INTERNATIONAL JOURNAL OF RESEARCH IN SCIENCE & TECHNOLOGY

e-ISSN:2249-0604; p-ISSN: 2454-180X

Leveraging the Big Data Tools and Techniques to Attempt Scientific Analytics for Drawing an Optimum Business Intelligence

Teesha Ahuja

Bharati College, University of Delhi, New Delhi, India

Paper Received: 19th July 2021; **Paper Accepted:** 26th September 2021; **Paper Published:** 17th October 2021

DOI: http://doi.org/10.37648/ijrst.v11i04.003

How to cite the article:

Teesha Ahuja, Leveraging the Big Data Tools and Techniques to Attempt Scientific Analytics for Drawing an Optimum Business Intelligence, IJRST, Oct-Dec 2021, Vol 11, Issue 4, 20-25, DOI: http://doi.org/10.37648/ijrst.v11i04.003

2021
We invite articles from Engineering & Science Faculties
Open Access, Peer Reviewed Journal
International Journal of Research in Science & Technology

ABSTRACT

In healthcare, big data analytics is becoming a promising field for gaining insight from large data sets, enhancing outcomes, and decreasing costs simultaneously. The benefits of big data analytics in healthcare are outlined, an architectural framework and methodology are outlined, examples from the literature are discussed, the challenges are briefly addressed, and conclusions are provided in this paper.

INTRODUCTION

Due to patient care, record keeping, and regulatory and compliance requirements, the healthcare sector has historically generated a significant amount of data [1]. The majority of data is still stored in hard copy, but the current trend is to rapidly digitize these vast amounts of data. Clinical decision support, disease surveillance, and population health management are just a few of the medical and healthcare functions that could benefit from these enormous amounts of data, which are referred to as "big data" [2–5]. They are motivated by mandatory requirements and the potential to reduce costs while simultaneously enhancing healthcare delivery quality. In 2011, 150 exabytes of data were stored by the healthcare system in the United States alone, according to reports. At this rate of growth, big data for healthcare in the United States will soon reach zettabyte (1021 gigabyte) and yottabyte scales [6]. Between 26.5 and 44 petabytes of potentially rich data from EHRs, including annotations and images, are thought to exist in the health network Kaiser Permanente, which is based in California and has more than 9 million members.

The vast quantity and variety of data are available to the big data scientist. By identifying associations, patterns, and trends in the data, big data analytics has the potential to improve care, save lives, and reduce costs.

As a consequence of this, applications of big data analytics in the healthcare sector make use of the explosion of data to extract insights that enable better-informed decisions to be made [10–12]. Given the context, big data analytics in healthcare is referred to as [13–15] as a research category.

METHODOLOGY

A. Big Data Analytics in Healthcare

The volume of health data is anticipated to increase significantly in the coming years [6]. Healthcare reimbursement models are also evolving. In today's healthcare environment, meaningful use and performance pay are emerging as crucial new factors. although profit is not the primary factor and should not be. Healthcare organizations must acquire the tools, infrastructure, and methods necessary to effectively utilize big data in order to avoid losing millions of dollars in revenue and profits.

B. Architectural Framework

A traditional health informatics or analytics project's conceptual framework is comparable to a big data analytics project's in the healthcare sector. The manner in which processing is carried out is the primary distinction. A typical health analytics project can be analysed using a business intelligence tool on a stand-alone system, like a desktop or laptop. Because big data is, by definition, large, processing is broken up and performed across multiple nodes. Distributed processing is a concept that has been around for a long time. It is relatively new to analyse large data sets as healthcare providers tap into their vast data repositories to gain insight for better-informed health-related decisions. Open-source cloud platforms like Hadoop and MapReduce have also encouraged big data analytics in healthcare.

