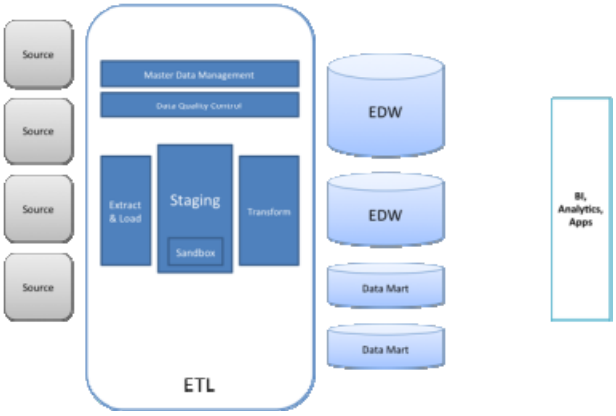

Fig. 1 Typical Enterprise Data Warehouse (EDW) architecture (source: http://cognilytics.com/blog1-php/)

A modern EDW architecture would necessitate replacing of staging area with a data lake based on Hadoop. A data lake (http://en.wiktionary.org/wiki/data_lake) is a term which defines a massive repository for storing big data. Unlike data marts designed to perform analysis and storing only some attributes of data, a data lake is designed to keep all the data attributes. It is designed to be easy accessible and made from relatively cheap hardware for storing massive data.

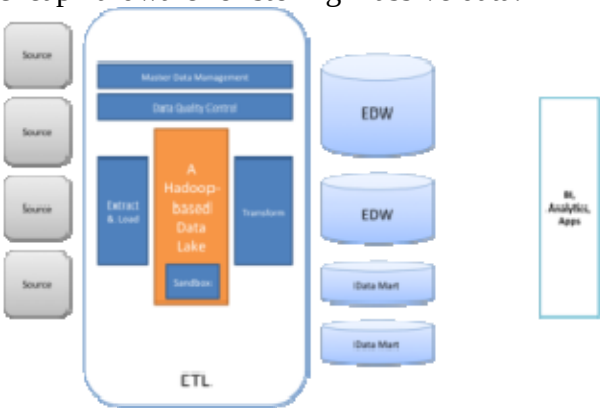

Fig. 2 Changes in EDW architecture caused by big data (source: http://cognilytics.com/blog1-php/)

Hadoop (https://hadoop.apache.org/) is an open source software developed by Apache for distributed computing. It is a framework that manages processing large volume of data in clusters of computers utilizing simple programming models, designed to scale to multiple servers and detect and handle failures

at application layer to deliver high-availability. It is written in Java and was developed after Google's MapReduce and Google File System (GFS)[1].

According to [11] some characteristics of a data lake are:

- Raw data can be stored permanently not just temporary
- It include tools used to make analysis on raw data
- Can store semi structured or unstructured data
- Provides effective and extensible platform for sandboxes
- Reduces costs of storage and processing

Apart from above advantages, Hadoop used in a data lake could solve one important problem of traditional architectures, the lack of agility.

In the figure below we present a general infrastructure for Big Data [14]. It includes layers for data management usually solved by cloud and analytics, requiring computing clusters.

