Outlooks of Big Data and its Analytics

Rajkumar. R
Assistant Professor
Department of BCA & MSc SS
Sri Krishna Arts & Science College, Tamil Nadu

Giri Pai. U
Student
Department of BCA & MSc SS
Sri Krishna Arts & Science College, Tamil Nadu

Gokul Balaji. S
Student
Department of BCA & MSc SS
Sri Krishna Arts & Science College, Tamil Nadu

Abstract

Today firms are starting to realize the importance of using more data in order to support decision for their strategies. It was stated and proved through study cases that "More data usually beats improved algorithms". With this statement firms started to realize that they can select to invest more in processing bigger sets of data rather than investing in expensive algorithms. The big measure of data is improved used as a whole because of the possible connections on a bigger amount, connections that can never be found if the data is analysed on separate sets or on a smaller set. A bigger amount of data gives an improved output but also working with it can become a challenge due to processing limitations. This article proposes to define the concept of Big Data and stress the importance of Big Data Analytics.

Keywords: Big Data and its Analytics, Database, Internet, Hadoop scheme

I. INTRODUCTION

Nowadays the Internet characterizes a big space where boundless amounts of data are added every day. The IBM Big Data Overflow Infographic shows that 2.7 Zettabytes of data occur in the digital universe today. Also according to this study there are 100 Terabytes efficient daily through Facebook, and a lot of commotion on social networks this leading to an estimation of 35 Zettabytes of data generated per annum by 2020. Just to have an idea of the amount of data being made, one zettabyte (ZB) eguals 1021 bytes, meaning 1012 GB. [1]

We can subordinate the importance of Big Data and Big Data Analysis with the culture that we live in. Today we are living in a Dataal Society and we are affecting towards a Knowledge Based Society. In order to extract sound knowledge, we need a bigger amount of data. The Society of Data is a society where data plays a major role in the economic, cultural and political phase.

In the Knowledge society the inexpensive profit is grown through understanding the data and guessing the evolution of facts based on data. The same happens with Big Data. Every organization needs to gather a big set of data in order to support its decision and extract relationships through data analysis as a base for decisions.

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

II. BIG DATA CONCEPT

The term "Big Data" was first presented by Roger Magoulas from O'Reilly media in 2005 to the computing world in order to define a boundless amount of data that traditional data achievement techniques cannot achieve and course due to the complexity and size of the data.

A study on the Development of Big Data as a Research and Scientific Topic shows that the term "Big Data" was existing in research starting with 1970s but has been embraced in publications in 2008. [2] Nowadays the Big Data concept is preserved from different points of view casing its consequences in many fields.

According to MiKE 2.0, the open source typical for Data Achievement, Big Data is defined by its size, including a big, complex and independent group of data sets, each with the potential to relate. In addition, a vital feature of Big Data is the fact that it cannot be handled with normal data achievement practices due to the inconsistency and randomness of the possible combinations. [3]
In IBM's view Big Data has four features:

1) Volume: refers to the measure of data gathered by a company. This data must be used further to obtain vital knowledge;
2) Velocity: refers to the time in which Big Data is processed. Some activities are very vital and need immediate replies, that is why fast processing maximizes effectiveness;
3) Variety: Refers to the type of data that Big Data can embrace. This data can be structured as well as can be unstructured;
4) Veracity: refers to the degree in which a leader trusts the used data in order to take decision. So getting the right connections in Big Data is very vital for the firm future. [4]

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

According to Ed Dumbill chair at the O'Reilly Strata Conference, Big Data can be defined 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, we must choose an alternative way to process it." [6] In a simpler definition we consider Big Data to be an expression that embraces different data sets of very big, highly complex, unstructured, organized, stored and processed using definite methods and techniques used for firm processes.

There are many definitions on Big Data circulating around the world, but we consider that the most vital one is the one that each leader gives to its one firm's data. The way that Big Data is demarcated has implication in the policy of a firm. Each leader has to describe the concept in order to bring competitive profit for the company.

A. The importance of Big Data

The key importance of Big Data consists in the potential to improve effectiveness in the context of use a big volume of data, of different type. If Big Data is defined properly and used accordingly, organizations can get an improved view on their firm therefore leading to effectiveness in different areas like sales, improving the manufactured artefact and so forth.
Big Data can be used effectively in the following areas:

- In data technology in order to improve security and troubleshooting by analysing the designs in the existing logs;
- In client service by using data from call centres in order to get the client pattern and thus enhance client satisfaction by adapting services;
- In improving services and artefacts through the use of social media content. By knowing the potential client's preferences, the company can modify its artefact in order to address a bigger area of people;
- In the detection of fraud in the online dealings for any industry;
- In risk assessment by analysing data from the transactions on the financial market.

In the future we propose to analyse the potential of Big Data and the power that can be allowed through Big Data Analysis.

B. Big Data challenges

The *understanding of Big Data* is keenly very vital. In order to determine the best stratagem for a firm it is essential that the data that you are counting on must be properly analysed. Also the time span of this analysis is vital because some of them need to be performed very frequent in order to determine fast any change in the firm environment.

