Big Data: present and future

Mircea Răducu TRIFU, Mihaela Laura IVAN
University of Economic Studies, Bucharest, Romania
trifumircearadu@yahoo.com, ivanmihaela88@gmail.com

The paper explains the importance of the Big Data concept, a concept that even now, after years of development, is for the most companies just a cool keyword. The paper also describes the level of the actual big data development and the things it can do, and also the things that can be done in the near future.

The paper focuses on explaining to nontechnical and non-database related technical specialists what basically is big data, presents the three most important V's, as well as the new ones, the most important solutions used by companies like Google or Amazon, as well as some interesting perceptions based on this subject

Keywords: *Big Data, domains, risks, resources, information*

1 Introduction

Today world is totally and continuously connected to at least a big data informatics system, from support services, social network or a service that provide GPS localization. Every time you connect the phone to internet, or even pay with a card to get a soda you leave behind yourself traces full of information, using those traces you leave behind, any marketing department can and will know what is your destination and what are your habits.

Think yourself like a dear in the woods, you have many need and you what to satisfy them, you do so using at least one informatics tool, so when you search for the perfect bar it is like you shout in the woods and explore for the near water source. And like in any woods, there are some kind of friends, or a smells or sounds that can help you to find the best water source, in the real word that friend is can be any map provider, and the smell of the water is represented by any commercial you see.

The Big Data concept represents a in essence an "ocean of data" [5], lots of information and the means to analyses them.

In the present days most of the humans can access more information than most of our ancestors in a lifetime. Nowadays we double, in every year, the amount of data that we create. According to the global market intelligence firm IDC [5], in 2011 we played, swam, wallowed, and drowned in 1.8 zeta bytes of data.

Big data has many characteristics that made this term unique, big data is a concept that integrates all kinds of data, not just some basic ones like in a normal data warehouse, from text to pictures, sounds, movies, music, satellite coordinates and basically all kinds of input or output data that came from different types of sensors

Comparing with the actual flow of data, even Alexandria great library could be held at most 70.000 scrolls.

There are many ways to describe the concept. You can define it like the ability to extract meaning from this "ocean of data": to sort through masses of data and to find the hidden patterns and unexpected correlations.

The key of big data is not just to overflow servers with data, but use different types of algorithms that can use text and graphical

2. Defining Big Data

Forbes defines big data like this:

Table 1. Business Intelligence VS BigData

Traditional BI	**Reporting Big Data Analysis**
Reporting tool like Cognos, SAS, SSIS, SSAS	Visualization tool like QlikView or Tableau
Sample data or specific historical data	Huge volume of data
Data from the enterprise	Data from external sources like social media apart from enterprise data
Based on statistics	Based on statistics and social sentiment analysis or other data sources
Data warehouse and data mart	OLTP, real-time as well as offline data
Sequential Computing	Parallel Computing using multiple machines
Query languages SQL, TSQL	Scripting Languages Java script, Python, Ruby and SQL
Specific type of data : txt, xml, xls, etc	Multiple kinds of data : pictures, sounds, text, map coordinates

“Big data is a collection of data from traditional and digital sources inside and outside your company that represents a source of ongoing discovery and analysis” [5] (Fig. 1)

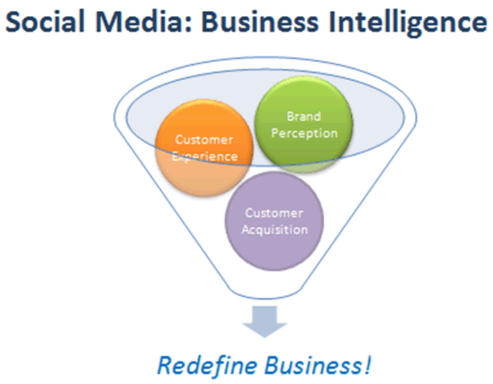

Fig. 1. Big data usage in BI

This is a more business approach, more practical and more business orientated than other definitions.

Big data is a mix of unstructured and multi structured data, those types of data are analyzed together to get more knowledge and information to company than could be get using the usual methods and infrastructure.