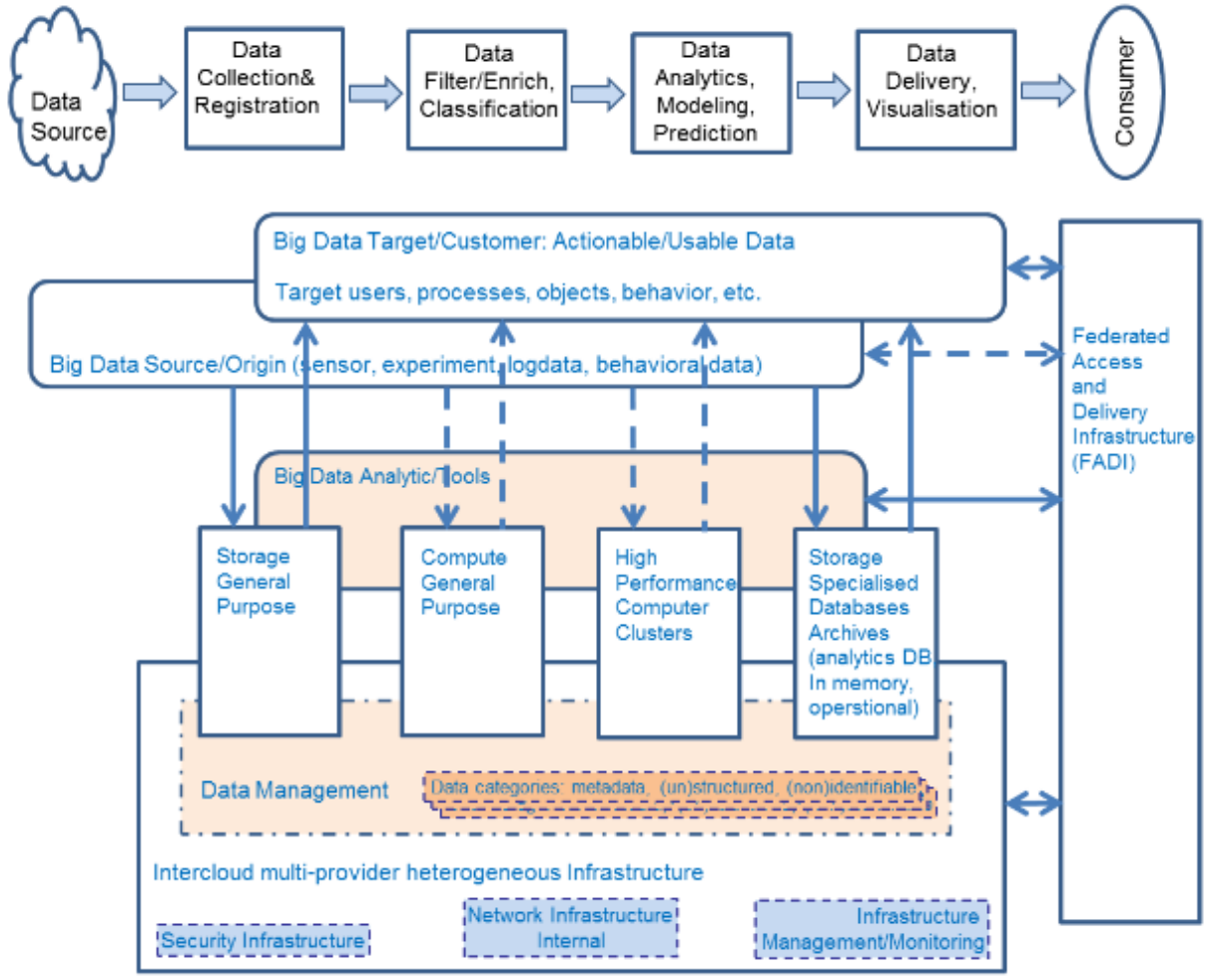

Fig. 3 Big Data infrastructure functional components (source: http://www.uazone.org/demch/worksinprogress/sne-2013-02-techreport-bdaf-draft02.pdf)

The infrastructure include: big data management tools, registries, indexing, semantics, namespaces, security infrastructure, collaborative environment, cluster services, Hadoop framework, Data analytics tools, Sql and nosql databases, Parallel processing databases.

A study [12] about Big Data architectures outline that all include the following areas:

- Big data analytics
  - Descriptive, Predictive and Spatial, Real-time, Interactive, Batch Analytics, Reporting, Dashboard
- Big Data Management/Data Store
  - Structured, semi-structured and unstructured data, Velocity, Variety and Volume, SQL and noSQL, Distributed File System
- Big Data Infrastructure
  - In Memory Data Grids, Operational Database, Analytic Database, Relational Database, Flat files, Content Management System, Horizontal scalable architecture

4. DATA SOURCES

The range of data used in big data projects varies from internal companies' operational data extended with sensors and instruments data, to external data from syndicated data providers and public data. An important source of data is social media from which could be extracted the sentiment regarding products or services, consumers perception about using their products or competition ones.

