ISSN 2250-3153

Analysis on big data over the years

R.Devakunchari

Computer Science Department,
Arul Murugan College of Engineering & Technology, Karur.

Abstract- Big Data is characterized by increasing volume and velocity of data.IBM estimates that every day 2.5 quintillion bytes of data are created – so much that 90% of the data in the world today has been created in the last two years. The traditional data-intensive sciences such as astronomy, high energy physics, meteorology, genomics, biological and environmental research in which peta- and Exabyte of data are generated are common domain examples. Here even the capture and storage of the data is a challenge. Google implemented hundreds of special-purpose computations that process large amounts of raw data, such as crawled documents, Web request logs, etc., to compute various kinds of derived data, such as inverted indices, various representations of the graph structure of Web documents, summaries of the number of pages crawled per host, and the set of most frequent queries in a given day. In this paper big data that is navigating in years from the past to present and to the future is analyzed. To address the problem space of unstructured analytics, Map Reduce with Hadoop distributed File System (HDFS) is also discussed. To process terabytes of data efficiently on daily basis some of tools and techniques available and challenges, issues and benefits of big data is also listed.

Index Terms- Hadoop Distributed File System (HDFS), MapReduce, NoSQL,NLP, Big Data, Name Node(NN)

I. INTRODUCTION

The amount of Digital data being produced, in real time, has been exploding at an unknown rate, even across the developing world, just as we all go about our daily lives. Today, 98 percent of all stored data is in digital form whereas storing in paper goes hand in hand along with digital form some 15 years before. The size of the databases has been growing at exponential rates in today's enterprises. The need to process and analyzethese large volumes of data for decision making in businesses has also increased along with it. There is also a need to process petabytes of data in efficient manner on daily basis in several business and scientific applications.

The interactions of billions of people using mobile devices and Internet every day, generates a flood of data. The increasing volume of enterprise information, genomics, medical records, information- sensing mobile devices, multimedia and social media will fuel exponential growth in data in the future. This has given rise to the big data problem due to the inability of conventional database systems and software tools to manage or process the big data sets within tolerable time limits by the industries.

The International Data Corporation (IDC) study predicts that overall data will grow by 50 times by 2020, driven in large part more by embedded systems such as sensors in clothing, medical devices and structures like buildings and bridges. This study also determined that unstructured information - such as files, email and video - will account for 90% of all data created over the next decade [12].

Analyzing and making intelligent decision out of these large data sets comprising of unstructured, semi structured and structured big data—will become a key basis of competition in business and technology.

II. BIG DATA –WHAT AND WHY?

Big data is a term that came from the need of big companies like yahoo, google, Facebook, Etc. and in many enterprises and R&D to analyze big amounts of unstructured data they are generating every second.

Big Data, in general, falls into 3 categories:

- Business application data (e.g. CRM, SAP or Oracle ERP)
- Human-generated content (e.g., Internet text, social media traffic etc.) and
- Machine data (e.g., M2M, RFID, Log Files, sensors etc.).

The most common definition for big data which is used by many others is "Big data refers to the large data sets which are very difficult to store, analyze and manage due to their size as well as complexity. Their size ranges fromthousands of terabytes to peta-, and exa- bytes."[10]

A. Big Data Characteristics

The most common characteristics of big data arises from 3v's by Gartner namely

- Volume – The size of data is very large and in terabytes and petabytes.
- Velocity –The pace at which data flows in from sources. The time plays a key role. The reasons for data getting generated faster includes,
    a. Increasingly automated processes
    b. Increasingly interconnected systems
    c. Increasing social interaction by people
- Variety –It includes structured,semi-structured and unstructured data of all varieties: text, audio, video, posts, log files etc

In addition many papers propose new v's other than the above 3v's by Gartner to characterize big data. They are,

- Veracity –It refers to the biases, noise and abnormality in data. Is the data that is being stored, and mined meaningful to the problem being analyzed.
- Value –Measures the usefulness of data in making decisions. The purpose of computing is insight and not numbers

B.Big data Trends

"Information is one of the four powerful forces changing the way business is done," said Regina Casonato, managing vice president at Gartner, Inc. and they have identified Big data as one of the top technology trends that will play key roles in modernizing information management (IM) in 2013 and beyond.

