IEEE International Conference on Advances in Engineering & Technology Research (ICAETR - 2014), August 01-02, 2014, Dr. Virendra Swarup Group of Institutions, Unnao, India

Big data for Better Health Planning

Prof. Jigna Ashish Patel
Assistant Professor, CE Department,
Institute Of Technology
Nirma University, Ahmedabad
Gujarat, India
jignas.patel@nirmauni.ac.in

Dr. Priyanka Sharma
Associate Professor , MCA Department
Institute of science & technology for Advanced Studies & Research, VV Nagar
Gujarat, India
pspriyanka@yahoomail.com

***Abstract*—**The exponential evolution of data in health care has brought a lot of challenges in terms of data transfer, storage, computation and analysis. For healthcare usage and applications, ample patient information and historical data, which enclose rich and significant insights that can be exposed using advanced tools and techniques as well as latest machine learning algorithms. Though, the size and rapidity of such great dimensional data requires new big data analytics framework. This paper introduces the thought of data in healthcare and the results of various surveys to show the impact of big data. Few case studies of big data analytics in healthcare is presented. Last section is about the tools and techniques comprising Hadoop, Storm, Spark and HPCC - big data solution offered to solve big data issues and challenges

Keywords- Big data, health care,Hadoop

I. INTRODUCTION

The term "Big Data" become popular in last few years, as it represents the hard work of researchers to achieve business intelligence by processing tremendously large amount of data. To collect, store, manage and analyse it is very difficult for typical dataset software tools. Of course big data is too large to load into memory and store on a hard-drive and fit in a standard database.[1] Big data extents three dimensions: Volume, Velocity and Variety. Organizations from retail to wholesalers and Enterprises are overburdened with growing data of all types and in petabytes of information. Which leads the rapid increase in big data size called volume. Time –sensitive processes leads to the timely data response of big data called velocity. All Variety of data such as text, audio, video, click streams, sensor data, log files and news will lead to classify the data in three category. Structured, Semi-structured and unstructured. Researchers find the insights by analyzing these data types together.[7]

From banking to retail many sectors have already embraced big data traditionally. Industries taking advantage of big data by gaining deeper insights. It is very important as the emergence of cloud computing and the demand of analysing massive data arisen. Data increase in terms of peta bytes to thousands of gigabytes the ability of handling big data become so essential that nobody can mistreat it. [2] The definition of big data may vary depending upon the application and kind of tools and techniques available. Big data may be considered as "digital breadcrumbs", which we left behind our every communication transaction and then we need to analyze it for better business. Sometimes big data is less about the data and more about the analytics. [9]The information that we filter from the data will drive you to proper solution. To acquire an actual meaning of the face-book status is important than the just an updated status.

Even though industry hype, most of institutions and organizations have yet to develop, implement a big data policy [4]. SAS (leading business analytics Software Company) had carried out one survey in July 2013. Survey was conducted on 339 companies with thorough experience in data management strategy and execution. They were asked few questions based on the usage of tools and technology for data management in their existing organization. For total sample size, 95% confidence level was set and thus marginal error was +/- 5.3. The survey clearly mention (as shown in figure:1) 39% of the organizations are currently exploring big data environment. It is found that 12 percent of organizations have considered big data strategy and they are in Execution/Implementation phase. Unexpectedly, by eliminating those who don't know or do not consider and not already implementing, testing or planning[7].

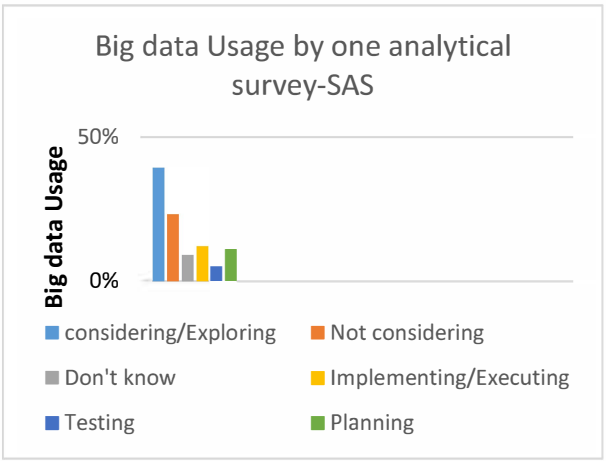

Figure: 1

There are various methods that big data can be used to generate value through zones of the world-wide economy. McKinsey Institute has shown the abundant interest through great surveys to build the potential of big data:

978-1-4799-6393-5/14/$31.00 ©2014 IEEE

- Healthcare: medical decision support structures, precise analytics applied for patient profile, revised and reformed medicine, performance based pricing for personnel, analyze disease forms, improve public health, health care forums, health insurance issues, Research issues to decrease the un-natural death
- Public sector: producing transparency by available data, discover requirements, expansion of business and improved performance, convert activities for suitable products and amenities, decision making with automated systems to reduce risks, revolutionizing unusual products and services, public sector and creating important economic additional for consumers.
- Retail: in store performance research and analysis, variability and value optimization, product design, increase performance, work contributions optimization, delivery techniques , Online efficient marketing strategies
- Manufacturing: better demand estimating, supply chain management, sales maintenance, advanced production operations, online search based requests
- Personal location data: Efficient routing, disaster or alternative reaction, town planning, novel commercial models

II. BIG DATA TRENDS AND HEALTHCARE

Dr.Joel Selanikio says "The surprising seeds of a big data revolution in the health care". The healthcare industry is one of the extreme leading and rising industry .With plenty of challenges big data opportunities transformed into healthcare. Here, big data refers the compilation of data from different people. The data consists of disease, varying symptoms, medicines, diet, exercises, prescriptions, lab reports, treatment schedule, allergy, insurance data, all records of Physicians, nurses and patients. Health-care investors are more pointing towards benefit and improved healthcare. With older tools and technology they couldn't take much benefit by processing big data. Figure2 shows the survey conducted in year 2012, whole 34% processes use the big data applications in improvement in performance of management, budget ,planning and forecasting issues, which increased by 11% in year 2013.A comparative survey of increased usage of big data applications was conducted by Mc. Kinsey global institute is imagined out in figure:2

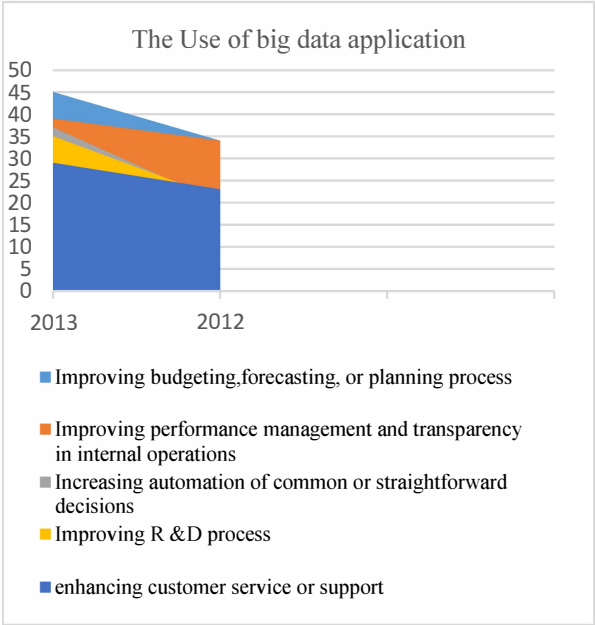

Figure 2

The big data Plan which include three core elements data, analytic models and tools to get more business value. Critical planning is needed to assemble and integrate the data. Incorporating data only does not produce value. Progressive logical models are needed to permit data-driven plan. A plan must recognize where models will create surplus business value, who will need to use them. The output from the business models should be understood by every users, employee and managers. Here the equation used by[19] is

Interlinked data input + analytic models +decision support tools = Business value

How to reduce the hospital cost data by including their primary care and increase business value is suggested in [19]. It clearly mention the project hospitalization risk for individual patients by calculating the patient risk calculator.

III. BIG DATA EDGES IN HEALTHCARE

There are several case studies with clear vision in various parks of healthcare field.

Case-Study 1: DiabeticLink

By utilizing advanced data,text and web mining algorithms and other computational techniques, the founders of DiabeticLink made a real time system in both Taiwan and US markets. The system is capable of addressing the needs of patients, care takers, nurse educators, physicians, researchers, pharmaceutical companies. It provides all features that encourage social connection, data sharing and educational opportunities. It started with aim to provide unique research

978-1-4799-6393-5/14/$31.00 ©2014 IEEE

opportunities in healthcare decision support and patient empowerment. There are other social media sites such as DailyStrength and PatientsLikeMe provides unique discussion space for chronic disease like cancer, Alzheimer's dieses, Parkinson's disease, Diabetes. Internet usage in US and in Taiwan is around 70% to 80% for health social media sites. [21]

The current system provide the health portal which uses advanced algorithms and techniques developed in AI lab. Fig:3 shows all its unique features.