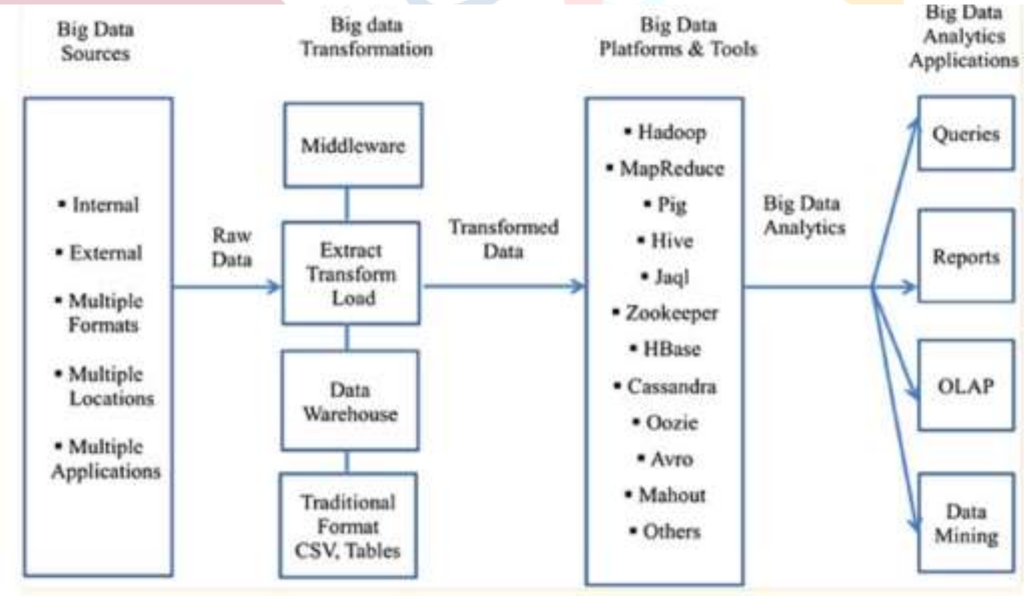


Figure 1. An applied conceptual architecture of big data analytics

The complexity begins with the data themselves, as shown in Figure 1. Even though the algorithms and models are the same, the user interfaces of traditional analytics tools and big data tools are completely different. Tools for traditional health analytics are now very easy to use and transparent. On the other hand, big data analytics tools are extremely difficult to use, require extensive coding, and call for a wide range of skills. They have appeared independently, mostly as open-source development platforms and tools. As a result, they require more assistance and ease of use than vendor-driven proprietary tools do.

In the healthcare industry, big data can come from within (e.g., clinical decision support systems, CPOE, etc.). and external sources (government sources, labs, pharmacies, health maintenance organizations, etc.), frequently in multiple formats, including ASCII/text, relational tables, flat files, and.csv. And residing in numerous legacy and additional applications (such as transaction processing applications, databases, and others) at multiple locations (both geographically and at the sites of various healthcare providers). Types of data and sources include:

1) Data from the web and social media: Data on interactions and clickstreams from sites like Facebook, Twitter, LinkedIn, and blogs. Websites for health plans, smartphone apps, and other things are examples. [6].

2) Data from machine to machine: readings from remote meters, sensors, and other devices that monitor vital signs [6].

3) Big data on transactions: Health care claims and other billing records are becoming increasingly accessible in semi- and unstructured formats [6].

4) Data from biometrics: fingerprints, genetics, handwriting, retinal scans, x-ray and other medical images, and other similar types of data, as well as readings of blood pressure, pulse, and pulse-oximetry [6]