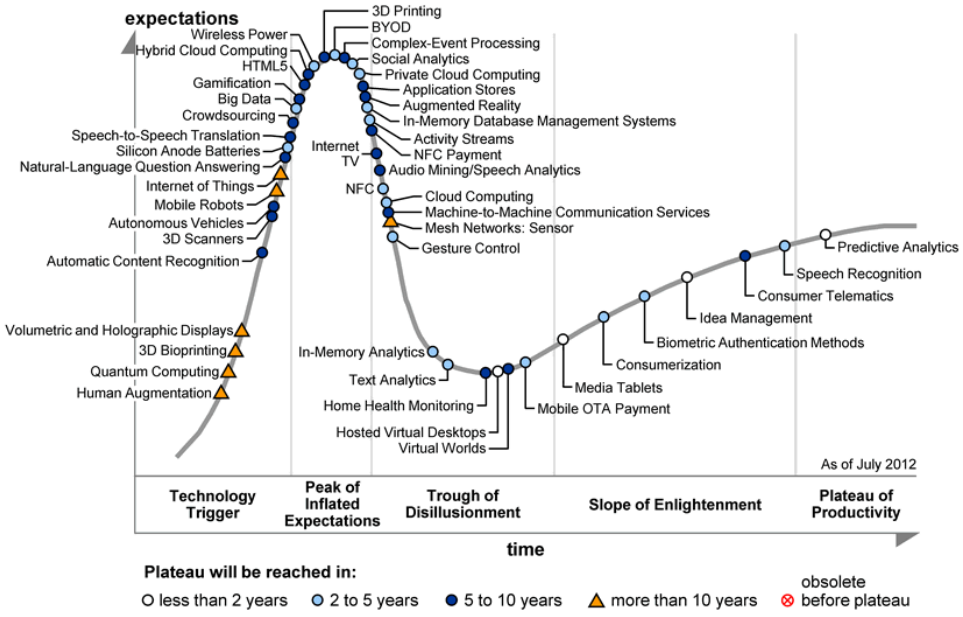

Figure 1: Gartner Hype cycle [11]

III. EVOLUTION OF BIG DATA

Wesleyan University Librarian estimated and published in a book "The Scholar and the Future of the Research"[1] that American university libraries were doubling in size every sixteen years

1956 FICO, then called Fair, Isaac, and Company, a leading provider of credit scoring, decision management, fraud detection and credit risk score services was founded which works on the principle that data, used intelligently, can improve business decisions.

1958 FICO builds its first credit scoring system for American Investments.

1961 Derek Price publishes "Science since Babylon", in which he charts the growth of scientific knowledge by looking at the growth(doubling every fifteen years) inthe number of scientific journals and papers. Price calls this the "law of exponential increase," explaining that "each [scientific] advance generates a newseries of advances at a reasonably constant birth rate, so that the number of births is strictly proportional to the size of the population of discoveries at any given time."

1967 B. A. Marron and P. A. D. de Maine publish "Automatic data compression" in the Communications of the ACM, stating that"The 'information explosion' noted in recent years makes it essential that storage requirements for all information be kept to a minimum." The paper describes"a fully automatic and rapid three-part compressor which can be used with 'any' body of information to greatly reduce slow external storage requirements and to increase the rate of information transmission through a computer."

1971 Arthur Miller writes in "The Assault on Privacy"[1] that "Too many information handlers seem to measure a man by the number of bits of storage capacity his dossier will occupy."

1975 The Ministry of Posts and Telecommunications in Japan starts conducting the Information Flow Census, tracking the volume of information ("amount of words" as the unifying unit of measurement) circulating in Japan. The 1975 censusalready finds that information supply is increasing much faster than information consumptionand in 1978 it reports that "the demand for information provided by mass media, which are one-way communication, has become stagnant, and the demand for information provided by personal telecommunications media, which are characterized by two-way communications, has drastically increased. Our society is moving toward a new stage… in which more priority is placed on segmented, more detailed information to meet individual needs, instead of conventional mass reproduced conformed information."

1980 [April] I.A. Tjomsland gives a talk titled "Where Do We Go From Here?"at the Fourth IEEE Symposium on Mass Storage Systems, in which he says"Those associatedwith storage devices long ago realized that Parkinson's First Law may be paraphrased to describe our industry—'Data expands to fill the space available'…. I believe that large amounts of data are being retained because users have no way of identifying obsolete data; the penalties for storing obsolete data are less apparent than are the penalties for discarding potentially useful data."

1980 The CPG / Retail industry transitioned from bi-monthly audit data to scanner data changed the dynamics of the industry.