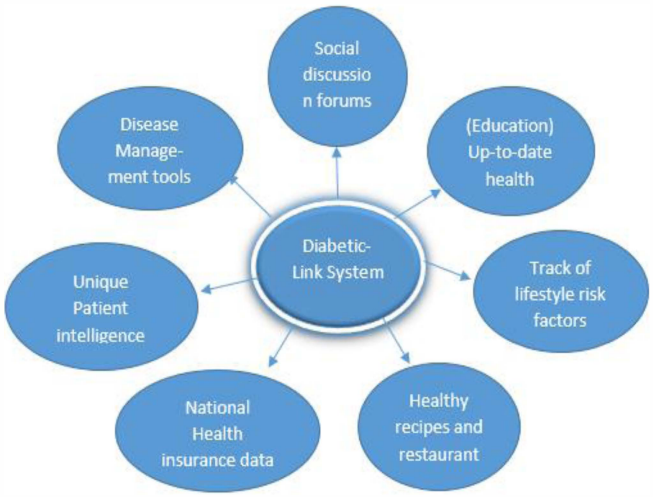

Figure 3: Diabetic Link System

The system is capable of handling near 2.5 million electronic health records for nearly 8,93,000 patients. Tool clearly shows the management of big data which lead to better lifestyle choice and better disease outcome. Modules associated with the system is shown in figure. **Unique Patient intelligence tools** are used to tell the relationship between the factors of lifestyle and treatment. By using the efficient data mining tools and techniques here they provided the EHR datasets and social media stuffing for user to follow new methods and services for protective interferences. **Disease management tools** keep track of serious diabetes parameters like Blood pressure, Blood glucose etc. Additionally it keep track of diet/nutrition, physical movement and exercises, medicine, insulin quantity and disease regression or progression. **Social discussion forums** are provided to all the users to have a very healthy discussion on treatment, disease, experiences, success and challenges. It also provide the safe and secret platform for users who can get benefit from available social support. **Up-to-date education** offers the health statistics, facts and information, videos, news are provided from well-known diabetes sites. Creation of digital representation of the disease is very significant as it would be more trajectory and manageable by providing the **track of diabetes and life style risk**. By providing 652 **healthy recipes and restaurant** to the users to have the knowledge of how to prepare healthy meal, nutrient charts so that increase in glucose level can be manage. **National Health Insurance Data**, the objective of this module is to reply several frequently asked questions which the diabetes patients and their caretakers may have with statistical reports extracted from the Taiwanese National Health Insurance database. All the questions are sorted and put in the four categories: prevalence of Diabetes in Taiwan, Out-patient Related Questions, Inpatient Related Questions, and Medication Related Questions. Actually sampling is done here in this module to facilitate patients to choose the questions, based on the answers patients were categorised as having diabetes and included in the analysis which will lead to escape growth of misdiagnosis.[21]

Case study 2 : Bringham Women's Hospital

Brigham Women's hospital is a nonprofitmaking and schooling hospital .It is 793-bed hospital associated with teaching in Harvard Medical School. Hospital is an internationally prominent hospital, well-known for its quality in patient attention, its status for excellent biomedical research as well as teaching. The tasks and challenges for the hospitals are extensive follow-up and additional processes with limited analysis, data pre-processing with missing values, lack of efficient reporting, considerable IT resources required for changes. By developing IBM Pure data system for analytics BWH delivered a solution that could use its huge data dimensions and conduct several medication studies concurrently consenting researchers to design, experiment and put on novel algorithms to rapidly identify drug risk warning signals.[22]. The result similarly allows the section to conduct simple analytic processing three times prior speeds, though their new high-dimensional tendency scoring algorithms can run 20 to 30 times faster than its previous environment. The IBM big data platform covers the value of current technology by patter into all traditional and non-traditional sources of data and effectively handling, assimilating, governing and performing analytics on this torrent of complexity and data growth.[22]

Case study 3: Asthmapolis

Asthmapolis launched in 2010 to assist people to discover a solution with the help of sensor technology and the mobile data to help people manage their asthma more commendably. Expenditure of Asthmapolis reduced the cost of healthcare and the number of patients in US. The best available help for patients by doctors using the digital health technology. Similar type of systems are now available for patients suffering from sinusitis, diabetes, too.

