VOL. 11, NO. 19, OCTOBER 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

A JOURNEY FROM BIG DATA TOWARDS PRESCRIPTIVE ANALYTICS

S. Poornima and M. Pushpalatha
Department of Computer Science and Engineering, SRM University, Tamilnadu, India
E-Mail: pushpalatha.m@ktr.srmuniv.ac.in

ABSTRACT

This survey paper addresses the concept of big data and various types of analytics that can be undergone by the researchers using big data. Since the data size is big and mixture of different data type, the time to store, retrieve and process the big data is a challenging issue, so an outline to improve the performance of big data processing is also given in this paper. Finally, to make use of big data analytics in a positive and beneficial manner, this paper explores the importance of predictive and prescriptive analytics for any type of applications which is very useful for the people to come to a conclusion at various stage, when they are in need to take best decision.

Keywords: big data analytics, descriptive analytics, predictive analytics, prescriptive analytics.

1. INTRODUCTION

Big data is not a specific technique or technology, which means there was nothing called big data, other than the data size grows beyond our storage capacity with the combination of variety of data together. As the data size grows, some of the issues arises such as additional storage capacity, software to manage them i.e. mechanism to store, retrieve, search etc. new processing methods to handle with considerable cost and time, and it must guarantee for the successful completion of task. Based on these issues a lot of technological developments have been arised in past few years which will be discuss in debrief.

Figure-1. 3V's of Big data.

In 2001, META group analyst Doug Laney (now Gartner) defined the data growth [1] in three dimensions as data volume, velocity and variety which is shortly represented as "3Vs" as shown in Fig 1. In 2012, he updated his definition as "Big data is high volume, high velocity and high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization" [2]. Some organizations added one more V called "Veracity" to describe big data. Veracity means the level of accuracy of generated data. But Gartner's definition was accepted by many industries and it was followed till today. The definitions of 3Vs are given below as:

Data volume: Volume means the quantity of data generated for a unit of time, it may be a year or a month or a day. Five billion gigabytes of data was generated until 2003, from the beginning of recorded time. The same amount of data was generated every two days in 2011 and every ten minutes in 2013. International Data Corporation (IDC) research forecast that the data growth might be 8000 exabytes in 2015 and in 2020; data growth may cross 40,000 exabytes [3] as shown in Figure-2.

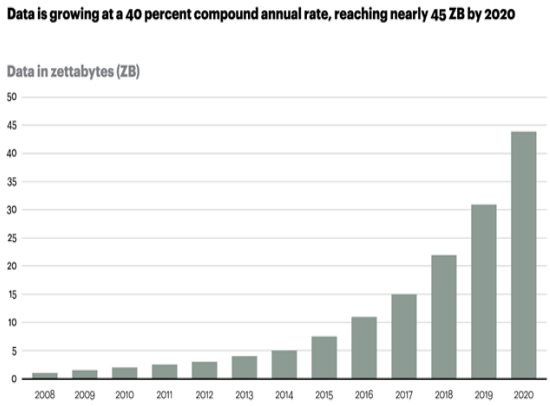

Figure-2. Big data growing rate.

Data velocity: It refers to the interaction speed that includes the speed of data generation, travel along network, processing and storing it back. To manage the flow of high velocity data, Complex Event Processing (CED) [4] [5] is used nowadays which includes Java Messaging Service (JMS) and Apache Kafka. For processing such high velocity data, hadoop is a good distributed, parallel computing environment. NoSQL, In-memory DBMS, ScaleDB are used for persistent of high velocity data.

Data variety: Variety is meant for various types of data generated from various applications around the world as shown in Figure-3. These data may be structured, semi-structured or unstructured. Structured data are those which can be tagged and accurately identified. It is easy to store/retrieve the structured data in/out of database through queries. Examples of structured data are numbers, dates, strings etc. Relational databases are most commonly used database model to store the structured data which can be accessed and manipulated through SQL queries efficiently. Semi-structured data do not follow any specific structure of data associated with any form of data tables but contains tags and markers to separate semantic elements

[6]. Email is a best example for semi-structured data which contains the fields like sender, recipient, date, time etc. whereas the content of the email may have text, images, audio, video in it. XML and other markup languages. Unstructured data does not have a predefined data model or does not fit into relational tables. Examples for unstructured data are multimedia files like images, audio, videos, web pages, PDF files, PowerPoint presentations, blog entries, wikis and word processing documents. Octarian *et al*. [6] mentioned the steps to convert the unstructured data to structured format and the steps for Knowledge Discovery in Databases (KDD). Knowledge discovery is useful in the areas like marketing, finance, fraud detection, manufacturing, telecommunications and internet agents.

