JaeHwuen Jung

MIS 2502 Data Analytics-Section 3

Jianjin Liao

April 27, 2019

Hadoop

Apache Hadoop is an open-source software that used to store and process huge amount of big data. It makes up of four important components which are Hadoop Distributed File System(HDFS), YARN, MapReduce and Hadoop Common. Specifically, HDFS allows to store any kinds of data through cluster and YARN plays the core role to process the data in HDFS. YARN also stands for another resource negotiator and as well as the middle layer between MapReduce and HDFS. With YARN, Hadoop could greatly extend and cooperate with other Apache data science tools such as Spark, HIVE, Storm and Kafka. Moreover, Hadoop was developed for big data distributed analysis and help large organizations to make decisions in minimizing cost, transactional and threat analysis(bmc.com).

Compared to the materials we have covered in MIS2502, for example, we have learned how to retrieve and store structure data across MySQL, however, Hadoop is the replacement of SQL and can perform large size of big data set with both structured or unstructured, instead, MySQL can do only limited amount of data manipulation. With the using of HDFS and YARN, it dramatically increases the storage capacities and scalability. Similar with R, we have done a lot of practices by using RStudio through R programming language for computing and graphics analysis. Hadoop also need programming language such as java, Python to finish the data distribution process.

Based on the research of “ARPN Journal of Engineering and Applied Science,” it states that with the growing numbers of clusters and data store into HDFS, the chance of hardware and software failures will also increase. At this situation, Hadoop’s scalability and fault tolerance play an important role to deal with these failures. The first method the passage mention is data replication, HDFS stands for distribution file system and includes one master node “NameNode” and three slaves node “DataNode”(computing nodes), therefore, when the file input into NameNode, it automatically makes three copies on different slaves nodes to prevent either one loss. Secondly, checkpoint and recovery method, is kind of rollback way that when failure occurs, it will go back to the last fixed saved point to begin the transaction again. In conclusion, Hadoop is a great distributed data technology that utilized a series of techniques to accomplish big data analysis to help company to do the business.

Reference

2016. Benefits&Advantages of Hadoop. http://www.bmc.com/guides/hadoop-benefits-business-case.html

T. Cowsalya and S.R. Mugunthan. 2015. ARPN Journal of Engineering and Applied Sciences: http://www.arpnjournals.com/jeas/research_papers/rp_2015/jeas_0415_1837.pdf