Unstructured data is information that is not organized or easily interpreted by traditional *data models* or *databases*, and usually is text-heavy. Good examples are posts from twitter, LinkedIn and other social media services. [5]

Multi-structured data is represented by a variety of data formats that came from interaction between peoples and machines, such as web applications and social services. Those include text and multimedia formats, like photos and videos, with structured data. [5]

Maybe the most use approach in defining Big Data is the one that was made by Gartner in 2001.According to Gartner Big data *is high volume*, *high velocity*, and/or high *variety* information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization. This approach will be discussed in the next chapter [2].

3. Big data dimensions

The three V's of Gartners definition are: Volume, velocity and variety [9].

Volume: big data is that "Ocean of data" that we talk about in the rows above. It Is represented bay information that can came from every possible sensor, and some even say that we people are also sensors and data gatherers for *big data* [9].

The challenges of having such a big quantity of data is that is very hard to sustain it, to store it, to analyze it and ultimately to use it.

Velocity: is all about the speed of data traveling from one point to another and the speed of processing it. Sometimes it is crucial for the manager to be able to decide in a very little time on a variety of issues [9].

The most important issue is that the resources that analyses data is limited compared to the *volume* of data, but the requests of information is unlimited and usually information gets through at least one bottleneck.

Variety, the third characteristic is represented by the types of data that are stored. Because there are many types of sensors and sources, the data that came from them is varying very much in size and type. It is very complicated to analyze text, images and sounds in the same context and get a result that can be relied on.

And then is the issue of dark data, data that sits in the organization and is unused and also is not free.

There are one new dimension that were added to the existing ones: *Veracity* (Fig. 2)

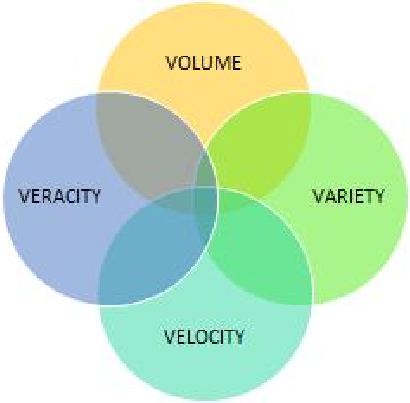

Fig. 2. Big data four V's

Veracity is the hardest thing to achieve with big data, because due to the *Volume* of information and the *variety* of its type is hard to identify the useful and accurate data from the "dirty data". The biggest problem is that the "dirty data" can lead very easy to an avalanche of errors, incorrect results and can affect the *Velocity* attribute of Big Data. The main purpose of the Big Data can be corrupted and all the information can lead to a useless and very expensive Big Data environment if there is not a good cleaning team.

The *Veracity* attribute is in its self also an objective for the Big Data developers. If the data cannot be accurate, is redundant or is unreliable, the whole Company can have a big problem, especial the companies that use big data to sell information like the marketing ones, or the ones that make market studies.

A lot of social media responses to campaigns could be coming from a small number of disgruntled past employees or persons employed by competition to post negative comments.

4. Big Data infrastructure

For a medium size or even big company that is not necessary making a living from renting space and processing power to clients, the construction of a Big Data infrastructure is often as expensive as inefficient.

So Big Data is not the answer for every type o company, is very expensive and hard do make on your own, and you need a specialized human resources. In the current labor market the Big Data specialists are very few and also the means to train programmers, architects and business analysts are few and very expensive.

The architecture of a Big Data solution is rather different from other data storage solution like Data Warehouse. The difference

It is basically represented by the four V's that characterizes the concept of big data.

One of the most important roles in the Big Data infrastructure is the NoSql Databases.

NoSql Databases

The term it means "Not Only SQL" rather than "No SQL", and it represents in essence a different kind of database approach, where the databases are not build with the relational databases structure, but use wide column store, document, key value structures or other types of structure that often are more easy to manage, and customize.