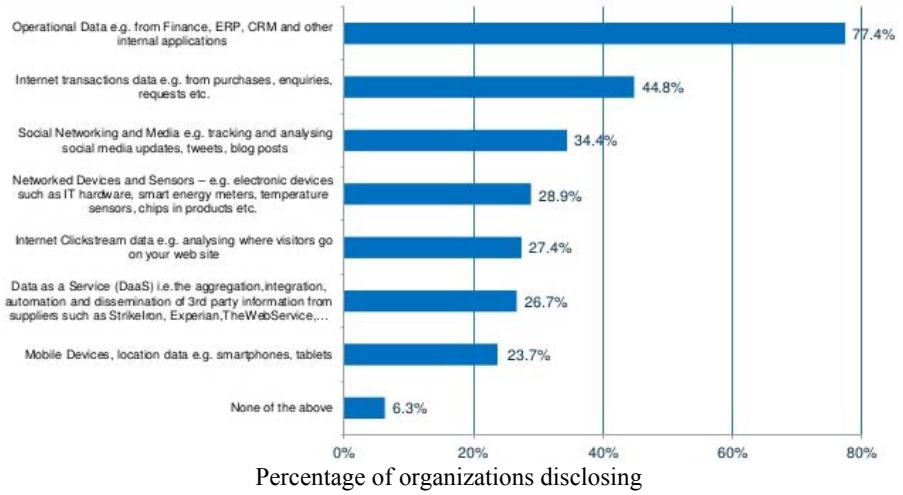

Figure-3. Variety of data sources generating various types.

Xindong *et al.* [7] compared big data with a number of blind men trying to size up a giant elephant. Their goal is to draw the picture of an elephant by touch and feel. Every individual's point of view differs such that they may feel like a rope, a hose, a wall etc. Thus big data definition may vary on every person's point of view. Some of the additional characteristics of big data are variability and complexity. Variability refers to data whose meaning is constantly changing and complexity refers to connect and correlate the data from multiple sources.

Rest of the paper is sectioned as follows: section 2 is about big data analytics with its phases, types of analytics and performance improvement, section 3 gives an idea about prescriptive analytics with its related work along with the applications of it, finally the conclusion of the survey.

2. BIG DATA ANALYTICS

Big data analytics defines the process of collecting, arranging and analyzing large sets of data (called big data) to discover patterns and other useful information. Big data analytics can help organizations to understand the useful patterns in the data and will also help to identify the data that is most important to the business and decision making in business. Big data analysts want the knowledge that is discovered by analyzing the data. The aim in analyzing the data is to uncover hidden patterns and their relationships that might not be visible to all, and they may provide valuable insights which help us to make superior decisions. In general, big data comprises of four different phases [3] as shown in Figure-4 and the details are given in the following section.

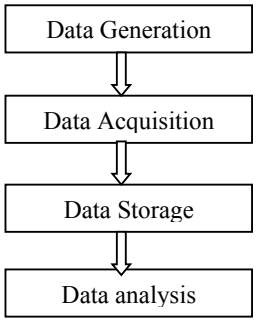

Figure-4. Phases of big data.

2.1 Phases of big data system

The four phases of big data are listed below:
Data Generation
Data Acquision
Data Storage
Data Analysis

2.1.1 Data generation

In the first phase, generated data may be categorized into business data, scientific data and networking data. Business data are those data which is generated due to business-to-business and business to customer transactions such as online buying and selling of products, bank transactions etc. for example Walmart serves more than a million customer transactions every

hour and just imagine the number of transactions per day and per year. Sources that generate business data are all the companies and organizations who use internet across the world. Scientific data are the data generated within the context of scientific investigation by observation and recording. Sources that generate scientific data are biological instruments and scientific instruments and scientific instruments. Example for scientific data is GenBank database of National Centre for Biotechnology Innovation generates data double in size every ten months. Networking data includes data generated by search, Social Networking Service (SNS), websites and click streams. Sources of these data are mobile devices, social media, Internet Service Providers etc.

2.1.2 Data acquisition

The second phase aggregates information in a digital form for storage and analysis. The subtasks of data acquisition are:

Data Collection: It is the process of retrieving raw data from real world objects. The data is collected from sensors, log files and web crawlers.

Data Transmission: The collected raw data is transferred into data storage infrastructure like data centers for subsequent processing through high speed fiber optic cables with the help of IP backbone.

Data Pre-processing: the collected data sets may have different levels of quality in terms of noise, redundancy, consistency which increases the cost of storage. So pre-processing has to be carried out to improve data quality. Some of the pre-processing techniques are data integration, cleansing and redundancy elimination.

2.1.3 Data storage

The acquired data is organized in a convenient format for analysis and value extraction. The features needed for data storage are storage infrastructure and data management framework. Storage infrastructure such as Hard Disk Drive (HDD), Solid State Drive (SSD), Storage Area Network (SAN), Network Attached Storage (NAS) and Direct Attached Storage (DAS) can be used. Data management framework like file systems, database technologies, column oriented database, document database can be used. Examples of file systems are Google's Google File System (GFS), Apache's Hadoop Distributed File System (HDFS), Microsoft's Cosmos, Facebook's Haystack, Tao File System (TFS) and FastDFS. Google's Bigtable, Facebook's Cassandra, Apache's HBase and Hyper-table are the examples of Column oriented databases. Some of the document databases used in real time are MongoDB, SimpleDB and CouchDB. NoSQL is the most common database technology used for storing bigdata which contains several data models like key-value stores, column oriented and document oriented database.

Figure-5. An Amazon Web Services data center.

