BIG DATA IN BUSINESS ENVIRONMENT

Logica BANICA[1], Alina HAGIU[2]

[1] Faculty of Economics, University of Pitesti, Romania
olga.banica@upit.ro
[2] Faculty of Economics, University of Pitesti, Romania
alina.hagiu@upit.ro

Abstract: *In recent years, dealing with a lot of data originating from social media sites and mobile communications among data from business environments and institutions, lead to the definition of a new concept, known as Big Data. The economic impact of the sheer amount of data produced in a last two years has increased rapidly. It is necessary to aggregate all types of data (structured and unstructured) in order to improve current transactions, to develop new business models, to provide a real image of the supply and demand and thereby, generate market advantages. So, the companies that turn to Big Data have a competitive advantage over other firms. Looking from the perspective of IT organizations, they must accommodate the storage and processing Big Data, and provide analysis tools that are easily integrated into business processes.*
This paper aims to discuss aspects regarding the Big Data concept, the principles to build, organize and analyse huge datasets in the business environment, offering a three-layer architecture, based on actual software solutions. Also, the article refers to the graphical tools for exploring and representing unstructured data, Gephi and NodeXL.

Key words*:* Big Data, business environment, analysis software

JEL Classification Codes: C82, M21, C88.

1. INTRODUCTION

Increasingly, nowadays, companies realize the importance of using Big Data in business environment, because of their confrontation with huge amounts of data that originate in social networks and mobile communications, in addition to the traditional databases.

The economic impact of the sheer amount of data produced in a last two years has increased rapidly. It is necessary to aggregate all types of data (structured and unstructured) in order to improve current transactions, to develop new business models, to provide a real image of the supply and demand and thereby, generate market advantages.

Big Data already penetrated businesses processes and in few years will impact all economic and social sectors, even there are several technical challenges. The influence of Big Data analysis on management and decision making is too valuable to be ignored. A major issue for companies leveraging Big Data is the processing time for data analysis. After gathering and storing data, the massive parallel processing and the usage of analytics in order to better understand the dynamics of business is essential.

Looking from the perspective of IT organizations, they must accommodate the storage and processing Big Data, and provide analysis tools that are easily integrated into business processes. Also, the business analysts are faced with a challenge regarding data flow filtering, and implementation of Business Intelligence or Forecasting strategies for their activity (Banica et al., 2014).

Starting from the current studies and implementations concerning the Big Data concept, and their integration in actual software systems, we have concentrated our efforts around the principles to build, organize and analyse huge datasets in the business environment, offering a three-layer architecture.

This paper is organized as follows: Section 2 presents the concept of Big Data, summarizing the state-of-the-art, and proposes several software solutions to store, process and analyze massive volume of data. Section 3 develops a system able to accommodate Big Data in business environment, by presenting the software platforms for each of the three levels of the designed architectural model.

Conclusions states our confidence in obtaining successful results by the companies using Big Data and suggest ways of improving our research in this domain, firstly by building and testing such an experimental system.

2. STATE OF THE ART

2.1 Big Data concept

„Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured. And big data may be as important to business – and society – as the Internet has become. Why? More data may lead to more accurate analyses." (Davenport & Dyché, 2013).

An important volume of semi-structured and unstructured data is generated by transaction systems, social networks, server logs, sensors and mobile networks. Even the ratio of data coming from social media sites, mobile networks and relational databases is clearly in favor of the first group, the ratio of useful information collected is reversed.

The main characteristics of Big Data are described by experts through the five V's as follow (Bowden, 2014) (Dijcks, 2013):

- **Volume** – refers to the scalability as the most important aspect for every domain of application. The data volume build-up may be from unstructured sources like the social media and from traditional databases. The relevance of the volume of data collected may be obtained by filtering using analytic tools in order to identify important patterns and metrics that are found in business field.
- **Velocity** – the increasing flows of data need hardware and software solutions to process data streaming in a paced as fast as possible; the actors of the market need answers to their questions in real time.
- **Variety** – concerns the combination of all types of formats and the differing meanings attached to the same forms, Big Data must cover every opportunity of connecting the business with the customers in a virtual marketplace.
- **Veracity** - refers to the trustworthiness of the information; the data may be not significant, also there could be discrepancies in the sample of data collected, filtered and processed.
- **Value –** is the most important V of Big Data because it turn the increased amount of data into commercial or scientific value. The final target of processing Big Data is to develop the business, to obtain a stronger competitive position and a high level of knowledge, to find new solutions in all areas (economic, social, health and education).

