SSARSC

Scientific Society of Advanced Research and Social Change
SSARSC International Journal of Information and Communication Technology
Volume 1 Issue 1, January-June 2018

Perspectives on Big Data and Big Data Analytics- A Review

Dr. Amit Asthana[1] , Khushboo[2]
[1]Dept. of Computer Science & Engineering#,
[2]M.Tech. Research Scholar Dept. of Computer Science & Engineering
Subharti Institute of Engineering & Technology, Uttar Pradesh, INDIA.

Abstract : Nowadays companies are starting to realize the importance of using more data in order to support decision for their strategies. It was said and proved through study cases that "More data usually beats better algorithms". With this statement companies started to realize that they can chose to invest more in processing larger sets of data rather than investing in expensive algorithms. The large quantity of data is better used as a whole because of the possible correlations on a larger amount, correlations that can never be found if the data is analyzed on separate sets or on a smaller set. A larger amount of data gives a better output but also working with it can become a challenge due to processing limitations.

This article intends to define the concept of Big Data and stress the importance of Big Data Analytics.

Keywords: Big Data, Big Data Analytics, Database, Internet, Hadoop project.

1. Introduction

Nowadays the Internet represents a big space where great amounts of information are added every day. The IBM Big Data Flood Infographic shows that 2.7 Zettabytes of data exist in the digital universe today. Also according to this study there are 100 Terabytes updated daily through Facebook, and a lot of activity on social networks this leading to an estimate of 35 Zettabytes of data generated annually by 2020. Just to have an idea of the amount of data being generated, one zettabyte (ZB) eguals 1021 bytes, meaning 1012 GB [1]. We can associate the importance of Big Data and Big Data Analysis with the society that we live in. Today we are living in an Informational Society and we are moving towards a Knowledge Based Society. In order to extract better knowledge we need a bigger amount of data. The Society of Information is a society where information plays a major role in the economical, cultural and political stage.

In the Knowledge society the competitive advantage is gained through understanding the information and predicting the evolution of facts based on data. The same happens

with Big Data. Every organization needs to collect a large set of data in order to support its decision and extract correlations through data analysis as a basis for decisions.

In this article we will define the concept of Big Data, its importance and different perspectives on its use. In addition we will stress the importance of Big Data Analysis and show how the analysis of Big Data will improve decisions in the future.

2. Big Data Concept

The term "Big Data" was first introduced to the computing world by Roger Magoulas from O'Reilly media in 2005 in order to define a great amount of data that traditional data management techniques cannot manage and process due to the complexity and size of this data.

A study on the Evolution of Big Data as a Research and Scientific Topic shows that the term "Big Data" was present in research starting with 1970s but has been comprised in publications in 2008. [2] Nowadays the Big Data concept is treated from different points of view covering its implications in many fields.

According to MiKE 2.0, the open source standard for Information Management, Big Data is defined by its size, comprising a large, complex and independent collection of data sets, each with the potential to interact. In addition, an important aspect of Big Data is the fact that it cannot be handled with standard data management techniques due to the inconsistency and unpredictability of the possible combinations [3].

In IBM's view Big Data has four aspects:

Volume: refers to the quantity of data gathered by a company. This data must be used further to obtain important knowledge;

Velocity: refers to the time in which Big Data can be processed. Some activities are very important and need immediate responses, that is why fast processing maximizes efficiency;

Variety: Refers to the type of data that Big Data can comprise. This data can be structured as well as unstructured; Veracity: refers to the degree in which a leader trusts the used information in order to take decision. So getting the right correlations in Big Data is very important for the business future. [4]

In addition, in Gartner's IT Glosarry Big Data is defined as high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making [5].

According to Ed Dumbill chair at the O'Reilly Strata Conference, Big Data can be described as, "data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from

this data, you must choose an alternative way to process it." [6]

In a simpler definition we consider Big Data to be an expression that comprises different data sets of very large, highly complex, unstructured, organized, stored and processed using specific methods and techniques used for business processes.