2.1.4 Data analysis

Some of us may have a little confusion regarding the difference between analysis and analytics. Analysis extracts the useful information from raw data otherwise analysis tells what happened. Analytics discover computational methods for finding useful models from massive amount of data otherwise analytics tells what will happen. The purpose of undergoing for data analysis are:

- to determine how to use the data
- to validate the data
- to identify the reasons for fault
- whereas the purpose of data analytics are,
- to assist decision making
- to predict the future occurrences
- to give suggestions

There are wide range of big data analytics based on the applications. Those are listed below:

- Structured data analytics relies on DBMS, data warehousing, Online Analytical Processing (OLAP) and Business Process Management (BPM).
- Text analytics is also called as text mining extracts hidden information and knowledge from unstructured text. Natural Language Processing [8] is most common technique for text analytics which leads to information extraction, topic modeling, summarization, categorization, clustering, opinion mining, and sentiment analysis. Some of the text mining tools [10] which are used for real time text analytics are SAS [9], IBM Text Analytics, SAP, Lexalytics, Smartlogic etc.
- Web analytics retrieve, extract and evaluate information for knowledge discovery from web sites and and their related documents. Page Rank, CLEVER and focused crawling [11] methods are used to find web pages related to a already established topic.
- Multimedia analytics refers to extracting interesting knowledge and to understand the semantics captured in multimedia files like audio, images and video. Research is undergoing for multimedia analytics in google, yahoo, Alta Vista, IBM, etc. but the playback performance is poor.
- Social analytics investigates social networks [12] through the use of network and graph theories. The

research direction in network analytics are linkage based structural analysis and content based analysis.

- Mobile analytics studies the behavior of mobile users regarding their usage of various applications. Examples are mobile phone, sensors and RFID information collected from them are useful for object tracking and monitoring the system status.

2.2 Processing and performance of big data

Due to generation of huge data every second around the world, high storage capacity is needed. Social networks like Facebook, Twitter and Whatsapp maintains racks of disks and cluster of systems to manage the stored data. Thus a distributed environment with parallel computing methodologies are set up. The reason to set up distributed environment is to connect a group of systems as cluster so that to ensure availability of data and fault tolerance. The reason to set up parallel computing is to get quick response by processing bulk amount of data simultaneously. Few techniques are addressed in forthcoming section.

2.2.1 Hadoop framework

It is an open source software project that implements the distributed processing of large dataset across clusters of commodity servers [13]. Hadoop is the most commonly used framework now a days to handle big data. In 2003, Google found that it is very difficult process to index the entire Internet on a regular basis. So Google engineers invented the technology called Map-Reduce which breaks large sets of data into little chunks and processed them in parallel across thousands of computers which allows search software to run faster on the cheaper, less reliable computer with lower capital costs [14].

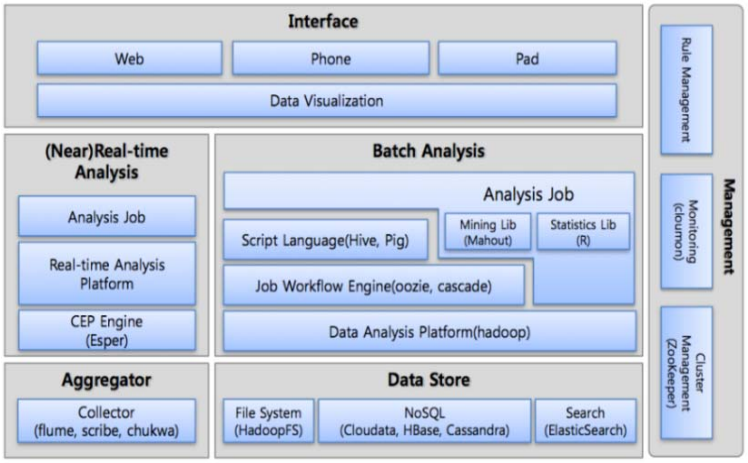

Figure-6. Hadoop architecture.

In 2005, Doug Cutting and Mike Cafarella introduced Hadoop which was named after Cutting's son's stuffed elephant. Cutting was working at Yahoo at that time and Cafarella was a graduate student in University of Washington. They developed hadoop after the release of Google File System and MapReduce paper which they felt as a better way to handle large amount of data for their search engine "Nutch" [15].

Initial version of hadoop i.e. Hadoop 1.x consists of two main modules namely MapReduce and HDFS. Hadoop 2.x versions [16] was appended with two more modules YARN which does a major role in hadoop for resource and node management and HDFS to ensure the availability of data and fault tolerance on the distributed environment. Mappers read the data from HDFS process it and generate an intermediate result which is given to the reducers. Reducers aggregate these results to generate final output which is again written onto HDFS [17]. A number of mapreduce modules are executed in parallel by splitting the related data into clusters, so that huge amount of data is processed in small amount of time. Basically hadoop is developed using Java and its applications can be coded using Python also. The main drawback of hadoop is its disk I/O transfer time.