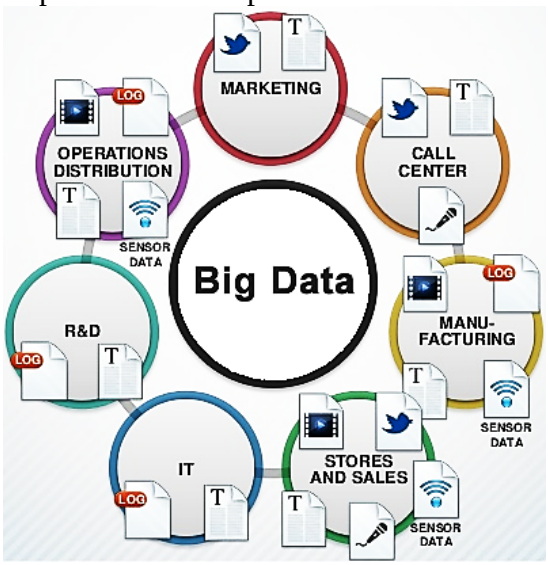

Fig. 4 Big Data sources

(source: adapted from http://www.hds.com/solutions/big-data/?WT.ac=us_mg_sol_bigdat)

There is also a less used source of data that now become a subset of big data - the dark data. Dark data is usually human generated enterprise data, produced by employees such as emails, internal reports, contracts and other document in readable format that has been archived. This data could be analyzed for patterns and correlations.

In a survey conducted by IDG [16] in the first trimester of 2015 is revealed which are the most common sources of data used in big data systems. The percent shows the proportion of systems that use the resource:

- customer databases - 63%
- emails - 61%
- transactional data - 53%
- worksheets - 51%
- word document - 48%

The survey notices that from last year, the percent of systems that use customer databases increased from 49% to 63% and a more notable increase is in case of transactional data, from 33% to 53%.

5. CHALLENGES

The main challenges companies have to manage in dealing with Big Data are [17]:

- Having a strategy to discover valuable information in multiple sources and being capable of collecting data. The companies must be ready to adapt to business transformations imposed by information trend. Information will improve the decision making process and help determine the investment with bigger profit. Gartner predicts that in 2015 will be around 4.4 millions of jobs in domains related to big data.
- Extract knowledge from data - finding new ways to analyze the collected data for predicting trends and outcomes. Business intelligence software platforms perform now predictive and real time analysis in multiple business domains, and should be capable to use all kind of unstructured information, and also new typed of data like clickstream, video, images, sentiment data. It is important to know

how to use all data to gain more competitiveness.

- Enterprise information management - the information volume is constantly growing and companies need to adapt to new requirements for storing and processing it. Also companies must manage the need to extend the access of employees to big data software platforms and this must be made fast and cost effective. This has great impact on companies' data centers.

Due to these challenges not all the companies will be able to benefit from the advantages of big data solutions. A Gartner study shows that through 2015, only 15% percent of Fortune 500 companies will be able to benefit, the other 85% percent would not be capable to gain competitive advantage from exploiting such technologies.

But despite this situation the companies are interested in investing in big data systems. According to IDG survey [16], 27% of respondents already implemented a big data solution and 14% are in the process of implementing/testing the solution. Also 12% have plans to implement in the next 12 month, 8% in the next 13-24 month and 8% are likely to implement a solution in future, but with no specific timeframe. Only 30% of respondents said they have no plans in implementing a big data solution.

The growing trend is also revealed in 2015 Big Data study made in [15], which states that in the last year the number of companies that adapted data analytics solutions increased by 125%. The study was made on 1139 respondents and was conducted online on IDG Enterprise websites.