2.2 Introducing Big Data in business environment

Business environment is interested in collecting information from unconventional data sources, in order to analyse and extract meaningful insight from this maze of data, be it security related or simply behavioural patterns of consumers.

We will specify several ways by means of which the companies using Big Data could improve their business (Rosenbush & Totty, 2013):

1. Great software companies, like Google, Facebook and Amazon, showed their interest in processing Big Data in the Cloud environment many years ago. They are collecting huge amounts of information, analyze traditional measures like sales using comments on social-media sites and location information from mobile devices. This information is useful to figure out how improve their products, cut costs and keep customers coming back.
2. Product development for online companies – Big Data could help to capture customer preferences and use the information in designing new products. For example, Zynga Inc., the game maker, uses the collected data for customer service, quality assurance and designing the features for the next generation of games. Also, Ford Motor Co. designed a common set of components that would be on Ford cars and trucks by using algorithms that summarize more than 10,000 relevant comments.
3. Human resources - some companies are using Big Data to better handle the health care of their employees. For example, Caesars Entertainment Corp. analyzes health-insurance data for its 65,000 employees and their family members, finding information about how employees use medical services, the number of emergency-room visits and whether they choose a generic or brand-name drug.
4. Marketing – is the enterprise department made to understand the customers and their choices. Using Big Data analytics the information is better filtered and the forecasts are more accurate. An important company, InterContinental Hotels Group PLC has gathered details about the 71 million members of its Priority Club program, such as income levels and whether they prefer family-style or business-traveler accommodations.
5. Manufacturing companies, as well as retailers begin to monitor Facebook and Twitter and analyse those data from different angles, e.g. customer satisfaction. Retailers are also collecting large amounts of data by storing log files and combine that growing number of information with other data sources such as sales data in order to forecast customer behaviour.

3. METHODOLOGY

In order to build a Big Data infrastructure for an enterprise, there are required major hardware resources and specific software applications. The Cloud Computing environment is a solution to provide the resources required to store and access important data volumes.

In this chapter we will briefly present a new type of database - NoSQL, used for storing Big Data, and a series of successful implementations, such as the software that allows for massively parallel computing – Apache Hadoop, and several applications for collecting structured and unstructured data, such as Apache Flume and Apache Sqoop.

Also, we consider that Gephi and NodeXL are two representative applications for SNA software, so we made a short presentation of their facilities.

NoSQL databases are a new type of database that manages a wide variety of unstructured data described by several models: key value stores, graph, and document data (Mohamed et al., 2014). NoSQL data are implemented in a distributed architecture, based on multiple nodes, as Hadoop software.

Hadoop is an open source software platform that enables the processing of large data sets in a distributed computing environment.

Its core concept is distributing the data in the system to the individual nodes, which can work on data locally and without transferring data over the network for processing.

Hadoop is compatible with any of operating system families (Linux, Unix, Windows, Mac OS) and can be run on a single or on multi-node cluster.
The most widespread solution is the open source Apache Hadoop distribution, including several components (Frank, 2014):

- *Hadoop Distributed File System (HDFS)* - the storage component
- *MapReduce engine* - the processing component
- *Hadoop Common* - a module for libraries and utilities
- (*Hadoop YARN*) – the component responsible for managing and scheduling cluster resources

A Hadoop Cluster is a set of computers running HDFS and MapReduce, one of them being master node and the others slave nodes (Figure 1).

