

BigData - Data to Insight

Amarendra Mohanty¹, Dr.D.Venkata Subramanian²

¹ Research Scholar, Dept. of MCA

Hindustan Institute of Technology & Science, Chennai, India

² Professor & Dean, School of Computing Sciences

Hindustan Institute of Technology & Science, Chennai, India

*Corresponding author E-Mail: ¹ mohanty.amarendra@gmail.com, ² deancs@hindustanuniv.ac.in

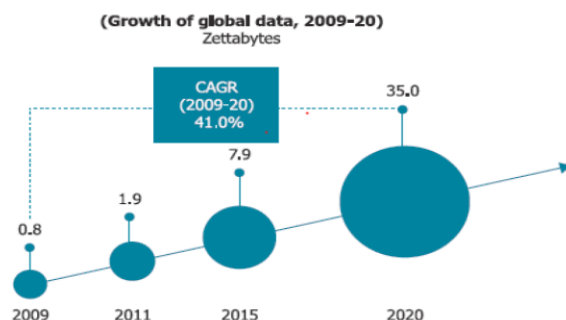
Abstract

Sign language is a language that uses manual communication to convey meaning. This includes simultaneously employing hand gestures, movement, orientation of the fingers, arms or body, and facial expressions. RASPBOT has been designed to understand not only human voice but hand signs language as well. It has been designed in such a way that it can be used by a wide variety of individuals. RASPBOT allows the users to: Ask a question normally or ask a question using sign language. RASPBOT recognizes the Hand Signs and produces the recognized Alphanumeric Character as output. It forms meaningful sentences and give responses to the queries by recognizing the characters thus breaking down the barriers between normal people and specially disabled people. These features have been represented by functions in the program. RASPBOT then understands the question and then replies appropriately.

Keywords: Chatbot, RNN, Convolutional Neural Networks, Sign Language ,Deep Learning..

1. Introduction

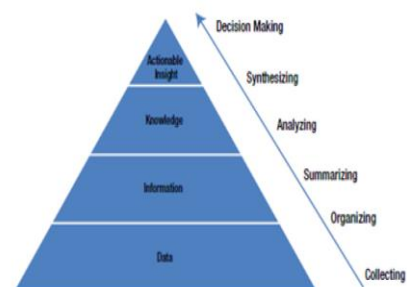
We are sinking in data and starving for knowledge. We are having oceans of data, rivers of information, streams of knowledge and a drop of understanding. Data has always been important in the business world, for decade's we are using outdated data storage techniques like file systems and databases. Technology created the need for new analytics capabilities and new technology platforms. With technology's increasing reach into people's lives data creation is occurring at a record rate and number of suppliers from where data is being generated has also undergone a change. Data are becoming a crucial part of every economy, industry, organization and individuals, it is being actively took by organizations to understand their customers, suppliers, partners and operators. The data is produced at a fast pace around 2.5 billion GB every day and more than 90% of the data is available today which is created in the past 3-4 years. According to IDC, data generated universally is likely to reach 35.0 Zettabytes witnessing a 41.0%increase cumulative aggregate rate of change (CAGR) between 2009 and 2020 as shown below.

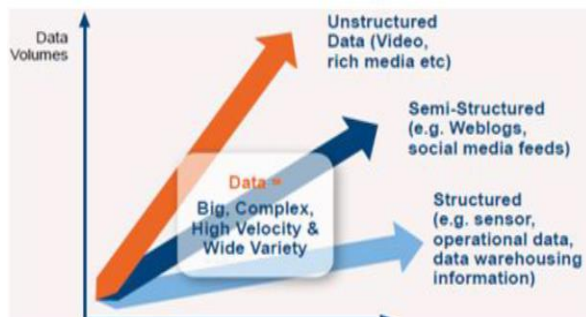


This paper explores the concepts like decision making issues in a non-big data spectrum, high level approach of Big Data and how big data analytics help business to take improved decisions and drive performance, challenging real-world problems and how big data provide solutions in various domains like health care, finance, automobile and retail, advantages and limitations of Big Data, discover the real world Business Benefits from Big data, Big Data tools and technologies which are used for grappling the skyrocketing amount of data.

