Volume 6, Issue 5, May 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Vision Entitled Transforming Big Data to Analytics: Methodology and Future Trends

S. Aarathi*, P. Srilakshmi, B. Kiranmayee
Dept of CSE, G. Pulla Reddy Engineering College, Kurnool,
Andhra Pradesh, India

Abstract— Big data is defined as large amount of data generated from various data sources that requires new technologies and architectures so that it becomes possible to extract value from it by capturing and analysis process. The big data phenomenon concentrates on collection and processing of the massive data sets. The analytics procedure describes the steps in the transformation of big data to analytics. This paper concentrates on steps involved in transforming big data to analytics, its applications in the real world. It revolves around the basic areas of data storage, data mining, types of analytics prescribed in the model development, visualisation and up gradation. We discuss various software resources applicable in the transforming process of big data to analytics at each step. Further we also refer to the evolving advanced technology cloud computing and its effect in efficient data storage and management of big data.

Keywords— big data, big data analytics, data storage (structured &unstructured), Hadoop , HDFS, visualisation

I. INTRODUCTION

The current day to day society is becoming more instrumented and the organisations are producing and storing massive amounts of data. Managing such massive data, analysing it, gaining insights from it is a challenging aspect. This massive amounts of data generated from various sources and forms like various private and government organisations, MNC's, the world's most popularly used social networking sites all around like the twitter, facebook, linkedin are generating massive amounts of data usually in exabytes. .The main difficulty in handling such large amount of data is because that the volume is increasing rapidly in comparison to the computing resources

The US healthcare system alone already reached 150 exabytes five years ago[1].Before long,we will be dealing with zettabyte (1021) and yottabyte (1024) data for countries with large populations including emerging economies, such as China and India. This trend is due to the fact that multiscale data generated from individuals is continuously increasing, particularly with the new high-throughput sequencing platforms, real-time imaging, and point of care devices, as well as wearable computing and mobile health technologies.

Table 1: Terms Used for Big Data Storage Capacity

Term	Size	Example of Capacity
Gigabyte (GB)	1,000,000,000 (one billion) bytes	1 GB = 2 hours of CD-quality audio or 7 minutes of HDTV
Terabyte (TB)	1,000,000,000,000 (one trillion) bytes	1 TB = 2,000 hours of CD-quality audio, or almost 5 days of HDTV
Petabyte (PB)	1,000,000,000,000,000 (one quadrillion) bytes	1 PB = 7 weeks of HDTV, or 1.5 million 64 GD iPods full
Exabyte (EB)	1,000,000,000,000,000,000 (one quintillion) bytes	1 EB = 16 months of HDTV, or 15 million 64 GB iPods full

Previously the storage of data was a great issue but now the processors are extended with a memory capacity of 1TB.Though the processor capacity is extended the volumes and the speed at which the data is ranging over the web is drastically increasing. Each time a new storage medium was invented, the amount of data accessible exploded because it could be easily accessed. Explosion of data has not been accompanied by a corresponding new storage medium. It requires the scalability of the resources and their efficient storage for their future processing. Thus this massively generated amount of data is often termed as the big data. Any data cannot be referred as big data, According to MC.DOUGLANEY for a data to be treated as big data it needs to satisfy the basic 3 V's viz:

A) VOLUME B) VARIETY C) VELOCITY slowly the updation on it for providing good QOS (Quality of Service) also added VERACITY and VALUE but the volume, variety and velocity are defined as the basic one's.

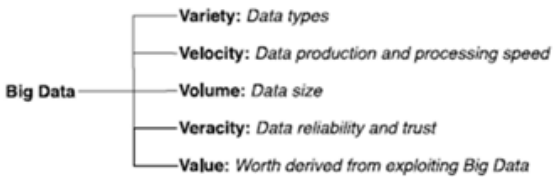


Fig 1: Basic V's of Big Data

A. Volume

The word BIG in Big data itself defines the volume. At present the data existing is in exabytes, petabytes, Zettabytes and is supposed to increase to yottabytes in nearby future. The social networking sites, health service providers, private and government organisations are themselves producing data in order of terabytes every hour out of which major content of data is produced through the social networking sites and this amount of data is definitely difficult to be handled using the existing traditional systems.

B. Variety

Data being produced is not of single category as it not only includes the traditional data but also the semi structured data from various resources like web Pages, Web Log Files, social media sites, e-mail, documents, sensor devices data ,health care providers, diagnostics, medical images both from active passive devices. All this data is totally different consisting of raw, structured, semi structured and even unstructured data which is difficult to be handled by the existing traditional analytic systems

C. Velocity

Velocity in Big data is a concept which deals with the speed of the data generated across sources. This characteristic is not being limited to the speed of incoming data but also speed at which the data flows. For example the data from the sensor devices would be constantly moving to the database store and this amount won't be small enough. Thus our traditional systems are not capable enough on performing the analytics on the data which is constantly in motion.

