International Journal of Scientific & Engineering Research, Volume 8, Issue 4, April-2017
ISSN 2229-5518

Big Data Analysis

D.Chandra, R.Monika, And E.Thilagavathi

Abstract-*Big data* is a term for data sets that are so large or complex that traditional data processing application software are inadequate to deal with them. Challenges include capture, storage, analysis, data curation, search, sharing, transfer, visualization, querying, updating and information privacy. The term "big data" often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. "There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem." Analysis of data sets can find new correlations to "spot business trends, prevent diseases, and combat crime and so on." Scientists, business executives, practitioners of medicine, advertising and governments alike regularly meet difficulties with large data-sets in areas including Internet search, finance, urban informatics, and business informatics. Scientists encounter limitations in e-Science work, including meteorology, genomics, connectomics, complex physics simulations, biology and environmental research.

Data sets grow rapidly - in part because they are increasingly gathered by cheap and numerous information-sensing mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 Exabyte's (2.5×10^{18}) of data are generated. One question for large enterprises is determining who should own big-data initiatives that affect the entire organization.

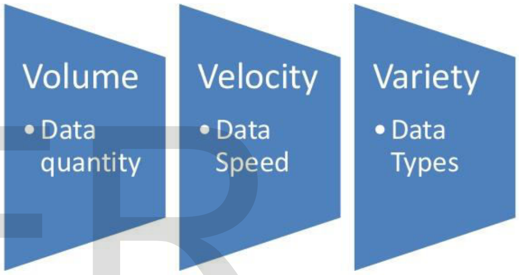
INTRODUCTION:

- Big data may well be the next big thing in the IT world
  big data burst upon the scene in the first decade of the 21st century.
- The first organizations to embrace it were online and startup firms. Firms like goggle, eBay, inkedln. and face book were built around big data from the beginning.
- Like many new information technologies, big data can bring about dramatic cost reductions, substantial improvements in the time required to perform a computing task, or new product and service offerings.

CHARACTERISTICS:

- Big data can be described by the following characteristics:

Three Characteristics of Big Data V3s



1) Volume (amount of data the size of the data set)

Volume Refers to the vast amounts of data generated every second. We are not talking Terabytes but Zettabytes or Brontobytes. If we take all the data generated in the world between the beginning of time and 2008, the same amount of data will soon be generated every minute. This makes most data sets too large to store and analyze using traditional database technology. New big data tools use distributed systems so that we can store and analyze data across databases that are dotted around anywhere in the world.

90% of all data ever created, was created in the past 2 years. From now on, the

Amounts of data in the world will double every two years. By 2020, we will have 50 times the amount of data as that we had in 2011. The sheer volume of the data is enormous and a very large contributor to the ever expanding digital universe is the Internet of

D.Chandra, Ii-M.C.A Gandipathy Tulsi's Jain Engineering College

R.Monika, Ii-M.C.A Gandipathy Tulsi's Jain Engineering College

E.Thilagavathi, Ii-M.C.A Ganadipathy Tulsi's Jain Engineering College

D.Chandra1096@Gmail.Com,Monikamalar10496@Gmail.Com

Things with sensors all over the world in all devices creating data every second. The era of a trillion sensors is upon us.

If we look at airplanes they generate approximately 2.5 billion Terabyte of data each year from the sensors installed in the engines. Self-driving cars will generate 2 Petabyte of data every year. Also the agricultural industry generates massive amounts of data with sensors installed in tractors. Shell uses super-sensitive sensors to find additional oil in wells and if they install these sensors at all 10,000 wells they will collect approximately 10 Exabyte of data annually. That again is absolutely nothing if we compare it to the Square Kilometer Array Telescope that will generate 1 Exabyte of data per day.

In the past, the creation of so much data would have caused serious problems. Nowadays, with decreasing storage costs, better storage solutions like Hadoop and the algorithms to create meaning from all that data this is not a problem at all.

2) Velocity (speed of data in and out or data in motion)

Velocity Refers to the speed at which new data is generated and the speed at which data moves around. Just think of social media messages going viral in seconds. Technology allows us now to analyze the data while it is being generated (sometimes referred to as in-memory analytics), without ever putting it into databases.

The Velocity is the speed at which the data is created, stored, analyzed and visualized. In the past, when batch processing was common practice, it was normal to receive an update from the database every night or even every week. Computers and servers required substantial time to process the data and update the databases. In the big data era, data is created in real-time or near real-time. With the availability of Internet connected devices, wireless or wired, machines and devices can pass-on their data the moment it is created.

The speed at which data is created currently is almost unimaginable: Every minute we upload 100 hours of video on YouTube. In addition, every minute over 200 million emails are sent, around 20 million photos are viewed and 30,000 uploaded on Flickr, almost 300,000 tweets are sent and almost 2.5 million queries on Google are performed.

The challenge organizations have is to cope with the enormous speed the data is created and used in real-time.

3) Variety (range of data types, domains and sources)

Variety Refers to the different types of data we can now use. In the past we only focused on structured data that neatly fitted into tables or relational databases, such as financial data. In fact, 80% of the world’s data is unstructured (text, images, video, voice, etc.) With big data technology we can now analyze and bring together data of different types such as messages, social media conversations, photos, sensor data, and video or voice recordings.