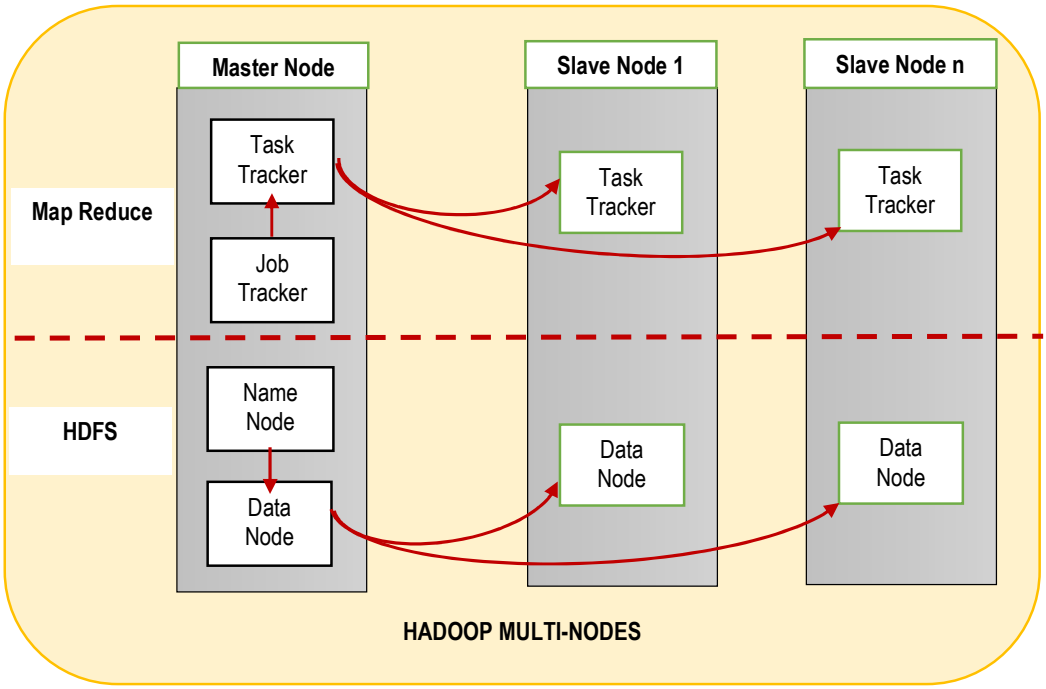

Figure 1. The structure of a Hadoop cluster *(Source: Gupta, 2014)*

HDFS splits data files into blocks and distributes them across the cluster, each block being replicated three times (by default), to ensure system reliability and security (Yahoo Developer Network, 2007).

The master node called the Name Node monitors distributed blocks, known as Data Nodes.
MapReduce is the component that Hadoop uses to distribute work around a cluster, so that operations can be run in parallel on different nodes (Frank, 2014). MapReduce is controlled by JobTracker software, that resides on the master node and it monitors the TaskTracker applications, running tasks at each node.
The MapReduce system is based on two components performed (Brust, 2012):

- Map – reads data in the form of key/value pairs and processes tasks on nodes, avoiding network traffic;
- Reduce – is waiting until Map phase is completed and combines all the intermediate values into a list, providing final key/value pairs as outputs that are written into HDFS.

Hadoop 2.0 is a new version, providing more capacity to be integrated with stack technology, adding support for running non-batch applications through the introduction of YARN (Yet

Another Resource Negotiator), a separate layer that includes resource management and job scheduling functions, according to Gartner researchers (Laskowski, 2014).
Hadoop 2.0 added new important features: support for Microsoft Windows and improvements of system availability and scalability.
In figure 2 is presented the architecture diagram of Hadoop 1.0 in comparison with Hadoop 2.0/YARN.