2. Decision Making

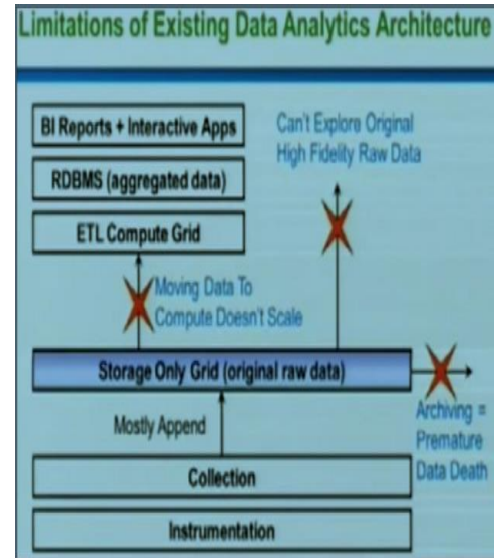
When we say "Data", what is the first thing that pops into our mind? Perhaps, it is a spreadsheet or a weekly status report or a data warehouse. While all are true, the term "data" means much more. Regardless the way, data is managed in an enterprise, if it is monitored and benefited properly, it can bring immense corporate values. Fig. no. 1, explains the process cycle of data, steps undergone from raw data to decision making. Data can be of any type like as shown in Fig. no.2, Structured, Semi-structured or Unstructured





In the initial 2000's, the reception of ideas like Business Intelligence (BI), Enterprise data warehouse (EDW), and analytics, aided enterprises to alter raw data collections into actionable height. Analytics applications such as Financial Analytics, Risk Analytics, Customer Analytics and Product Analytics became an essential part of the business applications architecture of any enterprises. But all of these applications were using structured data. Structured data is only 10 percentage of the whole existing data. Currently, the data is transactional hence, can be kept in rows and column with a recognizable structure. The data outline and reliability agree it to respond to simple requests to come up as a useful information, based on the parameters offered by organization's and operational needs and its restriction. The rest 10% of existing total data are semi-structured data which a form of structured data are but not confirm to the authorized structure of data models, so the question is about rest 80% of data. The 80% of data prevailing are unstructured data. The unstructured data have information that could aid them to make a improved business choices. Tactlessly, it is difficult to analyze unstructured data. To come up with the problem, organizations have curved to a numerous of different software keys designed. To find unstructured data and extract significant information. The main advantage of this is the skill to collect actionable information that can help a business prospers in a modest environment. There are certain limitations of the existing Data Analytics architecture as shown in figure 4, Thus, the daunt-

less of improved analysis, monitoring, and capitalize the unstructured data, gave natal to the era of Big Data. Big data is a dictate which presently appears to be popular. It is a formula that we exist in a world of Big Data, where data never sleep its keep on going and expanding. Every minute huge quantities are being produced from various ways example, every phone, website and application across the internet give continuous data. The curious thing is from where data is being created and where does it go.

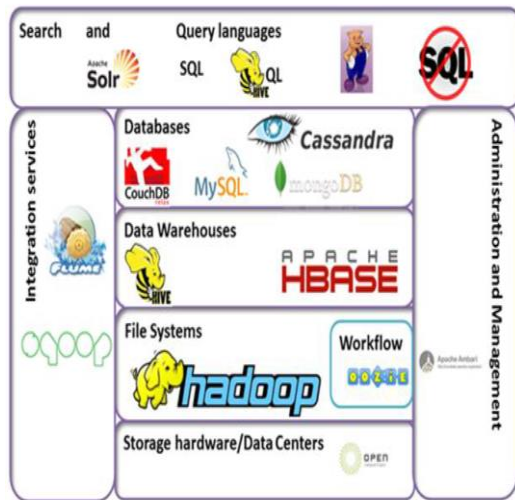


3. High Level Approach Using the Big Data Analytics Platform

Big data analytics is the process of applying advanced analytics and conceptual techniques to expose hidden patterns and unknown correlations for real decision making and answer questions. Previously, considered beyond reach big data analytics help business to take decisions, predict and identify change and new chances. Some of the techniques used to investigate big data are natural language processing (NLP), Visualization, Time series analyses Sentiment analysis, and Crowd sourcing which are described as Natural language processing (NLP). It is a set of methods from a sub field of computer science and linguistics that uses computer algorithms to examine human dialectal. Sentiment analysis, the key features of this analyses is detecting the feature, aspect, or product with which a feeling are being stated and determining the type, "polarity" (i.e., positive, negative, or neutral) and the degree and strength of the sentiment. Visualization: Techniques used for creating images, diagrams, or animations to communicate, understand, and improve the results of big data analyses. Some of the widely used big data Visualization tools are Tag cloud, Cluster gram, Heap map etc. Time series analysis is a set of techniques from both signal processing and statistics for analyzing sequences of data points, on behalf of values at following times, to retrieve meaningful features from the data. Crowd sourcing is a method for collecting succumbed data by a huge group of people or community through an open call, usually through networked media as web. A fuzzy based method can also be used to assess the same. Linguistic variables are used with a limit range of values. The fuzzy inference engine receipts the inputs as fuzzy sets and lastly defuzzification bounces a crisp output.

