A Study of Various Big Data Emerging Technologies

Harjender Singh*

Abstract: Big data is a new driver of the world economic and societal changes. The world's data collection is reaching a tipping point for major technological changes that can bring new ways in decision making, managing our health, cities, finance and education. While the data complexities are increasing including data's volume, variety, velocity and veracity, the real impact hinges on our ability to uncover the `value' in the data through Big Data Analytics technologies. Big Data Analytics poses a grand challenge on the design of highly scalable algorithms and systems to integrate the data and uncover large hidden values from datasets that are diverse, complex, and of a massive scale. In this paper, we explain the concept, characteristics & need of Big Data & different offerings available in the market to explore unstructured large data. Our analysis illustrates that the Big Data analytics is a fast-growing, influential practice and a key enabler for the social business. Analytics companies develop the ability to support their decisions through analytic reasoning using a variety of statistical and mathematical techniques.

Keywords: Deep Learning, Business Intelligence, Business Analytic, Hadoop, Cluster, NoSQL Database

Big data: The big data refers to huge volume of data which cannot be stored and processed using the traditional approach within the given time frame. Big data is a term that describe the large volume of data both structured and unstructured- that inundates a business on a day to day basis. But its not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves.

How huge this data need to be:

- Analyst predicts that by 2020, there will be 5200GB of data on every person in the world.
- On average, people send about 500 million tweets per day.
- The average U.S. customer uses 1.8 gb of data per month on his or her cell phone plan.
- Walmart processes one million customer transactions per hour.
- Amazon sells 600 items / second.
- On average each person who uses email receives 88 emails per day and send 34.
- That adds up to more than 200 billion emails each day.
- Master card processes 74 billion transactions per year.
- Commercial airlines make about 5800 flights per day.

Characterization of Big Data:

1. **Volume**: organizations collect data from a variety of sources including business transactions, social media and information from sensor or machine to machine data. In the past, storing it would've been a problem but new technologies such as hadoop have eased the burdens. By 2020 accumulated digital universe of data will grow from 4.4 zettabytes today to around 44 zettabytes or 44 trillion gigabytes.

2. **Variety**: Data comes in all types of formats- from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ficker and financial transactions. Different kinds of data is being generated from various sources like audio, video, PNG, Jfiles, emails, text files, you tube etc, these data can be structured, semi structured and unstructured. The structured data can be arranged in a proper schema i.e. it should be arrange in rows and columns or tabular format. In semi structured the schema is not defined properly.in un structured file we have log file, audio, video and png files.

3. **Velocity**: Data is being generated at an alarming rate. In every minutes 10000 tweets are generated. In FB 69500 status updated and 11millions messages generated in every minutes. Around 698445 Google searches in every minutes. Around 168000,000+ emails generated that is around 1820 TB data created. In every 60 second more than 217+ new mobile users added i.e. data streams in at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near future.

4. **Value**: Mechanism to bring the correct meaning out of the data. i.e. we can extract the meaningful information from the huge stored data.

5. **Veracity:** It is the degree to which data is accurate, precise and trusted. Some examples are:

**Asst. Professor, Maharaja Surajmal Institute, New Delhi); harjendersingh@msi-ggsip.org*

- Data Lineage: An organization gets data fromhundred of sources. It discover that one of the sources is extremely inaccurate but lacks the data lineage information to identify where the data has been stored in various databases.
- Bugs: A software bug causes data to be calculated or transformed incorrectly.
- Information Security: An organization's data is changed by an advanced persistent threat.
- Human Error: A customer enters their phone number incorrectly.

Types of Big Data:

Sensors: Data streamed by sensors such as video, cameraor GPS. Many types of sensors have become very cheap which can be used in our day to day life. These sensors plays a vital role in the growth of big data in coming years.

- **Machine Data**: The data which can be produced by machines such as commands, calculations and information streaming of any automation system.
- **Scientific and Medical**: The data produced by scientific instruments and medical equipments like weather forecasting system and satellite.
- **Communications**: communications trough emails, and other social networking sites.
- **Knowledge**: knowledge acquired through various blogs and other documentation.
- **User Interface**: User interface such as social media apps can produce large stream of user input data.
- **Transaction**: Commercial transactions such as using various ecommerce purchasing or using stock market.
- **Derived Data**: Data calculated or interpreted from other data.

