ISSN No: 2350-1278
Peer Reviewed & Refereed Journal
National Research Refereed Journal

PERFORMANCE EVALUATION OF HADOOP HBASE

Shiv Prasad

Research Scholar, MIT World Peace University, Maharashtra, India

ABSTRACT

With the unremitting advancement of internet and IT, tremendous growth of data has been observed. Data creation occurring at very fast pace, referred as big data, is a trending term these days. Addressing Big data is a challenging task that requires giant computational infrastructure for successful data processing and analysis. Hadoop, an open source distributed platform addresses big data challenges. Hadoop has many components including HDFS , Map Reduce, HBase and many more. But Hbase is pilfering much of the attention these days. HBase is Hadoop's column oriented NoSQL data base, sketched after Google's big table. This paper vindicates the efficiency of Hbase which is superior to the traditional relational database. Hbase is currently being used in many applications mainly adobe, facebook, twitter, yahoo.

***Keywords*-- Big Data, Hadoop, HDFS, HBase, NoSQL**

I. INTRODUCTION

In the last couple of decades huge advancement in the field of technology has been seen. This advancement has led to the huge increase in volume of data. Cloud Computing is a powerful technology which can perform very large scale computations. Most of the cloud applications used to process huge amount of data. Data creation occurring at very fast pace, referred as big data, is a trending term these days. As per International Data Corporation's (IDC) estimate, from the decade 2010 to 2020, digital data will grow 44 fold to 35 Zeta bytes (ZBs) per year. Big data is a data which is difficult to store, process and manage. Big Data is demanding new techniques to analyze and process the data.

Tackling these large-data problems require a distinct approach that is stronger than traditional models of storage and computing and provides good scalability and desired level of performance with insignificant or little cost.There are many projects have been developed as an alternative to traditional database system. Google's BigTable , Amazon's Dynamo, Apache's Cassandra, Hypertable, Apache's CouchDB, LinkedIn's Project Voldermort, MemcacheDB, MongoDB are just to name a few. The Apache Hadoop-based project - HBase is one such approach.

HBase is a distributed, fault-tolerant, highly scalable, no-SQL database, built on top of Hadoop Distributed File System (HDFS). Hbase is an Apache open source project and aims to provide a storage system similar to Bigtable in the Hadoop distributed computing environment.

Hadoop, a distributed processing framework addresses these demands. It is built across highly scalable clusters of commodity servers for processing, storing and managing data used in advanced applications. Hadoop has two main components-MapReduce and HDFS (Hadoop Distributed File System). MapReduce is a programming model for processing huge datasets .It was introduced in 2004 .MapReduce model breaks the big data into small portions called chunks and performs operations on those chunks of data. HDFS is a file system of Hadoop.

II. LITERATURE SURVEY

W.Peng,, Q. Yan, Y. Hua, [13] explain that with the rise of the Internet and the arrival of the trend of big data, NOSQL database is developing rapidly. The traditional relational database in solving large scale data is facing many problems. This paper introduces the concept of NOSQL database and data storage model. NOSQL (Not only SQL) refers to the non-relational, distributed, and does not provide the design mode of ACID database. At the same time, NOSQL database does not fixed pattern of data Taking the HBase database as an example, it describes the structure of system and data model of HBase, and it also demonstrates data query efficiency of the HBase database which is superior to the relational database.A comparison test of Hbase with traditional database has been performed and the result shows that NOSQL database performance in processing large data has obvious advantage over the traditional relational database.

Nance, C., Losser, T., Iype, R., Harmon, G [14] described that the relational database or RDBMS has been the dominant model for database management since it was developed by Edgar Codd in 1970. However, a new database model known as NoSQL is gaining attention in these days. NoSQL databases are non-relational data stores that are used in massively scaled web site scenarios, where traditional relational database comes out matter less, and the improved performance of retrieving relatively simple data sets matters most. The relational database model and the NoSQL database model both are each beneficial for specific applications. Depending on what kind of problem the organization is trying to solve, it will determine if a NoSQL database model should be used or relational database model should be used. Also, some organizations may use a hybrid mix of NoSQL databases and relational databases.

