Study of Big Data for Data Analytics: Challenges and Progress

Parminder Kaur
Department of Computer Science
Guru Nanak College, Ferozepur, Punjab
Parminder92sandhu@gmail.com

Abstract: - Big Data is commonly defined as data that contains greater variety arriving in increasing volumes and with ever higher velocity. Big data is a data or data sets so large or complex that traditional data processing applications are inadequate and distributed databases are needed. Firms like Google, eBay, LinkedIn, and Face book were built around big data from the beginning. Hadoop is the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes. Big Data is a data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. We need a different platform named Hadoop is an open source software project that enables the distributed processing of large data sets across clusters of commodity servers. It is designed to scale up from single server to thousands of machines, with a very high degree of fault tolerance

Keywords: - . Big Data, Big data Applications, Architecture, Categories, Hadoop, HDFS, MapReduce, NoSQL, Hive,spark,Kofta,Storm

I.INTRODUCTION

Big data means really a big data; it is a collection of large datasets that cannot be processed using traditional computing techniques. Big data is not merely a data; rather it has become a complete subject, which involves various tools, techniques and frameworks. Big Data is a phrase used to mean a massive volume of both structured and unstructured data that is so large it is difficult to process using traditional database and software techniques. In most enterprise scenarios the volume of data is too big or it moves too fast or it exceeds current processing capacity. Big Data has the potential to help companies improve operations and make faster, more intelligent decisions. This data, when captured, formatted, manipulated, stored, and analyzed can help a company to gain useful insight to increase revenues, get or retain customers, and improve operations. Big data is a blanket term for the non-traditional strategies and technologies needed to gather, organize, process, and gather insights from large datasets. While the problem of working with data that exceeds the computing power or storage of a single computer is not new, the pervasiveness, scale, and value of this type of computing has greatly expanded in recent years. Big data can be applied to real-time fraud detection, complex competitive analysis, call center optimization, consumer sentiment analysis, intelligent traffic management, and to manage smart power grids, to name only a few applications. Big data is data that becomes so large that it cannot be processed using conventional methods. The size of the data which can be considered to be Big Data is a constantly varying factor and newer tools are continuously being developed to handle this "Big Data".

II.CHARACTERSTICS OF BIG DATA

Big data is characterized by five primary factors:

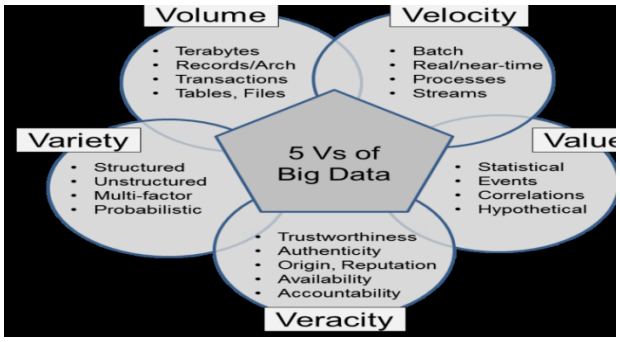

Figure 1: characteristics of big data

Volume: **-** refers to the vast amount of data generated every second. Just think of all the emails, Twitter messages, photos, video clips and sensor data that we produce and share every second. We are not talking terabytes, but zettabytes or brontobytes of data. On Facebook alone we send 10 billion messages per day, click the like button 4.5 billion times and upload 350 million new pictures each and every day. If we take all the data generated in the world between the beginning of time and the year 2000, it is the same amount we now generate every minute! This increasingly makes data sets too large to store and analyze using traditional database technology. With big data technology we can now store and use these data sets with the help of distributed systems, where parts of the data is stored in different locations, connected by networks and brought together by software.

Velocity: refers to the speed at which new data is generated and the speed at which data moves around. Just think of social media messages going viral in minutes, the speed at which credit card transactions are checked for fraudulent activities or the milliseconds it takes trading systems to analyze social media networks to pick up signals that trigger decisions to buy or sell shares. Big data technology now allows us to analyze the data while it is being generated without ever putting it into databases.