There are a lot of definitions on Big Data circulating around the world, but we consider that the most important one is the one that each leader gives to its one company's data. The way that Big Data is defined has implication in the strategy of a business. Each leader has to define the concept in order to bring competitive advantage for the company.

The importance of Big Data

The main importance of Big Data consists in the potential to improve efficiency in the context of use a large volume of data, of different type. If Big Data is defined properly and used accordingly, organizations can get a better view on their business therefore leading to efficiency in different areas like sales, improving the manufactured product and so forth.

Big Data can be used effectively in the following areas:

- In information technology in order to improve security and troubleshooting by analyzing the patterns in the existing logs;
- In customer service by using information from call centers in order to get the customer pattern and thus enhance customer satisfaction by customizing services;
- In improving services and products through the use of social media content. By knowing the potential customers preferences the company can modify its product in order to address a larger area of people;
- In the detection of fraud in the online transactions for any industry;
- In risk assessment by analyzing information from the transactions on the financial market.

In the future we propose to anayze the potencial of Big Data and the power that can be enabled through Big Data Analysis.

Big Data challenges

The understanding of Big Data is mainly very important. In order to determine the best strategy for a company it is essential that the data that you are counting on must be properly analyzed. Also the time span of this analysis is important because some of them need to be performed very frequent in order to determine fast any change in the business environment.

Another aspect is represented by the new technologies that are developed every day. Considering the fact that Big Data is new to the organizations nowadays, it is necessary

for these organizations to learn how to use the new developed technologies as soon as they are on the market. This is an important aspect that is going to bring competitive advantage to a business.

The need for IT specialists it is also a challenge for Big Data. According to McKinsey's study on Big Data called Big Data: The next frontier for innovation, there is a need for up to 190,000 more workers with analytical expertise and 1.5 million more data-literate managers only in the United States. This statistics are a proof that in order for a company to take the Big Data initiative has to either hire experts or train existing employees on the new field.

Privacy and Security are also important challenges for Big Data. Because Big Data consists in a large amount of complex data, it is very difficult for a company to sort this data on privacy levels and apply the according security. In addition many of the companies nowadays are doing business cross countries and continents and the differences in privacy laws are considerable and have to be taken into consideration when starting the Big Data initiative.

In our opinion for an organization to get competitive advantage from the manipulation of Big Data it has to take very good care of all factors when implementing it. One option of developing a Big Data strategy is presented below. In addition, in order to bring full capabilities to Big Data each company has to take into consideration its own typical business characteristics.

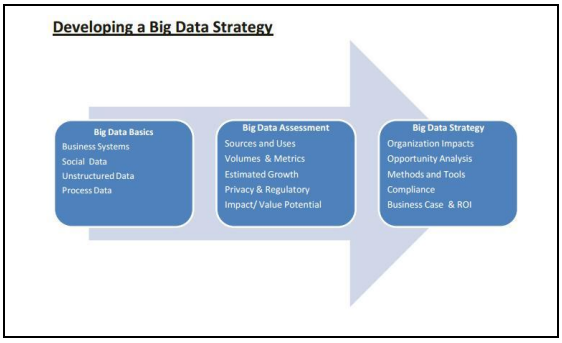

Fig 1. Developing a Big Data Strategy (Source http://www.navint.com) [7]

3. Big Data Analytics

The world today is built on the foundations of data. Lives today are impacted by the ability of the companies to dispose, interrogate and manage data. The development of technology infrastructure is adapted to help generate data, so that all the offered services can be improved as they are used.

As an example, internet today became a huge information-gathering platform due to social media and online services. At any minute they are added data. The explosion of data cannot be any more measured in gigabytes, since data is bigger there are used etabytes, exabytes, zettabytes and yottabytes.