In the past, all data that was created was structured data, it neatly fitted in columns and rows but those days are over. Nowadays, 90% of the data that is generated by an organization is unstructured data. Data today comes in many different formats: structured data, semi-structured data, unstructured data and even complex structured data. The wide variety of data requires a different approach as well as different techniques to store all raw data.

There are many different types of data and each of those types of data requires different types of analyses or different tools to use. Social media like Face book posts or Tweets can give different insights, such as sentiment analysis on your brand, while sensory data will give you information about how a product is used and what the mistakes are.

In summary, with better storage solutions, the challenge of rapid data creation, and the diverse tools to store and analyze data, there are practical approaches to performing analytics on data for making informed business decisions. I trust that Part One content will assist you with evaluating if a project requires a Big Data Solution.

What is big data?

- Big data’ is similar to ‘small data’, but bigger in size.
- But having data bigger it requires different approaches:

- Techniques, tools and architecture
- An aim to solve new problems or old problems in a better way
- Big data generates value form the storage and processing of very large quantities of digital information that cannot be analyzed with traditional computing techniques. Face book handles 40 billion photos its user base.
- Decoding the human genome originally took 10 years to process; now it can be achieved in one week

The structure of big data:

Structure:

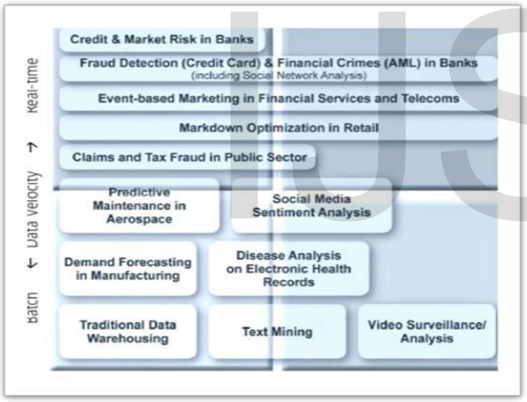
- **Most traditional data sources**

Semi structured:

- **Many sources of big data**

Unstructured:

Video data, audio data

- 

Credit & Market Risk in Banks
Fraud Detection (Credit Card) & Financial Crimes (AML) in Banks
(including Social Network Analysis)
Event-based Marketing in Financial Services and Telecoms
Markdown Optimization in Retail
Claims and Tax Fraud in Public Sector
Predictive Maintenance in Aerospace
Social Media Sentiment Analysis
Demand Forecasting in Manufacturing
Disease Analysis on Electronic Health Records
Traditional Data Warehousing
Text Mining
Video Surveillance/ Analysis
Batch ← Data velocity → Real-time

Why big data?

- **Growth of big data is needed**
  - ✓ Increase of storage capacities
  - ✓ Increase of processing power
  - ✓ Availability of data(different data types)

Every day we create 2.5 quintillion bytes of data; 90% of the data in the world today has been created in the last two years alone.

STORING BIG DATA:

- ❖ Analyzing your data characteristics
  - ➢ Selecting data sources for analysis
  - ➢ Eliminating redundant data
  - ➢ Establishing the role of NoSQL
- ❖ Overview of Big Data Stores
  - Data models: key value, graph, document, column family
  - Hadoop Distributed File System
  - Hbase
  - Hive

HOW BIG DATA IS DIFFERENT:

1) Automatically generated by a machine

(e.g... Sensor embedded in an engine)

2. Typically an entirely new source of data

(e.g... Use of the internet)

3. Not designed to be friendly

(e.g... Text streams)

4. May not have much values

Need to focus on the important part

REFERENCES:

1. Jung, G. (2013). World population one out of every four wireless internet, electronics information center day trend. http://203.253.128.6:8088/servlet/eic.wism.EICWeb

2. Gartner. (2010). Gartner identifies the top 10 strategic technologies for 2011. http://www.gartner.com

3. Kim, Y. C., Cha, M. H., Lee, S. M., & Kim, Y. G. (2009). Trends of storage virtualization technologies on cloud computing. Electronics and Telecommunications Research Institute, 24(4), 69–78. (In Korean).

4. Carol, S. (2009). Storage explained: Cloud storage defined. http://SearchStorage.com.

5. Kim, H. Y., Min, O. G., & Nam, G. H. (2010) The technology trend of mobile cloud. Electronics and Telecommunications Research Institute, 25(3), 43–44 (in Korean).

6. Chunlin, L., & LaYuan, L. (2015). Efficient market strategy based optimal scheduling in hybrid cloud environments. Wireless Personal Communications, 83(1), 581–602.

7. Yang, Z., Liu, X., Hu, Z., & Yuan, C. (2014). Seamless service handoff based on Delaunay triangulation for mobile cloud computing. Wireless Personal Communications. Doi:10.1007/s11277-014-2229-6.

8. NIA. (2009). ICT new technology paradigm: Cloud computing strategy. CIO Report, 17, 1–40. (In Korean).

9. O'Neal, J. (2009). NetApp, storage infrastructure for the cloud. http://www.netapp.com/us/ communities/tech-ontap/tot-cloud-storage-0509.aspx

10. Wang, Y., Chen, T., & Wang, D. (2015). A Survey of mobile cloud computing applications: Perspectives and challenges. Wireless Personal Communications, 80(4), 1607–1623.

11. Yang, H. D., & Hwang, S. W. (2010). A study on the improvement of mobile cloud computing services. DIGIECO focus. http://www.digieco.co.kr.