Asthma is a disease where patient's lungs airways get narrow down and it become compulsion to use inhaler. Asthmapolis provides the GPS sustained device which screen the usage of inhaler. The date, time and location of the inhaler usage with its analytics is tracked by sensor technology. With the help of latest mobile apps doctors are being informed about the patients status. Doctors or management staff is allowed to classify the patients according to their situation which lead to control the disease in emergency. By consuming advanced

data classification algorithms and machine learning approaches in Asthmapolis the number of people suffered from asthma are more controlled now.

Case study 4: Wearable monitors

To save the life of small babies the concept of wearable monitor is found. 8% Canadian babies who born prematurely, this become one of the reason which increase the infant death ration. By using the big data and its solutions for real time processing, new approaches to Physiological data analysis and cloud computing and its services Carolyn Mcgregor has given proper solution to the said issue.

IV. TOOLS AND TECHNIQUES:

(1)Hadoop

The father of all framework to organise large number of computation on distributed environment is Map Reduce, released by Google. By aiming the easier data processing on huge number of clusters Map reduce framework was introduced in 2004. Hadoop developed by apache for ascendable, reliable distributed processing. Hadoop mainly consists of HDFS and Map Reduce. The Hadoop distributed file system used to store and process the large amount of data in distributed manner.

Hadoop was generated by Doug Cutting, the maker of Apache Lucene, the extensively used to provide a framework for big data. To take advantage of the parallel processing that Hadoop provides, we need to express our query as a MapReduce job. After some local, small-scale testing, we will be able to run it on a cluster of machines. MapReduce works by breaking the processing into two phases: the map phase and the reduce phase. Each phase has key-value pairs as input and output, the types of which may be chosen by the programmer. The programmer also specifies two functions: the map function and the reduce function.[14]

(2) Spark

Hadoop provides cluster storage approach, whereas Spark provides scalable data analytics platform with in-memory computing. It has been proved that in-memory computing provides faster data access by eliminating the I/O overhead. Spark supports open source environment which increase the computing power which leads to superiority then Hadoop. Spark is designed for explicit applications like machine learning algorithms and natural language processing.

The Spark runs on Apache Mesos, a cluster manager by which Spark applications coexist with Hadoop. The drivers working in Spark uses two type of operation. (1) Action (2) Transformation. Action is similar as reduce and transformation is similar as map and cache operation. Spark is developed in scala and it supports the scala, which is functional programming language used to provide distributed and iterative environment.

(3)Storm

Storm is launched as an open source by Twitter in September, 2011.Storm is the implementation of Map-Reduce concept of Hadoop. It is implemented in Clojure language to support machine learning environment. Ruby and Python is supported to make applications in Storm. The key idea is it is used for streaming processes. Running of "topologies" would never end till you kill the process, contradictorily with Hadoop job. Storm uses no storage concept, which simply tells all about semi-structured, un-structured and structured data together.

(4) HPCC

High Performance Computing Cluster also uses the Map-Reduce framework for the analysis of large amount of data. It works with Enterprise control language (ECL), a declarative programming language. ECL provides entire programming paradigm in which highly parallelism is achieved. Mainly two clusters are concerned: Thor and Roxie, Thor provides the simplification of ETL (Extract-Transform-Load) process and Roxie provides data delivery by highly concurrent procedures. It uses HPCC distributed file system. Major two advantages provided by HPCC over Hadoop is scalability and speed. HPCC platform provides the intricate physics submissions and imagining of simulations in depth. In support of decision making HPCC is used by Elsevier to boost its logical and critical skills for SciVal.

(5) SAP-HANA

In-memory computing for big data SAP devised a new tool HANA, which processes on block of the data by using advanced parallel architecture and algorithms for faster speed.