Eg. Face book, you tube, twitter google+ and linkden. They can produce huge amount of data.

Why is big data Important:

The importance of big data doesn't revolve around how much data you have, but you do with it. You can take data from any source and analyze it to find answer that enable:

1. Cost reduction
2. Time reduction
3. New product development and offerings
4. Smart decision making

A. Applications:

Understanding and Targeting Customers: this is one of the biggest and publicized areas of big data use today. Here big data is used to better understand customers and their behaviors and preferences. Companies are keen to expand their traditional data sets with social media data, browser logs as well as text analytics and sensor data to get a more complete picture of their customers.

B. Understanding and Optimizing Business Processes: Big data is also increasingly used to optimize business processes. Retailers are able to optimize their stock based on predictions generated from social media data, web search trends and weather forecast. One particular business process that is seeing a lot of big data analytics is supply chain or delivery route optimization. Here geographic positioning and radio frequency identification sensors are used to track goods or delivery vehicles and optimize routes by integrating live traffic data.

C. Personal Quantification and Performance Optimization: Big data is not just for companies and governments but also for all of us. We can now benefit from the data generated from wearable devices such as smart watched or smart bracelets. Take the up hand from jawbones as an examples: the armband collects data on our calorie consumption, activity levels and our sleep patterns. While it gives individuals rich insights the real value is in analyzing the collective data.

D. Improving healthcare and public health: the computing power of big data analytics enables us to decode entire DNA strings in minutes and will allow us to find new cures and better understanding and predict disease patterns. Just think of what happens when all the individuals' data from smart watches and wearable devices can be used to apply it to millions of people and their various diseases. The clinical trials of the future won't be limited by small samples sizes but could potentially include everyone.

E. Improvingsports performance.

F. Improving science and performance.

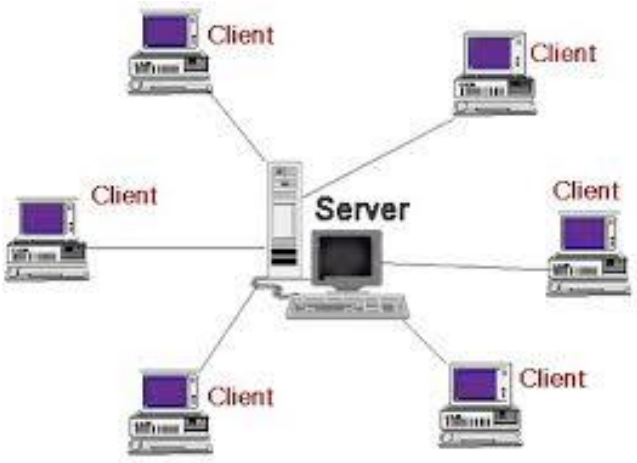
G. Optimizing machine and device performance.

H. Improving security and law enforcement.

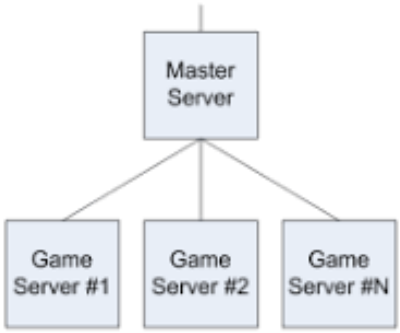
I. Improving and optimizing cities and countries.

Architecture :

Traditional data is processed with client server architecture.



Big Data is processed with Master Slave architecture.



