Hadoop in Healthcare

Lara L. Mechling

Bellevue University

DSC 200 Computer Systems for Data Science

Professor Neugebauer

October 18, 2021

Hadoop in Healthcare

Hadoop is an open-source, meaning free to use for all, software that is able to handle massive amounts of big data. Hadoop is considered to be highly scalable. The software "allows for the distributed processing of large data sets across clusters of computers, thereby removing the data ceiling by taking advantage of a "divide and conquer" method" (Nemschoff, n.d.). What this means is you can continue adding data without having to worry about maxing out your software's ability to store the information. Hadoop creates clusters of data stored over multiple computers allowing for a much greater storage and processing capacity. Added to the benefit of scalability is the cost-effectiveness of the platform. Since it is open source the software is free to utilize. Rowe demonstrates an added cost benefit is the ability of Hadoop to run on commodity servers, or servers which are generalized and usually made up of inexpensive, accessible hardware, which is less costly than specialized, dedicated storage area networks (Rowe, 2016). With these benefits it is no wonder many businesses turn to Apache Hadoop to manage their big data needs. One example of enterprises utilizing Hadoop is in hospital situations to assist in classifying patients and diagnosing diseases. "The combination between remote sensing devices and the big data technologies have been proven as an efficient and low cost solution for healthcare applications" (Harb, Mroue, Mansour, Nasser, & Motta Cruz, 2020). With so many devices recording and creating data in healthcare one of the issues with healthcare data is the collection and storage. "Biosensors continuously record vital signs of patients [which results in] massive data collection, high-speed generation, and heterogeneous nature" (Harb, Mroue, Mansour, Nasser, & Motta Cruz, 2020) making the data incapable of being processed by traditional storage methods. Hadoop is scalable so the massive amount of data can be captured

and stored continuously. In their journal article Harb et. Al utilize Spark, the Hadoop HDFS, and modules from the Hadoop ecosystem including Hive, SparkSQL, and Matplotlib (Harb, Mroue, Mansour, Nasser, & Motta Cruz, 2020). These Hadoop modules allow for ease of storage, processing, analysis, and visualization of the data. The modularity of Hadoop makes it a great choice for healthcare and a myriad of other enterprises. The combination of all aspects of the Hadoop architecture allow for the architects to effectively deploy a system of algorithms to assist in managing patients and disease classification needs.

References

Harb, H., Mroue, H., Mansour, A., Nasser, A., & Motta Cruz, E. (2020, March 30). A Hadoop-Based Platform for Patient Classification and Disease Diagnosis in Healthcare Applications. *Sensors (Basel)*. doi:10.3390/s20071931

Nemschoff, M. (n.d.). *How To Maximize Performance and Scalability Within Your Hadoop Architecture*. Retrieved October 18, 2021, from Smart Data Collective: https://www.smartdatacollective.com/how-maximize-performance-and-scalability-within-your-hadoop-architecture/

Rowe, W. (2016, July 7). *Advantages of Using Hadoop*. Retrieved October 18, 2021, from bmc blogs: https://www.bmc.com/blogs/hadoop-benefits-business-case/