Volume 6, Issue 1, January 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey on Processing Heterogeneous Datasets Using Hadoop Based Approach

[1]Patkar Komal R., [2]Gundecha Pooja R., [3]Awari Shital A., [4]Prof. Pratik Kalamkar.
[1, 2, 3] Department of Computer Engineering, GHRCOEM SPPU University, Maharashtra, India
[4] Associate Professor, Computer Engineering, GHRCOEM, Maharashtra, India

Abstract— Database is the high usage of industries and scientific datasets. This datasets require high processing power which can't be offered by traditional databases. As the datasets quantity become larger the handling capacity also needs to get larger at the processing end. Although one can use Map Reduce paradigm to solve this problem using java based framework that is Hadoop but it cannot provide us with maximum functionality. As Hadoop is distributed framework which allows multiple users to perform single task parallel. Suppose there are one thousand records and from that records one want to find occurrences of particular name then these records are distributed among multiple users ,they perform operations individually and at the end final solutions are collected by Hadoop to represent the result. Though Hadoop provides large number of advantages, people avoid using this framework due to its complex working like map and reducing model. They prefer to use NoSQL databases to perform their tasks. Drawbacks can be overcome using Hadoop streaming techniques that allow users to define non-java executable for processing this datasets. This paper mostly concentrates on a tool i.e. MARISSA (Hadoop streamer) and Hadoop streaming and processing of Cassandra datasets. This paper also considers Hadoop and NoSQL database related concepts which will be useful to enhance the functionalities of existing tool.

Keywords— Hadoop streaming, NoSQL, Hadoop framework, MapReduce .

I. INTRODUCTION

With the increased amount of data collection taking place as a result of social media interaction, scientific experiments, and even e-commerce applications, the nature of data as we know it has been evolving. This dataset require high processing power which is not provided by a traditional database. Thus for processing such a large data we use a technique called "Hadoop streaming". Using this technique we will not only process the large data but also process the data which is non-java executable. Thus, a tool can be developed which can execute native as well as non-java executable which results in fast and efficient processing of Big data. Hadoop streaming contains data transformation and data processing steps. As figure shows Data is downloaded at Cassandra dataset, Cassandra allows to export data from target source in JSON format, this data is stored at Hadoop file system. In Data transformation stage Cassandra dataset in JSON format is processed by map function and formatted data is generated. In data processing stage query is processed and user gets output at Hadoop

II. LITERATURE SURVEY

Dede E, Sendir B, Kuzlu P, Weachock J, Govindaraju M, Ramakrishna L:

The progressive transition in the nature of both scientific and industrial datasets has been the driving force behind the development and research interests in the NoSQL data model. Loosely structured data poses a challenge to traditional data store systems, and when working with the NoSQL model, these systems are often considered impractical and expensive. As the quantity of unstructured data grows, so does the demand for a processing pipeline that is capable of seamlessly combining the NoSQL storage model and a "Big Data" processing platform such as Map Reduce. Although, Map Reduce is the paradigm of choice for data-intensive computing, Java-based frameworks such as Hadoop require users to write Map Reduce code in Java. Hadoop Streaming, on the other hand, allows users to define non-Java executable as map and reduce operations. Similarly, for legacy C/C++ applications and other non-Java executable, there is a need to allow NoSQL data stores access to the features of Hadoop Streaming.[1]

Jin-Ming Shih, Chih-Shan Liao, Ruay-Shiung Chang:

Map Reduce is a programming model developed by Google for processing and generating large data sets in distributed environments. Many real-world tasks can be implemented by two functions, map and reduce. Map Reduce plays a key role in Cloud Computing, since it decreases the complexity of the distributed programming and is easy to be developed on large clusters of common machines. Hadoop, an open-source project, is used to implement Google Map Reduce architecture. It is wildly used by many applications such as Face Book, Yahoo, Twitter, and so on. However, it is difficult to decouple an application into functions of map and reduce for common users. In this paper, we develop a web based graphic user interface for ordinary users to utilize Map Reduce without the real programming. Users only have to know how to specify their tasks in target-value-action tuples. Real examples are provided for demonstration.[2]

Dede E, Endir B, Kuzlu P, Hartog J, Govindaraju M:

In the last decade, the increased use and growth of social media, unconventional web technologies, and mobile applications, have all encouraged development
of a new breed of database models. NoSQL data stores target the unstructured data, which by nature is dynamic and a key focus area for "Big Data" research. New generation data can prove costly and unpractical to administer with SQL databases due to lack of structure, high scalability, and elasticity needs. NoSQL data stores such as Mongo DB and Cassandra provide a desirable platform for fast and efficient data queries. This leads to increased importance in areas such as cloud applications, e-commerce, social media, bioinformatics, and materials science. In an effort to combine the querying capabilities of conventional database systems and the processing power of the Map Reduce model, this paper presents a thorough evaluation of the Cassandra NoSQL database when used in conjunction with the Hadoop Map Reduce engine.[3]

Fadika Z, Ramakrishna, L, Gunter D, Canon R:

Map Reduce has since its inception been steadily gaining ground in various scientific disciplines ranging from space exploration to protein folding. The model poses a challenge for a wide range of current and legacy scientific applications for addressing their Big Data challenges. For example: Best known implementation of Map Reduce, Apache Hadoop, only offers native support for Java applications. While Hadoop streaming supports applications compiled in a variety of languages such as C, C++, Python and FORTRAN, streaming has shown to be a less efficient Map Reduce alternative in terms of performance, and effectiveness. Additionally, Hadoop streaming offers lesser options than its native counterpart, and as such offers less flexibility along with a limited array of features for scientific software.[4]

Fadika Z, Dede E, Hartog J, Govindaraju M

Map Reduce has gradually become the framework of choice for "big data". The Map Reduce model allows for efficient and swift processing of large scale data with a cluster of compute nodes. However, the efficiency here comes at a price. The performance of widely used Map Reduce implementations such as Hadoop suffers in heterogeneous and load-imbalanced clusters. We show the disparity in performance between homogeneous and heterogeneous clusters in this paper to be high. Subsequently, we present MARLA, a Map Reduce framework capable of performing well not only in homogeneous settings, but also when the cluster exhibits heterogeneous properties. We address the problems associated with existing Map Reduce implementations affecting cluster heterogeneity, and subsequently present through MARLA the components and trade-offs necessary for better Map Reduce performance in heterogeneous cluster and cloud environments.[4]

Fadika Z, Dede E, Govindaraju M, Ramakrishna L:

Map Reduce is increasingly becoming a popular framework, and a potent programming model. The most popular open source implementation of Map Reduce, Hadoop, is based on the Hadoop Distributed File System (HDFS). However, as HDFS is not POSIX compliant, it cannot be fully leveraged by applications running on a majority of existing HPC environments such as Teragrid and NERSC. These HPC environments typically support globally shared file systems such as NFS and GPFS. On such resourceful HPC infrastructures, the use of Hadoop not only creates compatibility issues, but also affects overall performance due to the added overhead of the HDFS. This paper not only presents a Map Reduce implementation directly suitable for HPC environments, but also exposes the design choices for better performance gains in those settings. By leveraging inherent distributed file systems' functions, and abstracting them away from its Map Reduce framework, MARIANE (Map Reduce Implementation Adapted for HPC Environments) not only allows for the use of the model in an expanding number of HPC environments, but also allows for better performance in such settings.[6]

Lei Zhang; Kaiping Li; Bin Wu:

SQL as a database language has been widely used in the modern society. Its function mainly focuses on the data processing, which can be used in data-mining. Due to the rapid growth of data, large-scale data processing is becoming a focal point of information techniques. Though we can still use SQL, but where to store the data and how to get the data efficiently, cost effectively, can be a tricky problem. Cloud computing emerges to solve the problem. It is mainly to deal with large-scale data processing. In this paper, we design a data-mining system which can directly deal with SQL processing based on Hadoop, a parallel store and computing platform. Then we will have a discussion about running time's efficiencies.[7]

Yang C, Yen C, Madden S.R:

Map Reduce is a programming model developed for processing and generating large dataset in disseminated environment. Many real world tasks can be implemented by using two functions i.e. Map and Reduce. Many applications such as Facebook, twitter, yahoo n so on uses Hadoop framework. It is however difficult to decouple application into function of Map and Reduce. Web-based graphic user interface for ordinary users to utilize Map Reduce without real programming.[8]

Brito A, Martin A, Knauth T:

This paper present Stream Map Reduce, a data processing approach that combines ideas from the popular Map Reduce paradigm and recent developments in Event Stream Processing. This approach adopted the simple and scalable

programming model of Map Reduce and added continuous, low-latency data processing capabilities previously found only in Event Stream Processing systems. This combination leads to a system that is efficient and scalable, but at the same time, simple from the user's point of view. For latency-critical applications, our system allows a hundredfold improvement in response time. Notwithstanding, when throughput is considered, our system offers a ten-fold per node throughput increase in comparison to Hadoop. As a result, it show that this approach addresses classes of applications that are not supported by any other existing system and that the Map Reduce paradigm is indeed suitable for scalable processing of real-time data streams.[9]

Guoxi Wang, Jianfeng Tang:

This paper reveals the secret of NoSQL. The CAP theorem, the BASE theorem and the Eventual Consistency theorem construct the foundation stone of NoSQL Cassandra is one kind of NoSQL databases; it is used by Twitter, Facebook and some other famous corporations. Taking it for example, this online trading system is based on Cassandra database. This approach contains first step of designing the contrast the relational model and Cassandra-based model of this system, then construct the key space, the column family and do some other configuration.

III. FUTURE SCOPE

Creating a framework which would work on multiple datasets will give ease to the end-user to execute non-java executable and process the data in less amount of time. Thus it will help user execute non-java executable instead of migrating the application to another platform. As the data to be processed is large, Hadoop framework can be used to implement such model. As the research work is related to the Cassandra dataset we can make use of heterogeneous NoSQL dataset. Enhancement in such a framework would improve the performance by adding the different functionality and making the framework easier for the user.

IV. CONCLUSIONS

Various techniques are introduced for efficient management of large amount of complex data. NoSQL datasets are used instead of traditional datasets for data management. Hadoop streaming is used for processing NoSQL datasets and avoiding the disadvantages of these databases. MARISSA tool is used to perform the various operations on Cassandra dataset and takes advantages of Hadoop in automatic sense. To perform the operations other than Cassandra, there is a need to develop the tool.

REFERENCES

[1] Dede E, Sendir B, Kuzlu P, Weachock J, Govindaraju M, Ramakrishna L "A Processing Pipeline for Cassandra Datasets Based on Hadoop Streaming", *Big Data (Big Data Congress)*, 2014 IEEE International Congress, 2014.

[2] Jin-Ming Shih, Chih-Shan Liao, Ruay-Shiung Chang, "Simplifying MapReduce Data Processing", *Utility and Cloud Computing (UCC)*, 2011 Fourth IEEE International Conference, 2011

[3] Dede E, Endir B, Kuzlu P, Hartog J, Govindaraju M, "An Evaluation of Cassandra for Hadoop*", Cloud Computing (CLOUD)*, 2013 IEEE Sixth International Conference, 2013.

[4] Fadika Z, Ramakrishna, L, Gunter D, Canon R, "MARISSA: MapReduce Implementation for Streaming Science Applications", *E-Science (e-Science)*, 2012 IEEE 8th International Conference, 2012.

[5] Fadika Z, Dede E, Hartog J, Govindaraju M, "MARLA: MapReduce for Heterogeneous Clusters", Cluster, *Cloud and Grid Computing (CCGrid*), 2012 12th IEEE/ACM International Symposium, 2012.

[6] Fadika Z, Dede E, Govindaraju M, Ramakrishna L, "MARIANE: Map Reduce Implementation Adapted for HPC Environments", *Grid Computing (GRID)*, 2011 12th IEEE/ACM International Conference, 2011.

[7] Lei Zhang; Kaiping Li; Bin Wu, "The research and design of SQL processing in a data-mining system based on MapReduce",*Cloud Computing and Intelligence Systems (CCIS)*, 2011 IEEE International Conference, 2011.

[8] Yang C, Yen C, Madden S.R, "Osprey: Implementing Map Reduce-style fault tolerance in a shared-nothing distributed database", *Data Engineering (ICDE)*, 2010 IEEE 26th International Conference, 2010.

[9] Brito A, Martin A, Knauth T, "Scalable and Low-Latency Data Processing with Stream Map Reduce", *Cloud Computing Technology and Science (CloudCom)*, 2011 IEEE Third International Conference, Dec 2011.

[10] Guoxi Wang, Jianfeng Tang, "The NoSQL Principles and Basic Application of Cassandra Model*", Computer Science Service System (CSSS*), 2012 International Conference, 2012.

