International Journal of Multidisciplinary Research and Modern Education (IJMRME)
Impact Factor: 6.725, ISSN (Online): 2454 - 6119
(www.rdmodernresearch.org) Volume 3, Issue 1, 2017

EMERGING TRENDS IN BIG DATA ANALYTICS IN HEALTHCARE

J. Christina Bai Annapoorani* & C. Ruby Gnanaselvam**
* M.Phil Scholar, Bishop Appasamy College of Arts and Science, Coimbatore, Tamilnadu
** Assistant Professor, Bishop Appasamy College of Arts and Science, Coimbatore, Tamilnadu

Cite This Article: J. Christina Bai Annapoorani & C. Ruby Gnanaselvam, "Emerging Trends in Big Data Analytics in Healthcare", International Journal of Multidisciplinary Research and Modern Education, Volume 3, Issue 1, Page Number 143-147, 2017.

Abstract:

Everyday data is growing exponentially and it becomes necessary to analyze the massive amount of data to achieve meaning full results. Big data Analytics help in analyzing high volume of data to identify useful information patterns and trends using various Algorithms. Healthcare is one of the most important areas for developing and developed countries to facilitate the priceless human resource. The healthcare industry has produced enormous amount of data through patient information and record storage. This papers focus on the emerging trends in big data analytics in healthcare industry.

Key Words: Big Data Health Care, Hadoop & Analytics

1. Introduction:

Big data refresh to massive amount of structured, semi structured and unstructured data evolving from various sources like social data machine generated data healthcare data, traditional enterprise data which are so large that it becomes difficult to process with traditional database and software techniques. The need to process and analyze such massive data sets introduced analytics of data called big data analytics is mainly used to identify trends to determine research quality to prevent des etc. Big data analysis is used in various application like medicine, physics, simulation, biology etc. Healthcare data analysis involved analyzing data generated from wearable implantable biometrics sensors blood pleasures heart rate data is collected is analyzed.

Characteristics of Big Data: There are 6 characteristics of Big data which are called as 6 Vs. Fig 1 shows the 6Vs of Big data.

a. Volume: Big data produces massive amount of data sizes such as Terabytes (TB), approximately 1012 bytes, Petabytes PB) approximately 1015 bytes, Zeta bytes (ZB) approximately 1021bytes.

b. Velocity: Velocity refers to speed at which new data is generated. In healthcare industry sensors devices and wearable collect real times physiological data of patients at rapid space or velocity. This new data is generated every second and causes of big data challenged for data analytics.

c. Variety: Variety represents various types of data such as structured data like returned tables, semi structured data like key value web clicks and unstructured data like email message particles and streamed audio and videos.

d. Value: Value is defined as the amount of information, the collected data could bring.

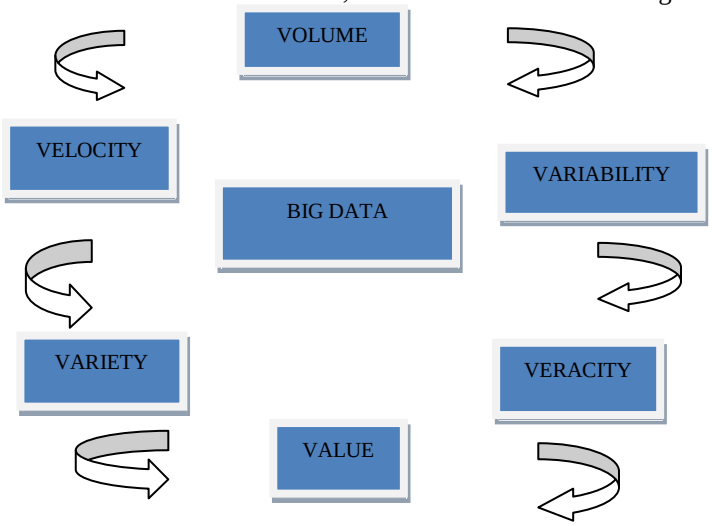

Figure 1: 6Vs of Big Data

e. Veracity: The veracity refers to the trustworthiness of data and data consistency. It deals with data that is unsure and vague.

f. Variability: It is refers to data changes during the processing and life cycle increase in variability increase the attractiveness of data and potentiality in providing unexpected hidden and valuable information.