In order to manage the giant volume of unstructured data stored, it has been emerged the "Big Data" phenomena. It

stands to reason that in the commercial sector Big-Data has been adopted more rapidly in data driven industries, such as financial services and telecommunications, which it can be argued, have been experiencing a more rapid growth in data volumes compared to other market sectors, in addition to tighter regulatory requirements and falling profitability. At first, Big Data was seen as a mean to manage to reduce the costs of data management. Now, the companies focus on the value creation potential. In order to benefit from additional insight gained there is the need to assess the analytical and execution capabilities of "Big Data".

To turn big data into a business advantage, businesses have to review the way they manage data within data centre. The data is taken from a multitude of sources, both from within and without the organization. It can include content from videos, social data, documents and machine-generated data, from a variety of applications and platforms. Businesses need a system that is optimised for acquiring, organising and loading this unstructured data into their databases so that it can be effectively rendered and analysed. Data analysis needs to be deep and it needs to be rapid and conducted with business goals in mind.

The scalability of big data solutions within data centres is an essential consideration. Data is vast today, and it is only going to get bigger. If a data centre can only cope with the levels of data expected in the short to medium term, businesses will quickly spend on system refreshes and upgrades. Forward planning and scalability are therefore important.

In order to make every decision as desired there is the need to bring the results of knowledge discovery to the business process and at the same time track any impact in the various dashboards, reports and exception analysis being monitored. New knowledge discovered through analysis may also have a bearing on business strategy, CRM strategy and financial strategy going forward. See figure 2

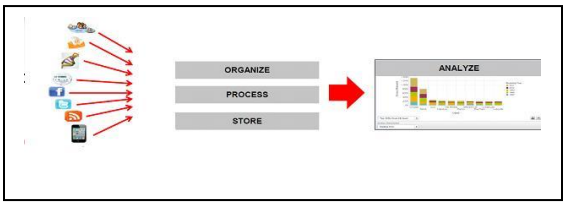

Fig 2. Big Data Management

Up until mid 2009 ago, the data management landscape was simple: Online transaction processing (OLTP) systems (especially databases) supported the enterprise's business processes; operational data stores (ODSs) accumulated the business transactions to support operational reporting, and enterprise data warehouses (EDWs) accumulated and transformed business transactions to support both operational and strategic decision making.

Big Data Management is based on capturing and organizing relevant data. Data analytics supposes to understand that happened, why

and predict what will happen. A deeper analytics means new analytical methods for deeper insights.[9]

Big data analytics and the Apache Hadoop open source project are rapidly emerging as the preferred solution to business and technology trends that are disrupting the traditional data management and processing landscape. Enterprises can gain a competitive advantage by being early adopters of big data analytics. Even though big data analytics can be technically challenging, enterprises should not delay implementation. As the Hadoop projects mature and business intelligence (BI) tool support improves, big data analytics implementation complexity will reduce, but the early adopter competitive advantage will also wane. Technology implementation risk can be reduced by adapting existing architectural principles and patterns to the new technology and changing requirements rather than rejecting them. [10]

Big data analytics can be differentiated from traditional data-processing architectures along a number of dimensions:

- Speed of decision making being very important for decision makers
- Processing complexity because it eases the decision making process
- Transactional data volumes which are very large
- Data structure data can be structured and unstructured
- Flexibility of processing/analysis consisting in the amount of analysis that can be performed on it
- Concurrency [9]

The big data analytics initiative should be a joint project involving both IT and business. IT should be responsible for deploying the right big data analysis tools and implementing sound data management practices. Both groups should understand that success will be measured by the value added by business improvements that are brought about by the initiative.