2.2.2 Spark framework

It is also an open source cluster computing framework developed by researchers of University of California at Berkeley. It can be interfaced with HDFS, limitations of hadoop through in-memory computation which caches the data in memory such that it is 100 times faster than hadoop for certain applications [18]. It supports Java, Scala and Python for its application development.

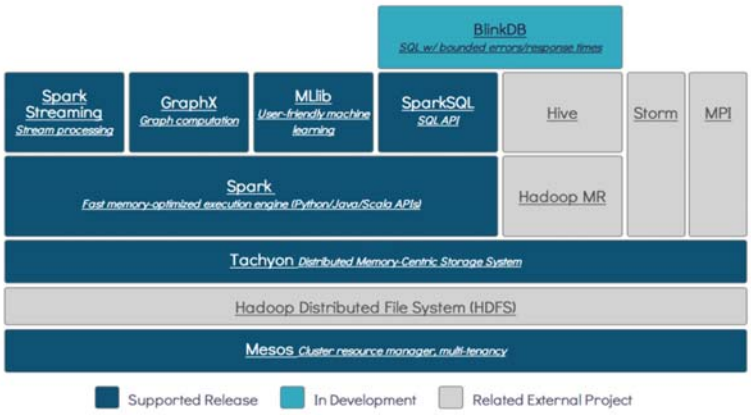

Figure-7. Berkeley spark architecture.

2.2.3 Other frameworks

Even though hadoop is most commonly used framework to process big data, some companies use their own framework to handle their huge data developed as a result of usage of their application by the users. For example Twitter's Storm is an open source system used for processing data in distributed manner and real-time streaming processing whose architecture is shown in Figure-8. Storm implements a data flow model in which data (time series facts) flows continuously through a topology (network of Cassandra, Openstack Swift and Amazon S3. Spark overcomes the problem of I/O transformation entities). The part of data analyzed at any time in an aggregate function is specified by a sliding window, a concept in CEP/ESP. A sliding window may be like one day or one hour, which is constantly shifting overtime. Data is inputted to Storm in distributed manner as messaging queues like Kafka, Kestrel, and even regular JMS. Trident is an abstraction API which is easy to use. Similar to Twitter Storm, Apache S4 is also a product for distributed, scalable, continuous and stream data processing.

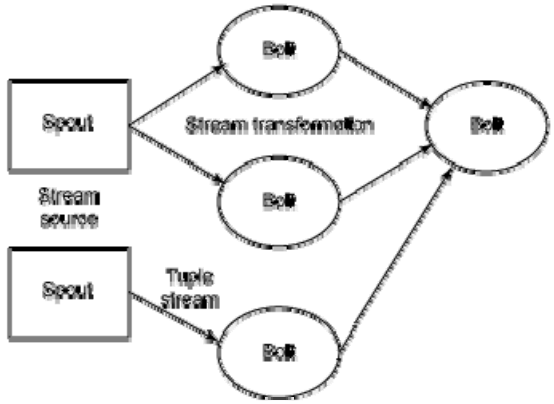

Figure-8. Conceptual architecture of a trivial Storm topology.

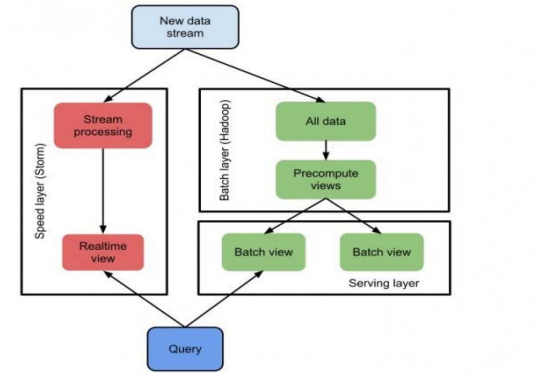

Figure-9. Lambda architecture.

Dremel is a distributed system coded by Google to query large datasets and powers Google's BigQuery service [19]. Cloudera Impala and Apache Drill are introduced after Dremel. These techniques run fast in that coordination, query planning, optimization, scheduling, and execution are all distributed throughout nodes in a cluster to maximize parallelization. Lambda Architecture [20] proposed by Nathan Marz as shown in Fig 9, takes a very unique approach from the three tools above. The problem of computing arbitrary functions on a big data set is solved by lambda in real-time by decomposing the problem into three layers: the batch layer, the serving layer, and the speed layer.

Batch layer is implemented as a Hadoop cluster with JCascalog the abstraction API. Hadoop cluster data are periodically updated by merging incremental changes. After each update, queries are re-computed from scratch. The results are called batch views. Serving layer saves the batch views in a Splout SQL (or ElephantDB). To access data fast, Batch views are indexed. Evidently, batch views are not real time. Storm (Trident) is implemented in speed layer, that computes ad-hoc functions on a data stream (time series facts) in real-time. The materialized batch view is merged with the result of incremental changes confined to a sliding window, by viewing from the serving layer to produce current analysis results. epiC is also an Extensible and Scalable System for Processing Big Data.

epiC introduces a general Actor-like concurrent programming model, independent of the data processing models, for specifying parallel computations[21]. Users process multi-structured datasets with appropriate epiC extensions, the implementation of a data processing model best suited for the data type and auxiliary code for mapping that data processing model into epiC's concurrent programming model. Like Hadoop, programs written in this way can be automatically parallelized and the runtime system takes care of fault tolerance and inter-machine communications.