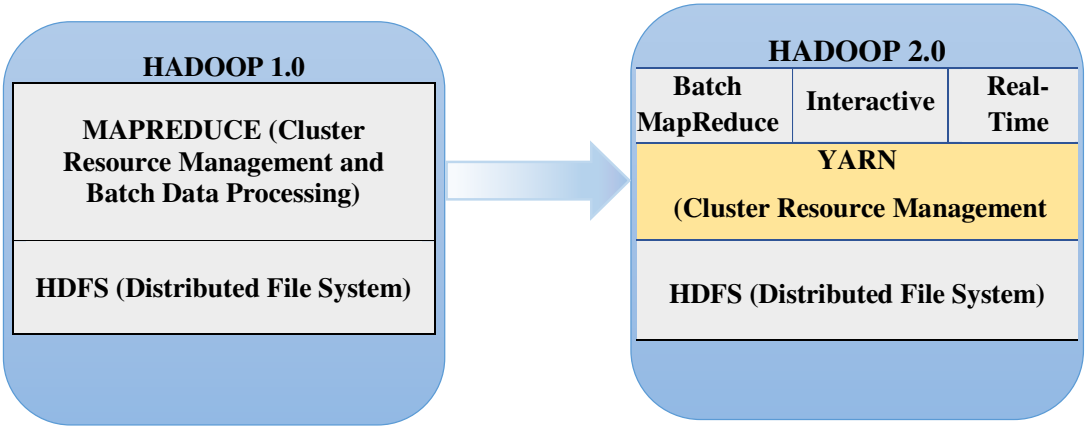

Figure 2. The architecture diagram of Hadoop 1.0 and Hadoop 2.0/YARN
Source: http://hortonworks.com/blog/how-to-plan-and-configure-yarn-in-hdp-2-0/

Nowadays, another open source technology emerged, called Spark, that can be applied for handling massive amounts of data and also deployed as a component running on the Hadoop 2.0 platform (Rouse, 2014). Spark's in-memory allows user programs to load data into a cluster's memory and query it repeatedly, providing performance up to 100 times faster for certain applications (Vaughan, 2014).
For implementing Big Data into a company, is required to extend the existing systems, usually based-on RDBMS (Relational DataBase Management System).
For this purpose, the information system of the company should accept the transfer of their structured data, into a consolidated Data Warehouse, by using a software solution, such as (Mohamed et al., 2014):

- ***ETL process*** (*Extract, Transform, Load)* - structured data are moved into the Data Warehouse;
- ***Apache Sqoop*** – a tool for moving data from relational databases into Hadoop and back.

Finally, the methodology includes Analytics layer, having different goals: finding correlations across multiple data sources, forecasting the indicators or analysing social networks (Krebs, 2013).
In this study, we suggest that the system can be further explored using social network analysis (SNA) software, such as Gephi and Node XL.

Gephi is a free, open source interactive visualization and exploration platform for all kinds of networks and complex systems, capable of accommodating networks up to 50,000 nodes and 1,000,000 edges (Bastian et al., 2009). Also, it generates metrics, identifies subgroups in a network, clusters of actors or individuals, or emphasizes isolated nodes of the network.
The application is successfully used to analyze pages and groups of Facebook, Twitter networks and e-mail. Gephi works with imported files from .csv, .gml, .gdf and .gefx format, which can be achieved with software converters (e.g. Facebook or Twitter to .gdf files) from unstructured data, by applying an algorithm that transforms key-value words on nodes and their connections on edges.

NodeXL is a free, open-source template for Microsoft Excel 2007, 2010 and 2013 that makes it easy to explore network graphs. Concerning metric calculations**,** it allows calculate degree, centrality, PageRank, clustering coefficient and graph density (Messarra, 2014)
NodeXL can have a direct connection to social networks: Twitter, YouTube, Flickr and email. requesting the user's permission before collecting personal data and focuses on the collection of publicly available data, such as Twitter statuses and follows relationships for users who have made their accounts public (Kennedy et al., 2013)
After using both SNA applications and making a comparison of their features, we believe that Gephi tool is more powerful than NodeXL.

4. A MODEL FOR INTEGRATION HADOOP IN AN ENTERPRISE INFORMATION SYSTEM

In this chapter is described a three-layer Big Data architecture for business environment, based on a Hadoop cluster.
The main task is not Hadoop itself, but integrating it with the existing ERP (Enterprise resource planning) system that the companies usually already own. Thus, our ***model*** refers ***to a unified architecture***, using Hadoop as a data integration platform (figure 3).