4. Big Data Ecosystem

Big Data is comprehensive and includes many trends and innovative technology development. Listed below are some of the more prominent technologies developed to support big data techniques in a gainful manner.



Hadoop

The Apache's hadoop software library is a framework that permits the dispersed processing of huge data sets across bunches of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, single of them offers a local computation and storage.

Ambari: Ambani is a web-based tool, which offer monitoring provisioning and supervision. Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig and Sqoop.

Flume is a available, distributed and reliable system for professionally accumulating, combining and moving huge amount of log data from different sources to a centralized data store.

Sqoop allows import and export of data from structured data stores example:- relational databases, enterprise data warehouse and NOSQL systems.

Pig is a stage for examining huge data sets that contains high-level language for stating data analysis programs, joined with infrastructure for estimating the programs.

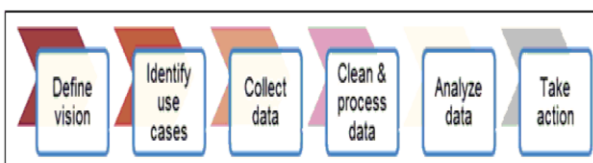
HBASE is a scalable, distributed database that support storage of structured data for huge tables.

HIVE eases questioning and handling large datasets exist in distributed storage. It facilitates a mechanism to show structure onto the data and query the data by means of an SQL-like language called HiveQL.

Zookeeper – It is used for distributed applications by providing high-performance coordination service

Oozie is used to manage Apache Hadoop jobs by workflow scheduler system.

5. Implementing Big Data strategy



Big Rewards of Big Data

Big Data has great profile and outstanding successes. The re-nowned company like Target, MasterCard, Tesco and Amazon have used Big Data for competitive advantages in retail section. Perhaps one of the strongest case for Big Data use is done by Google to sell and target advertising. Till this point, we have well-defined Big Data and its high-level method now we can see, how the technologies offered above in action in depth, exploring how organizations have successfully installed Big Data tools and visualized astonishing results. Few Big Rewards of Big Data are listed as explaining real time application to find solution.



one of the prior company who explored data analytics in the field of marketing. Data analytics and big data are beneficial under certain scenarios, first when there are clear questions and second where based on behavior models can be created and tested. One of the question asked by US Target is "Can we figure out if a customer is pregnant" and a model developed was quite predictive.

Target desired to confine a very attractive and profitable market for new parents as new parents devote lots of time and money in shopping and making new purchasing habits, and structuring loyalty among the audience, which can be very profitable. This idea is so appreciated that Target worked to identify the clients who might be pregnant. Beforehand a new parent purchases the first diaper, or registers for the baby shower. Since Target encapsulates and records massive number of consumer data, they analyse purchase patterns by observing trends and inspects the items parents inclined to buy before pregnancy, like vitamins, odorless lotion, hand towels, etc. By mathematical intrigues, Target determined the probability that couples were pregnant and used these visions to market to the couples before their child's birth, making client faithfulness and capturing an tremendously valuable market section.



They applied Big data in another applauded manner. They got an idea about buyer requirement and then they used big data techniques to report, insights customer data and forecast to the merchants. Per transaction the data get received concerning the amount of the deal, the merchant name, time, date and credit card number. Afterwards the amount number is stripped and the data is made unidentified. However, the problem faced was the disorder data as the name of the merchant on a point-of-sale is a free-text field, leads to diverse names for the similar merchants, retail chains or businesses. MasterCard sells information and spending pattern (all individual recognizable info is exposed keeping the private information safe) on detailed customer segments and spending patterns. MasterCard disintegrate all the data that can get hands on, including five years of historical data, and correspondingly creates segments about loyal customers to a certain product group This provides great vision.