2.2.4 Performance improvement of big data Processing

Some of the major issues identified in big data are,

- Sufficient speed for large data size
- Familiarize the data
- Exploring data quality
- Displaying meaningful results

According to the first issue, to improve the performance of data processing, data is stored among several racks of disks, hadoop uses two types of nodes called name node and data node as shown in Fig 10. Data nodes contains the actual data and name node contains the meta data (information about which data node contains what data and its duplicates), thus hadoop contains one name node and several data nodes are clustered as a group. Horizontal scaling and vertical scaling [18] is another method that can be implemented to improve the performance of not only hadoop but any parallel computing environment. Horizontal scaling (known as "scale out") in which multiple independent machines are added together to improve the processing capability through which the workload is distributed across many servers. Main advantage of horizontal scaling is less expensive. Vertical scaling is also called as "scale up" that involves more processors, more memory and faster hardware installed in a single server. This type of scaling is easy to manage and install but needs more financial investment when compared to horizontal scaling. One more problem with vertical scaling is up-gradation is constrained to certain limit.

HDFS Federation [22] improves the existing HDFS architecture through a clear separation of namespace and storage, in which block storage layer is enabled. Scalability and isolation is improved by supporting multiple namespaces in the cluster. Federation also opens up the architecture, in which HDFS cluster is expanded to apply to new implementations and use cases. In order to scale the name service horizontally, federation uses many independent namenode with namespaces. The namenodes are independent and does not expect coordination with each other. The "datanodes" are used as common storage for blocks by all the namenodes. All the namenodes in a particular the cluster are registered with its corresponding datanode. Datanodes send periodic signals and block reports that execute commands from the namenodes. This new architecture was depicted in Figure-11.

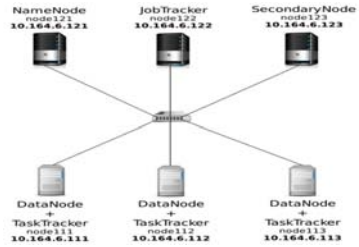

Figure-10. Name node and data nodes of hadoop.

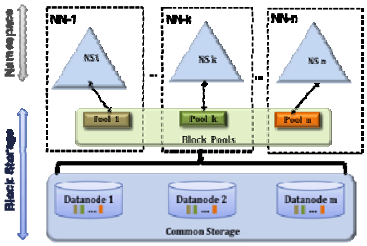

Figure-11. HDFS Federation architecture.

SP machine is a Simple and Powerful system developed based on SP theory of intelligence [23]. Benefits of using SP systems to be recommended for big data are overcoming the problem of variety in big data, learning and discovery, interpretation of data, speed and bulk, veracity and visualization. It uses a method called multiple alignment in which sentences are parsed in a specific pattern, that discover the structure in data. SP computer model needs less than 500KB of storage space [24] for its exec file and no additional programming is needed for processing of data. Hence, researchers of big data analytics may try to use SP computer model for efficient processing of data and knowledge discovery.

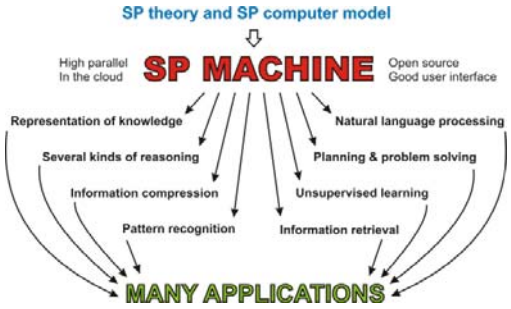

Figure-12. SP computer model and its applications.

To solve the second issue of big data, various tools available for various types of big data analytics in the market as mentioned in section 2.1.4, which will help us to extract needed information from the unstructured data. Lot of tools available for text mining to perform sentiment analysis, opinion mining etc. but to mine knowledge from multimedia data, many software developing companies are working under it including google, yahoo etc. to achieve

satisfiable results. Still there are some challenges which are yet to be solved in case of mining knowledge from images and video.

The third issue talks about the genuinity of data in correct time. When considering the sentiment analysis, the problem is to find whether the conveyed data by the user is genuine data or not. Researchers are currently working to identify the genuinity of the data based on various factors like sentence typed by the user, collecting their activity details from various social networks like facebook, twitter, whatsapp etc.

The problem with visualizing the results of big data analytics is to deal with large volume of variety of data, but nowadays tools like R, SAS (Statistical Analysis System), Data Applied, MEPX (Multi Expression Programming), etc. can be used to handle large amount of data. The method like binning is used to group the data together and make it easier for the user to view.