1981 The Hungarian Central Statistics Office starts a research project to account for the country's information industries, including measuring information volume inbits.

1983 Ithiel de Sola Pool publishes "Tracking the Flow of Information" in Science. Looking at growth trends in 17 major communications media from 1960 to 1977, heconcludes that "words made available to Americans (over the age of 10) through these media grew at a rate of 8.9 percent per year. In the period ofobservation, much of the growth in the flow of information was due to the growth in broadcasting.But toward the end of that period [1977] the situation was changing: point-to-point media were growing faster than broadcasting."

1986 Hal B. Becker publishes "Can users really absorb data at today's rates? Tomorrow's?" in Data Communications.[1] Becker estimates that "the recoding density achieved by Gutenberg was approximately 500 symbols (characters) per cubic inch—500 times the density of [4,000 B.C. Sumerian] clay tablets. By the year 2000, semiconductor random access memory should be storing 1.25X10^11 bytes per cubic inch."

1996 [October] The world's leading online travel company which is a small division within Microsoft launched online travel booking site Expedia.com®, giving consumers a revolutionary new way to research and book travel

1996 Digital storage becomes more cost-effective for storing data than paper according to R.J.T. Morris and B.J. Truskowski, in "The Evolution of StorageSystems," IBM Systems Journal, July 1, 2003.

1997 [October] The first article in the ACM digital library to use the term "big data" was published. Michael Cox and David Ellsworth publish "Application controlled demand paging for out-of-core visualization" [2] in the Proceedings ofthe IEEE 8th conference on Visualization. They start the article with"Visualization provides an interestingchallenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of big data."

1997 [October] Michael Lesk publishes "How much information is there in the world?"Lesk concludes that "There may be a few thousand petabytes of information all told; and the production of tape and disk will reach that level by the year 2000. So in only a few years, (a) we will be able [to] save everything–no information will haveto be thrown out, and (b) the typical piece of information will never be looked at by a human being."

1998 Most visited website in the world Google was founded. It has been estimated to run more than one million servers in data centers and to process over one billion search requests and about 24 petabytes of user-generated data each day.

1998 [April] John R. Masey, Chief Scientist at SGI, presents at a USENIX meeting a paper titled "Big Data… and the Next Wave of Infrastress."

1998 [October] K.G. Coffman and Andrew Odlyzko publish "The Size and Growth Rate of the Internet." They conclude that "the growth rate of traffic on the public Internet, while lower than is often cited, is still about 100% per year, much higher than for traffic on other networks. Hence, if present growth trends continue, data traffic in the U. S. will overtake voice traffic around the year 2002 and will be dominated by the Internet."

1999 comScore,an American Internet analytics company providing marketing data and analytics to many of the world's largest enterprises, agencies, and publishers was founded.[4]

1999 [August] Steve Bryson, David Kenwright, Michael Cox, David Ellsworth, and Robert Haimes publish "Visually exploring gigabyte data sets in real time" in the Communications of the ACM. It is the first CACM article to use the term "Big Data" (the title of one of the article's sections is "Big Data for Scientific Visualization"). The article opens with the following statement: "Very powerful computers are a blessing to many fields of inquiry. They are also a curse; fast computations spew out massive amounts of data. . Richard W.Hamming, mathematician and pioneer computer scientist, pointed out, the purpose of computing is insight, not numbers."

1999 [October] Bryson, Kenwright and Haimes join David Banks, Robert van Liere, and Sam Uselton on a panel titled "Automation or interaction: what's best for big data?" at the IEEE 1999 conference on Visualization.

2000 Peak of dot-com boom, also known as dot-com bubble, the Internet bubble and the information technology bubble. [5]

2000 [October] Peter Lyman and Hal R. Varian at UC Berkeley publish "How Much Information?" It is the first comprehensive study to quantify, in computer storage terms, the total amount of new and original information (not counting copies) created in the world annually and stored in four physical media: paper, film, optical (CDs and DVDs), and magnetic. The study finds that in 1999, the world produced about 1.5 exabytes of unique information, or about 250 megabytes for every man, woman, and child on earth. It also finds that "a vast amount of unique information is created and stored by individuals"and that "not only is digitalinformation production the largest in total, it is also the most rapidly growing.