Variety: - refers to the different types of data we can now use. In the past we focused on structured data that neatly fits into tables or relational databases such as financial data (for example, sales by product or region). In fact, 80 percent of the world's data is now unstructured and therefore can't easily be put into tables or relational databases—think of photos, video sequences or social media updates. With big data technology we can now harness differed types of data including messages, social media conversations, photos, sensor data, video or voice recordings and bring them together with more traditional, structured data.

Veracity:- refers to the messiness or trustworthiness of the data. With many forms of big data, quality and accuracy are less controllable, for example Twitter posts with hashtags, abbreviations, typos and colloquial speech. Big data and analytics technology now allows us to work with these types of data. The volumes often make up for the lack of quality or accuracy.

Value: refers to our ability turn our data into value. It is important that businesses make a case for any attempt to collect and leverage big data. It is easy to fall into the buzz trap and embark on big data initiatives without a clear understanding of the business value it will bring.

III.CATEGORIES OF BIG DATA

Structured: Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data. Over the period of time, talent in computer science have achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it. However, now days, we are foreseeing issues when size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabyte.

Unstructured: Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. Typical example of unstructured data is, a heterogeneous data source containing a combination of simple text files, images, videos etc. Now a day organizations have wealth of data available with them but unfortunately they don't know how to derive value out of it since this data is in its raw form or unstructured format.

Semi-structured: Semi-structured data can contain both the forms of data. Semi-structured data is a structured in form but it is actually not defined with e.g. in relational DBMS. Example of semi-structured data is a data represented in XML file.

IV.ARCHITECTURE OF BIG DATA

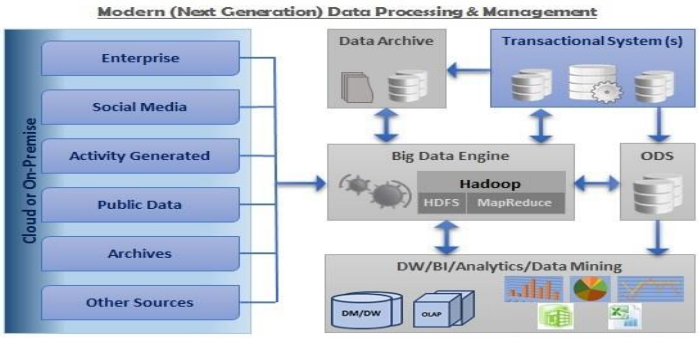

Figure 2: Data Processing & Management

Source Systems:-There are various different sources of Big Data including Enterprise Data, Social Media Data, Activity Generated Data, Public Data, Data Archives, Archived Files, and other Structured or Unstructured sources.

Transactional Systems:-In an enterprise, there are usually one or more Transactional/OLTP systems which act as the backend databases for the enterprise's mission critical applications. These constitute the transactional systems represented above.

Data Archive:-Data Archive is collection of data which includes the data archived from the transactional systems in compliance with an organization's data retention and data governance policies, and aggregated data (which is less likely to be needed in the near future) from a Big Data engine etc.

ODS: - Operational Data Store is a consolidated set of data from various transactional systems. This acts as a staging data hub and can be used by a Big Data Engine as well as for feeding the data into Data Warehouse, Business Intelligence, and Analytical systems.

Big Data Engine:-This is the heart of modern (Next-Generation / Big Data) data processing and management system architecture. This engine capable of processing large volumes of data ranging from a few Megabytes to hundreds of Terabytes or even Petabytes of data of different varieties, structured or unstructured, coming in at different speeds and/or intervals. This engine consists primarily of a Hadoop framework, which allows distributed processing of large heterogeneous data sets across clusters of computers. This framework consists of two main components, namely HDFS and MapReduce.

V.ROLE OF BIG DATA

1. BDA: Big Data Analytics Applications (BDA Apps)are a new type of software applications, which analyse big data using massive parallel processing frameworks (e.g.,Hadoop). Developers of such applications typically develop them using a small sample of data in a pseudo-cloud environment. Afterwards, they deploy the applications in a large-scale cloud environment with considerably more processing power and larger input data (reminiscent of the mainframe days)

2. Clustering: Using clustering (K-means algorithm) through a simple point and click dialog, users can automatically find groups within data based on specific data dimensions. With

clustering, it is then simple to identify and address groups by customer type, text documents, products, patient records, click path, behaviour, purchasing patterns, etc

***3.Data Mining*:** Decision Tree--Datameer's decision trees automatically help users understand what combination of data attributes result in a desired outcome. Decision trees illustrate the strengths of relationships and dependencies within data and are often used to determine what common attributes influence outcomes such as disease risk, fraud risk, purchases and online signups. The structure of the decision tree reflects the structure that is possibly hidden in your data.

