The Future of Big Data (Using Hadoop Methods)

[1] Alexander Adjei-Quaye,[2] Mohammed Mahfouz Alhassan, [3]Prof. Jianbin Wu
[1,2] Zhejiang Normal University,
Graduate School of MPI,
Jinhua, Zhejiang Province, CHINA
[3]Research interests (Formalization of Software Engineering& Information Retrieving)
[1] adjeiquayealexander@gmail.com, [2]mmalhassan@tamalepoly.edu.gh, [3]395230397@qq.com

Abstract— **Big Data is one of the most important and exciting career paths in today's world. Data Science – Recently, "Data Scientist" has become a popular job title for companies looking for technical experts with interdisciplinary background, a data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician. It is a general term that can be used regardless of the form the data may take (e.g. electronic, physical data, with knowledge of information security we are confident that our details protected and also assured of the safety of our data and ensure that the value of our organization's maintained.**

Keyword — **The future of big data in our hands now and forever with Hadoop.**

I. INTRODUCTION

Undeniably, every organization operates on data. Besides, the statistics are quite essential as how they are handled dictates the success or failure of those behind the information. However, despite being quite useful to the entity concerned, such facts may impose many challenges dealing with. As a result, any future-oriented 21st-century business would opt for a tool that it can utilize to evade such encounters. Simply put, the increased complexity of dealing with big data continue to force many enterprises globally into using software programming frameworks, especially "Hadoop," so as to diminish cost and amount of time spent on loading such voluminous information into an interactive database for analysis.

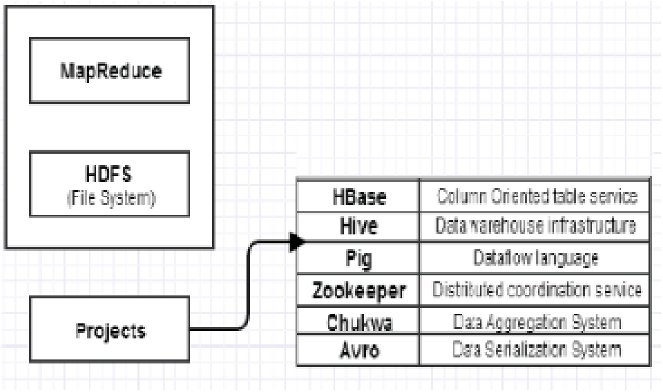
II. CHALLENGES AND SOLUTIONS

A. Big Data

The term "Big data" denotes a universal jargon used when describing the vast quantity of statistics that is both semi-structured and unstructured (Kleespies& Fitz-Coy, 2016). As mentioned earlier, such pieces of information are often generated by the company but consume a lot of money and time to load into an appropriate database for scrutiny. In other words, due to the nature of such data to enlarge so rapidly, many enterprises always find it not easy to tackle using the regular analysis tools. Therefore, such pieces of facts must be partitioned prior to examining them.

III. ARCHITECTURE PLATFORM

Essentially, Hadoop is comprised of two principal components, namely HDFS and MapReduce as displayed in the diagram that follows. While the latter is the processing section concerned with job management, the former, which denotes "Hadoop Distribution File System," is charged with the role of storing all facts redundantly needed for computations (Lin & Dyer, 2010). At the same time, projects are set of tools managed by Apache to offer support in the task correlated to Hadoop. Therefore, the diagram below denotes the entire architecture.

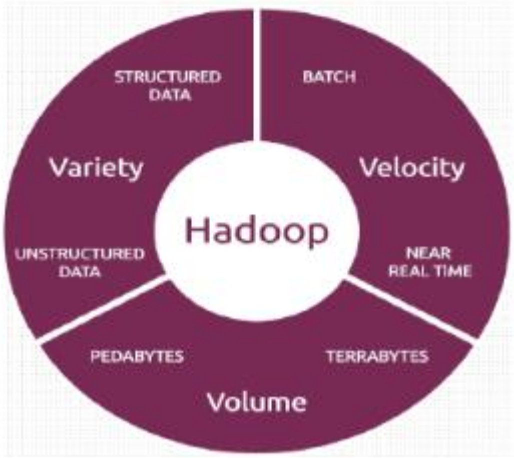


HBase	Column Oriented table service
Hive	Data warehouse infrastructure
Pig	Dataflow language
Zookeeper	Distributed coordination service
Chukwa	Data Aggregation System
Avro	Data Serialization System

Adopted from Singh & Kaur, 2014.