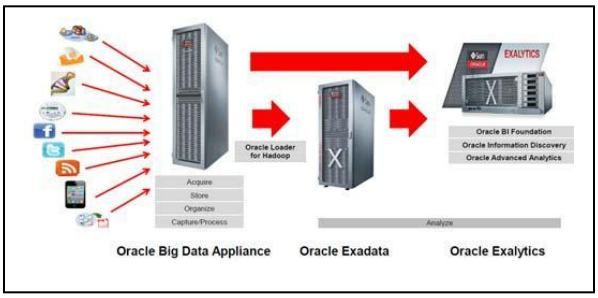

Fig 3. Oracle Big Data Solution (Source: myoracle.com)

In terms of Big Data Management and analytics Oracle is offering Engineered Systems as Big Data Solutions (Fig.3), such as Oracle Big Data Appliance, Oracle Exadata and Oracle Exalytics. Big Data solutions combine best tools for each part of the problem. The traditional business intelligence tools rely on relational databases for storage and query execution

and did not target Hadoop. Oracle BI combined with Oracle Big Data Connectors. The architecture supposes to load key elements of information from Big Data sources into DBMS. Oracle Big Data connectors, Hive and Hadoop aware ETL such as ODI provide the needed data integration capabilities. The key benefits are that the business intelligence investments and skills that are leveraged, there are made insights from Big Data consumable for business users, there are combined Big Data with Application and OLTP data. [11]

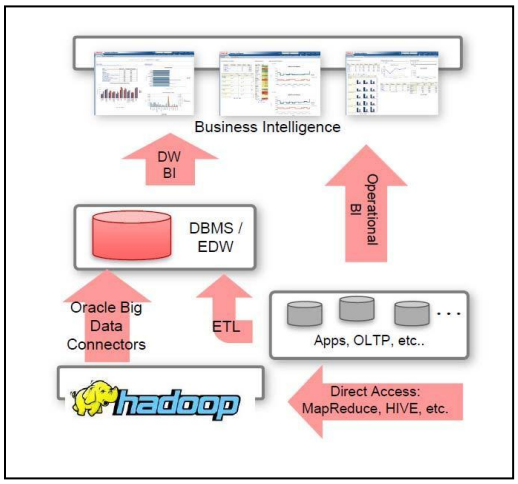

Fig 4. BI and Data Warehousing on Big Data (Source: myoracle.com)

Big Data provides many opportunities for deep insights via data mining:

- Uncover relationships between social sentiment and sales data
- Predict product issues based on diagnostic sensor data generated by products in the field
- In fact, the signal-to-noise issues often mean deep analytics to mine insight hidden in the noise is essential, as many forms of Big Data are simply not consumable in raw form

"Big Data" is a Data Management & Analytics market opportunity driven by new market requirements. In-Database Analytics – Data Mining there are used Big Data Connectors to combine Hadoop and DBMS data for deep analytics. Also there is the need to re-use SQL skills to apply deeper data mining techniques or re-use skills for statistical analysis. Everything is all about "Big Data" instead of RAM-scale data. This is how the predictive learning of relationships between knowledge concepts and business events is done. [12]

Big- Data presents a significant opportunity to create new value from giant data. It is important to determine appropriate governance procedures in order to manage development and implementations over the life of the technology and data. Failure to consider the longer term implications of development will lead to productivity issues and cost escalations.

On the face of it, the cost of physically storing large quantities of data is dramatically reduced by the simplicity by which data can be loaded into a Big-Data cluster because there is no longer required a complex ETL layer seen in any more traditional Data Warehouse solutions. The cluster itself is also typically built using low

cost commodity hardware and analysts are free to write code in almost any contemporary language through the streaming API available in Hadoop.

- The business logic used within an ETL flow to tokenise a stream of data and apply data quality standards to it must be encoded (typically using Java) within each Map-Reduce program that

processes the data and any changes in source syntax or semantics [8]

- Although the storage nodes in a Hadoop cluster may be built using low cost commodity x86 servers, the master nodes (Name Node, Secondary Name Node and Job Tracker) requiring higher resilience levels to be built into the servers if disaster is to be avoided. Map-Reduce operations also generate a lot of network chatter so a fast private network is recommended. These requirements combine to add significant cost to a production cluster used in a commercial setting. [8]

- Compression capabilities in Hadoop are limited because of the HDFS block structure and require an understanding of the data and compression technology to implement adding to implementation complexity with limited impact on storage volumes.