4. Banking: The use of customer data invariably raises privacy issues. By uncovering hidden connections between seemingly unrelated pieces of data, big data analytics could potentially reveal sensitive personal information. Research indicates that 62% of bankers arecautious in their use of big data due to privacy issues. Further, outsourcing of data analysis activities or distribution of customer data across departments for the generation of richer insights also amplifies security risks.For instance, a recent security breach at a leading UK-based bank exposed databases of thousands of customer files. Although this bank launched an urgent investigation, files containing highly sensitive information. Such as customers' earnings, savings, mortgages, and insurance policies ended up in the wrong hands10. Such incidents reinforce concerns about data privacy and discourage customers from sharing personal information in exchange for customized offers.

5. Sap: Sybase (now SAP) laid the groundwork for the analytical platform market when it launched the first columnar database in 1995. Tera data was also an early forerunner, shipping the first analytical appliance in the early 1980s. Netezza kicked the current market into high gear in 2003 when it unveiled a popular analytical appliance and was soon followed by dozens of start-ups. Recognizing the opportunity, all the big names in software and hardware. They are Oracle, IBM, HP, and SAP subsequently jumped into the market, either by building or buying technology, to provide purpose-built analytical systems to new and existing customers.

6. Stock: A private stock exchange in Asia uses in-database analytics to establish a comprehensive system to detect abusive trading patterns to detect fraud.

7. Credit Cards: Credit card companies rely on the speed and accuracy of in-database analytics to identify possible fraudulent transactions. By storing years' worth of usage data, they can flag atypical amounts, locations, and retailers, and follow up with cardholders before authorizing suspicious activity

***8. Enterprise**:* For enterprises around the world, in many industries, in-database analytics are providing a competitive advantage. When data doesn't have to commute to work and back, it can deliver faster insights that help businesspeople make informed decisions in real time for less expense than traditional data analysis tools.

9. Consumer Goods: A maker of consumer products collects consumer preference and purchasing data extracted from surveys, purchases, web logs, product reviews from online retailers, phone conversations with customer call centres, even raw text picked up from around the Web. Their ambitious goal: to collect everything being said and communicated publicly about their products and extract meaning from it. By doing this, the company develops anuanced understanding of why certain products succeed and why others fail. They can spot trends that can help them feature the right products in the right marketing media.

10. Hadoop: In every vertical there are data tasks with which Hadoop can assist. These tasks have different terms depending on the industry but they all come down to either advanced analytics or data processing

11. Agriculture: A biotechnology firm uses sensor data to optimize crop efficiency. It plants test crops and runs simulations to measure how plants react to various changes in condition. Its data environment constantly adjusts to changes in the attributes of various data it collects, including temperature, water levels, soil composition, growth, output, and gene sequencing of each plant in the test bed. These simulations allow it to discover the optimal environmental conditions for specific gene types

12. Finance: A major financial institution grew wary of using third-party credit scoring when evaluating new credit applications. Today the bank performs its own credit score analysis for existing customers using a wide range of data, including checking, savings, credit cards, mortgages, and investment data.

13. Economy: Designed from the ground up to deal intelligently with commodity hardware, Hadoop can help organizations transition to low-cost servers.

14. Conservation: Keeping data in a merged, isolated system provides business intelligence benefits and is both financially and ecologically sound.

15. Marketing: Marketers have begun to use facial recognition software to learn how well their advertising succeeds or fails at stimulating interest in their products. A recent study published in the Harvard Business Review looked at what kinds of advertisements compelled viewers to continue watching and what turned viewers off. Among their tools was "a system that analyses facial expressions to reveal what viewers are feeling." The research was designed to discover what kinds of promotions induced watchers to share the ads with their social network, helping marketers create ads most likely to "go viral" and improve.

