

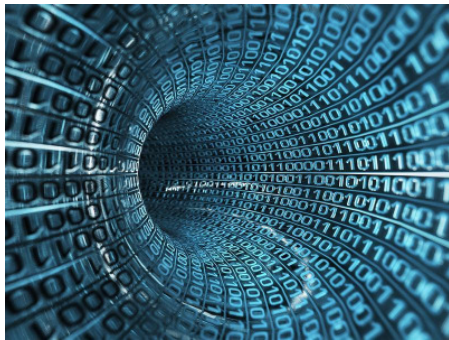
Committee/Council: Youth Assembly

Issue: Big Data

Head: Vasiliki (Vanessa) Pasalidi-Chantzi

Co-Head: Kyriaki (Sandy) Papadimitriou

Introduction

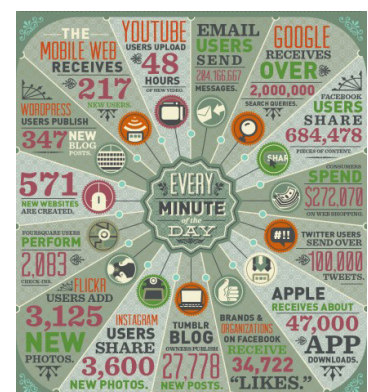


“Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone.” According to IBM, data, ranging from that collected for climate information to that accrued in multinational companies, has been accumulating very rapidly during the last years. Therefore, new technologies are expected not only to handle large amounts of data, but also to guarantee and purvey storage facilities for it. This approach is incumbent and the only solution to the issue of the amassing portions of big data.

In this conference, delegates will be asked to debate on the issue of Big Data, which divulges and entangles more demanding and complex problematics. They will be encouraged to express their ideas on new technologies, the appropriate policies monitoring the data storage process, as well as the ensuing consequences, in the instance that they apply. Furthermore, they will be able to expand their global knowledge on such crucial an issue and obliged to answer fortuitous questions that were previously considered beyond their reach.

Definition of Key-Terms

Various definitions have been provided for the term big data. Amidst the numerous definitions stipulated, the most upfront, direct and reliable are the ones of IBM (International Business Machines Corporation), SAS (Statistical Analysis System), The Meta Group and Gartner, Inc. . The concept of Big Data has its roots back in the 1980s, but it was in 2001 when Doug Laney of

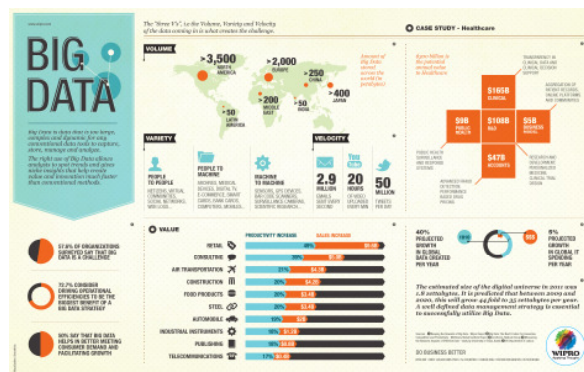


Meta Group clearly defined big data as data sets where the *three Vs*—volume, velocity and variety—demonstrate distinguishing peculiar challenges in regulating and organizing them.

In the definition provided by Doug Laney, a Gartner analyst, in compliance with those of IBM and SAS, three main characteristics of Big Data are examined, namely its volume, velocity, and variety (the three V's). The **volume** of data of all types amassing at such rapid rates indicates its vast magnitude, which ranges from thousands of terabytes to petabytes (1,024 terabytes each) of information. For instance, it has been estimated that 12 terabytes of Tweets are created every day. Due to the enormous volume of data the question raised is how to distinguish and ascertain relevance amongst them. **Velocity** is a very crucial aspect of data, and, according to Gartner, often the most

misunderstood, as it determines both the speed with which data is created and the speed with which data is processed in order to meet the requirements. The rate of data production has been increasing dramatically; for example, “each minute, 48 hours of video are uploaded to YouTube, Twitter users send 100,000 tweets and

Instagram users share 3,600 photos.” As far as the second aspect of velocity is concerned, it empowers data’s management. To cite an instance, Big Data may be used in identifying and impeding fraud, which requires the swift scrutinization of the information. Finally, **variety** of data is another feature signifying that nowadays data can take the form of all types of format, spanning every dimension, from databases over documents and emails to video and audio files as well. This practically means that big data can be text, sensor data, audio, video, click streams, or log files in diverse formats. In accordance with recent polls and records, 80% of data of organizations are not numeric; yet they still have to be taken into consideration upon scrutinizing and decision-making. As claimed by Mario Bojilov, founder of the Meta Business Systems, several “intelligent” devices, such as water- or electricity-meters, produce types of data particular to the industry, application or design. Thus, a new challenge rises concerning the storage and access to those multifarious databases.



