The Inescapable Application of Big Data To Healthcare

Alkesh Verma[1] , Shubham Vishwakarma[2] , Vijay Shankar Rai[3] ,Purwa Maheshwari[4]
[1, 2, 3, 4] Dept of Computer Science & Engineering
[1, 2, 3, 4] ABES Institute of Technology,Ghaziabad, Uttar Pradesh, India

Abstract*- The main purpose of this study was to explore the use of the big data in healthcare, and design the patterns and trends for the disease in future such as the selection of appropriate treatment paths, improvement of healthcare systems through the big data analytics.By providing an overview of the current state of the application of the big data in the healthcare environment, this study has explored the current challenges that governments and healthcare stakeholders are facing as well as the opportunities presented by big data In this paper, the goal is to forecast the prediction through analyzing the healthcare for those clients who are interested in healthcare investment which is done via Big Data technology using Hadoop, HDFS, MapReduce, Sqoop, and Hive. This is a process of financial decision making for the investments. This model will help the shareholders to know the current scenario of any company and the market.*

Keywords- Big Data, Healthcare, Big Data Analytics,Apache Spark,Jupyter Notebook.

I. INTRODUCTION

The healthcare industry historically has generated huge amounts of data, driven by record storing, compliance & requirements, and patient data [1]. While most of the data is stored in hard copy form, the current trend is toward rapid digitization of these huge amounts of data. Driven by mandatory requirements and the potential to improve the quality of healthcare delivery meanwhile reducing the costs, these massive quantities of data (known as to be ‘big data’) hold the promise of supporting a wide range of medical and healthcare functions, including among others clinical decision support, disease surveillance, and population’s health management. A report says data from the U.S. healthcare system exclusively reached, in 2011, 150 Exabyte’s. At this rate of growth, data for healthcare industry will soon reach to the zettabyte (1021 gigabytes) scale and, not long after, the yottabyte (1024 zettabyte)

Healthcare: Health care or healthcare is the maintenance or improvement of health via the prevention, diagnosis, and treatment of disease, illness, injury, and other physical and mental impairments in human beings. Healthcare is delivered by health professionals (providers or practitioners) in allied health fields. Physicians and physician associates are a part of these health professionals[2].Dentistry, midwifery, nursing, medicine, optometry, audiology, pharmacy, psychology, and other health professions are all part of healthcare. It includes work done in providing primary care, secondary care, and tertiary care, as well as in public health. **Big Data:** Big Data is the technology to analysand the large and massive data sets which are having a huge amount of data, this may be structured or unstructured. The data can be retrieved from Facebook, Twitter, or real-time data. Most of the data sets are analysand on the singer server environment but whenever the data set increases there is a need for increased infrastructure to handle the data sets with high memory speed and storage drives[1]. The data sets are in Hera-bytes, Pentax-bytes or ea-bytes .

Apache Spark :Apache Spark is a lightning-fast cluster computing technology, designed for fast computation. It is based on Hadoop MapReduce and it extends the MapReduce model to efficiently use it for more types of computations, which includes interactive queries and stream processing. The main feature of Spark is its in-memory cluster computing that increases the processing speed of an application.Spark is designed to cover a wide range of workloads such as batch applications, iterative algorithms, interactive queries and streaming. Apart from supporting all these workload in a respective system, it reduces the management burden of maintaining separate tools[3].

I. **Components of Spark:**The following illustration depicts the different components of Spark.
II. **Apache Spark Core:**Spark Core is the underlying general execution engine for spark platform that all other functionality is built upon. It provides In-Memory computing and referencing datasets in external storage systems.
III. **Spark SQL:**Spark SQL is a component on top of Spark Core that introduces a new data abstraction called SchemaRDD, which provides support for structured and semi-structured data.
IV. **Spark Streaming:**Spark Streaming leverages Spark Core's fast scheduling capability to perform streaming analytics. It ingests data in mini-batches and performs RDD

(Resilient Distributed Datasets) transformations on those mini-batches of data.

V. **MLlib (Machine Learning Library):**MLlib is a distributed machine learning framework above Spark because of the distributed memory-based Spark architecture. It is, according to benchmarks, done by the MLlib developers against the Alternating Least Squares (ALS) implementations. Spark MLlib is nine times as fast as the Hadoop disk-based version of Apache Mahout (before Mahout gained a Spark interface).