Other aspects to consider include the true cost of ownership of pre-production and production clusters such as the design build and maintenance of the clusters themselves, the transition to production of Map-Reduce code to the production cluster in accordance with standard operational procedures and the development of these procedures. [8]

Whatever the true cost of Big-Data compared to a relational data storage approach, it is important that the development of Big-Data strategy is consciously done, understanding the true nature of the costs and complexity of the infrastructure, practice and procedures that are put in place.

4. Big Data Analytics Software

Currently, the trend is for enterprises to re-evaluate their approach on data storage, management and analytics, as the volume and complexity of data is growing so rapidly and unstructured data accounting is for 90% of the data today.

Every day, 2.5 quintillion bytes of data are created — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from various sources such as: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals, web and software logs, cameras, information-sensing mobile devices, aerial sensory technologies and genomics. This data is referred to as big data.

“Legacy systems will remain necessary for specific high-value, low-volume workloads,

and compliment the use of Hadoop - optimizing the data management structure in the organization by putting the right Big Data workloads in the right systems"[14].

As it was mentioned in the Introduction Big data spans four dimensions: Volume, Velocity, Variety, and Veracity

- Volume: Enterprises are awash with ever-growing data of all types, easily amassing terabytes - even petabytes - of information(e.g. turn 12 terabytes of Tweets created each day into improved product sentiment analysis; convert 350 billion annual meter readings to better predict power consumption);

- Velocity: For time-sensitive processes such as catching fraud, big data flows must be analysed and used as they stream into the organizations in order to maximize the value of the information(e.g. scrutinize 5 million trade events created each day to identify potential fraud; analyze 500 million daily call detail records in real-time to predict customer churn faster).

- Variety: Big data consists in any type of data - structured and unstructured data such as text,

sensor data, audio, video, click streams, log files and more. The analysis of combined data types brings new aspect for problems, situations etc.(e.g. monitor 100's of live video feeds from surveillance cameras to target points of interest; exploit the 80% data growth in images, video and documents to improve customer satisfaction);

- Veracity: Since one of three business leaders don't trust the information they use to make decisions, establishing trust in big data presents a huge challenge as the variety ad number of sources grows.

Apache Hadoop is a fast-growing big-data processing platform defined as "an open source software project that enables the distributed processing of large data sets across clusters of commodity servers"[15]. It is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance.

Rather than relying on high-end hardware, the resiliency of these clusters comes from the software's ability to detect and handle failures at the application layer.

Developed by Doug Cutting, Cloudera's Chief Architect and the Chairman of the Apache Software Foundation, Apache Hadoop was born out of necessity as data from the web exploded, and grew far beyond the ability of traditional systems to handle it. Hadoop was initially inspired by papers published by Google outlining its approach to handling an avalanche of data, and has since become the de facto standard for storing, processing and analyzing hundreds of terabytes, and even petabytes of data.

Apache Hadoop is 100% open source, and pioneered a fundamentally new way of

storing and processing data. Instead of relying on expensive, proprietary hardware and different systems to store and process data, Hadoop enables distributed parallel processing of huge amounts of data across inexpensive, industry-standard servers that both store and process the data, and can scale without limits.

In today's hyper-connected world where more and more data is being created every day, Hadoop's breakthrough advantages mean that businesses and organizations can now find value in data that was recently considered useless.

Hadoop can handle all types of data from disparate systems: structured, unstructured, log files, pictures, audio files, communications records, email - regardless of its native format. Even when different types of data have been stored in unrelated systems, it is possible to store it all into Hadoop cluster with no prior need for a schema.

By making all data useable, Hadoop provides the support to determine inedited relationships and reveal answers that have always been just out of reach.

In addition, Hadoop's cost advantages over legacy systems redefine the economics of data. Legacy systems, while fine for certain workloads, simply were not engineered with the needs of Big Data in mind and are far too expensive to be used for general purpose with today's largest data sets.