2001 Paper on 3D Data management [6] by Doug laney explaining about 3 v's.

2004 Facebook is a social networking service was launched. Google published a paper on MapReduce

2005 Apache Hadoop, an open-source software framework for storage and large scale processing of data-sets on clusters of commodity hardware, was created by Doug Cutting and Mike Cafarella.

2006 [July] Twitter called as "the SMS of the Internet" an online social networking and microblogging service was launched.

2007 The first generation iPhone (smart phone from Apple inc) was released

2008 Facebook reaches 100M users.

2010 Special report on Data, data everywhere by "The Econonmist", EMC buys Greenplum,IBM buys Netezza

2011 Mckinsey report on big data, oracle buys endecea,Hp buys vertica.

2012 Big Data becomes buzz word after Gartner prediction, Facebook user hits 1B

2013 Fast Data era, YouTube hits 1B users

V. SQL VS NOSQL

A big truth about big datain traditional databases: it's easier toget the data in than out. Database gets overloaded and "Time out error occurs on inserting into database".The trouble comes when we want totake that accumulated data, collectedover months or years, and learn somethingfrom it—and naturally we wantthe answer in seconds or minutes. The pathologies of big data are primarilythose of analysis. Traditional methods of database shard, scaling with queue, RDBMS-based dimensionalmodeling and cube-based OLAP (onlineanalytical processing) turn outto be either too slow or too limited tosupport big data.SQL based databases are data warehouses and data marts where dimensional and normalized approaches of storing is done.

To solve several needs of big data a variety of "NoSQL (NOT ONLY SQL) databases have appeared.

1. For storing and managing unstructured data (non-relational data).
2. Focus on high-performance scalable data storage, and provide low-level access to a data management layer (data validity and integrity). Also called key-value stores, schema-free massive scaling on-demand Databases.
3. NOSQL databases [8] separate data management and data storage.
4. Relaxes the consistency requirement. Relaxing consistency is often called eventual consistency.
5. Also has ACID properties and follows CAP theorem [3] (consistency, availability and tolerance of network partition), customized replication, high availability and greater flexibility in storing heterogeneously structured data.

VI. BIG DATA TECHNIQUES AND TOOLS

.
Big data is spawning new tools that are mix of significant processing power, parallelism and statistical, machine learning, or pattern recognition techniques. A wide variety of techniques and technologies has been developedand adapted to aggregate, manipulate, analyze, and visualizebig dataincludes massively parallel processing (MPP) databases, data mining, grids, distributed file systems, distributed databases, cloudcomputing platforms, the Internet, and scalable storagesystems

- Natural Language Processing (NLP) techniques (Lexical/morphological analysis,Syntactic analysis,Semantic analysis) to extract information from unstructured data
- CBIR (Content-Based Image Retrieval) enable us to pave the way toward new accessibility for large-volumemultimedia collections.
- Sentiment analysis uses semantic technologies
- SAP HANA
- Hadoop - reliable data storage and high-performance parallel data processing
- Cloud is extensible, flexible, scalable, elastic, self-healing, on-demand, etc. and provides the inexpensive hardware/software platform with all applications with lower capital cost requirements
- For streaming data it includes IBM's InfoSphere Streams and emerging Twitter's Storm, and Yahoo S4

- Using multi/many cores, wide SIMD and dynamic optimization of the applications requiring exascale computing
- IBM Infosphere Big Insights
- WX2 kognitio Analytical Platform(fast and scalable in-memory analytic database)
- SAND Analytic Platform(columnar analytic database)
- IBM Infosphere Streams(analysis of massive volumes of streaming data in sub-millisecond)
- New parallel programming models and programming languages such as Map Reduce, Software Transactional Memory,Galois,CUDA,X10,Chapel,Fortress,POSIX Threads,C++ Thread Support Library,MPI,OpenMP,OpenCLTask Parallel Library, Threading Building Blocks, CilkPlus

VII. UNSTRUCTURED DATA ANALYSIS

A. Hadoop and Map Reduce

Hadoop is both a distributed file system (HDFS) modeled on GFS (2004), a distributed processing framework, using Map Reduce concepts and a distributed database called HBase. It is a framework for distributed computing and large datasets on a scale-out shared-nothing architecture to address processing of large unstructured data sets.It is open-source software, reliable and scalable. The principle here is "Moving computation is cheaper than Moving data" [7]

HDFS (Hadoop Distributed File System):

- fault tolerance
- run on commodity hardware
- high throughput access
- store data across thousands of servers
- running work (Map/Reduce jobs) across those machines,
- running the work near the data
- Master/slave architecture.