16. Smart Phones: Perhaps more impressive, people now carry facial recognition technology in their pockets. Users of I Phone and Android smart phones have applications at their fingertips that use facial recognition technology for various tasks.For example, Android users with the remember app, can snap a photo of someone, then bring up stored information about that person based on their image when their own memory lets them down a potential boon for salespeople. I Phone users can unlock their device with recognize me, an app that uses facial recognition in lieu of a password. If deployed across a large enterprise, this app could save an average of $2.5 million a year in help-desk costs for handling forgotten passwords.

17. Telecom: Now a day's big data is used in different fields. In telecom also it plays a very good role. Service providers are trying to compete in the cutthroat world of telecom services. Where mare and more subscribers rely on over-the-top (OTT) players as providers of value-added services are focused on increasing revenue, reducing opex, chum and enhancing the customer experience as key business objectives.

***18.Healthcare*:**-Big data analytics has helped healthcare improve by providing personalized medicine and prescriptive analytics, clinical risk intervention and predictive analytics, waste and care variability reduction, automated external and internal reporting of patient data, standardized medical terms and patient registries and fragmented point solutions. This includes electronic health record data, imaging data, patient generated data, sensor data, and other forms of difficult to process data.

VI.CHALLENGES OF BIG DATA

***Information Growth*:** Over 80 percent of the data in the enterprise consists of unstructured data, which tends to be growing at a much faster pace than traditional relational information. This massive information threatens to swamp all but the well-prepared IT organizations.

Processing power: The customary approach of using asingle, expensive, powerful computer to crunch information just doesn't scale for Big Data. As we soon see, the way to go is divide-and-conquer using commoditized hardware and software via scale-out.[2]

Physical storage: Capturing and managing all this information can consume enormous resources, outstripping all budgetary expectations

***Data issues*:** Lack of data mobility, proprietary formats, and interoperability obstacles can all make working with Big Data complicated.[3]

Cost: Extract, transform, and load (ETL) processes for Big Data can be expensive and time consuming, particularly in the absence of specialized, well-designed Software. [4]

VII.BIG DATA TECHNOLOGIES AND TOOLS-HADOOP AND NOSQL ECOSYSTEM

Hadoop is used when you have data in the terabyte or petabyte range—too large to fit on a single machine. It's made up of HDFS, which lets you store data on a cluster of machines, and MapReduce, which lets you process data stored in HDFS. It lets you treat a cluster made up of hundreds or thousands of machines as a single machine. HDFS is the disk drive for this large machine, and MapReduce is the processor.Hadoop's use is widespread for processing Big Data, though recently Spark has started replacing MapReduce. Hadoop is used in maintaining, scaling, error handling, self healing and securing large scale of data. These data can be structured or unstructured.

High-availability distributed object-oriented platform or "Hadoop" is a software framework which analyse structured and unstructured data and distribute applications on different servers. Below is an overall Hadoop architecture –

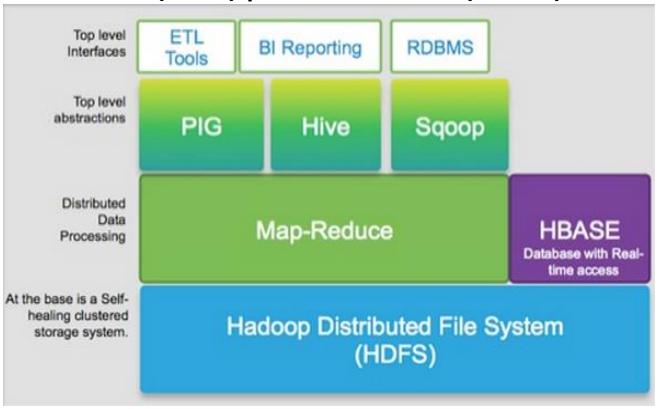

Figure 3: Hadoop architecture

A. Below are some basic features of Hadoop :-

- Hadoop maintains and secures the data by storing and keeping its replica.
- It is focused on scaling according to data usage.
- It can detect and delete the failed task and as well as failed transaction of data.
- It not only recovers the data but also automatically restores the data at its place.