Apache Hadoop has two main subprojects:

- MapReduce - The framework that understands and assigns work to the nodes in a cluster. Has been defined by Google in 2004 and is able to distribute data workloads across thousands of nodes. It is based on "break problem up into smaller sub-problems" strategy and can be exposed via SQL and in SQL-based BI tools;

- Hadoop Distributed File System (HDFS) - An Apache open source distributed file system that spans all the nodes in a Hadoop cluster for data storage. It links together the file systems on many local nodes to make them into one big file system. HDFS assumes nodes will fail, so it achieves reliability by replicating data across multiple nodes

HDFS is expected to run on high-performance commodity hardware; it is known for highly scalable storage and automatic data replication across three nodes for fault tolerance. Furthermore, automatic data replication across three nodes eliminates need for backup (write once, read many times).

Hadoop is supplemented by an ecosystem of Apache projects, such as Pig, Hive and Zookeeper, that extend the value of Hadoop and improve its usability. Due to the cost-effectiveness, scalability and streamlined architectures, Hadoop changes the economics and the dynamics of large scale computing, having a remarkable influence

based on four salient characteristics. Hadoop enables a computing solution that is:

- Scalable: New nodes can be added as needed, and added without needing to change data formats, how data is loaded, how jobs are written, or the applications on top.

- Cost effective: Hadoop brings massively parallel computing to commodity servers. The result is a sizeable decrease in the cost per terabyte of storage, which in turn makes it affordable to model all your data.

- Flexible: Hadoop is schema-less, and can absorb any type of data, structured or not, from any number of sources. Data from multiple sources can be joined and aggregated in arbitrary ways enabling deeper analyses than any one system can provide.

- Fault tolerant: When you lose a node, the system redirects work to another location of the data and continues processing without missing a beat.

Text mining makes sense of text-rich information such as insurance claims, warranty claims, customer surveys, or the growing streams of customer comments on social networks.

Optimization helps retailers and consumer goods makers, among others, with tasks such as setting prices for the best possible balance of strong-yet-profitable sales. Forecasting is used by insurance companies, for example, to estimate exposure or losses in the event of a hurricane or flood.

Cost will certainly be a software selection factor as that's a big reason companies are adopting Hadoop; they're trying to retain and make use of all their data, and they're expecting cost savings over conventional relational databases when scaling out over hundreds of Terabytes or more. Sears, for example, has more than 2 petabytes of data on hand, and until it implemented Hadoop two years ago, Shelley says the company was constantly outgrowing databases and still couldn't store everything on one platform.

Once the application can run on Hadoop it will presumably be able to handle projects with even bigger and more varied data sets, and users will be able to quickly analyze new data sets without the delays associated with transforming data to meet a rigid, predefined data model as required in relational environments.

From architectural point of view, Hadoop consists of the Hadoop Common which provides access to the filesystems supported by Hadoop. The Hadoop Common package contains the necessary JAR files and scripts needed to start Hadoop. The package also provides source code, documentation, and a contribution section which includes projects from the Hadoop Community.

For effective scheduling of work, every Hadoop-compatible filesystem should provide location awareness: the

name of the rack (more precisely, of the network switch) where a worker node is. Hadoop applications can use this information to run work on the node where the data is, and, failing that, on the same rack/switch, reducing backbone traffic. The Hadoop Distributed File System (HDFS) uses this when replicating data, to try to keep different copies of the data on different racks. The goal is to reduce the impact of a rack power outage or switch failure so that even if these events occur, the data may still be readable.

