The BIG DATA – Introduction

1. BIG DATA - Definition

Big Data, a buzz word of recent past, signifies the quantum of data (ranging from structured to unstructured) not readily handled by the conventional data processing tools. It also signifies the latest processing technologies evolve to handle massive volume of organizational data/ information. In simplified form it is normally called as “BIG DATA ANALYTICS”.

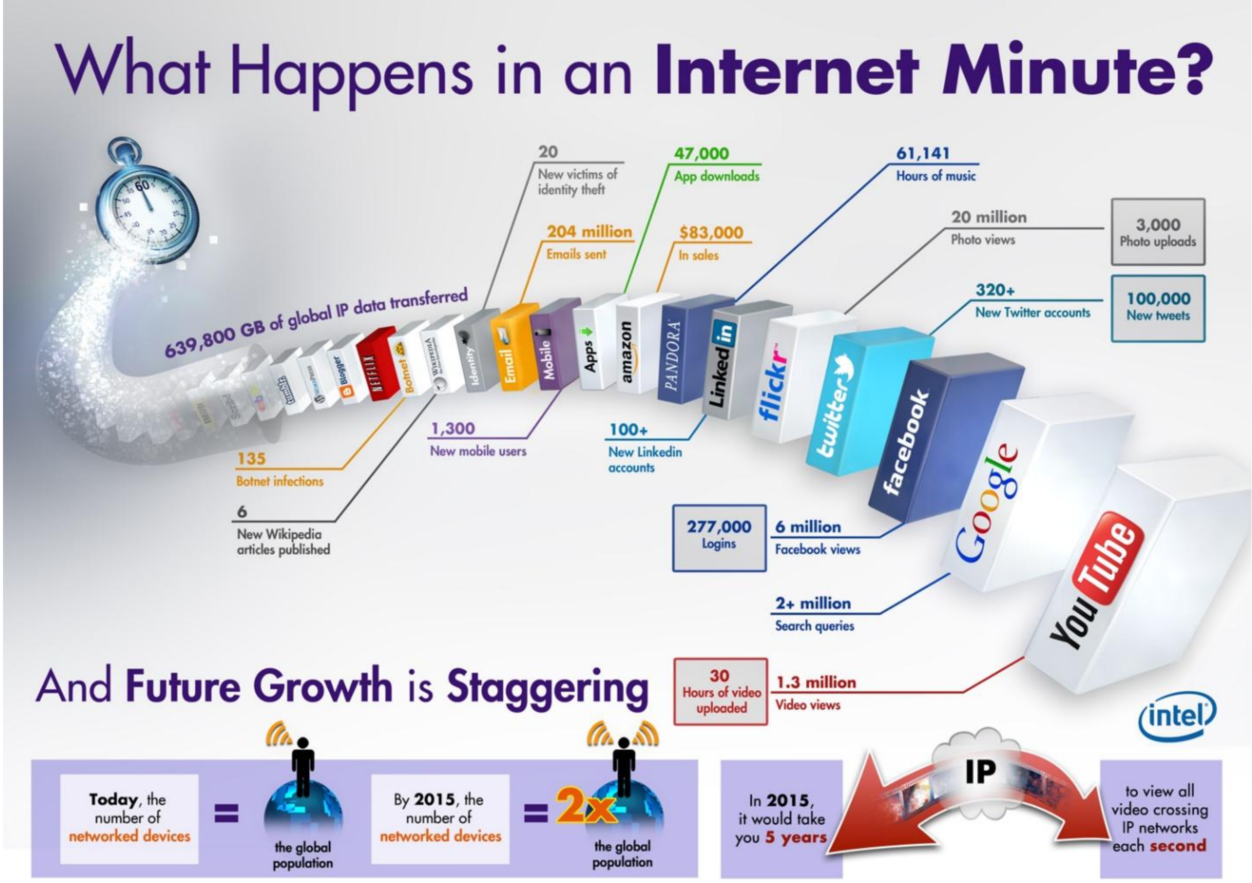
The term "big data" was used for the first time in an article by NASA researchers Michael Cox and David Ellsworth. The pair claimed that the rise of data was becoming an issue for current computer systems. This was also known as the "problem of big data".

> *Information is the oil of the 21st century and analytics is the combustion engine.*
> Peter Sondergaard, SVP, Gartner Research

2. BIG DATA – Quantum

As per IBM (IBM Big Data) 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. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few.

The Infographic, presents graphically the quantum of data/ information accumulation on internet termed as “What Happens in an Internet Minute?”