B. Typical Hadoop Platform Stack – HDFS + Hive + HBase + Pig and other tools

- ***HDFS*** (Hadoop Distributed File System) – is part of Hadoop and is known as a special file system which deals with distribution and storage of large set of data. HDFS stores file as sequence of same size of block except the last block. It also deals with hardware failure and smoothen the data handling.
- ***Hive*** – Hive was initiated by Facebook. Hive is data warehouse tool which is based on Hadoop and converts query language into MapReduce jobs. It deals with the storage , analysis and queries of large set of data. Query language in hive used as HQL statement. Hive Query Language is similar to standard SQL statement.
- ***Hbase*** – Hbase is a Hadoop application which runs on top of HDFS. Hbase system represents set of table but Hbase is column oriented database management system i.e. different from the row oriented database management system. Generally if we talk about database then we think of relational database system but unlikely Hbase is not relational database at all and also it doesn't support Structured Query Language like SQL. Java is prefered language use for Hbase application. One most important feature of Hbase is to real time read or write to large set of data.
- ***Pig*** – initiated by Yahoo, became open source in 2007. Do you know why it is named as Pig? It is because it can handle any type of data!! Strange but true. Pig is a high level procedural programming platform developed for simplifying large data sets query in Hadoop and MapReduce. Pig has two components- one is PigLatin which is programming language and the other is run time environment where PigLatin programs are executed.

- ***Spark***-Spark was created by Matei Zaharia at UC Berkeley's AMPLab in 2009 as a replacement for MapReduce. Like MapReduce, Spark lets you process data distributed across tens or hundreds of machines, but Spark uses more memory in order to produce faster results. Spark also has a simpler and cleaner API.The only cases where MapReduce is still used are either because someone has an existing application that they don't want to rewrite, or if Spark is not scaling. This would be because Spark is a newer technology, and it sometimes can fail on extremely large data sets.
- ***Kafka*:-** Kafka handles the case of real-time data, meaning data that is coming in right now. Most other technologies handle batch scenario, which is when you have data sitting in a cluster. Kafka represents a different way of looking at data. Where as Hadoop and HDFS look at data as something that is stationary and at rest, Kafka looks at data as in motion. Kafka is like TiVo for real-time data. It can buffer the data when it spikes so that the cluster can process it without becoming overwhelmed. If data is coming in faster than it can be processed, Kafka will store it.In this way, Kafka is like other queuing systems, such as RabbitMQ and ActiveMQ. But Kafka can store a lot more data (it can store Big Data) because it is distributed across many machines. Kafka is also used for fault-tolerance. It can store data for a week (by default), which means if an application that was processing the data crashes, it can replay the messages from where it last stopped. It can also be used as a multiplexer. When the same data needs to be consumed by different applications in the system, Kafka can take incoming data and send it to all the applications that have subscribed
- ***Storm*:-**Storm is used for real-time processing. While Kafka stores real-time data and passes it onto systems that want to process it, Storm defines the logic to process events.Storm processes records (called events in Storm) as they arrive into the system. For example, every time a credit card transaction is sent into a bank, a Storm application can analyze it and then decide whether to approve it or deny it.Storm was the first system for real-time processing on Hadoop, but it has recently seen several other open-source competitors arise. Spark Streaming is the primary competitor, which offers exactly-once semantics—meaning each message is processed exactly one time. Storm only offers at-least-once semantics, meaning a message may be processed more than once if a machine fails.Storm is used instead of Spark Streaming if you want to have the event processed as soon as it comes in. Spark Streaming processes incoming events in batches, so it can take a few seconds before it processes an event. When immediate processing is essential, Storm is superior to Spark Streaming. Other new systems that provide real-time processing are Flink and Apex.

C. ***NoSQL*:** As the term says NoSQL, it means non relational or Non-SQL database, refer to Hbase, Cassandra, MongoDb, Riak, CouchDB. It is not based on table formats and that's the reason we don't use SQL for data access. A traditional database deals with structured data while a relational database deals with the vertical as well as horizontal storage system. NoSQL deals with the unstructured, unpredictable kind of data according to the system requirement.NoSQL Technologies HBase (part of the Hadoop ecosystem), Cassandra, MongoDB, Riak, CouchDB.

D. ***Cassandra*** database is used to handle the large set of data when we need to scale the database with high performance. Cassandra deals with the fault tolerance and replication of the data. With this we can go deeper in columns, supercolumns and more. It is a partial relational database system, supports best query capability but don't have joins feature. It follows the column family model map with two dimensional and 3 dimensional. 2D model includes column family with some column in it, while 3D model created by associating super column in column family.