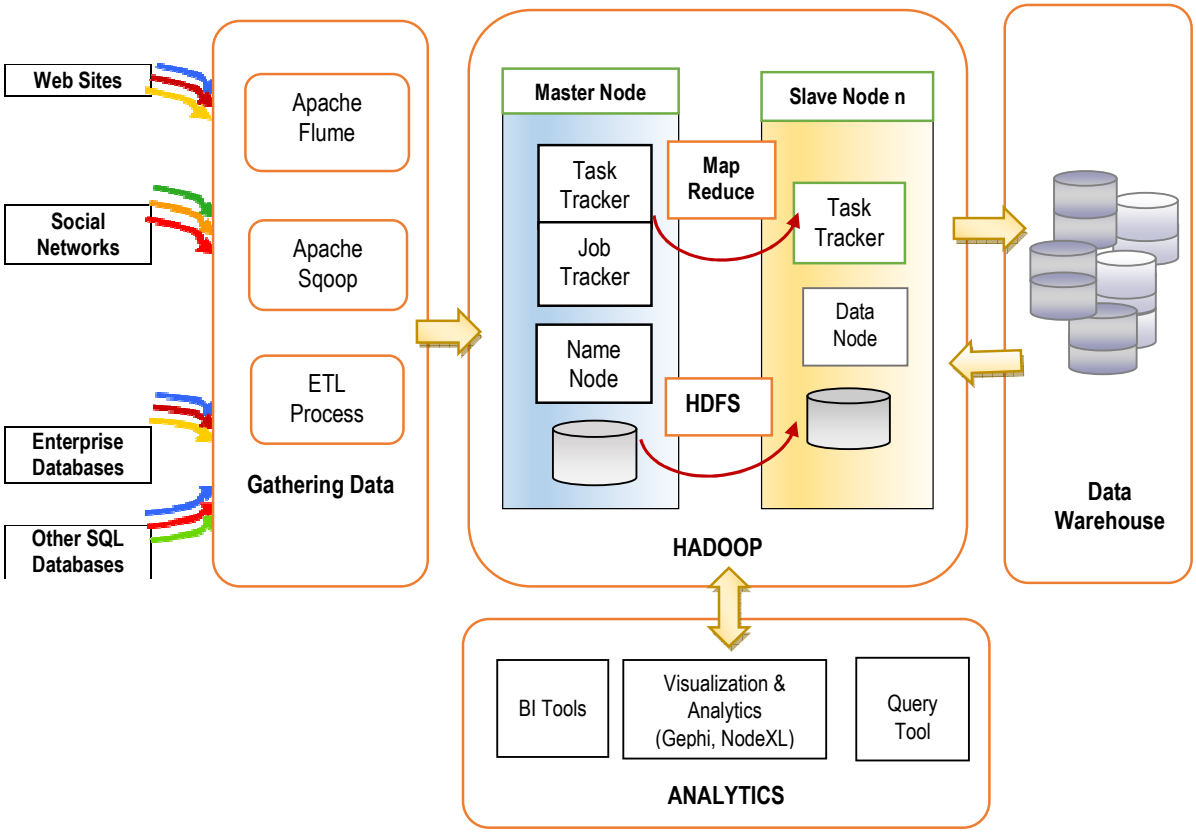

Figure 3. An architecture for Big Data in business environment

The three-layer architecture includes:

1) Gathering structured and unstructured data interesting for enterprise - The first layer is designated for collecting any type of data and is based on software tools Apache Flume (unstructured data), Apache Sqoop (structured data) and ETL (structured data).

2) Parallel Processing of Big Data using Hadoop - At the second level is implemented Apache Hadoop that aggregates and process data;
3) Analyzing Big Data - The third layer provides analytics tools, including business and modeling tools, exploring and visualization applications, such as Gephi and NodeXL.

For many enterprises, identifying the Web presence through the hyperlinks network and the components of social media such as Twitter, Facebook and YouTube is increasingly important for their business. Therefore, they are interested in investing for development of IT tools capable to mapping and deciphering organisational presence on the Web.

5. CONCLUSIONS AND FUTURE WORK

Big Data will radically change the current business models for gaining important benefits on marketplace. The evolution of open-source software for NoSQL databases will allow small and medium enterprises to benefit from this new trend that empowers today's business.