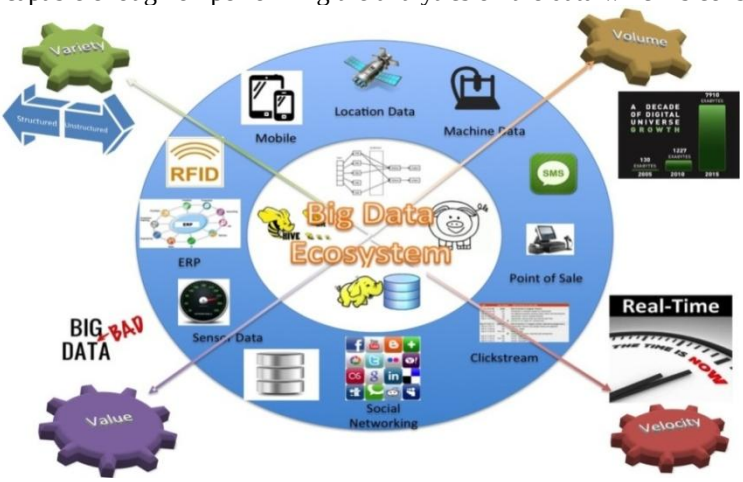


Fig 2: Data Sources of Basic V's

The other V's of big data are the Variability (consistency of data over time), Veracity (trustworthiness of the data obtained), and Value. Veracity is important for big data as, for example, personal health records may contain typographical errors, abbreviations, and cryptic notes[1]. Ambulatory measurements are sometimes taken within less reliable, uncontrolled environments compared to clinical data, which are collected by trained practitioners. The use of spontaneous unmanaged data, such as those from social media, can lead to wrong predictions as the data context is not always known. Furthermore, sources are often biased toward that young, internet savvy and expressive online. Last but not least, real value to both patients and healthcare systems can only be realized if challenges to analyse bigdata can be addressed in a coherent fashion.

II. ORGANISATION OF THE PAPER

This paper deals with the evolution of big data in the I).Introduction, then covers III.)Related work under which we cover , How traditional analytics is different from Big data analytics, IV).Steps in transforming big data to analytics V.) Methodologies, tools used in processing of data to gain insights. VI.) Good practices of big data, finally focus on VII.) how cloud computing can be effectively used on top of data storage in the future trends and enhancements .

III. RELATED WORK

Based on a survey conducted over the Volume at which data is generated across various sources. Each day, Face book operates on nearly 500 terabytes of user log data and several hundreds of terabytes of image data[3]. Every minute, 100 h of video are uploaded on to YouTube and upwards of 135,000 h are watched [98]. Over 28,000 multi-media (MMS) messages are sent every second [3]. Roughly 46 million mobile apps were downloaded in 2012, each app collecting more data. Twitter [87] serves more than 550 million active users, who produce 9100 tweets every second. eBay systems

process more than 100petabytes of data every day [64]. In other domains, Boeing jet engines can produce 10terabytes of operational information for every 30min of operation. This corresponds to a few hundred terabytes of data for a single Atlantic crossing, which, if multiplied by the 25,000 flights each day, highlights the data foot print of sensor and machine-produced information. These examples provide a small glimpse into the rapidly expanding ecosystem of diverse sources of massive datasets currentlyinexistence.Datacanbestructured(e.g.,financial,electronic medical records, government statistics), semi-structured(e.g., text, tweets, emails),unstructured(e.g., audio and video),and real-time (e.g.,networktraces,genericmonitoringlogs).Alloftheseapplicationssharethepotentialforprovidinginvaluableinsights,if organized and analysed appropriately.