VI. **GraphX:**GraphX is a distributed graph-processing framework on top of Spark. It provides an API for expressing graph computation that can model the user-defined graphs by using Pregel abstraction API[4]. It also provides an optimized runtime for this abstraction.

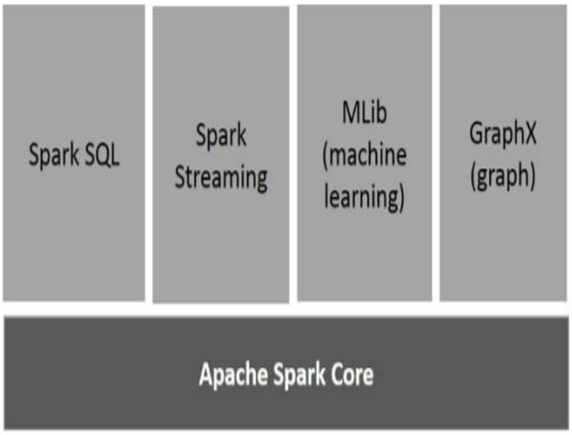


Fig2. Components of Apache Spark

VII. **Jupyter Notebook**:The notebook extends the console-based approach to interactive computing in a qualitatively new direction, providing a web-based application suitable for capturing the whole computation process: developing, documenting, and executing code, as well as communicating the results.The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.The Jupyter notebook combines two components:

A web application: a browser-based tool for interactive authoring of documents which combine explanatory text, mathematics, computations and their rich media output. Notebook documents: a representation of all content visible in the web application, including inputs and outputs of the computations, explanatory text, mathematics, images, and rich media representations of objects.

VIII. Resilient Distributed Datasets: Resilient Distributed Datasets (RDD) is a fundamental data structure of Spark. It is an immutable distributed collection of objects.Each dataset in RDD is divided into logical partitions, which may be computed on different nodes of the cluster. RDDs can contain any type of Python, Java, or Scala objects, including user-defined classes.Spark makes use of the concept of RDD to achieve faster and efficient MapReduce operations.

Data Sharing using Spark RDD:Data sharing is slow in MapReduce due to replication, serialization, and disk IO. Most of the Hadoop applications, they spend more than 90% of the time doing HDFS read-write operations.Recognizing this problem, researchers developed a specialized framework called Apache Spark. The key idea of spark is Resilient Distributed Datasets (RDD); it supports in-memory processing computation. This means, it stores the state of memory as an object across the jobs and the object is sharable between those jobs. Data sharing in memory is 10 to 100 times faster than network and Disk.The illustration given below shows the iterative operations on Spark RDD. It will store intermediate results in a distributed memory instead of Stable storage (Disk) and make the system faster.

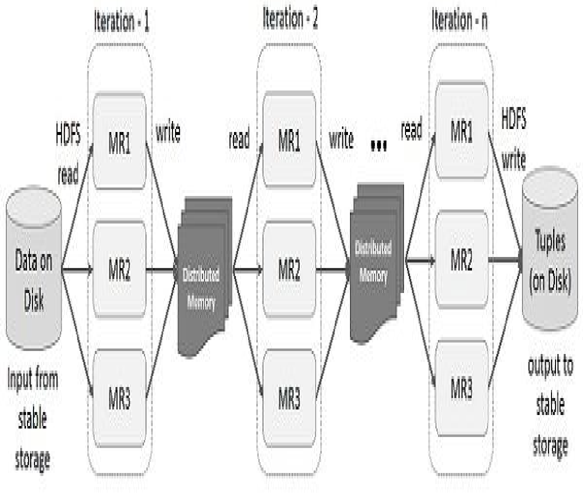


Fig 3.Iterative Operations on MapReduce

IX. Interactive Operations on Spark RDD This illustration shows interactive operations on Spark RDD. If different queries are run on the same set of data repeatedly, this particular data can be kept in memory for better execution times.