Feature wise comparison for Hadoop, SAP-HANA and Spark is shown in following table

Tech-nique	Language(s) Used	Components	Processing Of Data	Features
Hadoop	Java, C++ (MapReduce)	-HDFS -MapReduce -YARN	Clusters	• Distributed File System • Open source • Low cost • Simple Coherency Model • "Moving Computation is Cheaper than Moving Data" • Data Replication • Processing has only 2 steps: Map and Reduce Fault Tolerance Persisting and Check-pointing intermediate results Drawbacks Complexity of code
In-Memory Computing (SAP HANA)	SQL, SQL Script, ABAP (Advanced Business Application Programming), C++	• MDX (MultiDimensional Expressions) • Text Analytics • Application Function Libraries • Parallel Calculation Engine • Relational Stores • Object Graph Stores • Managed Appliances	Block	• Column Storage • Compression techniques • Parallelization • Data locality(data is placed in RAM) • No aggregate tables • High Availability of data(because large tables are distributed across multiple servers Fault Tolerance • Data Logs are kept in shared memory • If the SAP HANA system detects a failover situation, the work of the services on the failed server is reassigned to the services running on the standby host
Spark	Scala, Java, Python	• Shark • Spark Streaming • GraphX	Batch	• Event Driven Architecture • Better Parallelism • Resilient Distributed Datasets (RDDs) • Runs 100 times faster than Hadoop • Uses in-memory computing **Fault Tolerance** • It remembers the sequence of operations which led to a certain data set, so when a node fails, Spark reconstructs the data set based on the stored information **Drawbacks** • It's not really well suited in which require to change only a few entries of data set at the time due to the immutable nature of the RDDs.

V. REFERENCES:

[1] Big data survey research brief ,2013 www.sas.com/resources/whitepaper/wp_58466.pdf

[2] Avita katal,Mohammad wazid,R H Goudar," Big data: Issues,Challenges,Tools and Good practices" IEEE 2013.

[3] Xin Cheng,Chungjin Hu,Yang Li, Wei Lin, Haolei Zuo., "Data Evolution of virtual dataspace for managing the big data lifecycle ",IEEE 2013

[4] S. Seref and S. Duygu, "Big Data : A review",IEEE 2013

[5] Antonia Azzini, Paolo Ceravolo , "Consistent Process Mining Over Big data Triple Stores"

[6] IEEE Big data congress,2013

[7] "Big data without big database", http://www.slideshare.net/katemats/big-data-cachesurge2012?from_search=1, Last access 21st december,2013

[8] D.Yuri,G.Paola,Cees de Laat, "Addressing Big data Issues in Scientific Data Infrastrcuture" june 2013.

[9] "Heading Towards Big Data" R.Gardiner Goss,K.Veeramuthu.,Manufacturing technology, Globalfoundries, ASMC 2013

[10] "How is Big Content Different From Big Data" http://www.slideshare.net/jmancini77/movingthe-mountain-evanta-cio-presentation-on-big-data-and-big-content?from_search=2, July 2013.

[11] "Using Big data and Predictive Machine learning in Aerospace Test Environments" , Tom Armes, Mark Refern IEEE,2013.

[12] "Big data Processing in Cloud Computing Environments" by Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, Keqiu Li,2012 international symposium on pervasive systems , Algorithms and Networks

[13] P.Agawal "Approximate Incremental Big data harmonization", IEEE International conference on Big Data,2013.

[14] http://hadoop.apache.org/ last access 17.01.2014

[15] http://hpccsystems.com/ last access 17.01.2014

[16] E. Begoli and J. Horey, "Design Principles for Effective Knowledge Discovery from Big Data", Software Architecture (WICSA) and European Conference on Software Architecture (ECSA) Joint Working IEEE/IFIP Conference on, Helsinki, August 2012

[17] C. Tankard, "Big Data Security", Network Security Newsletter, Elsevier, ISSN 1353-4858, July 2012

[18] V. Borkar, M.J. Carey and C. Li, "Inside "Big Data Management": Ogres, Onions, or Parfaits", EDBT/ICDT 2012 Joint Conference Berlin Germany, 2012

[19] Basel Kayyali, David Knott, and Steve Van Kuiken The big-data revolution in US health care: Accelerating value and innovation, April 2013

[20] Kyuseok Shim, MapReduce Algorithms for Big Data Analysis,

[21] Hsinchun Chen, Sherri Compton, and Owen Hsiao , DiabeticLink: A Health Big Data System, © Springer-Verlag Berlin Heidelberg 2013

[22] Varun Chandola, Sreenivas R. Sukumar , Jack Schryver, Knowledge Discovery from Massive Healthcare Claims Data, KDD'13, August 11–14, 2013, Chicago, Illinois, USA., Copyright 2013 ACM

978-1-4799-6393-5/14/$31.00 ©2014 IEEE