Main component of Big Data:

- HDFS (Digital Data)
- MR (Map Reduce –write business logic to process (written in core java)
- SQOOP (SQL+HADOOP: can export or import SQL data in hadoop or vice versa).
- HIVE (It is a data ware house)
- OOZIE (workflow)
- FLUME (contnous streaming data like twitter, facebook etc)
- PIG (predefined component used for processing like MAPReduce)

Where is big data trend going:

Gartner says Big data is the new oil. Big data market is growing 7% more than IT companies. IBM says big data is not just a technology but also considered a business strategy. Analytics is poised to become a USD $203 billion industry by 2020, becoming one of the most fastest growing industries forever. So here we can enhance our knowledge and skills in one of the fastest growing sectors with great learning. These are the some big data analytics

- **Internet of Things** (IOT): around 20.8 billion devices are connected by 2020 through IOT.
- **Artificial Intelligence** (AI): by 2020 85% customers will interact with AI instead of human.
- **Augmented and Virtual Reality**: An estimated 64.8 million AR devices will be shipped in 2020.
- **Digital Assistant**: 40% of mobile interactions will be handled by digital assistants by 2020.
- **Security Analytics**: A fortune 500 company generates 10TB of data every month to look for security flaws.

Where does big data from:

Big data just keeps growing and growing, according to forrester research. The average organization will grow their data by 50% in the coming year. Overall corporate data will grow by a staggering 94% and databse systems will grow by 97%. Server backups for disasters recovery and continuity will expand by 89%. There are 3 major challenges in big data/

- Storing
- Processing
- Managing it efficiently.

by reducing the data footprint, virtualizing the reuse and storage of data and centralizing the management of the data set. Big data is ultimately transformed into small data and managed like virtual data.. now that the data footprint is smaller, organizations will dramatically improve data management in three key areas:

- Less time is required by applications to process data.
- Data can be better secured since the management is centralized even though access is distributed.
- Result of data analysis is more accurate since all copies of data are visible.

Why it needs attention:

1. By 2020 at least 1/3rd of data will be passed through a cloud server.

2. Every minute 300 hours of video are uploaded on YouTube.
3. 1 trillions photos (80% by smart phones) will be shared online.
4. By 2020 1.7 MB will be created every second for each human.
5. Digital universe of data will grow from 4.4 Zettabytes to 44 Zettabytes (44 Zettabytes=44 trillion GB)
6. 1.2 trillion searches in google every year.
7. White house has already invested $200 million in big data projects.

Big data tools:

Big data tools are used for saving time, money and recovering management insights. Some of the tools are

A. Data storage and management: these are some popular data storage tools like mangoDB, Cassandra, neo4j, Apache HBASE and ZooKeeper, talend, hadoop and Microsoft.

B. Data Cleaning: data should be cleaned, reshaping and well structured by using the MS Excel and open Refine tools.

C. Data Mining: it is a process of discovery insight in the databse. The TERADATA and Rapidminer tools are used for this purpose.

D. Data Visualization: it is used to combine complex data in a well tabular form. The pictorial representation is the best method to understand efficiently. Some tools are tableau, IBM Watson analytics and plotly.

E. Data Reporting: the power BI tools are used.

F. Data Ingestion: is process of gathering the data in hadoop form which can be done by scoop, flume and apache storm.

G. Data analysis: requires asking questions and finding the answers in questions. It can be done through HIVE, Pig, MapReduce and Spark.

H. Data Acquisition: is used for acquiring the data through scoop, flume and storm.

Advantages of Big Data Tools:

- Provide the analyst with advanced analytics algorithms and models.
- Can run on big data platforms such as hadoop or any high performance analytics systems.
- Can work with structured and unstructured data from multiples sources.
- Easy to visualize the analyzed data.
- Easy to integrate with other technologies.

Big data technologies

Big data technologies are used to perform accurate analysis to generate conclusions and predictions so as to minimize the risk in future.

Types of Big Data Technologies: There are two types:

1. **Operational Big-**Data: is normal day to day data which we generate, the data that the organization produced which might be online transactions like social media or the data may produced from school, college etc. it is kind of raw data which can be used for analytical studies.