But, there is another problem that is required to be solve: the integration between Hadoop and the existing ERP system of the organizations. Consequently, a possible scenario refers to a unified architecture that integrates Big Data technologies in actual systems.

In our future research we intent to implement a single node Hadoop structure and evaluate its performance working with unstructured data from Social media.
Also, considering Analytics a fundamental part of Big Data value, we will make a comparative analysis of ***several tools***, testing them against workloads with query, graphic interpretation and dynamic approach.

REFERENCES

1. Banica, L., Paun, V., Stefan, C., 2014, Big Data leverages Cloud Computing opportunities, International Journal of Computers & Technology, Volume 13, No.12, available at *http://cirworld.org/journals/index.php/ijct/article/view/3036*
2. Thomas H. Davenport, Jill Dyché, Big Data in Big Companies, 2013, available at: *http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/bigdata-bigcompanies-106461.pdf*
3. Jason Bowden, 2014, The 4 V's in Big Data for Digital Marketing, available at *http://www.business2community.com/ digital-marketing/ 4-vs-big-data-digital-marketing-914845*
4. Dijcks, J., 2013, Big Data for Enterprise, *http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf*
5. Rosenbush, S., Totty, M., 2013, How Big Data Is Changing the Whole Equation for Business, available at *http://www.wsj.com/articles/SB10001424127887324178904578340071261396666*
6. Mohamed, A., M., Altrafi, O.,G., Ismail, M., O., 2014, Relational vs. NoSQL Databases: A Survey, International Journal of Computer and Information Technology, Vol.3, Issue 3, pp. 598-601
7. Frank Lo, 2014, Big Data Technology, available at *https://datajobs.com/what-is-hadoop-and-nosql*
8. Gupta, R., 2014, Big Data With SQL, available at *https://www.linkedin.com/pulse/20140801134428-260299242-big-data-with-sql*
9. Yahoo Developer Network, 2007, Hadoop Tutorial, available at *https://developer.yahoo.com/hadoop/tutorial/*
10. Brust, A., 2012, *CEP and MapReduce: Connected in complex ways,* *http://www.complexevents.com/2012/03/10/cep-and-mapreduce-connected-in-complex-ways/*

11. Laskowski, L., 2014, Hadoop 2.0's Deep Impact on Big Data and Big Data Technologies, available at *http://searchcio.techtarget.com/opinion/Hadoop-20s-deep-impact-on-big-data-and-big-data-technologies*
12. Vaughan, J., 2014, Hadoop and Spark are Coming of Age Together, *http://searchdatamanagement.techtarget.com/podcast/Hadoop-and-Spark-are-coming-of-age-together*
13. Rouse, M., 2014, HADOOP 2 , available at *http://searchdatamanagement.techtarget.com/definition/Hadoop-2*
14. Mohamed, A., M., Altrafi, O.,G., Ismail, M., O., 2014, *Relational vs. NoSQL Databases: A Survey*, International Journal of Computer and Information Technology, Vol.3, Issue 3, pp. 598-601
15. Krebs, V., 2013, Social Network Analysis, A Brief Introduction, available at *http://www.orgnet.com/sna.html*
16. Bastian M., Heymann S., Jacomy M., 2009, Gephi: an open source software for exploring and manipulating networks, *International AAAI Conference on Weblogs and Social Media*, San Jose, USA, available at *https://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154/1009*
17. Messarra, N., 2014, Introduction to Social Graph and NodeXL, available at *http://nasri.messarra.com/introduction-to-social-graph-and-nodexl/*
18. Kennedy, H., Moss, G., Birchall,C., Moshonas, S., 2013, Digital Data Analysis: Guide to tools for social media & web analytics and insights available at *http://www.google.ro/url?sa=t&rct=j&q=&esrc=s&source=web&cd=31&ved=0CB4QFjAAOB4&url=http%3A%2F%2Fwww.communitiesandculture.org%2Ffiles%2F2013%2F02%2FDigital-data-analysis-guide-to-tools.pdf*