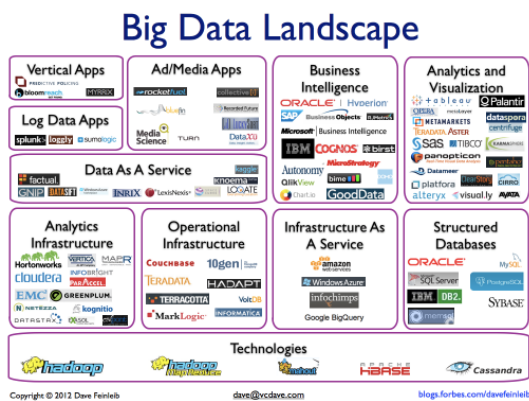
Additionally, another aspect of Big Data that should be debated is **veracity** and was introduced by the IBM. The content of veracity is namely

whether and how much leaders should rely on the information they use to make important decisions. It is recorded that one in three business directors mistrust their information. Furthermore, how trust can be established nowadays, since Big Data presents a huge challenge with the overgrowing variety and number of sources that continue to multiply, is another question that has yet to be answered.

Two further facets of big data conceived by SAS are **variability** and **complexity**. As far as **variability** is concerned, besides the escalating numbers of the velocity and the varieties of data, data streams many a time are of erratic, unreliable and unstable nature with recurrent culminations. Such sporadic and event-triggered zenith points of data loads turn out to be quite challenging to handle, particularly in the case that the media become involved.

Last but not least, **complexity** is best interpreted in the way of dealing with vast capacities of data that mainly stem from different sources and being able to identify and discern data. That is considered by many an endeavour. Notwithstanding, it is of paramount importance to link and correlate relationships between information, otherwise data will coil out of control.

According to Gartner, Inc., the three V's are three trademarks of Big Data that constitute part one (out of three) of the definition and are the main forces of Big Data Quantification. However, they are inclined to believe that human's ability to preserve merely one third of the information they read often leads to misunderstandings or truncation of a part of the definition. Gartner's Big Data definition consists of the following tweet, twenty three words, 181 characters:

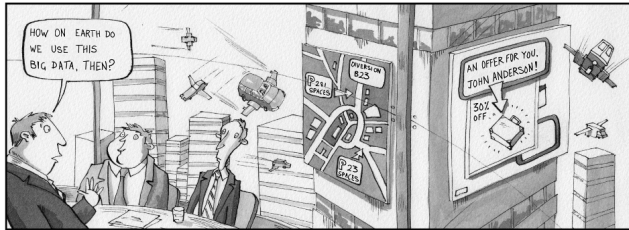


“Big data” is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making”.

From the above emerges that part one is composed of the three V's as explained above, part two of cost-effective and inventive ways to analyze information and lastly part three, which must be regarded as the ultimate objective for organizations, involves evaluating big data problems under the

Uses

Entering the era of big data, people are challenged how to embrace this great amount of information. Big Data is used predominantly by organizations to utilize relevant data so as to reach more auspicious decisions. Technology has facilitated not only owning wide collections and storages of information, but



also comprehending the content and nature of big data. According to SAS, the leader in business analytics software and services, there are many useful applications of Big Data. Specifically, it is

employed to:

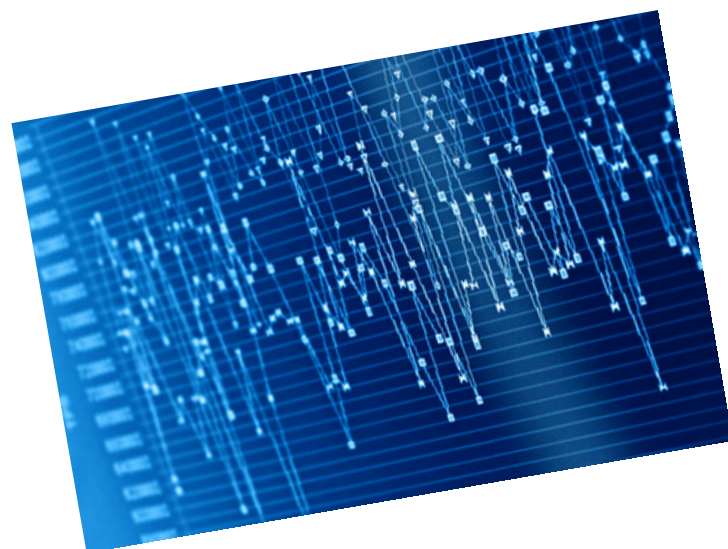
- analyze hundred of SKUs (Stock Keeping Units) in order to regulate optimal prices that maximize the earnings
- instantly determine the customers of utmost significance.
- revalue the state of perilous portfolios and unveil further potentiality that mitigate risk
- based upon the client's current and previous purchases configure retail coupons
- send customized proposals to mobile phones at the time, when customers are in the position to avail themselves of the offers
- evaluate data from social media to disclose possible future market trends
- via data mining reveal/expose punctually deceitful behaviour on the side of a client
- detect deleterious causes of failures.