M.N. Vora[1] describes that today the world is flooded with digital data. It is becoming very difficult to store and analyze data efficiently and economically using conventional database management tools. HBase, as an open source substitute to traditional database management systems, is highly scalable, fault-tolerant, reliable, noSQL, distributed database that operates on a cluster of commodity machines to handle large data volumes. In this paper, an evaluation of hybrid architecture where HDFS contains the non-textual data like images and location of such data is stored in HBase. The paper aims at evaluating the performance of random reads and random writes of data storage location information to HBase and retrieving and storing data in HDFS respectively. A comparative study of HBase-HDFS architecture with MySQLHDFS architecture has also been performed. The results showed that as the load increases with number of users, HBase-HDFS performance was superior to MySQL-HDFS. Overall, the experimental results were in favor of the HBase approach.

J.Nandimath, A.Patil, E.Banerjee, P.Kakade, S.Vaidya [3]describe that the rapid growth of internet has expanded the amount of data generation in a drastic manner. Many organizations have shifted large data sets from centralized environment to distributed architecture. As the enterprises faced pitfalls in gathering large chunks of data. Also other issues like poor efficiency, dropped performance, elevated infrastructure cost were seen.The authors took an example of location based problem and solved it using Apache Hadoop.Thus operations were performed in optimal time, less user effort and more efficiently.

Rabl, T., Sadoghi, M., Jacobsen, H.A.[4] explains that the new wave of big data analytics imposes new challenges especially for the application performance monitoring systems. In this work, the authors present their experience and a brief of performance evaluation of six modern (open-source) data stores. They evaluated these systems with data and workloads that can be found in application performance monitoring, as well as, on-line advertisement, power monitoring, and many other use cases. The authors observed linear scalability for Cassandra, HBase, and Project Voldemort in most of the tests. Cassandra's throughput dominated in all the tests, however, its latency was in all tests abnormally high. Project Voldemort exhibits a stable latency that is much lower than Cassandra's latency. HBase had the least throughput of the three but exhibited a low write latency at the cost of a high read latency.

Huang.W et.al. [5] described that Mobile Internet describes a huge opportunity to transform a telecom operator to become a Big Data operator. To achieve this, some hurdles need to be overcome. China Unicom takes the lead to embrace the Mobile Internet Explosion, and builds a big data platform to solve the challenges of data acquisition,data analysis, and data value-added services. Compared to oracle database, the open source solution Hbase adopted by China Unicom offers us more advantages to optimize data storage, speed up database transactions, and achieve better performance.

A.Pradeepa, Dr. A.S. Thanamani [11] explains that many fields in data mining like knowledge discovery and data processing dealing with huge volume of data faces many challenges. Hadoop's mapreduce implementation has been successful to manage large scale data computation. Pradeepa, and Thanamani firstly introduced HDFS, a storage system for Hadooop.HDFS creates multiple replicas of data block and distribute them among different nodes of clusters for fast computation. Mapreduce has been taken into account by many scientific organizations and industries for bid data analysis. It has shown significantly better performance in data analysis and related fields. In this paper, the authors introduced a MapReduce based on computing rough set approximations in data mining. This algorithm was successfully designed. Rough set theory (RST) is a powerful mathematical tool to describe the dependencies between attributes, estimates the significance of attributes, and derive decision rules. Many rough sets based-approaches have been successfully applied in machine learning and data mining. It is certainly an inexpensive and simple, yet powerful solution Parallel Processing.