MongoDB is a document orientated, based on JSON, database that can handle large number of data sets with a low maintenance and that is easy to work with. [4]

Cassandra was originally a Facebook project, and after it was release as open source. It's one of the most important solutions and it has a huge community support. [4]

Cassandra is key and column orientated and is in many ways similar to the classic databases. It is also very close to the Google's Big Table, offering column indexes, strong support for denormalization and materialized views. [10]

BigTable is the solution used by Google, it is defined like a distributed store system used for managing structured data that is designed to a very large scale. [8]

BigTable achieves several goals some of them are wide applicability, scalability, high performance and high availability, and it is used today in more than sixty Google projects, like Google Analytics, Google Finance, Orkut, Personalized Search and Google Earth [8].

As a data model, Bigtable uses a sparse, distributed and persistent multidimensional sorted map. This map is indexed by row key, column key and timestamp so that every value in the map is an uninterpreted array of bytes. [8]

HBase is designed as an open soured clone to the BigTable, and is very similar in most of its models and designs, supports the same data structures tables.

HBase is integrated in the Hadoop project, so is easy to work using the database from a Map Reduce job.

MapReduce model

It is a programming model that has the purpose to process large data sets in parallel. [11]

The MapReduce model is using a pipeline that reads and writes to arbitrary file formats, with intermediate results been passed between stages as files, using computational spread across many machines, unlike the relational tables where all processing happens after the information has been loaded into a store, using specialized query language.[11]

Hadoop

It is a MapReduce system developed by Yahoo after the Google's MapReduce infrastructure (Fig. 3).

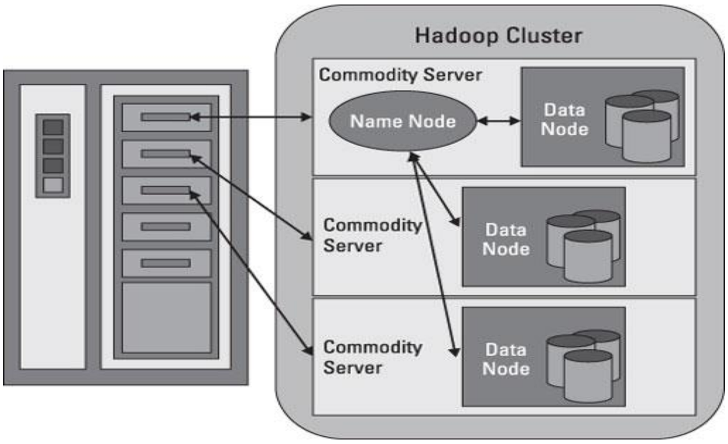

Fig.3. Hadoop Cluster [2]

To fully understand the capabilities of Hadoop MapReduce, we need to differentiate between MapReduce (the algorithm) and an implementation of MapReduce. Hadoop MapReduce is an implementation of the algorithm developed and maintained by the Apache Hadoop project. It is helpful to think about this implementation as a MapReduce engine, because that is exactly how it works. You provide input (fuel), the engine converts the input into output quickly and efficiently, and you get the answers you need.

Big data brings the big challenges of volume, velocity, and variety. As covered in the previous sections, HDFS resolves these challenges by breaking files into related collections of smaller blocks. These blocks are distributed among the data nodes in the HDFS cluster and then are managed by the Name Node. Block sizes are configurable and are usually 128 megabytes (MB) or 256MB, meaning that a 1GB file consumes eight 128MB blocks for its basic storage needs. HDFS is resilient, so these blocks are replicated throughout the cluster in case of a server failure. [2]

The HDFS keep track of all the pieces using metadata. The HDFS metadata provide detailed information like the following:

- Date of the creation, modification, execution of a file;
- Date where the blocks of the file are stored on the cluster;
- The rights to view or modify the file;
- Number many files are stored in one cluster;
- Number many data nodes exists in the cluster;
- Location address of the transaction log of the cluster.

JSON, we cannot continue the exemplification of some of the technologies that are used in the Big Data infrastructure without presenting one of the most popular formats for data processing.

Most of its popularity came from the easiness of reading and writing of both humans and machines. It is based on a JavaScript and is built on two structures: a collection of name/value pairs and an ordered list of values.