An additional use of Big Data is called Civil Engineering and has direct impact on society and government. Its purpose, alongside predicting disease upsurges or improving traffic control, is to interweave big data with social issues initially in major cities and gradually spread this method until local authorities to eliminate more efficiently problems pertinent to society. Furthermore, making information attainable to all citizens is crucial. As an

example of that serve companies and applications such as Siri, Google Now, Watson and Wolfram Alpha.

Hence, corporations make use of Big Data in a variety of fields, some of which are:

- Tech start ups/application developing
- General manufacturing
- Utilities (oil & gas)
 - o drilling observations
- Car makers
 - o condition superintendence
- Retail and marketing
 - o price optimization
- Governance
 - o crime forestallment
- Insurance
 - o catastrophe prediction
- Policing
 - o fraud, abuse or theft detection
- Healthcare and pharmaceutical
 - o involve citizens in healthcare
- Sport
- And even Gambling.



Benefits

There are many benefits arising from the use of Big Data. When organizations allocate and store efficiently their massive information, they take advantage of the exponential growth in data for the foreseeable future. To illustrate the benefits further, it should be highlighted that this process makes information transparent and usable at much higher frequency and efficiency. Besides, as more and more data is stored in digital form, we can have access to more precise and thorough performance information, such as much more accurately tailored products or services. As a result, this boosts performance and increases real output. Furthermore, by having more precise data, we can improve decision-making in a personal, local, national or even international level. Finally, our society needs the most efficient storage and use of data, in order to set the basis of future technologies of the next generation of products and services, and this can be achieved by Big Data.



What should also be considered is that after using Big Data, we will have access to affordable large-memory abilities, such as Hadoop, profuse storage and faster processing capacity. On top of that, this process will solve the problem of large unstructured data, through the practice of “parallel processing, clustering, virtualization, large grid environments, high connectivity and high throughputs.”



Apart from that, Big Data allows entrepreneurs to create their own website and customize its content to satisfy every customer from every corner of the world, relying on their gender or nationality. Striking examples of the above is Amazon's Item-Based Collaborative Filtering (IBCF), with features like 'Frequently bought together', and 'Customers who bought this item also bought' or LinkedIn, proposing 'People you may know' or 'Companies you may want to follow'.

Furthermore, factories which traditionally had to replace every piece of equipment that deteriorated can discard such methods. Using massive amounts of data in business one can predict when devices will fail resulting in a cost effective strategy and diminish the expenditures for preservation.

Big data tools in combination with human genome mapping will bring about medical records with genes mapped as a part of it. This heralds a new era where medicine will be able to determine the genes of a patient that trigger a disease and thus develop drugs tailored to the individuals' needs.

Not only do industries utilizing big data have a higher percentage in uncovering economic fraud and ensuring rehabilitation after one, but they can also evaluate risky portfolios and avert meaningless investments.

As a result of all the above benefits, new technology used to handle Big Data stands for a very powerful tool for both economic and social growth, innovation and differentiation. It will reinforce new waves of productivity growth and consumer surplus, and therefore it should be bolstered by all nations. Additionally, Big Data constitutes a solution to the rapid increase of unemployment of people with talent and specialized technical knowledge.

Challenges

Despite its benefits, Big Data also faces some challenges. Precisely, the issues of privacy, security, intellectual property and even accountability will need to be addressed in a big data world. As companies will need to incorporate data from multiple sources, access to data will be critical.

Given that data is constantly growing, unearthing the most valuable information is constituted difficult for organizations. Until lately corporations were confined to simplistic analyses on the grounds that the overwhelming amounts of data would quell their processing machines. From the above two choices deduce for them:

1. coalesce massive data quantities into analysis by implementing technologies such as grid computing, in-database processing and in-memory analytics. In consequence, high-performance analytics will be enacted to analyze massive data.
2. identify the relevance amongst the big data beforehand. Formerly it was common storing information and only when one input a query, was relevance determined. Nowadays, technology that determines which information must be incorporated in analytical processes and which must be placed into storage is available for enterprises to implement.