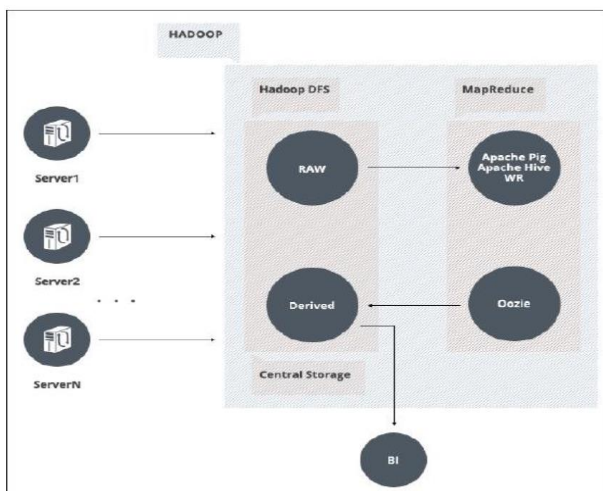
The data revolution is transforming the business landscape. Big Data at Ford has formed a entire host of novel potentials for the automaker. Ford is going in the correct way with its big data approach to be ahead of his competitors. The amount of data produced internally from numerous business actions, vehicle research activities, data from customers demand specialized techniques or platforms to manage data. Ford is depending on open source tools (Hadoop) for managing huge data and using big data analytics like data mining and text mining to leverage big data effectively. This helped to initiate planned choices for emerging innovative products. It helps to mitigate gate business risks, progress customer relations and drive income growth.

Telecom Sector:

Big data was also used for raw data processing in Telecom vertical. Customers starts making usage on a network as soon as they starts using the products and services sold by the operator. An event is a solitary bill occurrence of product usage, which is captured by a network electronically.

Problem: Most telecom operators don't have the hardware and software capacities to work effectively with raw UDRs because of the enormous number of events and their attributes that are generated every minute. Often operators skip raw UDRs and operate with formatted UDRs, which have lost some unique and significant inputs.

Solution: Raw UDRs in telecom are a classic example of Big Data to solve the problem, specialists propose to use Apache Hadoop, an open-source Big Data framework that provides a cost-effective and innovative approach to advanced analytics and can be used as a core Big Data technology for enterprises.

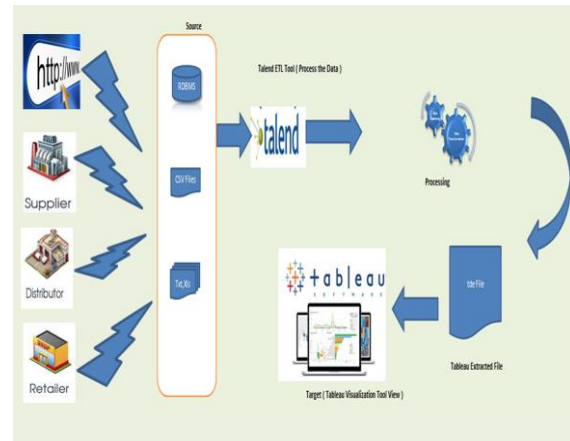


Benefits: Above solution led to benefits like Significant cost savings because of Hadoop is license-free distribution, ultimate scalability, commodity servers in use, and reduced TCO. Faster processing times – data storage and processing at least 5 times faster and becomes exponentially faster greater amounts of data. There is an opportunity to exploit data that was inaccessible before. Numerous dimensions for predictions and also more detailed reporting.

Research on Healthcare:

To improve health care and preventing diseases, big data can also be Leveraged. Big data can ease action on the adaptable risk factors that contribute to a great fraction of the chronic disease, such as physical activity, diet, tobacco use, and exposure to pollution.

Here is a proposed architecture to get insight from Data.



6. Pros and Cons

Till now we covered “progressive inclusion” to describe the evolution of big data and its wide variety of techniques and technologies usage to provide solutions to various real time problems. Now let's consider Pros and Cons that are associated with big data.

The biggest problem in the world can be thought of solving by taking the advantage of big data

Data accumulated from past joined with genome mapping can help to discover solutions for diseases or genetic disorders.

- Ecological data can assist to forecast climate changes and used to better farming methods.

- Industries can use big data to help make improved business choices. By predicting market conditions business can be re-stored and how well a product is doing and decide to scale up or scale down production. It may be able to discover flaws much earlier within their construction line.

Big data has also potential of decisions making. Big data can aid us fight crime by collecting intelligence and evidence in contradiction of criminal action. It might help to take a practical method to security threats by choosing trends that points illegal or terrorist movement.

At a communal level, big data along with troop obtaining can help to solve ordinary problems.

Big Data can help checking Food Security matters through Media and Analytics.

Every day following up of vital product prices through data mining and analytics

City crime pattern analysis, unemployment, education, social grading and financial connections.

Twitter and insights of catastrophic related stress

Analytics and understanding of social discussions through Big Data.