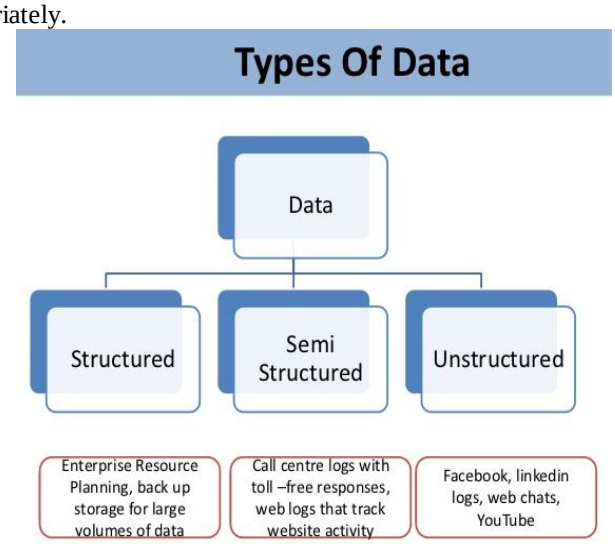


Fig 3: Types of Data

The massive amount of data collected doesn't make sense if we couldn't make real sense out of it. To make sense of the data we need to perform analytics i.e analysing the data performing pre-processing of data wherever necessary. The analytics of the data need to view data collected across heterogeneous sources. This data could be either structured, semi structured or Unstructured. Most of the real time data gathered could be either Semi structured or Unstructured

A. How Is Big Data Analytics Different From Traditional Analytics:

Traditionally Before analysing since RDBMS is structured in a proper fashion i.e we could identify common columns relations and develop relational models to predict its insights and gain knowledge. But in case of big data unlike traditional data here most of the incoming data streams are either Semi Structured or UnStructured where management of such data becomes quite complex since it doesn't have: 1. Fixed structure 2. Data model and 3. Data collected from various sources. Big data is flat in structure,

There are certain analytic challenges with big data:

- **Traditional RDBMS** fails to handle big data due to its limitations.
- Big data cannot fit into the memory of a single processor.
- Processing of this massive amount of data is time consuming.
- Scaling with traditional RDBMS is expensive.

The real world is using big data analytics in various sectors of life like medical images, diagnostics, EHR's(Electronic Health Records)[4],sensor data, trading statistics, web info, web logs, social media etc..There are various applications of big data in the day to day life. Analysis of data is a very crucial step in big data analytics to gain knowledge thus develops insights on data.

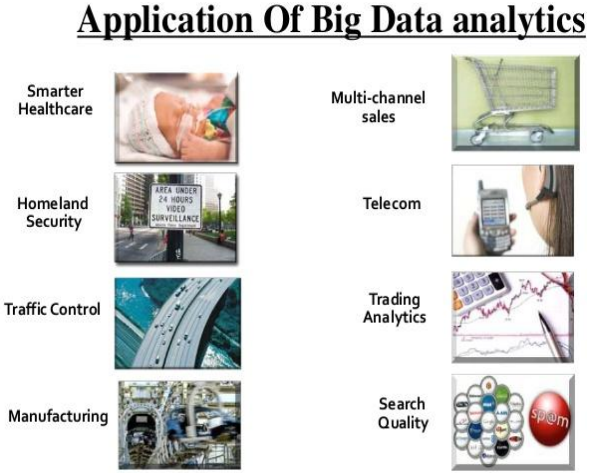


Fig 4: Big data Applications

B. Why Analytics Is Important:

The analytics of the gathered information would make considerable results in issues like[8]

- Cost reductions, time reductions, new product development, optimized offerings, smart decision making.
- Used for fraudulent behaviour detection.
- Determining the root causes of failure.

Fig 5: Benefits of Big Data Analytics

There are basically four types of analytics viz,
i.) Descriptive (or) fact analysis, ii.) Diagnostic analysis, iii) Predictive analysis and iv.) Prescriptive Analysis