The majorities of companies expects to gain business values from these solutions and are investing in them. Yet, more importance is accorded to structured data as 32% percent of subjects do not intend to use unstructured data in their solutions in near future. Also large companies invest over ten times more than SMBs in data solutions ($13.8 million compared to $1.6 million). In the picture below are presented the most important criteria, according to customers, in evaluating an offering from a data and analytics vendor:

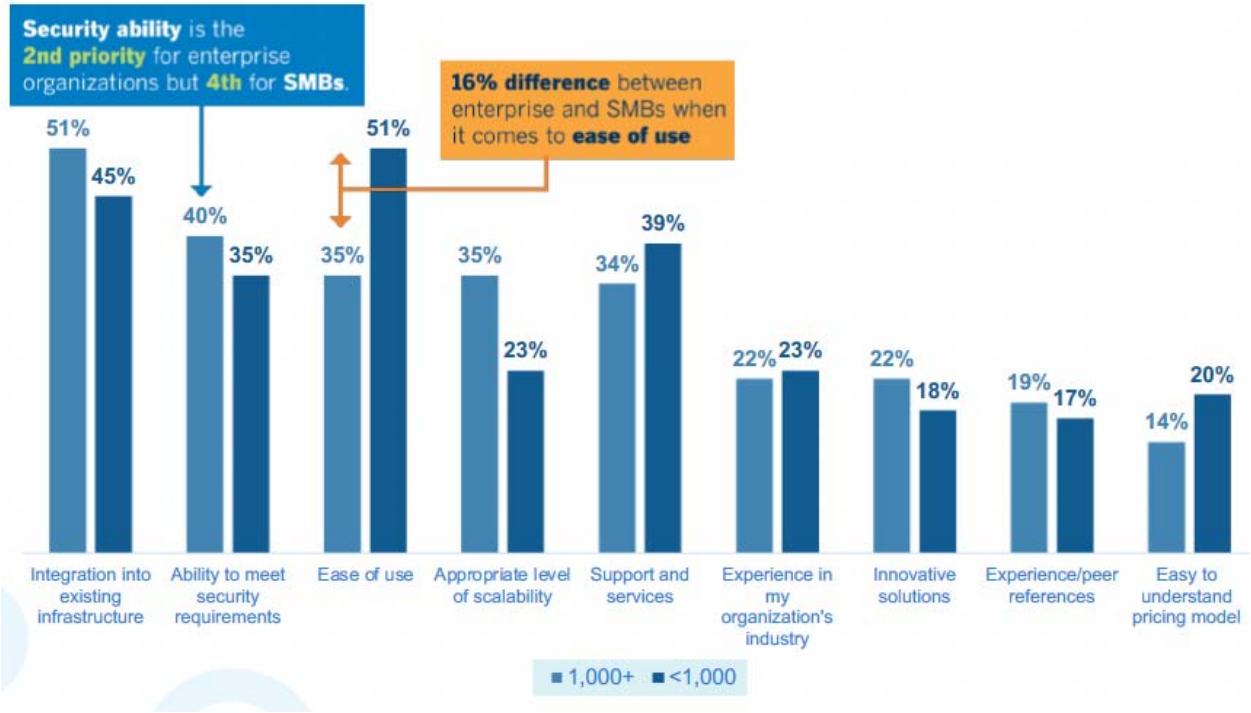

Fig. 5 Top criteria for acquiring a big data solution in case of enterprises and SMBs (source: http://www.idgenterprise.com/report/2015-big-data-and-analytics-survey)

We can see that the most important criteria in acquiring a data and analytics solution are integration into existing infrastructure, ability to meet security requirements, the ease of use, scalability and support and services. There are significant differences in importance of these criteria between large enterprises and small companies. The ease of use is one of the top criteria for small companies but only third for enterprises (16% difference), also there is an important difference in scalability which is not as important for SMBs as for large companies. Support and service is more important for SMBs even than security, being the 3rd priority and 4th security while in case of large companies security is 2nd priority. Moreover the study shows that the confidence in security solutions for company data increased from 49% to 66% in the last year.

