Assignment-3 Submission

Student Name	Amritha Jeyarathan
Student ID	45345473
Chosen Question Number	3
Workshop Number	25
Workshop Day and Time	Monday 11:00 AM
Tutor Name	Hadieh Ranjbartabar

Answer

Hadoop is an open-source software framework created on the foundation of MapReduce and Google File System papers, designed to take advantage of Big Data. It is intended for tasks that can be executed at a particular time or under particular conditions without the need for human interaction. Hadoop is an execution framework of *MapReduce*, employing its algorithm to help overcome real-world organisational challenges (Hoffer, Venkataraman & Topi, 2016).

MapReduce is an algorithm that allows for the automation of parallel processing a diverse range of large-scale tasks by worker nodes. The divide-and-conquer method i.e. the notion of breaking a considerable computing task into smaller sections and apportioning them to several commodity nodes in a cluster so that each node can efficiently process the same task at the same time, is what underpins its algorithm. MapReduce is purposed to be executed in a fault-tolerant way such that in the event that a portion of the system fails to operate, the remaining nodes will able to maintain its functionality. The algorithm is designed to enable users without any experience in parallel processing or distributed systems to make use of the resources of a sizeable distributed system with ease, so that rather, users – programmers and non-programmers, can focus on the problem at hand (Hoffer et al., 2016). MapReduce functions with data stowed in relational databases, local disk files or both; the data can be structured or unstructured and comprises of binary, text or multi-line records. Unlike data stored in the form of tables, commonly found in relational databases, data in MapReduce is shown in key/value pairs for instance, "Last name/Anderson" (Schneider, 2012). The distribution process begins with the "map" phase, in which a computing task is parallelly performed on several subsets of the data, in the form of key/value pairs, and then returns a result for each subset independently. In the "reduce" phase, the results of each of the map processes are integrated, producing a final result (Hoffer et al., 2016). This customary MapReduce pattern utilises a distributed file system referred to as the Hadoop Distributed File System.

The *Hadoop Distributed File System* (HDFS) is a file system that manages numerous sizeable amounts of data sets in a highly distributed environment. The data is stored on a local disk and is then processed locally on the computer. HDFS works by breaking data into small "blocks", with a default size of 128MB, and replicating them on several nodes throughout the Hadoop cluster, with a default of three nodes. This achieves reliability and enables data to be processed simultaneously by several nodes in an attempt to exploit the inconveniences of Big Data. An HDFS cluster comprises of a *NameNode*, a single master node administering the file system namespace and coordinating access to files by a client, and countless low-cost, low-performing commodity nodes known as *DataNode*, which processes the computing tasks as illustrated in figure 1 (Hoffer et al., 2016). Many of Hadoop operations contain multiple master nodes, as having two or more master nodes lessens the possibility of a single point of failure as, without the NameNode, all data on the filesystem would no longer exist (Schneider, 2012). Although it is through HDFS that Hadoop is able to process considerable amounts of data on a large scale, HDFS has its limitations: HDFS cannot tackle ongoing updates as well as a conventional relational database management system, and cannot be installed directly onto the existing operating system, which can make inputting and obtaining data in and from the HDFS difficult (Taylor, 2010).