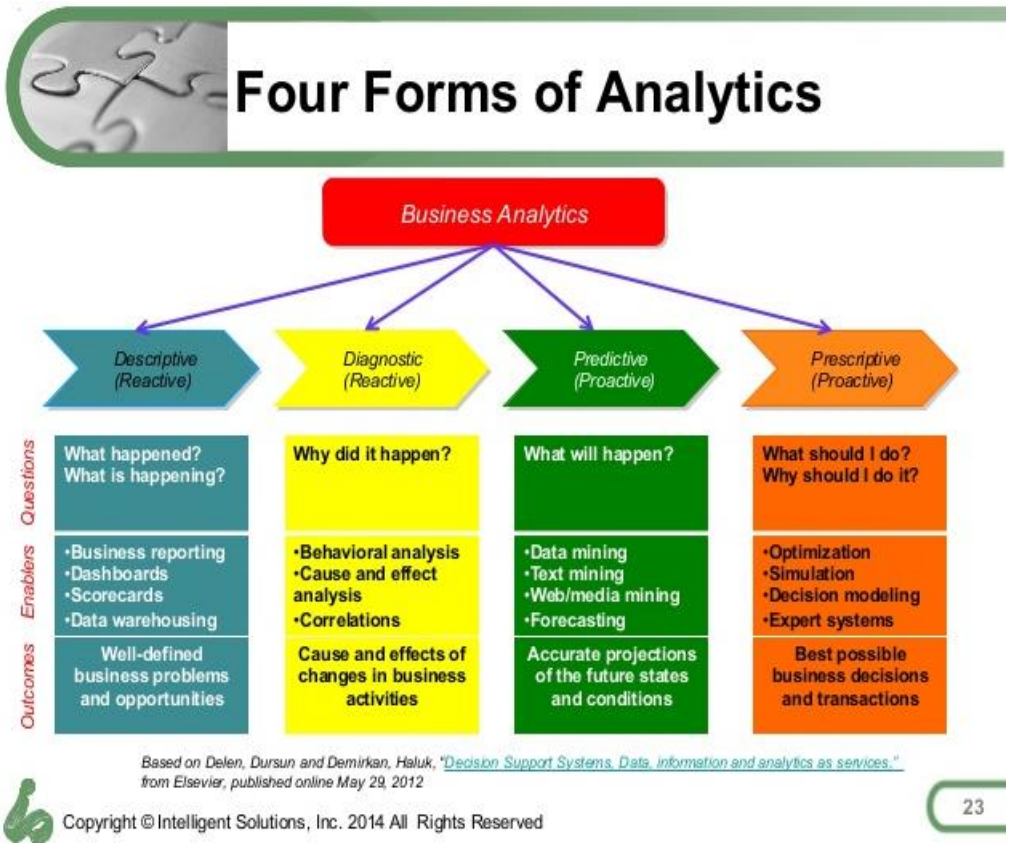


Fig 6: Types of Analytics

Descriptive or fact analytics concentrates on what has happened, where diagnostic analytics focuses on why did it happen and its root causes of happening, predictive analytics tends to concentrate on the upcoming issues with the currently available data i.e. what is likely to happen these type of analytics are generally used in trading, healthcare, marketing, business sectors, finally the prescriptive analytics focuses on what can we do further with the happening from our perspective.

As an example if we consider the Student record in an university where we are intended to learn how many failures, distinctions it can be made clear with the available sets of data where it clearly specifies how many failed, how many distinctions, this itself represents the fact analysis or Descriptive analytics.

If we raise an issue as why did they fail this leads to the concept called Diagnostic analytics.

If we want to learn about how many among the students are likely to pass in the class it would relate to the concept of Predictive analytics.

If we start analysing as what could we do about the no of failures i.e. suggestions and recommendations this leads to the Prescriptive analytics.

IV. STEPS IN TRANSFORMING BIG DATA TO ANALYTICS

To gain insights from the massive collected data, develop model, visualize it we need to process the data step by step[6].

- Identify the various data sources.
- Select right tools to collect, store and aggregate the data.
- Understand the appropriate business domain.
- Identify the tools and technology to process the data.
- Build models for the analytics.
- Visualize and validate your result.
- Learn, adapt and rebuild your model.