2. Traditional and Big Data Analytics:

The architecture of Big data analytics is different from Traditional data analytics. In traditional analytics, the sources are internal and are structured. Data integration tools are employed to extract and transform and load the data from transactional databases. Data quality and Data normalization are applied and the data is modelled into rows and columns. The modelled data is then loaded into enterprise data warehouse. Big data is data that is so extensive to process using normal and traditional ways. As the data volume explodes enterprise would need analytic tools which are reliable, robust and capable of being automated. Traditional data warehousing is not able to handle the amount of big data that is come from various sources like social media, videos etc. This kind of data grows at a very fast pace. Hence the database requirements are extremely different in case of big data. Data can be anywhere and will be in large volume in case of Big data. Big data provides meaningful information by discovering hidden data patterns. It concentrates on unstructured data. Technologies like Hadoop, NoSQL and Map reduce frameworks are used for big data analytics. In big data analytics, the Hadoop framework captures datasets from various sources and performs functionalities like storing, cleansing, distributing, indexing, transforming, searching, accessing, analyzing and visualizing. Through these functions, the unstructured data is converted into structured format. The working principle behind Hadoop and all big data is to move the query to the data to be processed and not the data to the query processor. Various languages used in big data analytics are Java, Oracle Java script etc. Big data requires many different approaches to analysis depending on the problem. Big data technologies work faster than traditional data warehousing techniques.

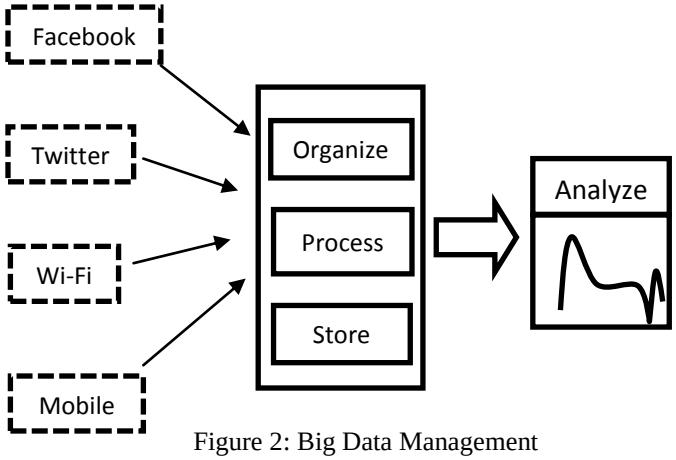

Figure 2: Big Data Management

The above diagram shows the Big data management. Data from various sources are collected together. These data could be either audio, video, images or data from social media etc. Since these data sources are unstructured, it is necessary to understand and categorize and analyze this large volume of data. Hence we need a system which would organize process and store the data into a database so that it can be analyzed efficiently.

Table 1: Difference between Traditional Data and Big Data Analytics

Traditional Data warehouse Analytics	**Big Data Analytics**
Traditional Analytics analyzes on the known data terrain that too the data that is well understood. Most of the data warehouses have an elaborate ETL processes and database constraints, which means the data that is loaded inside a data warehouse is well under stood, cleansed and in line with the business metadata.	The biggest advantages of the Big Data is it is targeted at unstructured data outside of traditional means of capturing the data. Which means there is no guarantee that the incoming data is well formed and clean and devoid of any errors. This makes it more challenging but at the same time it gives a scope for much more insight into the data.
Traditional Analytics is built on top of the relational data model, relationships between the subjects of interests have been created inside the system and the analysis is done based on them.	In typical world, it is very difficult to establish relationship between all the information in a formal way, and hence unstructured data in the form images, videos, Mobile generated information, RFID etc. have to be considered in big data analytics. Most of the big data analytics databases are based out Columnar databases.
Traditional analytics is batch oriented and we need to wait for nightly ETL and transformation jobs to complete before the required insight is obtained.	Big Data Analytics is aimed at near real time analysis of the data using the support of the software meant for it

Parallelism in a traditional analytics system is achieved through costly hardware like MPP (Massively Parallel Processing) systems and / or SMP systems	While there are appliances in the market for the Big Data Analytics, this can also be achieved through commodity hardware and new generation of analytical software like Hadoop or other Analytical databases.