A. Hadoop

Hadoop describes a platform developed to help in tackling the challenges of processing and analyzing Big data. According to Singh and Kaur (2014), Hadoop is "An open source cloud computing platform of the Apache Foundation that provides a software programming framework called MapReduce and distributed file system, HDFS" (686). Having been written in Java, the framework supports the running of software on voluminous statistics. Consequently, it can address primary encounters produced by Big data. Therefore, the diagram below is a depiction of the biggest challenges associated with huge statistics.



Adopted from Singh & Kaur, 2014.

B. Volume

As an entire ecosystem of projects, Hadoop works to deliver a standard set of facilities. Unlike the traditional approach, huge data is first subdivided into relatively smaller segments so as to ensure effective and efficient handling of statistics. Undeniably, the tool transforms product hardware to coherent services, which then store petabytes of figures steadfastly (Mayer-Schönberger & Cukier, 2014). Furthermore, as data gets segregated, likewise, the software breaks the computation into smaller pieces. Besides, such information is subsequently processed efficiently via vast deliveries. Therefore, mentioned platform offers a framework that is capable of scaling out horizontally to subjectively bulky records to tackle bulks of statistics.

IV. Velocity and variety

Unlike other tools, Hadoop is widely applauded for its ability to partition data and compute across numerous hosts consistently. Undeniably, big data is often propagated with excessive swiftness and multiplicity. Besides, the instrument is capable of performing application calculations in matching close to their mandatory figures. Given the existence of an upper boundary to an amount of facts that can be processed, it is uneatable that scaling up such data is quite a challenge (Mayer-Schönberger & Cukier, 2014). Worth noting is the case that Hadoop is both reliable and redundant; thus, it can automatically replicate facts in the absence of operator’s intervention in the event of any failure. Therefore, the framework is designed to tackle gigantic pieces of information regardless of their furious incoming rate .Furthermore, being exceptionally powerful in accessing raw information, Hadoop is “primary batch processing centric and makes it easier for distributed application with the aid of MapReduce platform model” (Kshetri, 2014). Subsequently, its commodity hardware can reduce the cost associated with purchasing unique lavish hardware structures. In other words, challenges of velocity and variety linked with dealing with massive datasets are top priorities in as far as this software is concerned. Simply put, as a framework, Hadoop has the capacity work with multifaceted tasks to support any diversity of unstructured statistics. Thus, as soon as all the sub-reckonings are completed, the outcome is joined and translated back to the application.

CONCLUSION

In conclusion, many issues were addressed concerning the fundamental motive of challenges brought by handling Big data in as far as Hadoop is concerned; however, a few cases did stand out. Firstly, it was discovered that Hadoop shadowed a diverse approach when matched with the traditional methodologies utilized by some organizations. Secondly, unlike any other data analysis tool, Hadoop is competent when it comes to dealing with problems associated with velocity, variety, and volume of big data. So, it is advisable that every organization that would wish to escape challenges linked to working with enormous statistics through its management to ensure that Hadoop becomes a part and parcel of itself as it will be able to reduce cost and save time.

References

[1] Kleespies, J., & Fitz-Coy, N. (2016). Big impacts and big data: Addressing the challenges of managing DebriSat's characterization data. 2016 IEEE Aerospace Conference. doi:10.1109/aero.2016.7500889

[2] Kshetri, N. (2014). The emerging role of big data in key development issues: Opportunities, challenges, and concerns. Big Data & Society, 1(2). doi:10.1177/2053951714564227

[3] Lin, J., & Dyer, C. (2010). Data-intensive text processing with MapReduce. San Rafael, Calif.: Morgan & Claypool Publishers.

[4] Mayer-Schönberger, V., &Cukier, K. (2014). Big data: A revolution that will transform how we live, work, and think. Boston: Houghton Mifflin Harcourt.

[5] Singh, K., & Kaur, R. (2014). Hadoop: Addressing challenges of big data. 2014 IEEE International Advance Computing Conference (IACC). doi:10.1109/iadcc.2014.6779407