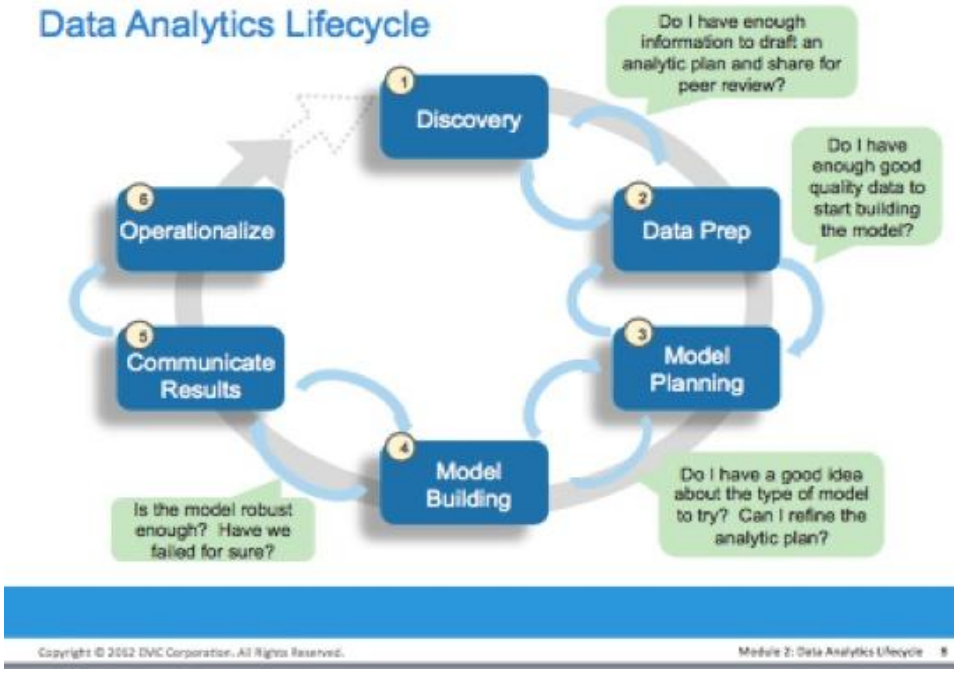


Fig 7: Phases of Data Analytics

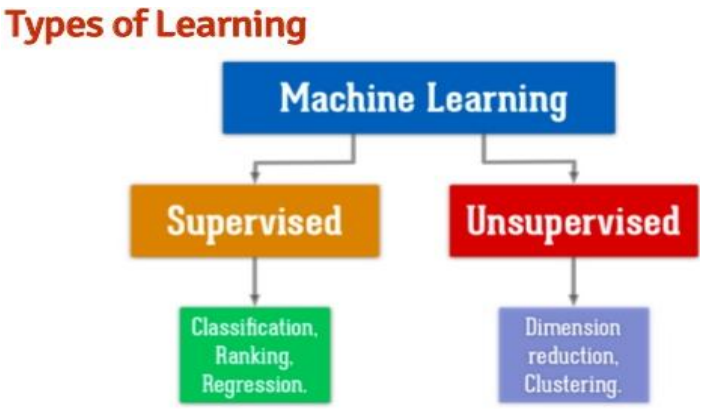
Decision makers like to base their decisions, actions and insights gained from this collected data and make a sense out of it. Extraction of non obvious patterns and using them to predict the future are no new. Here is where we deal with the KDD(Knowledge Discovery from Data) which does careful extraction of non obvious information using careful and detailed analysis and interpretation whereas Data Mining aims at discovering the previously unknown inter relations among the apparently unrelated attributes of the data by using several methods from various areas like Machine Learning, Database systems and Statistics.

Analytics contains techniques of KDD, Data Mining, Text Mining, Statistical and Quantitative Analysis, Explanatory and Predictive Models and Advanced interactive Visualisation to drive the decisions and actions[9]

Common phases of Traditional Analytics Workflow include:

- Data from various sources like databases, streams, Marts, Data warehouse.
- Collected different types of data may infer Pre processing of data before integrating, cleaning and filtering it.
- Prepared data is used to develop and estimate a model
- Validation of the model before consumption
- After validation is passed through model is consumed and apllied to data as it arrives(Model Scoring).

In the phases of analytics the last phase of learn, adapt and rebuild model[7] here we focus on the concept of Machine Learning, it refers to a field of study that gives computer ability to learn without explicitly being programmed. Ex: Playing chess.



There are **two types of Machine learning** namely Supervised and Unsupervised. To build any data/ mathematical model as part of our steps of analytics we need to have machine learning[9].

- ✓ Under **Supervised** Learning, we would have previous data sets often treated as the experience E, whenever we face a problem, we try to check out if there is any matching scenario and take the necessary action. This again has two types of learning namely REGRESSION and CLASSIFICATION.
  1. **Regression** helps out in prediction of a single value outcome, Ex: Saying a particular student would get >75 marks
  2. **Classification** predicts a class of values. Ex: Saying among a class of students some X, Y, Z would fail in the exam or secure border marks.
- ✓ Under **Unsupervised** Learning, we would be given the data where no previous datasets would be available , We do not have an idea as to how the data actually looks like, we need to make sense of it, we need to derive a structure from the given data
  1. Here we apply the mechanisms of **Dimension Reduction.**
  2. **Clustering** technique.

V. BIG DATA TOOLS AND TECHNIQUES

There are thousands of Big data tools in the market which help out in saving time, money and help in uncovering the never seen business insights[8]. To save our time all that we need to do is pick up a relevant tool for our corresponding domain. Here we state few best tools for

- ➢ Data Extraction
- ➢ Data Storage
- ➢ Data cleaning
- ➢ Data Mining
- ➢ Data Visualization
- ➢ Analysis and Integration

1) Data storage and management: The storage of big data couldn't be handled with the traditional Databases. An Open source good infrastructure called HADOOP[8], [1] is widely used for storage of large datasets on computer clusters. Big Data analytics and the open source Apache HADOOP are rapidly emerging as the preferred solution to address business and technology trends that are disrupting the traditional Data Management and Processing.