3. Technologies Used for Big Data Analytics:

Unstructured and unpredictable data can be handled by NoSQL database. Data stored in NoSQL database is of high variety. NoSQL database provides a mechanism for storage and retrieval of data that is modelled in means other than tabular relations other than relational databases. Relational and NoSQL databases are highly different. Relational database model takes the data and separates them into many interrelated tables that contains rows and columns. But document oriented NoSQL database takes data into documents using JSON format. JSON is the abbreviation for JavaScript Object Notation. Another major difference between the relational database and NoSQL is that relational database contains more rigid schemas while NoSQL database is schema less. Many NoSQL databases have excellent integrated caching capabilities. Frequently used data is kept in system memory.

NoSQL database types are:

- ✓ Document Database: Document database pair each key with complex data structure known as Document. Document may contain nested documents. This type of database store unstructured (text) or semi-structured documents which are hierarchical in nature.
- ✓ Graph Stores: Graph database is based on graph theory. It is used to store information about network.
- ✓ Key Value Stores: Every single items is stored as an attribute name together with its value.
- ✓ Wide Column Stores: They are optimized for queries over large datasets and store columns of data together instead of rows.

Apache Hadoop is a one of the fastest growing big data processing open source platform. Hadoop can handle all type of data like structured, unstructured, pictures or audio. It can run on Linux, OSX, Windows and Solaris. Hadoop is scalable, flexible and fault tolerant. It contains HDFS. Hadoop HDFS is scalable, distributed file system written in Java. Hadoop creates clusters of machines and coordinates work among them. If any of the cluster fails, then Hadoop continues to operate the cluster without losing data. Map Reduce is a programming model and software framework first developed by Google. It works like a UNIX pipeline. A Map Reduce job divides the input dataset into independent subsets that are processed by map tasks in parallel. This step of mapping is then followed by a step of reducing tasks. These reducing tasks use the output of maps to obtain the final results. Map Reduce framework consists of a single master Job Tracker and one slave Task tracker per cluster node. The master is responsible for scheduling the job's component tasks on the slave, re-executing the failed task. The slave executes the task as directed by the master.

A few of Hadoop oriented projects are given below:

- ✓ Pig: It is a Scripting language and run time environment. It allows users to execute Map Reduce on a Hadoop cluster. Pig's language layer currently consists of a textual language called Pig Latin.
- ✓ Hive: It provides SQL access for data in HDFS. Hive's query language, HiveQL, compiles to Map Reduce. It also allows user-defined functions.
- ✓ HBase: A scalable, distributed database that supports structured data storage for large tables. It is column based rather than row based.
- ✓ Mahout: Library of machine learning and data mining algorithm. It has four types of algorithm.
- ✓ Oozie: Oozie is a Java Web-Application that runs in a Java servlet-container – Tomcat. It is job coordinator and workflow manager.
- ✓ BigTop: It is used for for packaging and testing the Hadoop ecosystem.

4. Big Data in Health Care:

Big data in Healthcare refers to electronic health data sets so large and complex that it is difficult to manage with traditional or common data management methods and traditional software. Mobile applications, health monitors, personal fitness trackers, various sensors, social media and electronic medical records are all vast sources of data. In healthcare, there has been a massive initiative to digitize patient's health records. New tools and platforms comes out every day to help healthcare leverage the ever expanding set of data in a way that was ever imagined. As these tools evolve and are adopted, it will help to understand the current state and future state of healthcare in a much wider perspective. By leveraging Big data properly and delivering it in the context of user workflows doctors will be able to determine who is at risk for diseases like diabetes or certain type of cancer and provide preventive care. Data type can be classified into genomics data, clinical data, clinical notes, behavior data and patient sentiment data, Health publication and clinical reference data, administrative business and external data.

- ✓ Genomic Data: It represents significant amounts of new gene sequencing data; It refers to gene typing gene expression and DNA sequence.
- ✓ Clinical Data and Clinical Notes:

- Structured data (eg: laboratory data structured EMR/HER
- Unstructured data (eg: post-op notes, diagnostics testing reports patient discharge summary, x ray images, and radiological images.
- Semi structured data (eg: copy paste from other structured source.