E. ***MongoDB*** is an agile NoSQL document database, unlike the traditional database which store the data in rows and column, MongoDB stores the document data in binary form of JSON document which is also known as BSON format. It is used for high scalability, availability and performance. In MongoDB dynamic schemas are the unit of database, which found in document where set of documents are found in collection while set of collection makes the database.

F. ***Riak*** is open source NoSQL database system which is designed for availability, fault tolerance, scalability and high performance. It provides three kind of storage key/value store, document oriented store and web shaped store. It also stores documents in the JSON format. When we talk about data modeling, we will see that there is no 'Master', only nodes are there. All nodes are same and don't have different responsibility.

G. ***CouchDB*** is open source NoSQL database ,distributed, and schemaless.It stores the document data in the JSON format. It also provides feature related to web, like access of document from the web browser through HTTP. Javascript can also be use to modify the document. In CouchDB document is combination of strings "keys" and "values".

Conclusion: Big data is the large and complex datasets and it is generate from various sources like social media comments, playing a videogame, email attachments etc. There is complexity in big data such as velocity, variety and volume. These three terms are more challenging for big data analytics. There are variations possible while generating and storing data whether data is in audio, video, images and text. Research on typical big data application can generate profit for businesses, improve efficiency of government sectors. These technical challenges must be addressed for efficient and fast processing of Big Data. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of

application domains, and therefore not cost-effective to address in the context of one domain alone. The paper describes Hadoop which is an open source software used for processing of Big Data.

REFERENCES

[1] Sameera Siddiqui, Deepa Gupta," Big Data Process and Analytics : A Survey", International Journal Of Emerging Research in Management & Technology,ISSN: 2278-9359, Volume 3, Issue 7, July 2014.

[2] Han Hu, Yongyang Nen, Tat Seng Chua, Xuelong Li,"Towards Scalable System for Big Data Analytics: ATechnology Tutorial", IEEE Access, Volume 2, Page No653, June 2014.

[3] Bhatia Thakur, Manish Mann," Data mining for big data:A Review", International journal of advanced Research in Computer Science and Software Engineering, ISSN:2277 128x, Volume 4, Issue 5, May 2014.

[4] Ch.Sai Krishna manohar,"Analytics of Big Data Science using Big Data"Department of Information Technology, Tirumala Engineering College, Narasaraopet, India, IOSR Journal Of Computer Engineering(IOSR-JCE),e-ISSN:2278-0661,p-ISSN:2278 Volume 10,Issue2(Mar-Apr)

[5] Md.Salahuddin,Shaikh Muhammad Allayear,Sung Soon Park,"The Architectural Design Of Healthcare Services with big data storing Mechanism",IOSR Journal Of Computer Engineering(IOSR-JCE),e-ISSN:2278-0661,p-ISSN:2278-8727,Volume 16,Isseue 5,Ver.VIII(Sep-Oct.2014),PP 81-94

[6] Gemson Andrew Ebenezer J.1 and Durga S.2," BIG DATA ANALYTICS IN HEALTHCARE: A SURVEY" ARPN Journal of Engineering and Applied Sciences ©2006-2015 Asian Research Publishing Network (ARPN). All rights reserved. VOL. 10, NO. 8, MAY 2015, ISSN 1819-6608

[7] Manish Kumar Kakhani1, Sweeti Kakhani2 and S.R. Biradar3," Research Issues in Big Data Analytics",International Journal of Application or Innovation in Engineering & Management (IJAIEM),Volume 2, Issue 8, August 2013, ISSN 2319 –4847

[8] Nishchol Mishra1, Dr.Sanjay Silakari2,"Predictive Analytics: A Survey, Trends, Applications,Oppurtunities & Challenges", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (3) , 2012, 4434-4438,ISSN:0975-9646

[9] Dr Saravana kumar N M, Eswari T , Sampath P & Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing (ISBCC‟15)

[10] Nishchol Mishra1, Dr.Sanjay Silakari2,"Predictive Analytics: A Survey, Trends, Applications,Oppurtunities & Challenges", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (3) , 2012, 4434-4438,ISSN:0975-9646

[11] Kuchipudi Sravanthi, Tatireddy Subba Reddy," Applications of Big data in Various Fields", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (5) , 2015,4629-463