Big data infrastructure has multiple levels according to Michael Driscoll [3], who wrote about stack level of big data as can be seen in Fig. 4.

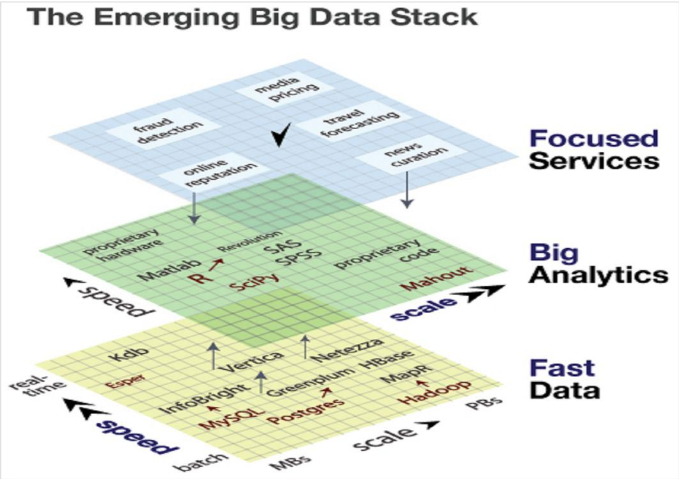

Fig. 4. Emerging big data stack [3]

Fast Data: At the base of the big data stack, where data is stored, processed, and queried the dominant axis of competition was once scale. But as cheaper commodity disks and Hadoop have effectively addressed scalable persistence and processing, the focus of competition has shifted toward speed. [3]

Big analytics: At the second tier of the big data stack, analytics is the brain to cloud computing. Here, however, the speed is less of a challenge; given an addressable data set in memory, most statistical algorithms can yield results in seconds. The challenge is scaling these out to address large datasets, and rewriting algorithms to operate in an online, distributed manner across many machines.

Focused services: The top of the big data stack is where data products and services directly touch consumers and businesses. For data start-ups, these offerings more frequently take the form of a service, offered as an API rather than a bundle of bits [3].

5. Uses of Big data

Scientific research has been revolutionized by Big Data. The Sloan Digital Sky Survey has today become a central resource for astronomers the world over. The field of Astronomy is being transformed from one where taking pictures of the sky was a large part of an astronomer's job to one where the pictures are all in a database already and the astronomer's task is to find interesting objects and phenomena in the database

Big Data has the potential to revolutionize not just research, but also education. A recent detailed quantitative comparison of different approaches taken by 35 charter schools in NYC has found that one of the top five policies correlated with measurable academic effectiveness was the use of data to guide instruction. Imagine a world in which we have access to a huge database where we collect every detailed measure of every student's academic performance. This data could be used to design the most effective approaches to education, starting from reading, writing, and math, to advanced, college-level, courses. We are far from having access to such data, but there are powerful trends in this direction. In particular, there is a strong trend for massive Web deployment of educational activities, and this will generate an increasingly large amount of detailed data about students' performance.

Big Data in Medical industry

In the medical industry we can see big changes represented by the way data is now used not only to sell better medicines, or to make more profit but to increase the population access to hospitals and medical help as well. [5]

There are more and more cases where people like Yasmine Delwari Johnson, decides to take medical testes to find out what will be the characteristics of their child. They can now find out what are the odds that her child will have green eyes, or if he has big chances to develop lactose intolerance. [5]

Another use in the medical industry is to predict the medical condition of the patients based on their medical records, or their family.

Big Data Marketing

Is a process of collecting, analyzing, and executing on insights you've derived from big data to encourage customer engagement, improve marketing results, and measure internal account-ability.

Companies are focused on harnessing new data types and utilizing data to drive customer experience. Social media is driving most text analytics initiatives:

43% of respondents expected to focus budget on "Customer data integration"

44% expected to focus budget on "Social media monitoring software" [5]

Future focus is on improving online customer experience.

77% of respondents stated "Improving online Customer Experience" as major objective for 2012 (Fig. 5).