Another feature is represented by the *new technologies* that are developed every day. Considering the feature that Big Data is new to the organizations nowadays, it is necessary for these organizations to learn how to use the newly developed technologies as soon as they are on the market. This is a vital feature that is going to bring competitive profit to a firm.

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 limit for innovation, there is a need for up to 190,000 more workers with analytical capskill and 1.5 million more data literate achievers only in the United States. These statistics are a proof that in order for a company to take the Big Data initiative has to either hire experts or train existing workers on the new field.

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

In our judgement for an organization to get competitive profit from the manipulation of Big Data it has to take very moral care of all issues when implementing it. One option of developing a Big Data policy is presented below. In addition, in order to bring filled capabilities to Big Data each company has to take into deliberation its own typical firm characteristics.

C. Big Data Analytics

The world today is constructed on the foundations of data. Lives today are obstructed by the skill of the firms to dispose, interrogate and achieve data. The development of technology infrastructure is adapted to help generate data, so that all the offered services can be better as they are used.

As an example, internet today became a huge data-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 petabytes, Exabyte, zettabytes and yottabytes.

In order to achieve the giant volume of unstructured data stored, it has been emerged the "Big Data" phenomena. It stands to aim that in the profitable sector Big-Data has been adopted more fast in data obsessed industries, such as economic services and telecommunications, which it can be claimed, have been experiencing a more rapid growth in data sizes compared to other market sectors, in addition to tighter regulatory requirements and falling profitskill. At first, Big Data was seen as a mean to succeed to reduce the costs of data achievement. Now, the firms focus on the value formation potential. In order to profit from additional insight gained there is the need to assess the analytical and execution capabilities of "Big Data".

To bring big data into a firm profit, firms have to review the way they achieve data within data centre. The data is taken from a horde of sources, both from within and without the organization. It can contain content from videos, social data, documents and machine-generated data, from a variety of applications and stages. Firms need a scheme that is optimised for acquiring, organising and loading this shapeless data into their databases so that it can be effectively rendered and analysed. Data analysis needs to be profound and it needs to be quick and conducted with firm goals in mind.

The scalskill of big data solutions within data centres is a crucial consideration. Data is vast nowadays, and it is only going to get bigger. If a data centre can only manage with the levels of data expected in the short to medium term, firmes will quickly spend on system refreshes and upgrades. Forward scheduling and scalskill are therefore vital.

In order to make every decision as anticipated there is the need to bring the results of knowledge discovery to the firm 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 firm policy, CRM policy and financial policy going forward.

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

Big Data Achievement is based on capturing and organizing relevant data. Data analytics supposes to understand that happened, why and predict what will happen. A unfathomableer analytics means new analytical methods for unfathomableer insights. [9]

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

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

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

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

In terms of Big Data Achievement and analytics Oracle is offering Engineered Systems as Big Data Solutions (Fig.3), such as Oracle Big Data Appliance, Oracle Exudate and Oracle Analytics. Big Data solutions combine best tools for each part of the problem. The traditional firm 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 data 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 profits are that the firm intelligence investments and skills that are leveraged, there are made insights from Big Data consumable for firm users, there are combined Big Data with Application and OLTP data. [11]

Big Data offers many opportunities for unfathomable insights via data mining:

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

"Big Data" is a Data Achievement & Analytics market opportunity obsessed by new market requirements. In-Database Analytics – Data Mining there are used Big Data Connectors to combine Hadoop and DBMS data for unfathomable analytics. Also there is the need to re-use SQL skills to apply unfathomable 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 prognostic learning of relationships between knowledge concepts and firm events is done. [12]

Big-Data presents an important opportunity to create new value from giant data. It is vital to determine appropriate governance procedures in order to achieve development and the operations over the life of the technology and data. Failure to consider the longer term implications of development will lead to reactivity issues and cost escalations.

On the face of it, the cost of physically storing big 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 firm 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 composition 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 tragedy is to be avoided. Map-Reduce operations also generate a lot of network chatter so a fast private network is optional. These requirements combine to add important cost to a rarefaction cluster used in a profitable setting. [8]
- Compression capabilities in Hadoop are restricted because of the HDFS block structure and require an understanding of the data and compression technology to contrivance adding to the operation complexity with limited impact on storage volumes.

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

Whatever the true cost of Big-Data compared to an interpersonal data storage approach, it is vital that the development of Big-Data policy is consciously done, understanding the true nature of the costs and complexity of the infrastructure, practice and actions that are put in place.

D. Big Data Analytics Software

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

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

"Legacy systems will rekey necessary for specific high-value, low capacity workloads, and compliment the use of Hadoop - optimizing the data achievement structure in the organization by pushing the right Big Data workloads in the right systems" [14].