Hadoop Components in detail

- ➢ **Hadoop Distributed File System(HDFS):** Hadoop comes with a distributed File System called HDFS, which stands for Hadoop Distributed File System. HDFS is a File System designed for storing very large files with streaming data access patterns, running on clusters on commodity hardware, its block size is much larger than that of normal file system i.e. 64 MB by default, its mainly large sized blocks to reduce the number of disk seeks[1], [13]. A HDFS cluster has two types of nodes i.e. namenode (the master) and number of datanodes (workers). The name node manages the file system namespace, maintains the file system tree and the metadata for all the files and directories in the tree. The datanode stores and retrieve blocks as per the instructions of clients or the namenode. The data retrieved is reported back to the namenode with lists of blocks that they are storing. Without the namenode it is not possible to access the file. So it becomes very important to make namenode resilient to failure.
  These are areas where HDFS is not a good fit: Low-latency data access, Lots of small file, multiple writers and arbitrary file modifications.
- ➢ **MapReduce:** MapReduce is a programming paradigm which allows massive scalability. The MapReduce basically performs two different tasks i.e. Map Task and Reduce Task.[1], [13]
  A map-reduce computation executes as follows: Map tasks are given input from distributed file system. They produce a sequence of key-value pairs from the input and this is done according to the code written for map function. These value generated are collected by master controller and are sorted by key and divided among reduce tasks. The sorting basically assures that the same key values ends with the same reduce tasks. The

Reduce tasks combine all the values associated with a key working with one key at a time. Again the combination process depends on the code written for reduce job. The Master controller process and some number of worker processes at different compute nodes are forked by the user. Worker handles map tasks (MAP WORKER) and reduce tasks (REDUCE WORKER) but not both. The Master controller creates some number of map and reduce tasks which is usually decided by the user program. The tasks are assigned to the worker nodes by the master controller. Track of the status of each Map and Reduce task (idle, executing at a particular Worker or completed) is kept by the Master Process

On the completion of the work assigned the worker process reports to the master and master reassigns it with some task. The failure of a compute node is detected by the master as it periodically pings the worker nodes. All the Map tasks assigned to that node are restarted even if it had completed and this is due to the fact that the results of that computation would be available on that node only for the reduce tasks. The status of each of these Map tasks is set to idle by Master. These get scheduled by Master on a Worker only when one becomes available. The Master must also inform each Reduce task that the location of its input from that Map task has changed.

- **HIVE:** It is a Data Warehousing tool mainly used by the popular social networking site FaceBook for its storage of data. It provides SQL type of infrastructure[6]
- **HBASE**: It is a HADOOP DB, where we can store large volumes of data without compromising the performance for large scale and large volume db.
- **ZOOKEEPER**: It provides meta information for the HBASE.
- **RHADOOP**: R contains lot amount of mathematical and statistivcal formula built in. R is a programming language to do analytics on traditional data. RHADOOP is one of thepopular tools in market that helps out in the development of Statistical model.
- **SQUOOP**: It is actually a tool that allows us to actually export/ import the data from Traditional RDBMS to HDFS.
- **OOZIE**: It is a flow orchestration engine or mechanism to actually allow us to process the dat inflow and control the flow.
- **MONGO DB**: It is a modern start up approach to the databases[5]. They can be considered as an alternative to the relational databases. They help in managing the data changes frequently also help to handle the Semi Structured and Unstructured data as well as helps in storing mobile apps, Product Catalogue, Content Management and applications real time personization.
- **TALEND:** It is another Open source that offers no of data products. We focus on MDM(Master Data Management) which combines real time data applications and process integration with embedded Data Quality and stewardship.
- **GA(General Assembly class):** For starters of Big Data rather than the Db the GA helps a lot to know the overview of technologies empowering Big Data, History of Db,Storage and Differenes Between Relational and Document Db.

2) Data Cleaning: Before the process of Mining cleaning the data and offering it as a useful dataset is more important. We need to especially clean the Data Sets coming from the Web.

- OPENREFINE: It is formerly called as GOOGLE REFINE is an open source tool dedicated to Data Cleaning(Cleaning of Messy Data)[12]. Using this we can explore huge Datasets easily and quickly even if the data is Unstructured a little bit.

  We have a GITHUB REPOSITORY where we can find the open refine wiki necessary to answer out our doubts through the community.
- DATA CLEANER: It recognises the manipulation of the data is along and drawn process. The Data Visualization tools can only read the nicely Structured “clean” Datasets.