Fig.5. Customer Relationship Management

74% stated "improving cross-channel customer experience" as a major objective

As we grapple with the consumption challenges presented by this deluge of data, new publishing platforms are also empowering us to gather, refine, analyze and share data ourselves, turning it into information.

6. Future of Big Data

Clearly Big Data is in its beginnings, and is much more to be discovered. Now is for the most companies just a cool keyword, because it has a great potential and not many truly know what all is about.

A clear sign that there is more to big data then is currently shown on the market, is that the big software companies do not have, or do not present their Big Data solutions, and those that have like Google, does not use it in ca commercial way.

The companies need to decide what kind of strategy use to implement Big Data. They could use a more *revolutionary* approach and move all the data to the new Big Data environment, and all the reporting, modeling and interrogation will be executed using the new business intelligence based on Big Data. [1]

This approach is already used by many analytics driven organizations that puts all the data on the Hadoop environment and build business intelligence solutions on top of it.

Another approach is the *evolutionary* approach; Big Data becomes an input to the current BI platform. The data is accumulated and analyzed using structured and unstructured tools, and the results are sent to the data warehouse. Standard modeling and reporting tools now have access to social media sentiments, usage records, and other processed Big Data items. [1] One of the issues of the *evolutionary* approach is that even if it gets most of the capabilities of the Big Data environment, but also gets most of the problems of the classic Business intelligence solution, and in some cases can create a bottleneck between information that came from the Big Data and the power to analyze of the traditional BI or data warehouse solution

Another approach is the *hybrid* one, where some types of data are analyzed by the Big Data and other by the traditional BI Solutions.

One of the solutions that are now available is the Hana solution from SAP.

The work done in the real-time analytics, in-memory database, with memory becoming cheaper, multi-core architecture which allow parallel processing, compression techniques useful to keep more data in less memory, column and row storage being able to access data at amazing speed and real-time replication capability of data is the role of SAP HANA. These capabilities are illustrated in Fig. 6. [6]

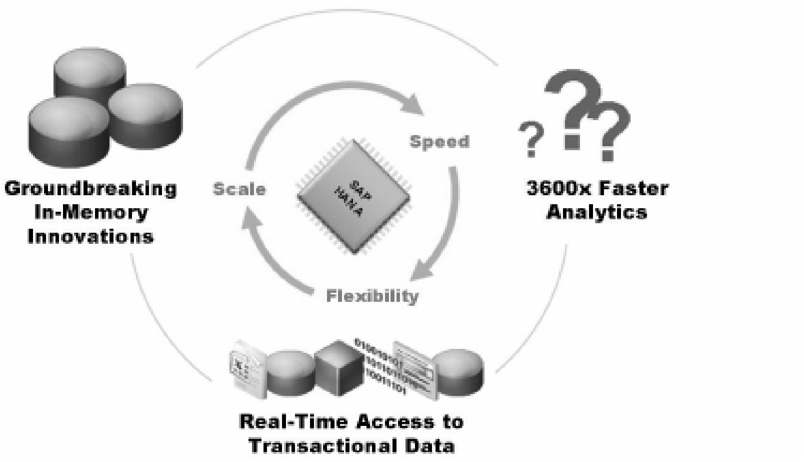
Groundbreaking In-Memory Innovations
Scale
Speed
SAP HANA
Flexibility
3600x Faster Analytics
Real-Time Access to Transactional Data

Fig. 6. The key capabilities of SAP HANA [6]

This technology itself brings business benefits by being leveraged across domains like Big Data, Business Intelligence and Analytics. These business benefits are:

- Speed and Accelerated performance: good query performance for improved decision making, boost of performance for data load processes for a low data latency, accelerated memory planning capabilities [7];
- New Business Insights: Self-Service BI and more flexible modeling capabilities;
- Faster Business Processes;
- Simpler Business Interaction empowering the decision maker to take action in a very short time [7].

SAP HANA as a database would bring performance with the use of in-memory database functionality. The role of SAP HANA in the future of big data is to move all the performance processing planning functions into the in-memory database. This enables planners to:

- Use more data for planning, in this case planning runs can be done daily or for multiple years [12];
- Plan at a detailed level of granularity. For example: Demand for products can be requested at a material and variant level than a product group level. Integrate planning activities across functions [12].

