Hadoop Architecture and its Usage at Facebook

Pranathi Priya Uppala

Wilmington University

Hadoop Architecture and its Usage at Facebook

**Iteration 1: Brainstorming and Gathering Requirements**

**Plan**

This first iteration is aimed at gathering useful information to understand the Hadoop data architecture and its usefulness in the analysis of Facebook data. To accomplish this purpose, the researcher intends to have a brainstorming session with the Hadoop and big data analysis experts so that he can get a general knowledge of the Hadoop data architecture and its use in the Facebook data analysis. He will write to the experts to inform them about the research and will then meet them for one week as they help him to understand the architecture and the use of the Hadoop platform. The experts are from Aero Century Corporation, working in senior positions as Hadoop data analysis. This brainstorming session will be conducted for two hours in one of the company’s hall every day, and the researcher will have one hour for revisiting what he has learned.

 The researcher also plans to research using scholarly materials to understand the requirements for accomplishing the project on Facebook data analysis with Hadoop. The researcher will first of all start with the gathering of the materials from the library and the Web and then proceed with mining the required information from these materials. These requirements will include the software and the hardware requirements required for the project. The researcher also plans to learn how to install the software on the computer and to make the necessary configurations as required by the project. The researcher plans to carry out this self-search for the second week of this iteration. After the iteration, the researcher then plans to document all the information he has acquired for future reference.

**Act**

The first thing that took place before embarking on the actual research was the writing of an invitation to inform the two Hadoop data analysts at Aero Century Corporation about the research and to have them prepared for the consultation. Having responded positively to the request, the research was ready to go. It took place in the company’s Technology Hall for two hours every day for the one week through which it ran. In understanding the architecture, the main components of Hadoop were covered in this brainstorming session including the Hadoop Distributed File System (abbreviated as HDFS), MapReduce, Pig, HBase, and Hive. The architecture of HDFS and MapReduce were also looked into detail, as these are crucial to Hadoop’s analysis of Facebook data. The research and consultation took place between 8 are, and 10 am throughout the first week.

After the consultation with the experts was completed for the first week, the researcher then took advantage of the weekend to plan and prepare for the second week’s research on the importance and the requirements for accomplishing the Hadoop Facebook data analysis. The researcher first gathered all the resources that included journals, books, magazines, and peer-reviewed articles on the topic, “Facebook Data Analysis with Hadoop.” The researcher then searched on the usefulness of Hadoop in analyzing big data, in this case, Facebook data. The researcher also searched for the software and hardware requirements for accomplishing the Facebook data analysis. That included how to install and configure the software and how to set up the project. Since the researcher was doing the search on his own, he did this in his house, taking two hours every data for the entire second week.

**Observation**

There were several observations, which were made from this first iteration on brainstorming and understanding the project requirements. The researcher observed that the Hadoop architecture is made up of two major components including the HDFS and the MapReduce (White, 2012). The HDFS is set of parallel and distributed files system that is designed to run on the commodity hardware. It is a highly fault tolerant system with its goals including the hardware failure, storage of large data sets, streamlining data access, and simple coherence model (Mishne et al., 2013). On the other hand, “the MapReduce is programming model. The use of this model is for executing large data sets using distributed and parallel algorithms in the cluster” (Ibrahim et al., 2009). The researcher also observed that MapReduce is the heart of the Hadoop system. The capacity of the data that can be stored and computed makes the Hadoop system useful for the analysis of Facebook data (Guo,2013).

From the personal search using scholarly resources the researcher also observed many things about the essence of the Hadoop in the analysis of Facebook data and how to set up the project. The researcher observed that Hadoop architecture could be useful in analyzing the Facebook data and fetching meaningful results. The Facebook data is being created at a huge speed, and it is the Hadoop system that can store, analyze, and carefully study this huge amount of data to make it useful. The Hadoop framework is preferred in analyzing data of large magnitude at a very great pace and as soon as the data is being created (Cowsalya & Mugunthan, 2015). The researcher also observed that setting up the project requires downloading and installing the Hadoop on the local machine, enabling the Facebook API Key, creating the client ID key, and following triggering steps to complete the project.

**Reflection**

The research has started very well with the researcher learning many things as a way of preparing for the project work ahead. It was a useful opportunity for learning from experts on what the Hadoop framework looks like and how useful it is in analyzing the Facebook data. Through the interaction with the experts, the researcher learned the functions of the various Hadoop framework components including the MapReduce, the HDFS, the Hive, HBase, and Pig. The experts moved at the pace that the researcher could easily catch up with and were always ready to answer all the questions the researcher asked without being bored. Time was also well used because every activity was completed with the set time and so all the research work was completed within the scope of the allotted time. The Proper preparation was also a very important ingredient for the success of this iteration.

The researcher, however, feels that some things did not go well with this research. For one, the inviting only two consultants restricted the researcher to only a little knowledge concerning the research in question. The researcher plans to ensure that he makes use of as many experts wit diverse skills and knowledge as possible to increase the scope of the knowledge gained from the research. The researchers also rushed through some concepts as they assumed that the researcher had a thorough knowledge of them, something that was a wrong assumption. The researcher will ensure that in the future he makes as much inquiry as possible from the experts to make sure that understands all the concepts without any concept being ignored. The researcher will also ensure that he draws a solid plan on all the concepts that researchers should handle and have them follow that.

References

Cowsalya, T. & Mugunthan, S.R. (2015). Hadoop architecture and fault tolerance based Hadoop clusters in geographically distributed data center. *ARPN Journal of Engineering and Applied Sciences, 10*(7), 2818-2820.

Guo, S. (2013). *Hadoop operations and cluster management cookbook*. Packt Publishing Ltd.

Ibrahim, S., Jin, H., Lu, L., Qi, L., Wu, S., & Shi, X. (2009, December). Evaluating mapreduce on virtual machines: The hadoop case. In *IEEE International Conference on Cloud Computing* (pp. 519-528). Springer Berlin Heidelberg.

Mishne, G., Dalton, J., Li, Z., Sharma, A., & Lin, J. (2013, June). Fast data in the era of big data: Twitter's real-time related query suggestion architecture. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data* (pp. 1147-1158). ACM.

White, T. (2012). *Hadoop: The definitive guide*. " O'Reilly Media, Inc."