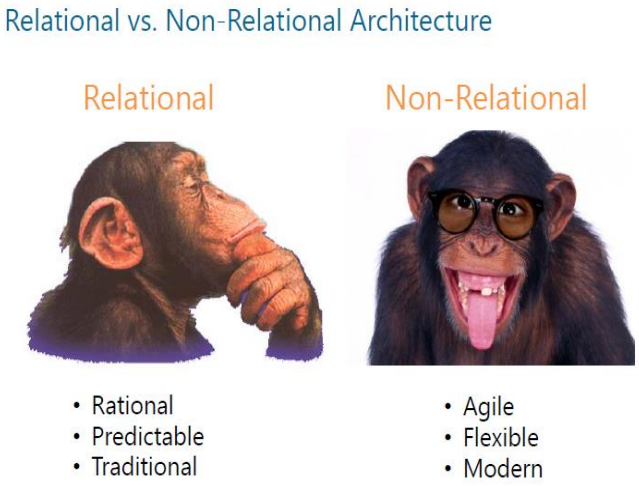
   Eg : Online booking like airline ticket, rail ticket and movie ticket.

   Eg : Online shopping like flipkart, walmart and amazon etc.

   Eg : social Media: like face book, twitter

   Eg : An employee details of multinational agencies.

2. **Analytical Big Data**: is an advanced part of big data technologies. The analytical data is little bit more complex than operational big data. The analytical data is where the actual performance comes into the picture.
   Eg : stock market and space mission, weather forecasting and medical fields.



Top Big-Data Technologies: these can be categorized into four category.

A. Storage:
B. Analysis
C. Data Mining
D. Data Visualization.

Big Data Technologies in data Storage:

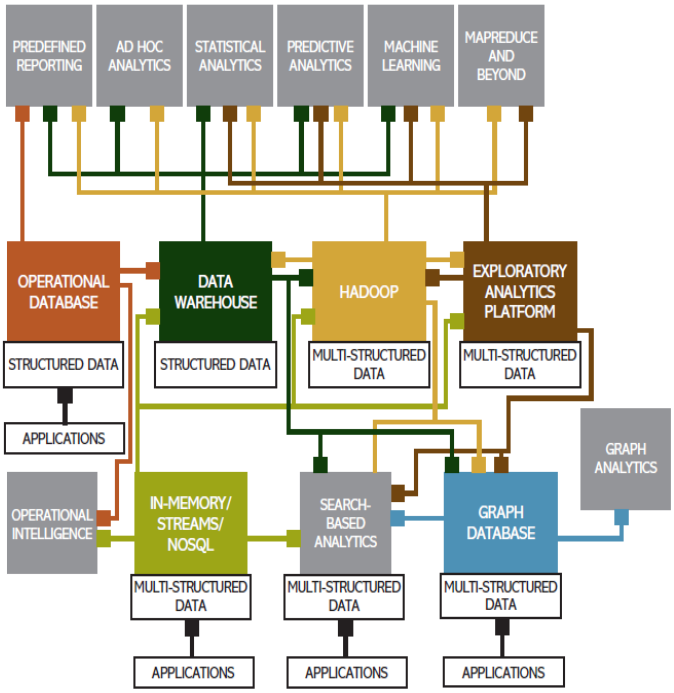
- **Hadoop** : Hadoop database were designed for distributed data processing environment where commodity hardware is used for programming model. Hadoop framework has the capabilityto store and analyze the data present indifferent machine and different locations with high speed. It was developed by Apache s/w foundation in the year 2011 10th of Dec. there code was written in java and current stable version is Hadoop 3.11. some of major companies that can use the hadoop are : Microsoft, Hartonworks, cloudera,MAPR, Intel and IBM.
- **MongoDB**: mongoDB provide the relational schema. It is a NoSQL document database. It was developed by MongoDB in the year 2009 11th of feb. The code is written in C++, Go, Java Script and python. The current stable version is MongoDB 4.0.10. some of major companies that will used the MongoDB are MS Access, MS SQL Server, My SQL and Mongo itself.
- **RainStor**: It is a s/w company that developed database management of the same name. the RainStor for analyzed and managed for large enterprises. It uses deduplication technology. It was developed by RainStor s/w company in the year 2004. It works like SQL and current stable version of RainStor is 5.5. some major companies are Barclays and Credit Suisse.