As challenges of large volumes of data appeared and more data in systems are requested, in-memory feature address some business drivers like:

- Manage big data and complexity in an efficient way [6];
- Remove constraints when analyzing big data - trends, data

mining or predictive analytics;
- Allow simulation capabilities for different solutions to do the best choose [6].

Accordingly, big data can be efficient optimized with the help of this revolutionary SAP Hana. One of the main reasons that SAP Hana appeared is big data requirements, which means large transaction volume.

7. Conclusions

In the present Big Data is at the beginning, is very rarely implemented in companies, most of them are getting there even if they don't realized it yet.

There are many definitions and visions about big data and there is no completely accepted one that everybody can agree on.

As explain earlier, the approach of Gartners Company is the mod popular, but not everyone agree with it because is more business orientated and less visionary.

Many imagine the future of big data like the central nervous system of the planet, with, for which the people are its sensors.

One thing all agree, the Big data concept is one that will revolutionize all businesses, because of the information it holds and the capability to interpret and analyze even the most volatile and non-related data.

I think that Big Data is the solution for many of the world problems, and is now been born to the most productive time in our history. Now is the time when the technology is developing so fast that even a super computer like the ones that are used by Google or Amazon, can be loan or bough at a lesser price than before.

Another advantage of the today technology level is the information that is free on the internet, or the multitude of sensors or companies that have those sensors that can capture valuable information about almost everything that is happening in the world.

The most important issue today is represented by the small number of Big Data specialists. There are very few courses and trainings available in the market, most of them are hold between close doors and most of them are very expensive.

Another issue in the process of training a specialist is the level of knowledge and know-how that is required. So there will be no IT interns learning about Big Data very soon.

In any case I think that Big Data is the next big thing, not only in IT market, but in the life an activity of every company.

References

[1] Dr. Arvind Sathi, *Big Data Analytics*, Ed. Distributive Technologies for Changing the Game, USA 2012.

[2] Judith Hurwitz, Alana Nugent, Dr. Fern Halper, Marcia Kaufman, *Big Data for Dummies, John Wiley & Sons Inc,* USA 2013

[3] Michael Driscoll *Big Data Now*, Ed. O'Reilly, USA 2012.

[4] Pete Warden, *Big Data Glossary*, Ed. O'Reilly, USA 2012.

[5] Rick Smolan, Jennifer Erwit, *The Human face of Big Data*, Ed. Against all odds production, Sausalito, CA 2012.

[6] SAP AG or an SAP affiliate company, SAP HANA Introduction, Participant Handbook, 2013

[7] SAP AG or an SAP affiliate company, SAP HANA Implementation and Modeling, Participant Handbook, 2013.

[8] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E Gruber, *Bigtable: A distributed Storage System for Structured Data,* Google Inc, http://static.googleusercontent.com/media/research.google.com/en//archive/bigtable-osdi06.pdf

[9] http://www.forbes.com/sites/lisaarthur/2013/08/15/what-is-big-data/

[10] https://cassandra.apache.org/#tab-presentation

[11] http://en.wikipedia.org/wiki/MapReduce

[12] DC READINESS SAP HANA. Building a Trusted SAP HANA Data

Center. Internet: http://www.cisco.com/assets/events/i/sapte-hana_whitepaper.pdf, September 2013.

Mircea Răducu TRIFU graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Bucharest University of Economic Studies in 2011 and the Informatics Security Master in 2013. He also finished the Faculty of Management in 2009, and the Master in Business Administration in 2011, both at the Bucharest University of Economic Studies. At present he is a System Support at Data warehouse team in the department of the Application Development of the ING Bucharest.

Mihaela Laura IVAN graduated from the Faculty of Cybernetics, Statistics and Economic Informatics of the Bucharest University of Economic Studies in 2011. She also finished the Master's degree in Economic Informatics in 2013, at the Bucharest University of Economic Studies. At the present she is a SAP Development Consultant at SAP Near Shore Center Romania.