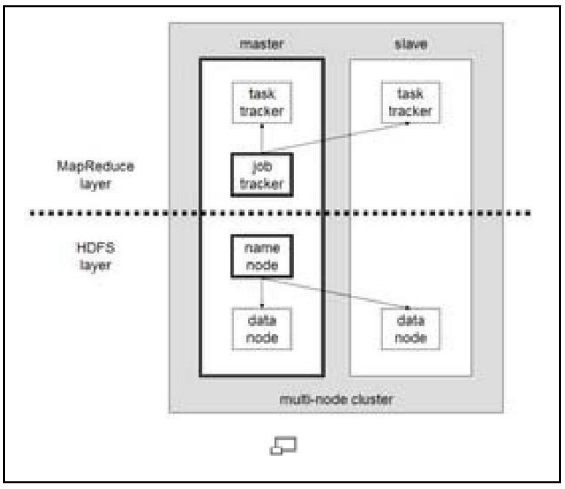

Fig 5. A multi-node Hadoop cluster[13]

As shown in Fig. 5, a small Hadoop cluster will include a single master and multiple worker nodes. The master node consists of a JobTracker, TaskTracker, NameNode, and DataNode.

A slave or worker node acts as both a DataNode and TaskTracker, though it is possible to have data-only worker nodes, and compute-only worker nodes; these are normally only used in non-standard applications.

Hadoop requires JRE 1.6 or higher. The standard startup and shutdown scripts require Secure Shell(SSH) to be set up between nodes in the cluster.

In a larger cluster, the HDFS is managed through a dedicated NameNode server to host the filesystem index, and a secondary NameNode that can generate snapshots of the namenode's memory structures, thus preventing filesystem corruption and reducing loss of data.

Similarly, a standalone JobTracker server can manage job scheduling.In clusters where the Hadoop MapReduce engine is deployed against an alternate filesystem, the NameNode, secondary NameNode and DataNode architecture of HDFS is replaced by the filesystem-specific equivalent.

One of the cost advantages of Hadoop is that because it relies in an internally redundant data structure and is deployed on industry standard servers rather than expensive specialized data storage systems, you can afford to store data not previously viable.

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 businesses more agile and to answer

questions that were previously considered beyond reach.

Enterprises who build their Big Data solution can afford to store literally all the data in their organization, and keep it all online for real-time interactive querying, business intelligence, analysis and visualization.

5. Conclusions

The year 2012 is the year when companies are starting to orient themselves towards the use of Big Data. That is why this articol presents the Big Data concept and the technologies associated in order to understand better the multiple beneficies of this new concept ant technology.

In the future we propose for our research to further investigate the practical advantages that can be gain through Hadoop.

References

[1] G. Noseworthy, Infographic: hic-big-flood-of-big-data-in-digital-marketing/

[2] H. Moed, The Evolution of Big Data as a Research and Scientific Topic: Overview of the Literature, 2012, ResearchTrends, http://www.researchtrends.com

[3] MIKE 2.0, Big Data Definition, http://mike2.openmethodology.org/wiki/Big_Data_Definition

[4] P. Zikipoulos, T. Deutsch, D. Deroos, Harness the Power of Big Data, 2012, http://www.ibmbigdatahub.com/blog/harness-power-big-data-book-excerpt

[5] Gartner, Big Data Definition, http://www.gartner.com/it-glossary/big-data/

[6]E. Dumhill, "What is big data?", 2012 ,http://strata.oreilly.com/2012/01/what-is-big-data.html

[7] A Navint Partners White Paper, "Why is BIG Data Important?" May 2012, http://www.navint.com/images/Big.Data.pdf

[8] Greenplum. A unified engine for RDBMS and Map Reduce, 2009. http://www.greenplum.com/resources/mapreduce/.

[9] For Big Data Analytics There's No Such Thing as Too Big The Compelling Economics and Technology of Big Data Computing, White Paper, March 2012, By: 4syth.com, Emerging big data thought leaders

[10] Big data: The next frontier for innovation, competition, and productivity. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. McKinsey Global Institute. May 2011

[11] Oracle Information Architecture: An Architect's Guide to Big Data, An Oracle White Paper in Enterprise Architecture August 2012

[12] Managing the Big Flood of Big Data in Digital Marketing,2012 http://analyzingmedia.com/2012/infograp

[13]http://www.oracle.com/us/corporate/press/1453796

[14]http://www.informationweek.com/softw