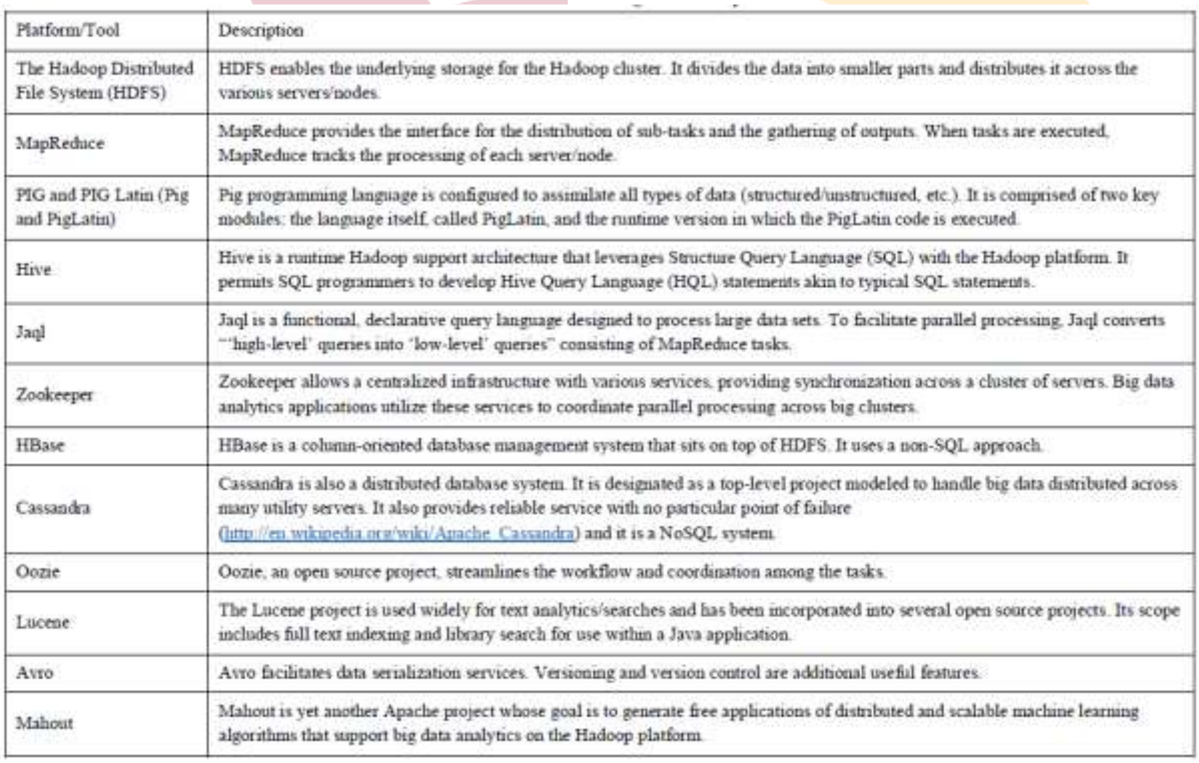
5) Data generated by humans: EMRs, doctor's notes, emails, and other paper documents are examples of semi- and unstructured data [6].

All of this data needs to be combined for big data analytics. The data in the second section needs to be altered or processed because it is "raw." At this point, there are a few options.

The most important platform for big data analytics is the open-source distributed data processing platform Hadoop (Apache platform), which was initially developed for routine tasks like aggregating web search indexes. It belongs to the "NoSQL" class. Hadoop can serve as both a data organizer and an analytics tool at the same time. It has a lot of potential for helping businesses make use of data that has been hard to manage and analyse up until now. In particular, Hadoop lets you process much data with different or no structure structures. However, it can be difficult to install, configure, and manage Hadoop, and it is difficult to locate individuals with Hadoop expertise.

Additionally, these factors prevent businesses from fully adopting Hadoop. The Hadoop distributed platform is supported by the ecosystem of additional platforms and tools in the surrounding area [30, 31]. Table 1 provides a summary of these.