6. CONCLUSION & ACKNOWLEDGEMENT

Big data refers to the technology an organization may require to deal with large amount of data and storage facilities. The first that had to deal with such problems were the web search companies that had to process very big volumes of distributed loosely structured data. Dealing with such datasets impose difficulties in manipulating and managing the big data. In current paper we presented an overview of the technologies involved and the use cases of big data systems. Also is presented a general architecture and how data analytics system evolved from traditional data warehouses. The main challenges a company is faced when implementing a big data system are presented and how the new technology is perceived by organizations. We concluded that the main factors involved in acquiring a data analytics solution in an organization are integration into existing infrastructure, ability to meet security requirements, the ease of use, scalability and support and services

This work was co-financed from the European Social Fund through Sectorial Operational Programme Human Resources Development 2007-2013, project number POSDRU/159/1.5/S/134197 „Performance and excellence in doctoral and postdoctoral research in Romanian economics science domain”.

REFERENCES

1. Art Lindsey, *The Origins of Hadoop*, [online], (March, 2015), http://siliconangle.com/ blog/2010/10/14/the-origins-of-hadoop/

2. Bucur Cristian (2014a), "*Aspects regarding detection of sentiment in web content*", International Journal of Sustainable Economies Management (IJSEM), Volume 3: 4 Issues (2014), p.24-32, ISSN: 2160-9659

3. Bucur Cristian (2014b), „*Opinion mining platform for intelligence in business*", Economic Insights – Trends and Challenges, Vol. III LXVI, No. 3/2014, ISSN 2284-8576

4. Cristian BUCUR, „*Implications and Directions of Development of Web Business Intelligence Systems for Business Community*", Economic Insights – Trends and Challenges, Vol. LXIV, No. 2/2012 p. 96 – 108, ISSN 2284-8576

5. David Floyer, *Enterprise Big-data*, [online], (March, 2015), http://wikibon.org/wiki/v/Enterprise_Big-data

6. Douglas Laney, Alexander Linden, Frank Buytendijk et all, *Answering Big Data's 10 Biggest Vision and Strategy Questions*, [online] 12 August 2014, http://www.gartner.com/doc/2822220

7. Krish Krishnan, *Data Warehousing in the Age of Big Data*, Morgan Kaufmann, 2013, ISBN: 0124058914
8. Matt Turk, *A chart of the big data ecosystem, take 2*, [online], accessed march 2015, http://mattturck.com/2012/10/15/a-chart-of-the-big-data-ecosystem-take-2/

9. Nik Bessis, Ciprian Dobre, *Big Data and Internet of Things: A Roadmap for Smart Environments*, Springer, 2014, ISBN: 3319050281

10. Rajendra Akerkar, "*Big Data Computing*", CRC Press, 2014, ISBN-10: 1466578378

11. Rob Klopp, *A Modern Data Warehouse Architecture Part 1 – Add a Data Lake*, [online], (March, 2015), http://cognilytics.com/blog1-php/

12. Sanjay Mishra, *Survey of Big Data Architecture and Framework from the Industry*, NIST Big Data Public Working Group, 2014

13. Vangie Beal, *Big data*, [online], (March, 2015), http://www.webopedia.com/TERM/B/big_data.html

14. Yuri Demchenko, Canh Ngo, Peter Membrey, *Architecture Framework and Components for the Big Data Ecosystem*, System and Network Engineering Group, UvA, 2013

15. ***, *2015 Big Data and Analytics Survey*, [online], (March, 2015), http://www.idgenterprise.com /report/ 2015-big-data-and-analytics-survey

16. ***, *Big Data and Analytics: The Big Picture*, [online], (March, 2015), http:// www.idgenterprise.com/report/big-data-and-analytics-the-big-picture

17. ***, *Big Data Management & Analytics*, [online], (March, 2015), http://www.gartner.com/technology/topics/big-data.jsp

18. ***. *Big data*, [online], accessed March 2015, http://www.sas.com/en_be/insights/big-data/what-is-big-data.html

19. ***, *Ten Practical Big Data Benefits*, [online], http://datascienceseries.com/stories/ten-practical-big-data-benefits