3) Data Mining: Data Mining refers to the concept of analysing the data from different data sources and summarising all that data into an useful information. Ex: Super Market like most of them purchase vegetables on Thursday also take Soft and Hot drinks for the weekend, Bread+Jam+Milk to be placed at one place to increase the overall revenue

- RAPIDMINER: It is a fantastic tool for Predictive Analysis[15], has a great community of clients like PayPal, Deloitte, ebay. We can also create our own Specialized algorithms and integrate them with appropriate API’s into the RapidMiner.
- IBM SPSS Modeller: It offers the solutions related to data mining.It includes text analysis, Entity Analysis, Decision Management and Optimization. It can virtually run on any type of database and we can also integrate it with other IBM SPSS[14].
- ORACLE DATAMINING: Another big hit in the process of Data Mining is Oracle. With their advanced Analytics DB Option they help their users to discover insights, predict and Leverage their Oracle data. We can also develop models to discover the customer behaviour, target our best customers and develop their profiles. Oracle Data Miner GUI helps the mining analytics to work with the features inside the Database. It also creates automated SQL/PL SQL Scripts.

- TERA DATA: Irrespective of the vast amount of data available what actually means is the way we make useful outcomes from the available Datasets [10].There is where actually the Teradata helps out. It also offers many host services including Implementation, Business Consulting, Training & Support.
- FRAMED DATA: In cases where we are specific about the type of Mining process on our project, there are many tools to satisfy such business needs. Framed Data helps to trace out when we are worried of the user churn(i.e. Which user wants to abandon our Product)
- Kaggle: It is the World's largest Data Science Community. It helps out in managing the situation whenever we are stuck with a Data mining issue[7]. Helps to post our problem and get solutions to develop Best Models.

4) Data Analysis: When Mining is done, we try to search through the data to recognise or trace out the un recognised patterns. Data Analysis is a process about breaking the data down and analysing the impact of those Patterns overtime. Even we can ask questions about what may happen to our data in future.

- QUBOLE:It helps in big data analytics on workloads stored on AWS, GOOGLE, AZURE Clouds. Once all these IT policies are effectively handled, no. of Data Analysts will be freed to collaboratively click to Query by using HIVE, Spark, Presto in the growing of the data processing engines. It is an Enterprise Level Solution.
- BIG ML: It is used to simplify the Machine Learning. It helps out in getting Predictions on our data by a easy to use interface, we can also generate models for that Predictive Analysis. To create Tasks<16mb they have their free version of the tool.
- STATWING: It helps in taking the Data Analysis from Visuals->Complex Analysis. They also have a beautiful Blog on NFL data. Its not a free Software and is around 50$ per month allowing until 50 mb of the Data Sources.

5) Data Visualization: It is a procedure of making our data come alive i.e. to convey the insights of our data in a visualized format. Ex: Apart from MySQL and SpreadSheets visualizing is a good way of conveying the data. The coding part required for this step is also comparatively less.

- TABLEAU: It is a visualization tool with primary interest on Business Intelligence. Using it we can create maps, charts, Scatter plots and more without much part of programming. Recently they also updated a web connector, thats helps to connect to the Web DB or API and get the live data.
  A Free Version of the software that helps the starters of visualization called the TABLE PUBLIC.
- SILK: It is much simpler visualization tool than the TABLEAU. Helps to bring our data on live using interactive maps and charts just with a few clicks of mouse. It also helps to collaborate on visualization with n. No of people. Latest feature of it is data visualization automatically.
- CARTO DB: This tool is specially for making maps. It can manage no of data diles and types, also has some sample Datasets. Initially it may not be that easy but once analysed would be a powerful tool for mapping locations.

VI. GOOD PRACTICES OF BIG DATA

- Creating dimensions of all the data being stored is a good practice for Big data analytics. It needs to be divided into dimensions and facts[1].
- All the dimensions should have durable surrogate keys meaning that these keys can't be changed by any business rule and are assigned in sequence or generated by some hashing algorithm ensuring uniqueness.
- Expectation of integrating both structured and unstructured data as all kind of data is a part of Big data which needs to be analyzed together.
- Generality of the technology is needed to deal with different formats of data. Building technology around key value pairs work.
- Analyzing data sets including identifying information about individuals or organizations privacy is an issue whose importance particularly to consumers is growing as the value of Big data becomes more apparent.
- Quality of the data has to be assured Different tasks like filtering, cleansing, pruning, conforming, matching, joining, and diagnosing should be applied at the earliest touch points possible.
- Limitations must be there on the scalability of the data stored.
- Business leaders and IT leaders should work together to yield more business value from the data. The decisions taken must be revised to ensure that the organization is considering the right data to produce insights at any given point of time.
- Investment in data quality and metadata is also important as it reduces the processing time.

VII. CONCLUSION

This paper described the new concept of Big data, its importance in the day to day world.The step by step procedure followed in transforming data to analytics. The types of analytics help out in gaining insights on the gathered data to make it useful information. Many software tools that could help out the processing steps of analytics like Data Storage, Data Cleaning, Data Mining are also discussed and also what are the good practices of the Big Data. For better development we need to have skilled experts who have knowledge on the domain.