Veracity, the fourth "V" as defined by IBM, has been proven to be quite a challenge for business leaders and may impede with correct decision making as they cannot reckon on their information due to the soaring number of sources. Therefore the question of granting validity to big data must be directly addressed immediately.



Additionally, many incidents of hacking have been reported earlier this year, the most profound being the one of Twitter, which accounts for 250,000 users losing their personal information. A linkage between big data and hacking is vital to ensure industries' ability to defend themselves against hacking attacks or attempts to break in and strengthen their security measures. For instance, they could identify hackers who use common or not approaches to enter and extinguish ongoing cyber attacks.

Furthermore, there have been a variety of reasonable questions emanating from the alarming rates of data compilations and the use of big data as a solution. For instance, should laws concerning the safe implementation of big data be embedded in corporations? Should standards on an international basis be enacted, which companies around the world have to act in accordance with? Should a policy regarding consumer privacy be embraced or a company-client privilege be established so as to forestall any possible sensitive, private data leaks? Should the government interfere in regulating Big Data use and if so, to what extent? These are questions that delegates will have to revise and propose solutions during the conference.

Lastly, another rational problematic of big data to be cogitated is whether the digitalization of enterprises using big data and making influx of information accessible to the public will lead to the empowerment of the affluent families and the control over the mass. Since big companies will own abundant information concerning people's everyday life and information can be leaked either deliberately or by accident, as in the case of Twitter.



Technologies

The constant advances in technology assist organizations to harvest every opportunity arises from big data. Such advancements are the following:

- cheap and plentiful storage
- faster processors

- economical and sustainable large memory facilities and programmes, such as Hadoop and Cassandra
- new technologies regarding storage and processing, especially modified for large amounts of information entailing unstructured ones
- parallel processing, clustering, MPP, virtualization, large grid environments, high connectivity and throughputs
- cloud computing

Notice: It is very significant to realize that not all big data is related nor substantial. The question raised that still troubles people worldwide is how to spot and distinguish the information of greater importance. The majority of businesses, though, have developed slow processing platforms that obtain value from big data.

Major Countries and Organizations Involved

- **Big Data Research and Development Initiative**

In the United States of America, the Obama government declared this project, in an attempt to discover ways, in which big data can be used to resolve some issues concerning the administration. The research was administered by 84 different big data programs spread across six departments.

- **Big Data Public Private Forum**

The European Commission is funding this 2-year-long research organization through their Seventh Framework Program to affiliate companies with issues concerning big data. The project's purpose is to enhance research and innovation to guide underpinning actions from the European Commission in the successful implementation of the Big Data economy, while “outcomes of this project will be used as input for Horizon 2020, their following framework program.”

Relevant Treaties & Events

Big Data Treaty

There has been no UN Treaties on the subject of Big Data signed yet. Regardless allegations underline that it is of major importance that the United States of America, as a country that has invested the most in Big Data technology and thus has a wider scope and insight in this field in combination to their economic status embroiled currently in Big Data, start drafting a resolution defining which aspects should be included, resulting in a Big Data Treaty.

Relevant Bodies

The following bodies have tried to curb the previously uncontrolled storage and usage of data. They collaborate narrowly with each other as they are closer related than it seems.

- EU Privacy Directive implemented an un-intrusive customer tracking through the web, which resulted in a Data Protection Investigation.
- EU Data Protection Regulation
- W3C Do Not Track (DNT) Initiative

Events

Apropos of the Big Data development many events and conferences have taken place within the business world. Below are some upcoming conventions that will occur in October 2013.

Date	Event	Location
6-9	DS-2013: The 16th Int. Conference on Discovery Science	Singapore
6-9	2013 IEEE International Conference on Big Data, disseminating information on the latest developments in Big Data Research, Development, and Application.	Silicon Valley, CA, USA

12-16	ACM International Conference on Recommender Systems (RecSys) 2013	Hong Kong, China
15-17	Big Data TechCon , Big Data TechCon emphasizes on practical tutorials for IT and Big Data professionals, inclusive of database and business analysts, developers, software architects, data scientists, and project managers	San Francisco, CA, USA
17-19	The Twelfth International Symposium on Intelligent Data Analysis (IDA 2013)	London, UK
21-22	SAS Analytics 2013 Conference	Orlando, FL, USA
20-23	The 14th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'2013)	Hefei, Anhui, China
23-24	PAW: Predictive Analytics World	London, UK
23-24	eMetrics	London, UK
28-29	Big Data World, referring to big data usage, management, analytics, protection and storage for all US enterprises	Las Vegas, NV, USA
28-30	Strata Conference + Hadoop World , Making Data Work	New York, NY, USA