Conclusion and Future Work

The advancement of internet and IT has led to the huge increase in volume of data. Data creation occurring at very fast pace, referred as big data, is a trending term these days. Cloud computing is providing the utility through Hadoop to manage the Big Data. Tackling these big data problems require a distinct approach that is stronger than traditional methods of storage and computing and provides good scalability and desired level of performance with very little cost. This led to the development of scalable, distributed, non-relational, No-SQL, column-oriented databases. Hbase, a sub project of hadoop is a column –oriented and NoSQL database. In this paper the various database such as oracle, MySQL, Cassandra, couchDB, mongoDb has been compared with HBase. The performance of HBase is seen to be much more better than MySQL and oracle. It has also been observed that throughput of Cassandra dominates Hbase. We believe that HBase is capable of dealing with large datasets and showed superior performance and scalability. In the future, the efforts must be made to improve the throughput of HBase in terms of latency .

REFERENCES

1. M.N. Vora "Hadoop HBase for large scale data" , International Conference on Computer Science and Network Technology, pp 601-605, 2011.
2. C.W.Lee,K.Hsieh, S.Hsieh, H.Hsiao "A dynamic data placement strategy for hadoop in heterogeneous environments" , Big Data Research ,Vol.1, 2014.

3. J.Nandimath, A.Patil, E.Banerjee, P.Kakade, S.Vaidya "Big Data Analysis Using Apache Hadoop",Information Reuse and Integartion ,IEEE 14th International Conference, pp.700-703,August 2013.

4. Rabl, T., Sadoghi, M., Jacobsen, H.A. 2012 "Solving Big Data Challenges for Enterprise Application Performance Management" , Proceedings of the VLDB Endowment, Istanbul, Turkey, Vol. 5, pp 1724-1735.

5. Huang, W., Chen, Z., Dong, W., Li, H., Cao, B., Cao, J.,2014 "Mobile Internet Big Data Platform in China Unicom" , Tsinghua Science and Technology, pp.95-101.

6. Xuan Wang, "Clustering in the cloud:clustering algorithms to hadoop map/reduce framework" Published by Technical Reports-Computer Science by Texas State University,2010.

7. L.Wang, J.Tao, R.Ranjan, H.Marten A.Streit, J.Chen, D.Chen"G-Hadoop: MapReduce across distributed data centers for data-intensive computing" , Parallel and Distributed Processing Symposium Workshops and Phd Forum ,2012 IEEE 26th International, pp.2004-2011.

8. B.T. Rao, N.V.Sridevi, V.K.Reddy, L.S.S.Reddy, "Performance Issues of Heterogeneous Hadoop Clusters in Cloud Computing", Global Journal of Computer Science and Technology ,Vol.11, Issue 8, May 2011.

9. J. Leverich, C. Kozyrakis "On the energy (in) efficiency of hadoop clusters", ACM SIGOPS Operating Systems review,Vol.44,pp.61-65, 2010.

10. M. Zaharia, A.Konwinski, A.D. Joseph, R.Katz, I.Stoica, "Improving Map Reduce Performance in Heterogeneous Environments", in:Proc 8th USENIX Symposium on Operating System Design and Implementation,OSDI,2008,San Diego,USA, pp.29-42.

11. A.Pradeepa, Dr. A.S. Thanamani " Hadoop file system and fundamental concept of mapreduce interior and closure rough set approximations**"**, International Journal of Advanced Research in Computer and Communication Engineering ,Vol. 2, Issue 10, October 2013.

12. K. Shvachko, H.Kuang, S. Radia, R. Chansler, "The Hadoop Distributed File System",MSST '10 Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST),pp.1-10.

13. W.Peng,,Q. Yan, Y. Hua,.,2014. "Analysis and Study on the Performance of Query based on NoSQL Database", Computer modelling & new technologies , pp.153-159 .

14. Nance, C., Losser, T., Iype, R., Harmon, G., 2013, "NoSQL vs RDBMS - Why There is Room for Both", Proceedings of

15. the Southern Association for Information Systems Conference, Savannah, GA, USA , pp.111-116.