3. BIG DATA - Characteristics

Big Data is a relative term and there is no yardstick available to categorize an organization data as BIG or not. Rather, experts define following characteristics of BIG DATA *(source: Wikipedia):*

a) **Volume** – The quantity of data that is generated is very important in this context. It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered Big Data or not. The name 'Big Data' itself contains a term which is related to size and hence the characteristic.

b) **Variety** - The next aspect of Big Data is its variety. This means that the category to which Big Data belongs to is also a very essential fact that needs to be known by the data analysts. This helps the people, who are closely analyzing the data and are associated with it, to effectively use the data to their advantage and thus upholding the importance of the Big Data.

c) **Velocity** - The term 'velocity' in the context refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development.

d) **Variability** - This is a factor which can be a problem for those who analyze the data. This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

e) **Veracity** - The quality of the data being captured can vary greatly. Accuracy of analysis depends on the veracity of the source data.

f) **Complexity** - Data management can become a very complex process, especially when large volumes of data come from multiple sources. These data need to be linked, connected and correlated in order to be able to grasp the information that is supposed to be conveyed by these data. This situation, is therefore, termed as the 'complexity' of Big Data.

4. BIG DATA ANALYTICS– Technologies

Unprecedented growth of electronic information halts the processing capabilities of existing IT infrastructure and demands innovative ways to process, handle and interpret the data for management decision making. Therefore, growing number of technologies are evolved to aggregate, manipulate, manage, and analyze big data.

> *You can have data without information, but you cannot have information without data.*
>
> *Daniel Keys Moran, an American computer programmer.*

Following are certain big players in the horizon of BIG DATA Analytics *(Source: Blog of Andy Patrizio Posted on www.datamation.com):*

a) HP

HP is a major hardware vendor and services provider, but its big analytics platform is Vertica, which it acquired in 2011. Vertica Analytics Platform is designed to manage large, fast-growing volumes of structured data and provide very fast query performance and petabyte scalability on commodity enterprise servers. It also has the Autonomy unit with its HAVEn software for analyzing and finding meaning from petabytes of structured and unstructured information.

b) EMC

EMC specializes in storage and its Big Data analytics are built around that. It has a Big Data group that covers hardware and software and a number of verticals, like high performance computing, enterprise and oil and gas exploration. EMC also has a Marketing Science Lab to help companies use Big Data analytics in their marketing department.

c) Teradata

Teradata's Aster platform has a mix of analytics, including the Discovery Platform, a database, a discovery portfolio with pre-built functions for a broad set of Big Data applications, the Aster SQL-GR next-generation graph analytics engine, SNAP Framework for integration and a unified SQL interface across multiple analytic engines and data sources and its own MapReduce.

d) Oracle

Oracle has its Big Data Appliance that combines an Intel server with a number of Oracle software products. They include Oracle NoSQL Database, Apache Hadoop, Oracle Data Integrator with Application Adapter for Hadoop, Oracle Loader for Hadoop, Oracle R Enterprise tool, which uses the R programming language and software environment for statistical computing and publication-quality graphics, Oracle Linux and Oracle Java Hotspot Virtual Machine.

e) SAP

SAP's best Big Data tool is its HANA in-memory database, which the company says can run analytics on 80 terabytes of data, integrate with Hadoop, search text content, harness the power of real-time predictive analytics, and more.

f) Microsoft

Probably not the first company you would think of, but Microsoft's Big Data strategy is fairly broad. It has a partnership with Hortonworks and offers the HDInsights tool based for analyzing structured and unstructured data on Hortonworks Data Platform. Microsoft also offers the iTrend platform for dynamic reporting of campaigns, brands and individual products.

g) IBM

In addition to its big iron, IBM offers DB2, Informix and InfoSphere database software, Cognos and SPSS analytics applications, and of course its well-known Global Services division. IBM also supports the Hadoop analytics platform.

h) Amazon

Amazon has a number of enterprise Big Data platforms, including the Hadoop-based Elastic MapReduce, DynamoDB big data database, and the Redshift massively parallel data warehouse. All of these services work within its greater Amazon Web Services offerings.

i) VMware

VMware is known best for its virtualization hypervisor, but it's building on that platform to offer Big Data software, such as its recent VMware vSphere Big Data Extensions, which lets vSphere control Hadoop deployments and make it easier for enterprises to launch Big Data projects.

j) Google

Google is more of a cloud services company but it is making a push into Big Data analytics by offering BigQuery, a cloud-based Big Data analytics platform for quickly analyzing very large datasets. Unlike most services, you send data up to BigQuery rather than store it in the cloud.