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

- Volume: Enterprises are awash with ever-growing data of all types, easily amassing terabytes - even petabytes - of data (e.g. turn 12 terabytes of Tweets created each day into improved artefact sentiment analysis; convert 350 billion annual meter readings to improved predict power consumption);
- Velocity: For time-sensitive procedures 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 data (e.g. scrutinize 5 million trade events created each day to identify potential fraud; analyse 500 million daily call detail records in real-time to predict client 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 feature for problems, situations etc. (e.g. monitor 100's of live video feeds from surveillance cameras to target points of interest; feat the 80% data growth in images, video and documents to improve client satisfaction);
- Veracity: Since one of three firm leaders don't trust the data they use to make decisions, establishing trust in big data presents an enormous challenge as the variety and number of sources grows.

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

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

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

Apache Hadoop is 100% open source, and founded a fundamentally new way of storing and processing data. Instead of relying on expensive, branded hardware and different systems to store and process data, Hadoop enables distributed parallel processing of enormous amounts of data across inexpensive, industry-standard servers that both store and process the data, and can scale without limits.

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

Hadoop can handle all types of data from dissimilar systems: structured, unstructured, log files, pictures, audio files, communications records, email - regardless of its innate 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 previous need for a schema.

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

In addition, Hadoop's cost profits over legacy systems redefine the economics of data. Legacy systems, while fine for certain workloads, only 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 biggest data sets.

Apache Hadoop has two key sub schemes:

- MapReduce - The framework that understands and allocates 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" policy and can be visible 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 numerous local nodes to make them into one big file system. HDFS assumes nodes will fail, so it achieves reskill by duplicating 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 schemes, such as Pig, Hive and Zookeeper that extend the value of Hadoop and improve its upskill. Due to the cost-effectiveness, scale skill and streamlined architectures, Hadoop changes the economics and the dynamics of big scale computing, having a remarkable influence based on four salient characteristics. Hadoop enables a computing solution that is:

- Scalable: New nodes can be added if required, and added without needing to change data formats, how data is laden, how jobs are written, or the applications on top.
- Cost effective: Hadoop brings massively parallel computing to product servers. The result is a sizeable decrease in the cost per terabyte of storage, which in turn makes it inexpensive 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 unfathomable 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 data such as insurance claims, guarantee claims, client surveys, or the growing streams of client remarks on social networks.

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

Cost will surely be a software selection factor as that's a big reason firms are adopting Hadoop; they're trying to recall and make use of all their data, and they're expecting cost savings over more conventionally 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 applied Hadoop two years ago, Shelley says the company was constantly outgrowing databases and still couldn't store everything on one stand.

Once the application can run on Hadoop it will presumably be able to handle schemes with even bigger and more mixed data sets, and users will be able to quickly analyse new data sets without the delays associated with converting data to meet a stiff, predefined data model as required in relational environments.

From architectural point of view, Hadoop consists of the Hadoop Common which offers 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 offers source code, documentation, and a contribution section which includes schemes from the Hadoop Community.

For effective forecasting of work, every Hadoop-compatible filesystem should provide location awareness: the name of the rack where a worker node is. Hadoop applications can use this data 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 line 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.

A small Hadoop cluster will include only one master and multiple worker nodes. The master node consists of a Job Tracker, Task Tracker, Name Node, and Data Node.

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

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

In a bigger cluster, the HDFS is achieved through a dedicated Name Node server to host the filesystem index, and a secondary Name Node that can generate snapshots of the name node's memory structures, thus preventing filesystem corruption and reducing loss of data.

Similarly, a standalone Job Tracker server can achieve job scheduling.

In clusters where the Hadoop MapReduce engine is deployed against an alternate filesystem, the Name Node, secondary Name Node and Data Node architecture of HDFS is replaced by the filesystem-specific equivalent.

One of the cost profits 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 firms 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, firm intelligence, analysis and visualization.

III. CONCLUSIONS

The year 2012 is the year when firms are starting to orient themselves towards the use of Big Data. That is why this article presents the Big Data concept and the technologies associated in order to understand improved numerous benefices of this new idea and technology. In the future we suggest for our research to further study the practical profits that can be gain through Hadoop.

REFERENCES

[1] G. Noseworthy, Infographic: Managing the Big Flood of Big Data in Digital Marketing, 2012 http://analyzingmedia.com/2012/infographic-big-flood-of-big-data-in-digitalmarketing/
[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/harnes s-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-isbig-data.html
[7] A Navint Partners White Paper, "Why is BIG Data Vital?" 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/mapre duce/.
[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 artefactivity. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. McKinsey Global Institute. May 2011
[11] Oracle Data Architecture: An Architect's Guide to Big Data, An Oracle White Paper in Enterprise Architecture August 2012
[12] http://bigdataarchitecture.com/
[13] http://www.oracle.com/us/corporate/pre ss/1453796
[14] http://www.informationweek.com/softw Database Systems Journal vol. III, no. 4/2012-13 are/business-intelligence/sas-gets-hip-tohadoop-for-big-data/ 240009035?pgno=2
[15] http://en.wikipedia.org/wiki/Apache_ Hadoop