Previous attempts to solve the issue

No previous attempts to solve the issue at hand have been registered due to the fact that Big Data is an emerging phenomenon in our society and businesses need time to acquaint themselves with its nature, its advantageous and detrimental sides as well as its usages. Not only is Big Data an arising occurrence, but it also promotes itself as the venture to elucidate and solve the subject of the plethora of information. Therefore, the need for the USA as a lead economy to take the initiative and set into action a global big data treaty, defining the basic principles of its uses and allowing

other countries to elaborate or disapprove of a number of clauses, is widely acknowledged and awaited.

Timeline

The concept of Big Data has been popular since the 2000s, although it originates back in the 1980s. Big Data has substantially affected the world of information and for this reason it is important to study key events over the past decades.

Date	Description of Event
1983	IBM releases DB2, its latest relational database management system using structured query language.(SQL)
1985	Object-oriented programming (OOP) languages, such as Eiffel, begin to grasp on the market and become the primary programming language.
1990	The creation of Archie, the first engine used for searching on the Internet.
1991	The World Wide Web, using HyperText Transfer Protocol (HTTP) and the HyperText Markup Language (HTML), seems to be a publicly accessible service for sharing information.
1993	The W3Catalog, the World Wide Web's first primitive search engine, is released.
1995	Sun releases the Java platform, whose language becomes one of the most widespread language in government, particularly in Web applications that will increasingly replace personal and paper transactions.

1997	Michael Cox and David Ellsworth of NASA's Ames Research Center publish a paper on visualization which includes the challenges of examining data sets too large for the computing resources provided. They state: "We call this the problem of big data."
2001	Doug Laney of Meta Group defines Big Data as data sets where the three Vs—volume, velocity and variety—demonstrate distinguishing peculiar challenges in regulating and organizing them.

Possible Solutions

According to SAS, key technologies to handle big data and distill valuable information from it are the following:

- **Information governance for big data**, a common, yet more comprehensive and efficient method by providing a plan.
- **High-performance analytics**, advanced form of former processing systems, particularly suitable for enormous data volumes.
- **High-performance visual analytics**, enables people to investigate large amounts of data in mere seconds for supplementary analysis
- **Pliable deployment options for big data**
- **Government interference, defining which actions in the matter of Big Data are legitimate and which are not.**

Bibliography

Big Data – What Is It? n.d. 17 June 2013. <<http://www.sas.com/big-data/>>.

Big data defined. 17 May 2013. 18 June 2013. <<http://www.isaca.org/Knowledge-Center/Blog/Lists/Posts/Post.aspx?ID=299>>.

Big data in practice - what it's used for sector-by-sector. 14 June 2012. 18 June 2013. <<http://www.thebigdatainsightgroup.com/site/article/big-data-practice-what-its-used-sector-sector>>.

James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers. *Big data: The next frontier for innovation, competition, and productivity.* n.d. 17 June 2013. <http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation>.

Tracking the evolution of big data: A timeline. Ed. GCN Staff. 28 May 2013. 17 June 2013. <<http://gcn.com/articles/2013/05/30/gcn30-timeline-big-data.aspx>>.

What is big data? 2013. 17 June 2013. <<http://www-01.ibm.com/software/data/bigdata/>>.

Should big data businesses be forced to prevent hacking <<http://www.xconomy.com/san-francisco/2013/03/19/should-big-data-businesses-be-forced-to-prevent-hacking/>>

Ten practical big data benefits <<http://datascienceseries.com/stories/ten-practical-big-data-benefits>>

<<http://www.crowdsience.com/2011/08/the-internet-in-60-seconds-694445-google-searches-per-minute/>>

<<http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/>>

<http://online.wsj.com/article/SB10001424127887323551004578121142661609474.html>

Hauth, Christian. *Big Data – The end of the Wild West of Data.* 22 June 2012. 15 August 2013. <<http://www.affilinet-inside.de/research/big-data-%E2%80%93-the-end-of-the-wild-west-of-data>>.

<http://www.kdnuggets.com/meetings/>

<http://www.ischool.drexel.edu/bigdata/bigdata2013/>

<http://www.bigdatatechcon.com/SanFrancisco2013/index.html>