5. BIG DATA – Benefits

> …the goal is to turn data into information and information into insight.
> *Carly Fiorina, ex-CEO HP*

According to Rick van der Lans (a leading Big Data Guru) new **d**ata systems can seriously increase the analytical capabilities of an organization. And that can lead to increased profits, lower costs, bigger market share, and so on.

Utilization of BIG DATA ANALYTICS can certainly emerged with following benefits due to variety of data sources and smart analytical capabilities:

a. Economize the production of goods/ rendering of services
b. Informed decision making based on multiple sources of relevant data
c. Opportunities to take preventive measures to manage control lapses
d. Placed in a better position to take effective corrective measures
e. Capable to detect abnormalities in the business process vigilantly
f. Ability to sort, extract, manipulate and analyze huge data repository
g. Better customer services in quest to enhance market share and strategy

6. BIG DATA – Challenges

We frequently heard the notion that our era is the age of information where the information is accessible and flooded all way round. How do we handle the information? is the real challenge.
Therefore, existing of BIG DATA, apart from emitting blessing may also embedded with risks which are as complicated as the BIG DATA concept itself.

a) Skill Shortage

The Big Data tools are complicated and require appropriate skills level to handle. Presently, the availability of skill resources are limited and it would take some time to breed the required skills for organizations to manage their Big Data technologies.

b) Privacy

The burning question of big data architecture is “How to protect user privacy?” The Big Data framework used to communicated multiple IT infrastructure, channels, data centers, cloud platforms, culture, nodes etc., therefore, it also expose the system to multiple information security vulnerabilities e.g. hacking, eavesdropping, privacy issues, phishing, passive attacks etc.

c) Investment in Technology

The technology architecture of modern era is evolving at a rapid pace. In order to reap the economies of Big Data the organization need to invest in latest technological tools and make sure to efficiently upgrade and transform the same continuously.

d) Right Data

The Big Data does not mean to digest/ absorb each and every record in the repository. One of the biggest challenges to attain Big Data benefits is the identification of right data for Big Data accumulation. Gertrude Stein, an American Author is very much right when he said that everybody gets so much information all day long that they lose their common sense.

7. BIG DATA – Real World Case Studies

(Source: Case Studies Business Intelligence.com)

a) Loss Angeles Police Department

The LAPD Foothill division is attempting to predict and prevent crime before it even happens. How do they do this? Big data, of course. UCLA Professor Jeff Brantingham and his team partnered with the LAPD to examine 13 million crimes recorded over 80 years in order to predict where crimes would occur in the future. By piecing together patterns of human behavior within the LAPD crime data, they believed they could determine which areas to more heavily patrol with police officers and when. By incorporating the 13 million arrests over 80 years into this model, the LAPD is able to use this algorithm to begin predicting the future of crime in the area. The early results from this test already show there has been a 12% decrease in property crime and a 26% decrease in burglary in the Foothill precinct alone. As methods develop and the prediction model is updated in real time, LAPD expects the data will become even more accurate and thus useful.

b) Tesco

Tesco, the largest retailer in the UK, was one of the first major companies to discover the endless benefits of big data analytics. Beginning in the mid-1990s, Tesco introduced its own loyalty program with the Clubcard. Many competitors used similar cards as a means to target discounts and coupons, however, Tesco realized the value of the insight it would give into its customers' behavior patterns. Tesco began processing the huge flood of data coming in from these cards, and was able to better target mailings of vouchers and coupons to customers, resulting in a huge increase from 3% to 70% in rate of coupon redemption. Seeing its analytics approach work, Tesco began applying it to other fields.

One of the company's most profitable uses of analytics, was observing historical sales and weather data and using predictive analytics to optimize their stock-keeping system. By being able to forecast sales by product for each store, Tesco was able to save 100 million pounds ($151,718,000 US dollars) in stock that would have otherwise expired and thus wasted.

Now following Tesco's lead, other competitive retailers are finding creative ways to use big data analytics in order to improve customer satisfaction and increase profits.

c) Netflix

According to Netflix quarterly earnings, the company currently has over 53 million streaming customers worldwide. With such a large amount of users, Netflix has the ability to gather massive amounts of data in order to make better decisions about which television series to buy, kick and keep as well as get a better sense for what original show types will do well among their audience.

For example, the Netflix series *House of Cards,* was not choses simply because it seemed like it had a good plot, but rather, the $100 million dollar show was given the green light because data proved subscribers would like it.

Using big data analytics, Netflix is able to track an endless amount of information, including: when you pause, rewind or fast forward, what day and time you watch content, where you watch it from (based on zip code), what device you use to watch, the ratings you give, the things you search for and your browsing and scrolling behavior.

(Contributed for the EACPE by Naj-Mus-Sahar)