B. HDFS System Architecture

- Hadoop cluster contains1 MN (master node) and no of slaves or WN (worker nodes).
- MN consist of
  a. Job tracker(JT)-schedules jobs across TT slaves
  b. Task tracker(TT)- Runs tasks within job
  c. Name node (NN) -contains metadata of DN,mapping of file blocks to DN(replication).
  d. Data node(DN)(acts also as TT) -stores and serves blocks of data
  e. Secondary NN (SNN) -snapshots of the name node's memory structures, preventing file system corruption and reducing loss of data.
- User data is stored in files.
- File is split into one or more blocks
- It stores each file as a sequence of blocks
- Blocks are stored in a set of Data Node
- Name Node determines the mapping of blocks to Data nodes.

C. Map Reduce Programming

- It is a Software framework introduced by Google in 2004
- Map step
  a. Master nodes(MN) takes the input, partitions it up into smaller sub-problems and distributes it to worker nodes(WN)
  b. Multi-level tree structure (WN distribute again) is used to sort.
  c. WN passes answer to MN
- Reduce step

  Merges answers to all sub-problems to form output.

VIII. BENEFITS AND CHALLENGES OF BIG DATA PROCESSING

Some bigger benefits that enterprises and organizations utilize includes,

- Making more informed decisions
- Increase productivity and reduce costs
- Increase transparency
- Improve citizen service and satisfaction for government
- Predicting trends
- Identify irregular patterns and activities that are often a sign of error or fraud
- Improve mission outcomes

- Ability to find, acquires, extract, manipulate, analyze, connect and visualize data with the tools of choice
- The capability of Hadoop for volumes to manage vast amounts of data, in or out of the Cloud, with validation and verification.
- Real-time monitoring and forecasting of events that impact either business performance or operation

Big data incurs management issue, transportissue, processing issue and storage issues. Some of the design and analytical challenges of big data include,

- Real time requirements
- Memory management
- Load balancing
- Support for Data Partitioning
- Latency-Throughput trade-off
- Multi-tenancy
- Data ownership
- Compliance & Security
- Data getting in is easier than getting it out
- Quality versus Quantity
- Need retrospective analysis due to expanding data
- Speed versus scale
- Distributed data and processing
- Turning straw into gold(processing large discrete data points into high valued data)
- Finding the needle in the haystack(finding key data among large)
- Need to address arising unpredicted effects due to data from diverse sources.

IX. CONCLUSION

The key skills in today's big data environments are dataintegration, triangulation, pattern recognition, predictive models and simulations. Big data has a lot to learn about projection, bias correction and sampling, which, when applied correctly, could yield even more important big data Insight...In a study by Mc Kinsey Global Institute (MGI) firm calculated that U.S faces shortage of 140,000 to 190,000 people with analytical expertise and 1.5 million managers and analysts with skills to understand and make decisions based on analysis of Big Data [9].But while the big data issues are fixable, big research's issues areendemic.To be competitive, organizations will require new technology with clear implementation strategies, iterative test-and-learn environments and data science talent.

REFERENCES

[1] http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/
[2] http://www.nas.nasa.gov/assets/pdf/techreports/1997/nas-97-010.pdf.
[3] https://blogs.oracle.com/MAA/entry/the_cap_theorem_consistency_and
[4] http://en.wikipedia.org/wiki/ComScore
[5] http://en.wikipedia.org/wiki/Dot-com_bubble
[6] http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management- Controlling-Data-Volume-Velocity-and-Variety.pdf
[7] http://hadoop.apache.org/docs/stable1/hdfs_design.html
[8] Deka Ganesh Chandra, Ravi Prakash and Swati Lamdharia "A Study on Cloud Database " in proc.Fourth Int'l Conference on Compu'nal Intelligence and Communication Networks,2012.
[9] http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation
[10] http://en.wikipedia.org/wiki/Big_data.
[11] Gartner Hype Cycle 2012 ,http://www.gartner.com/id=2065716
[12] World's data will grow by 50X in next decade, IDC study predicts http://www.computerworld.com/s/article/9217988/World_s_data_will_grow_by_50X_in_next_decade_IDC_study_predicts

AUTHORS

R.Devakunchari, B.Tech(I.T),M.Tech(CSE),Assistant Professor,Department of Computer Science and Enginering,
Arul Murugan College of Engineering and Technology,Karurand devakunchari.r@gmail.com.