3. IMPORTANCE OF ANALYTICS

Nearly 67% of big data generated by business applications belongs to different types of unstructured data as shown in Figure-13 and hence those companies are missing the opportunity to leverage the full value to obtain knowledge discovery from those data which might be useful to increase their profit, production etc. Schubert [25] says "When you get to a point where there is too much complexity for manual processes, you need to look at introducing new technologies to automate processes and simplify things". Thus analytics is a way of forecasting based on history which will prescribe us what we should be doing. Analytics is not only restricted to business applications, but also have wide usage in other applications like education, medical, entertainment, finance, marketing, communication, politics etc.

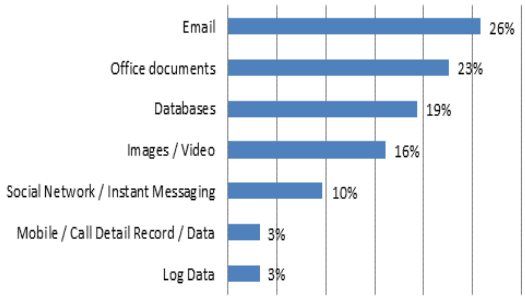

Figure-13. Type of unstructured data growing rate.

Analytics can be done in three levels on big data,

Descriptive analytics: Focuses on identifying the reasons behind the results by analyzing the current and history of data. For example, identification of disease from various diagnosis report of a patient. Descriptive analytics uses descriptive statistics and inferential statistics methods to perform the analysis. Some of the descriptive analytics tools are Excel descriptive statistics data analysis tool [26], ANOVA [27], and Tanagra [28].

Predictive analytics: Identifies the future outcome by analyzing the history of data, current data and newly generated possible combinations of existing data. Predictive analytics uses regression and machine learning methods to undergo for analysis. Some of the regression methods are linear regression, logistic regression, time series model, discrete choice model etc. For example, by analyzing the diagnosis report of different patients suffering with same disease may be useful to predict possible new symptoms for that disease. Tools used for predictive analytics are Waffles [29], Weka [30] and Rapid Miner [31].

Prescriptive analytics: After prediction, prescriptive analytics concentrates on how to take advantage of future opportunity or mitigate a future risk and implication of each decision option. For example, after predicting the new symptoms of the disease, prescriptive analytics suggests the precautions to be taken to avoid the risk of further spreading of that disease to others or to avoid damage of other parts of the body. It may also prescribe modification in the medicine preparation by adding or removing certain drugs. Predictive analytics does not produces one possible future outcome, it produces multiple future outcomes. Whereas prescriptive analytics tracks the consequences based on different choice of action and recommend the best course of action for any pre-specified outcome. Prediction alone does not help us to undergo for best decision making, prescriptive analytics plays a major role over there. Thus, the further discussion in this paper is about the survey of prescriptive analytics using big data.

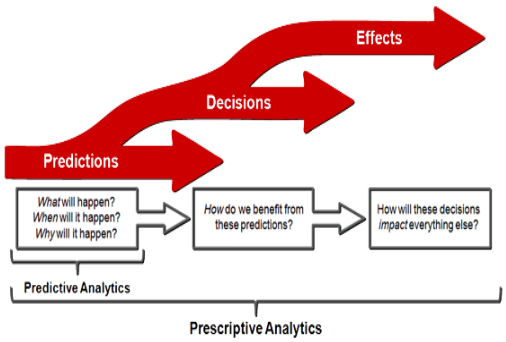

Figure-14. Inferring prescriptive analytics from predictive analytics.

3.1 Prescriptive analytics

Prescriptive analytics is the term introduced by IBM and later trademarked by Ayata [32] which is a software company in Australia. They have started their research in 2003 and release their first version of prescriptive analytics software in 2007.In 2013 they release version 4 of prescriptive analytics software which deals with hybrid data which is a combination of structured and unstructured data. Thousands of barrels of oil are lost every day due to unpredictable electronic submersible pump failures, so Ayata provide the software to predict not only what pumps might fail, when and why,

prescribes actions to minimize the production loss. Prescriptive analytics can make use of machine learning, natural language processing, pattern recognition, operations research, applied statistics, signal processing, computer vision, image processing and speech recognition methods to undergo for analysis as shown in Figure-15.

We need huge volume and variety of data (big data) to carry out prescriptive analytics because predicting the future and suggesting actions based on it, needs up-to date information(may be structured or unstructured) and tracking of the past occurrences and behaviors at various situations with respect to time. When the data size is large then it provides detailed and timely information about a specific application which results in better decision making but to obtain a solution, optimization must be carried out to find best solution to meet user's objectives. For the above mentioned example of predicting the new symptoms of disease and suggesting actions for it, large number of similar patient's records has to be analyzed from the big data, starting from the first diagnosis to latest diagnosis reports. Some of the reports obtained from internet or a cloud may not be continuous history of diagnosis report and the data in those reports may be images or video or may be combination of them with text also. Hence big data analytics such as text mining and multimedia analytics is also needed to undergo for predictive and prescriptive analytics. The advantage of applying prescriptive analytics on big data is data analytics can be carried out with reasonable cost by using framework like hadoop, reduce manual cost, provide better control and visibility.