Table1: Platforms & tools for big data analytics in healthcare



Platform/Tool	Description
The Hadoop Distributed File System (HDFS)	HDFS enables the underlying storage for the Hadoop cluster. It divides the data into smaller parts and distributes it across the various servers/nodes.
MapReduce	MapReduce provides the interface for the distribution of sub-tasks and the gathering of outputs. When tasks are executed, MapReduce tracks the processing of each server/node.
PIG and PIG Latin (Pig and PigLatin)	Pig programming language is configured to assimilate all types of data (structured/unstructured, etc.). It is comprised of two key modules: the language itself, called PigLatin, and the runtime version in which the PigLatin code is executed.
Hive	Hive is a runtime Hadoop support architecture that leverages Structure Query Language (SQL) with the Hadoop platform. It permits SQL programmers to develop Hive Query Language (HQL) statements akin to typical SQL statements.
Jaql	Jaql is a functional, declarative query language designed to process large data sets. To facilitate parallel processing, Jaql converts "'high-level' queries into 'low-level' queries" consisting of MapReduce tasks.
Zookeeper	Zookeeper allows a centralized infrastructure with various services, providing synchronization across a cluster of servers. Big data analytics applications utilize these services to coordinate parallel processing across big clusters.
HBase	HBase is a column-oriented database management system that sits on top of HDFS. It uses a non-SQL approach.
Cassandra	Cassandra is also a distributed database system. It is designated as a top-level project modeled to handle big data distributed across many utility servers. It also provides reliable service with no particular point of failure (http://en.wikipedia.org/wiki/Apache_Cassandra) and it is a NoSQL system.
Oozie	Oozie, an open source project, streamlines the workflow and coordination among the tasks.
Lucene	The Lucene project is used widely for text analytics/searches and has been incorporated into several open source projects. Its scope includes full text indexing and library search for use within a Java application.
Avro	Avro facilitates data serialization services. Versioning and version control are additional useful features.
Mahout	Mahout is yet another Apache project whose goal is to generate free applications of distributed and scalable machine learning algorithms that support big data analytics on the Hadoop platform.

CONCLUSION

Applications of big data analytics in healthcare are still in their infancy, but rapid advancements in platforms and tools may accelerate their development. Big data analytics can transform the use of sophisticated technologies by healthcare providers to gain insight from their clinical and other data repositories and make informed decisions. Healthcare organizations and the healthcare industry will soon implement big data analytics and use it extensively. As big data analytics becomes more mainstream, issues like ensuring privacy, protecting security, establishing standards and governance, and continuously improving tools and technologies will attract attention. As a result, it is necessary to address the aforementioned issues.

REFERENCES

1. Rocha A., Hauagge C.D., Wainer J. and Goldenstein S., (2010), Automatic fruit and vegetable classification from images, *Computers and Electronics in Agriculture*, Vol. 70, pp. 96–104.
2. Raghupathi W. (2010), Data Mining in Health Care. In: Kudyba S, editor. *Healthcare Informatics: Improving Efficiency and Productivity*. pp. 211–223.
3. Burghard C. (2012), *Big Data and Analytics Key to Accountable Care Success.*
4. Dembosky A. (2012), Data Prescription for Better Healthcare. *Financial Times*, December 12, 2012, p. 19.
5. Feldman B, Martin EM, Skotnes T. (2012), Big Data in Healthcare Hype and Hope. *Dr. Bonnie 360.*
6. Fernandes L, O'Connor M, Weaver V. J. (2012). Big data, bigger outcomes; *AHIMA* pp. 38–42.
7. IHTT . *Transforming Health Care through Big Data Strategies for leveraging big data in the health care industry*. 2013.
8. Frost & Sullivan: *Drowning in Big Data? Reducing Information Technology Complexities and Costs for Healthcare Organizations.* http://www.emc.com/collateral/analyst-reports/frost-sullivan-reducing-information-technology-complexities-ar.pdf
9. Bian J, Topaloglu U, Yu F, Yu F. (2012), Towards Large-scale Twitter Mining for Drug-related Adverse Events. Maui, Hawaii: *SHB*; 2012.
10. Raghupathi W, Raghupathi V. (2013), An Overview of Health Analytics. *Journal of Health and Medical Informatics*
11. Ikanow: Data Analytics for Healthcare: Creating Understanding from Big Data. http://info.ikanow.com/Portals/163225/docs/data-analytics-for-healthcare.pdf