VIII. FUTURE ENHANCEMENTS

Real world data produced from various sources like Data Streams, Marts, Social Networks, Web is massive in nature, To analyse and gain insights from that massive data there are many software tools available in market each used at a

different step of processing of the data into analytics but all these tools require consideration of various parameters, certain tools require Licensed Software, Up gradation, Domain experts in order to understand the Consumer needs, Expectations, Feedback and put them into development for the advancement of the industry. All these type of tools up gradations, Licensed ware are costly and less flexible. This could be overcome by linking up the Data Storage with the revolutionalized IT industry concept called CLOUD COMPUTING whose basic architecture contains SAAS, PAAS, IAAS[16] which mainly is cost effective, flexible as it's a pay per usage of resources. But there are practically many challenges that we need to overcome on clouds like fraudulent check, Risk Management, Security.

REFERENCES

[1] Big Data: issues, tools and challenges by Avita Katal, Mohd Wazid, R H Goudar at IEEE 2013. 978-1-4799-0192-0/13/$31.00 ©2013 IEEE

[2] Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money, "Big Data: Issues and Challenges Moving Forward", IEEE, 46th Hawaii International Conference on System Sciences, 2013.

[3] Big data analytics in healthcare: promise and potential Raghupathi and Raghupathi Health Information Science and Systems 2014, 2:3 http://www.hissjournal.com/content/2/1/3

[4] Big Data for Health Javier Andreu-Perez, Carmen C. Y. Poon, Robert D. Merrifield, Stephen T. C. Wong, and Guang-Zhong Yang, Fellow, IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 19, NO. 4, JULY 2015

[5] Big data analytics ,FOURTH QUaRTeR 2011,By Philip Russom.

[6] Design Principles for Effective Knowledge Discovery from Big Data, Edmon Begoli, James Horey, 2012 Joint Working Conference on Software Architecture & 6th European Conference on Software Architecture

[7] Trends in bigdata analytics Karthik Kambatlaa, ∗,Giorgos Kolli asb ,Vipin Kumar c, Ananth Grama a, E-mailaddresses:kkambatl@cs.purdue.edu(K.Kambatla),gkollias@us.ibm.com (G.Kollias),kumar@cs.umn.edu(V.Kumar),ayg@cs.purdue.edu(A.Grama),J.Parallel Distrib.Computing journal homepage: www.elsevier.com/locate/jpdc.

[8] Sachchidanand Singh, Nirmala Singh, "Big Data Analytics", IEEE, International Conference on Communication, Information & Computing Technology (ICCICT), Oct. 19-20, 2012.

[9] M. Cottle, W. Hoover, S. Kanwal, M. Kohn, T. Strome, and N. W. Treister, Transforming Health Care Through Big Data, Institute for Health Technology Transformation, Washington DC, USA, 2013.

[10] Dembosky A: "Data Prescription for Better Healthcare." Financial Times, December 12, 2012, p. 19; 2012. Available from: http://www.ft.com/intl/cms/ s/2/55cbca5a-4333-11e2-aa8f-00144feabdc0.html#axzz2W9cuwajK.

[11] IHTT: Transforming Health Care through Big Data Strategies for leveraging big data in the health care industry; 2013. http://ihealthtran.com/ wordpress/2013/03/iht%C2%B2-releases-big-data-research-reportdownload-today/.

[12] Matthew Smith, Christian Szongott, Benjamin Henne, Gabriele von Voigt, "Big Data Privacy Issues in Public Social Media", IEEE, 6th International Conference on Digital Ecosystems Technologies (DEST), 18-20 June 2012.

[13] Improving Map Reduce Performance in Heterogeneous Environments, USENIX Association, SanDiego, CA,2008, 12/2008.

[14] Azza Abouzeid, Kamil Bajda-Pawlikowski, Daniel Abadi, Alexander Rasin, Avi Silberschatz, HadoopDB: an architectural hybrid of MapReduce and DBMS technologiesforanalyticalworkloads,in:VLDB,2009.

[15] G. N. Forrest, T. C. Van Schooneveld, R. Kullar, L. T. Schulz, P. Duong, andM.Postelnick,"Use of electronic health records and clinical decision support systems for anti microbial stewardship,"Clin.InfectiousDis.,vol. 59, pp. 122–133, 2014.

[16] Big data computing and clouds: Trends and future directions by Marcos D. Assunc¸˜ao, Rodrigo N. Calheiros, Silvia Bianchi, Marco A.S. Netto, Rajkumar Buyya, Big data computing and clouds: Trends and future directions, J. Parallel Distrib. Comput. (2014), http://dx.doi.org/10.1016/j.jpdc.2014.08.003