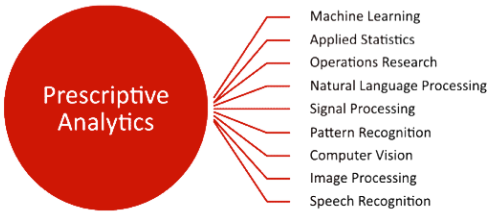

Figure-15. Technical disciplines of prescriptive analytics.

3.2 Related work in prescriptive analytics

Only 3% of the companies are currently using advanced modeling and optimization software to make use of prescriptive analytics. A real time case study for the predictive and prescriptive analytics is SBV (Standard Bank, Barclays, Volkskas) services Limited [24]. SBV is South Africa's largest cash management service provider and is jointly owned by the country's four major commercial banks. SBV's aim is to store cash at the lowest possible cost and simultaneously ensure the correct supply of notes in all denominations across all geographies. This means managing complex cost trade-offs between funding, logistics, processing and storage, insurance capacity, and Central Bank charges. SBV has implemented forecasting and business optimization technologies to create a planning platform for the South African cash management industry. SBV uses an accounting system called CISPro and vault management technologies, using a "best of breed" approach and use the Supply Planning Workbench (SPW) to obtain inventory position and Prophecy Forecasting to forecast for three months. SPW uses River Logic's Enterprise Optimizer (EO) to prescribe movements and balance activities such as deposit this cash in that vault, bring this cash to that cash center, for this denomination the best place to get it is the cash center down the road or the Central Bank, etc.

Madhav *et al* [33] undergone for predictive and prescriptive analytics for synthetic information and the immediate activities to be carried out after nuclear detonation. Synthetic information is the data created by combining data from multiple sources. They created synthetic information from the geographical map of certain region in Washington DC in which the people use to move from one place to another through bus network, metro network and automobile network. They also developed a social contact network which infers the locations visited by the people, the place where two or more people may interact and they meet at more than one location. From the above collected information, the authors predict the location of the individual, health level of the people, capacities of hospital in that region and suggest emergency broadcast and communication, recovery of power network for communication and various behavioral options. They also showed results comparison for saving the lives of people who receive the emergency broadcast and who do not. The proportion of death is high in case of not receiving emergency broadcast.

Jens Weber et all [34] explained about the software application called "Intelligence in Science and Technology" (InSciTe) given by Korea Institute of Science and Technology Information(KISTI) which uses prescriptive analytic methods in order to develop strategies and provide recommendations in order to improve research performance. InSciTe measures the research performance of the research scholar, finds the role model researcher, suggests research activities and generates report. Techniques like activity trend analysis and SWOT analysis has been used to determine the activity range and research capacity of the scholar. To recommend suggestions for the improvement of research performance 5W1H (what, who, where, when, why and how) is provided. This project will help the researchers to know their research level, their appropriate group to join with and actions to be taken to improve their ranking by comparing with other researchers in that group.

3.3 Applications of prescriptive analytics

The following lists out the applications where prescriptive analytics can be implemented.

- **Business:** The main aim of any business is to minimize the investment cost, maximize profit and reduce downtime. Prescriptive analytics helps the business people by suggesting whether to bring a product to market? Whether to move to new facility? If yes then which one? and finally to recall the product or not.

- **Finance:** Prescriptive analytics is helpful for the investors to select which investments to purchase? What maximum and minimum investment cost range? Which company product to purchase?
- **Medical:** As discussed in the earlier examples, prescriptive analytics recommends which drug to approve? Helps doctors to decide what type of treatment can be given for which patients?
- **Sports:** There are lot of recommendations given for sports application with the use of prescriptive analytics. Some of the important recommendations are, which player to draft or trade?, in case of cricket the bowler can decide his way of bowling after analyzing a specific batsman, likewise in all types of games the player may have an idea how to face his opponent player.
- **Systems engineering:** Prescriptive analytics helps the system engineers to choose best design for their system among various designs; also he may decide the better alternate design to replace existing?

4. CONCLUSIONS

This paper helps the new learners of big data to understand the basics about big data, its processing techniques and gives an idea to improve the performance of big data analytics. By having big data in hand, anyone can undergo for descriptive, predictive and prescriptive analytics which is helpful for the improvement of any type of business applications. In another perspective, now a day's people are in need of guidance and suggestions from expert members of various domain to take best decision among plenty of choices in any type of applications. Applications like business, education, online shopping, banking and healthcare needs prescriptive analytics to improve their standard and quality which helps both users and owners to be benefitted. Hence the researchers of big data analytics may concentrate on prescriptive analytics which plays a major role in this fast moving world to suggest the society with good recommendations.

REFERENCES

[1] Doug Laney: 2001. 3D Data Management: Controlling Data Volume, Velocity and Variety. META group Inc., File 949, 6th Feb.

[2] Doug Laney: 2012. The Importance of 'Big Data': A Definition. Gartner, 21 June.

[3] Han Hu, Yonggang Wen, Tat Seng Chua, Xuelong Li: 2014. Toward Scalable Systems for Big Data Analytics: A Technology Tutorial. IEEE Access. 2: 652-685.

[4] 2010. Oracle Complex Event Processing High Availability. An Oracle White Paper.