Big Data Technologies in Data Mining:

- **Presto**: It is an open source distributed SQL query engine designed for running analytical data for all sizes ranging from GB to PB. Presto is started by analyst who expects respond time ranging from sub second to minutes. Presto breaks the first choice between having fast analytical season expensive commercial solution or fre solution that require excessive hardware. It is an open source distributed SQL query engine. The presto was developed by apache foundation in the year 2013 in java language. The current stable version of Presto is 0.22. some major companies used presto are facebook, Repro, Checkr, Netflix and airbnb.
- **RapidMiner**: It is a robust graphical user interface. Rapidminer is a centralized solutions that has a powerful featured of rich interface. Rapidminer accepts many languages. It was developed by Rapidminer in the year 2001 and it uses java language. The current version of Rapidminer is 9.2. some of major companies are BCG, Infocus, slalom, Domino's and vivint.smartHome.
- **Elastic Search**: It is a search engine. It is a full text based search engine and based on Lucene library. It is developed by elastic NV in the year 2012 and use java language. The major companies are Netflix, Accenture, stackoverflow, medium and linkdin.

Big Data Technologies in Data Analytics:

- **Apache Kafka:**is a distributed streaming plateform. The streaming plate form has 3 capabilities publish, subscribe and consume. It was developed byApache S/w foundation in the year 2011 and use java language. The current version is Apache Kafka 2.2.0. the major companies are linkden, yahoo, Netflix and twitter.
- **Splunk**: is used to capture index and correlate the real time data in searchable repository which can generate graph, reports alerts, dashboards and visulalization. Splunk is a horizontal technology which can be used in application management, web analytics and security and compliance. It is developed by Splunk INC in the year 2014 and use AJAX, C++, Python and XML. The major companies are QRadar, QLabs, Trust waves and Splunk itself.
- **KNIME:** it is used to create Data flows and uses extensions mechanism. It was developed by KNIME in the year 2008 and use Java. Some of the major companies are paloalto, harnham and tyler.



R language: R is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Some of the major companies used R Language are Barclays, Bank of America and American Express.

Block Chain: Blockchain is a global online database that anyone with an internet connection can use, but it doesn't belong to anyone. Unlike traditional databases which are owned by central figures like banks and governments, a blockchain doesn't belong to any one and with the entire network looking after it, checking the system by faking documents, transactions and other information become near impossible. Blockchain store information permanently across a network of personal computers. This is not only decentralizes the information but distributes it too. This make it difficult for anyone person to take down the network or corrupt it. The many people who run the system use their own personal computers to hold bundles of records submitted by others. The records are known as blocks. Each block has a time timestamp and a link to a previous block, forming a chronological chain. It is like a giant google doc with one key difference. You can view it and add to it, but you can't change the information that's already there. The blockchain enforces this by using a form of math called cryptography which means records can't be counterfeited or altered by someone else. Bockchain most famous application is Bitcoin. It is digital currency that is created and held electronically and you can send it to anyone whether you know them or not.

CONCLUSION

Big data is important primarily because it is growing at an exponential rate. Over five exabytes is created every two days. The problem with Big Data is not just data analysis, but with discovering, harvesting, curating, storing and its management. Currently, there are massive amounts of data both structured and unstructured, that need to be analyzed in an iterative, as well as in a time sensitive manner. In response to this need, data analytical tools and services have emerged as a means to solve this problem. Big data analytics aims at deriving correlations and conclusions from data that were previously incomprehensible by traditional tools like spreadsheets. Big data analytics uses tools like Hadoop, SAS, R etc which are more powerful than previously used rows and columns. Big data analytics can help companies use data to influence not only future decisions but present decisions as well.

REFERENCES

[1] https://www.youtube.com/watch?v=8o9QxMxhTp8
[2] https://www.quora.com/What-is-an-abstract-of-data-analytics.
[3] https:/www.ugc.edu.hk/doc/eng/rgc/theme/hall/abs1.pdf
[4] https://resources.sei.cmu.edu/asset_files/Presentation/2014_017_101_89659.pdf
[5] https://link.springer.com/chapter/10.1007/978-1-4842-1910-2_16
[6] https://www.newgenapps.com/blog/what-is-big-data-analytics-benefits-challenges