✓ Behavior Data and Patient Sentiment Data:
- Web and social media data Search engine Internet consumer use and networking sites (Facebook, Twitter, LinkedIn, Health plan website and smart phone etc.)
- Health publication and clinical reference data Text based publication (journal article, clinical research, medical reference material and clinical text based reference practice guidelines.

✓ Health Publication and Clinical Reference Data: This include text- based publication Like Journal article clinical research and medical reference material and clinical Text based reference practice guidelines

✓ Administrative Business and External Data:
- Insurance claims and related financial data, billing and scheduling
- Biometrics data(finger prints Handwriting iris scans)
- Device data adverse events and patient feed back
- Content from portal or personal health records, messaging between patient and provider data generated in PHR

✓ Medical Image Processing: The importance information on anatomy and organ function to detecting diseases states is provided by medical techniques. There are used for organ delineation, tumor identification in lungs, spinal deformity diagnosis, artery stenosis detection and so forth. In these applications, image processing techniques such as enhancement, segmentation, and denoising in addition to machine learning methods are employed. Due to the increase in the size and dimensionality of data, understanding the dependency among the data and designing methods which are efficient and accurate demand new computer aided techniques and platforms.

5. Types of Analytics:

There are three types of Analytics:

✓ Descriptive Analytics
✓ Predictive Analytics
✓ Prescriptive Analytics

Descriptive analytics means looking into the historical data ranging from 1 minute to year ago. It can be compared as looking into the rear mirror while driving. Predictive analytics means using all the data to make a prediction about where to go. It is the navigation that tells you where to go and when you will arrive. Prescriptive analytics means exactly knowing the best path to take based on infinite data points and calculations. With prescriptive analytics it becomes possible to understand and grasp future opportunities or mitigate future risks as predictions are continuously updated with new data that comes in. Prescriptive analytics use patented software to predict what is going to happen, when it is going to happen and why it is going to happen. It is used under scenarios where there are too many variables, options, constraints and data sets. Also the healthcare industry deals with massive amounts of different data sets that need to be analyzed. When healthcare providers combine data sets such as patient records, medicine information, economic data, demographical and sociographical data, health trends, hospital data etc. they will be able to offer better healthcare for less money, they will be able to improve future capital investments for new facilities or hospital equipment and improve the efficiency of hospitals. Combining so many different data sets can also be used to offer doctors recommendations in the best possible treatment for a patient.

6. Conclusion:

Big Data Analytics is a very new concept and is growing at a very rapid pace. It provides new perspectives and opportunities to various businesses and government enterprises to come out with different varieties of analysis. Research and Development of Unstructured Data is highly a need and will contribute to extremely profitable opportunities. Big data Analytics in healthcare will be able to provide better healthcare services and will improve efficiency of hospitals. There are several challenges to Big data analytics as well. Highly advanced technological data handling is required and efficient tools and technologies are the need of the hour. Big data analytics techniques in healthcare helps in extracting new information which could add real value to healthcare analytics in future.

7. References:

1. Payal Malik, Lipika Bose," Study and Comparison oF Big Data With Relational Approach", International Journal of Advanced Research In Computer Science and Software Engineering, Volume 3, Issue 8, August 2013 pp 564-570
2. Wei Fan, Albert Bifet," Mining Big Data: Current Status, and Forecast to the Future", SIGKDD Explorations, Volume 14, Issue 2

3. Big Data Survey Research Brief", Tech. Rep.SAS, 2013
4. Elena Geanina ULARU, Florina Camelia PUICAN, Anca APOSTU, Manole Velicanu," Perspectives on Big Data and Big Data Analytics", Database Systems Journal vol. III, no. 4/2012
5. Srinivasan, "SOA and WOA Article, Traditional vs. Big Data Analytics, "Why big data analytics is important to enterprises", [Online].Available: http://soa.sys-con.com/node/1968472
6. Available: http://www.mongodb.com/learn/nosql
7. Dhruba Borthakur, The Hadoop Distributed File System: Architecture and Design [Online] Available: https://hadoop.apache.org
8. Online] Available: http://www.revelytix.com/?q=content/hadoopecosystem [10] Available: http://www.mckinsey.com [11] NESSI-Big Data White Paper," Big Data –a new world of opportunities" December 2012