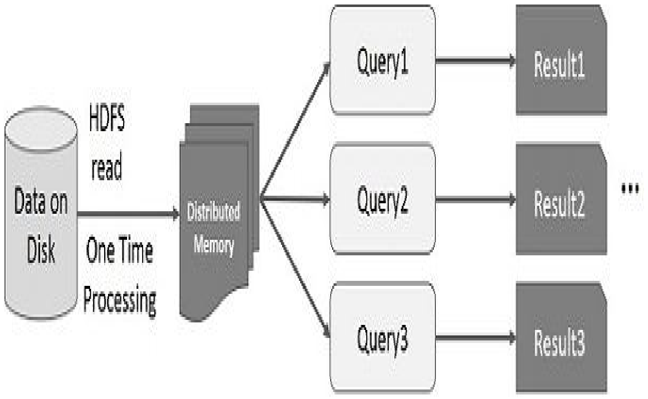


Fig 4- Iterative Operations on RDD

II. OBJECTIVE

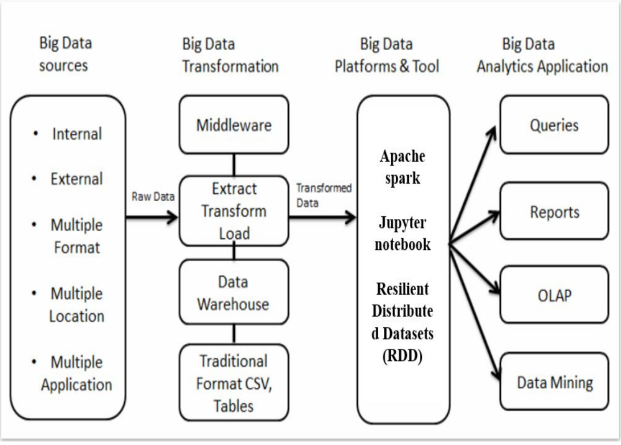
The objective of this paper is to provide a sight of vision towards the condition of disease in future on which an government and healthcare stakeholder is interested to invest in the field of the healthcare. In healthcare industry, every individuals should have the right to know all the ups and downs of a company through which the company is going, so that one can be prevented from the loss and also be benefited through the idea of analyzing.

Here the plan is to have such a framework which provides the transparency to the trends of diseases in future.

III. PROBLEM DOMAIN

According to the research analysis of the laboratory disease test data through big data, technology is only possible with the heavy data sets. The analysis of data is done as per the prices (low price, high price, etc.) of the healthcare which should correspondingly contribute in the analysis of a healthcare test data so that government, healthcare stakeholders would take actions in future before the condition got adverse.

IV. METHODOLOGY



Here we will take data from the different sources e.g. laboratories,hospials etc. and then we will then work on it,we will take about 2-4 years of data and we will generate the patterns on the basis of it.

As the big data is itself means the massive data that is why the data here is streamed for the model so that the analysis can be done the heavy data set.

V. CONCLUSION

Big data analytics has the potential to transform the way healthcare providers use sophisticated technologies to gain insight from their clinical and other data repositories and make informed decisions. In the future we'll see the rapid, widespread implementation and use of big data analytics across the healthcare organization and the healthcare industry. To that end, the several challenges highlighted above, must be addressed. As big data analytics becomes more mainstream, issues such as guaranteeing privacy, safeguarding security, establishing standards and governance, and continually improving the tools and technologies will garner attention. Big data analytics and applications in healthcare are at a nascent stage of development, but rapid advances in platforms and tools can accelerate their maturing process.

VI. FUTURE SCOPE

For the further consideration, the plan is to work on the live data streaming. So, that the government can plan for the controlling of the adversity of the diseases in future. The adversity of disease cannot be calculated before it goes adverse. Therefore for the better surely and support to the government and different healthcare stakeholders, they have the right to get to now about each and every possible effect that can happen in future by diseases.

REFERENCES

[1] Raghupathi W: Data Mining in Health Care. In Healthcare Informatics: Improving Efficiency and Productivity. Edited by Kudyba S. Taylor & Francis;2010:211–223.

[2] Burghard C: Big Data and Analytics Key to Accountable Care Su [2] Jeffrey and Sanjay Map-Reduce: Simplified processing on the large cluster, Google Research Publication 2004.

[3] "Spark Release 2.0.0". MLlib in R: SparkR now offers MLlib APIs [..] Python: PySpark now offers many more MLlib algorithms".

[4] Gonzalez, Joseph; Xin, Reynold; Dave, Ankur; Crankshaw, Daniel; Franklin, Michael; Stoica, Ion (Oct 2014). "GraphX: Graph Processing in a Distributed Dataflow Framework"