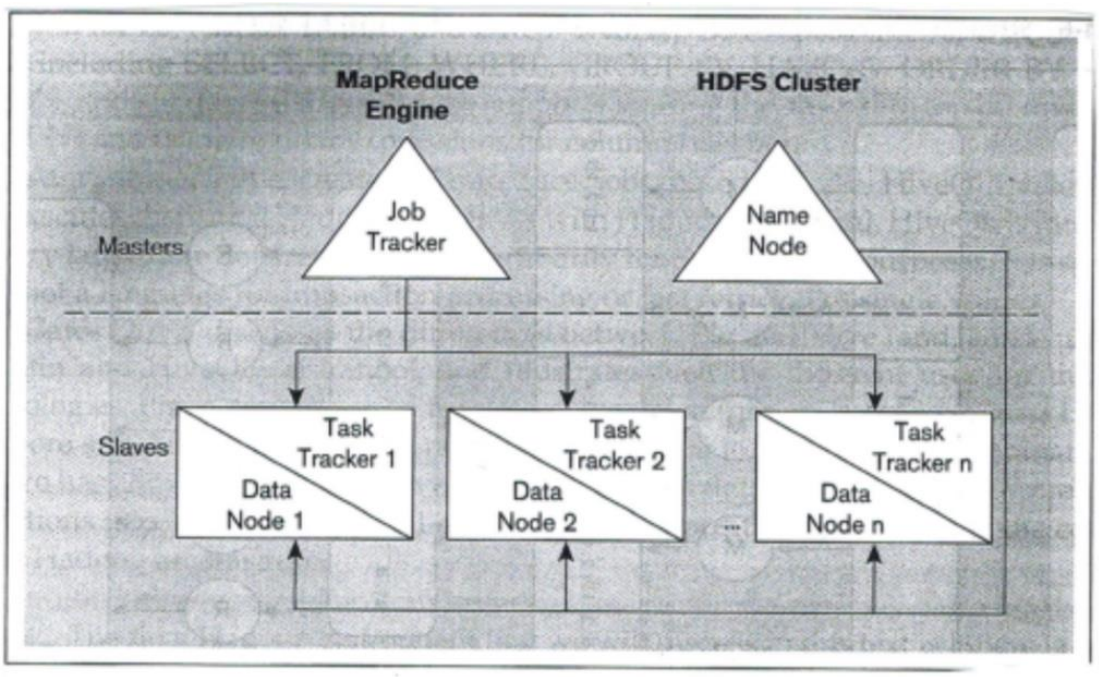

Figure 1. MapReduce and HDFS (Hoffer et al., 2016, p 491)

A wide array of supporting programming tools for the Apache Hadoop Project such as *Apache Pig* and *Apache Hive* allows users to obtain value from Big Data. Pig is an Apache subproject that incorporates *PigLatin*, a high-level scripting MapReduce language and an execution environment to operate PigLatin programs. It is intended to translate execution sequences, conveyed in PigLatin, into several sequenced MapReduce Programs and can automate vital data preparation tasks, transform data for processing, execute analytic functions, store results, define processing sequences etc. (Hoffer et al., 2016). Hive is another Apache subproject that strengthens the management of large data sets in HDFS. HiveQL a MapReduce language used by Hive similar to SQL offers a declarative interface for managing stored data. HiveQL consists of DML operations, DDL operations and SQL operations and subqueries. Using HiveQL, Hive queries the data, generating MapReduce jobs based on the HIVEQL statements and implements them on a cluster. Unlike Pig, which is intended for data preparation, Hive is utilised for data presentation (Hoffer et al., 2016).

In addition to MapReduce, HDFS, Pig, and Hive, the Apache Hadoop project is comprised of many other components such as Avro, Core, HBase, Zookeeper, and Chukwa, all of which cooperate with each other to achieve its purpose. These components allow Hadoop to operate at a low-cost with efficiency, flexibility, and scalability while maintaining a high fault tolerance. While Hadoop allows for storage of enormous amounts of data sets, immense processing power and the skill to manage various parallel computing tasks, it is not the solution for all problems regarding data management but rather a tool that makes solving for real-world problems easier.

References

Hoffer, J., Venkataraman, R., & Topi, H. (2016). Modern Database Management (12th ed.). Edinburgh Gate, Harlow: Pearson Education, Ltd.

McNulty, E. (24 June 2014) *Hadoop: The Components You Need To Know*. Retrieved from https://dataconomy.com/2014/06/hadoop-components-need-know/

Pierson, L (n.d.). *What is Hadoop?* Retrieved from https://www.dummies.com/programming/big-data/data-science/what-is-hadoop/

Schneider, R. D. (2012). Hadoop For Dummies, Special Edition. Mississauga, ON: John Wiley & Sons Canada, Ltd.

Taylor, R. C. (21 December 2010) *An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics*. Retrieved from https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-S12-S1

White, T. (2009). Hadoop: The Definitive Guide (1st ed.). Sebastopol, CA: O'Reilly Media, Inc.