[5] Sriskandarajah Suhothayan, Isuru Loku Narangoda, Subash Chaturanga: Siddhi: 2011. A Second Look at Complex Event Processing Architectures. In Proceedings of the 2011 ACM workshop on Gateway computing environments (GCE'11), 43-50.

[6] Octavian Rusu, Ionela Halcu: 2013. Converting unstructured and semi-structured data into knowledge. Roedunet International Conference (RoEduNet), IEEE publisher. pp. 1-4.

[7] Xindong Wu, Xingquan Zhu, Gong-Qing, wei Ding: 2014. Data Mining with Big Data. IEEE transactions on knowledge and data engineering. 26(16): 97-107.

[8] Sharvari Tamane: 2015. Text Analytics for Big Data. International Journal of Modern Trends in Engineering and Research. 2(3): 213-218.

[9] Goutam Chakraborty, Murali Krishna Pagolu: 2014. Analysis of Unstructured Data: Applications of Text Analytics and Sentiment Mining. SAS Institute Inc, Paper 1288-2014.

[10] K.L.Sumathy, M.Chidambaram: 2013. Text Mining: Concepts, Applications, Tools and Issues - An Overview. International Journal of Computer Applications. 80(4): 29-32.

[11] Meenu, Rakesh Batra: 2014. A Review of Focused Crawler Approaches. International Journal of Advanced Research in Computer Science and Software Engineering. 4(7): 764-767.

[12] Evelien Otte, Ronald Rousseau: 2009. Social network analysis: A powerful strategy, also for the information sciences. Journal of Information Science. 28(6): 441-453.

[13] Hadoop Basics: [Online], Available: http://en.wikipedia.org/wiki/Apache_Hadoop.

[14] Vance, Ashlee: 2009. Hadoop, a Free Software Program, Finds Uses beyond Search. The New York Times Published on 16 March.

[15] Hadoop History: [Online], Available: https://gigaom.com/2013/03/04/the-history-of-hadoop-from-4-nodes-to-the-future-of-data/.

[16] Hadoop Versions: [Online], Available: http://wiki.apache.org/hadoop/Roadmap.

[17] Das, Arati Mohapatro: 2014. A Study on Big Data Integration with Data Warehouse. International

Journal of Computer Trends and Technology (IJCTT). 9(4): 188-192.

[18] Dilpreet Singh, Chandan k Reddy: 2014. A survey on platforms for big data analytics. Journal of Big Data. 1(8): 1-20.

[19] Sergey Melnik, Andrey Gubarev, Jing Long, Geoffrey Romer, Shiva Shivakumar, Matt Tolton, Theo Vassilakis. Dremel: 2010. Interactive Analysis of Web-Scale Datasets. Google Inc., Proceedings of the VLDB Endowment. 3(1): 330-339.

[20] Lamda Architecture: [Online], Available: http://jameskinley.tumblr.com/post/37398560534/the-lambda- architecture-principles-for.

[21] Dawei Jiang, Gang Chen, Beng Chin Ooi, Kian-Lee Tan, Sai Wu: 2014. epiC: an Extensible and Scalable System for Processing Big Data. In Proceedings of the VLDB Endowment. 7(7): 541-552.

[22] HDFS Federation: [Online], Available: https://hadoop.apache.org/docs/r2.4.1/hadoop-project-dist/hadoop-hdfs/Federation.html#Background.

[23] Gerard Wolff: 2014. Big Data and the SP Theory of Intelligence. IEEE Access. 2: 301-315.

[24] SP Computer Model exec file: [Online], Available: www.cognitionresearch.org/sp.htm.

[25] Thomas Thompson, Philip Higginbotham: Predictive and prescriptive analytics. Grant Thornton LLP, Financial Executives Research Foundation. (2014).

[26] Excel descriptive statistics data analysis tool: [Online], Available: http://www.real-statistics.com/descriptive-statistics/descriptive-statistics-tools/.

[27] ANOVA: [Online], Available: http://home.business.utah.edu/mgtdgw/stat5969/Analysis_Tools.pdf.

[28] Tanagra: [Online], Available: http://tanagra.software.informer.com/.

[29] Waffles: [Online], Available: https://en.wikipedia.org/wiki/Waffles_(machine_learning) http://wafflesoftware.net/

[30] Weka: [Online], Available: https://en.wikipedia.org/wiki/Weka_(machine_learning) http://www.cs.waikato.ac.nz/ml/weka/downloading.html.

[31] Rapid miner: [Online], Available: http://docs.rapidminer.com/studio/getting-started/.

[32] Ayata: [Online], Available: http://ayata.com/about-us/.

[33] Madhav V.Marathe, Henning S. Mortveit, Nidhi Parikh, Samarth Swarup: 2014. Prescriptive Analytics Using Syntetic Information. Emerginh Methods in Predictive Analytics: Risk Management and Decision Making, IGI Global, 1-20.

[34] Jens Weber, Minhee Cho, Mikyong Lee, Sa-Kwang Song, Michaela Geierhos, Hamnim Jung. System Thinking: 2014. Crafting Scenarios for Prescriptive Analytics: In Proceedings of First International Workshop on Patent Mining and Its Applications (IPAMIN).