Effective Traffic Signal Management

The main drawback is privacy. One of the utmost fears, which is now gaining larger recognition, is the danger of theft of data. Though the data may not have name precisely but with correct analytics an eerily precise outline of somebody which can have a lot of personally identifiable information (PII). For instance when the NSA surveillance news was broken out previously last year, the NSA specified that only phone metadata was being logged, as phone metadata has no names, but numerous persons went on to show how that data can tell more that we ponder. Further example of Facebook's tagging feature, that lets a friend to either take a photo of individual and tag individual or check in to a location and tag individual as being with them. The image and location of individual can be stored and mined.

7. Business benefits

The introduction of Big Data provided many business benefits to organizations, to access their data more than ever before. Previously unstructured data was considered as 'dead' and of no worth, but by the upcoming of Big Data it can be collected and analyzed to profit and uplifting organizations. The prosperity of data available through Big Data also allows marketing strategies to be enhanced and more precisely targeted. This can aid to significantly rise in customer base, and push the organization ahead of the competition. Developments in these areas of business can eventually lead to an rise in income to the organization. It also able to both cut costs and attract more clients. Big data delivers business visions beyond traditional data from transactional systems. Effective ways and usage of correct analytics tools help in spinning "big data" into "big insights". Businesses are using the power of insights with the use of big data to instantly found who did what, when and where.

Inferring valuable insights from very big amounts of structured and unstructured data from dissimilar sources in dissimilar formats require the proper structure and the proper tools. Some of the possible business benefits from applying an effective big data insights methodology include:

1. Appropriately insights from the massive amounts of data. This comprises those previously kept in company databases, from external third-party sources, the Internet, social media and remote sensors.
2. forecasting of events and Real-time monitoring of events that influence moreover business performance or operation.
3. Capability to discover, obtain, extract, manipulate, examine, attach and imagine data with the tools of choice.
4. The ability of Hadoop for sizes to manage huge quantities of data, in or out of the Cloud, with authentication and confirmation.
5. Classifying important information that can improve decision quality.
6. Modifying danger by enhancing the complex verdicts of unplanned actions more quickly.

8. Conclusion & Next Steps

Enterprises are using various Open source and Commercial tools and technologies for grappling with the skyrocketing amount of data they have to handle as that data proliferates into Petabyte, Exabyte, Zettabyte, or a Yottabyte stage.

The data can be used to aid people to resolve the world's complications or instead can be used to further greed and devastation. There are limitless options to utilize Big Data. Taking the positive side, that can originate from examining big data, it would be better to use proper security and governance frameworks to safeguard that only good things occur. A common, horizontal big data analytics platform is necessary to support a diversity of analytics applications. Such a platform examines inward data in real time, makes associations, crops visions and discloses those insights to various applications. Storing and processing huge amounts of data is no longer the issue. The challenging part is to know what needs to be done within the big data analytics platform in order to generate better worth. Big data analytics platform can be used to study the behavioral analytics of students, drivers, customers etc. Big data has develop possible as lucrative methods have appeared to trained the volume, velocity and variability of enormous data. The challenge is to know appreciated forms and information within this data which is earlier concealed because of the quantity of effort required to extract them.

References

- [1] 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", McKinsey Global Institute, May 2011.
- [2] David Turner, Michael Schroeck and Rebecca Shockley, "Analytics: The real-world use of big data in financial services", IBM Global Business Services Business Analytics and Optimization Executive Report, May 2013.
- [3] Vincent Charles, Tatiana Gherman, "Achieving Competitive Advantage Through Big Data", Middle-East Journal of Scientific Research 16 (8): 1069-1074, 2013.
- [4] Thomas H. Davenport, Jill Dyche, "Big Data in Big Companies", International Institute for Analytics, May 2013.
- [5] Hsinchun Chen, Roger H. L. Chiang, Veda C. Storey, "BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT", Business Intelligence Research
- [6] Elena Geanina ULARU, Florina Camelia PUICAN, Anca APOSTU, Manole VELICANU, "Perspectives on Big Data and Big Data Analytics", Database Systems Journal vol. III, no. 4/2012 3
- [7] Bernice Purcell, "The emergence of "big data" technology and analytics", Journal of Technology Research.
- [8] Anthony G. Picciano, "The Evolution of Big Data and Learning Analytics in American Higher Education "
- [9] Eaton, Deroos, Deutsch, Lapis, Zikopoulos, " Understanding big data: Analytics for enterprise class Hadoop and streaming data ",2012.
- [10] Richard M. Burton, Dolly Mastrangelo, Fabrizio Salvador, " Big Data and Organization Design ", 2014
- [11] Dr. D.Venkata Subramanian, Dr. K.Pradheep Kumar "Fuzzy Based Modeling for an Effective IT Security Policy Management", SAI Computing Conference 2016.