ISSN(Online) :2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

A Survey on Network-Levitated Merge for Hadoop Acceleration

Madhav D.Ingle[1], Vaibhav Pathare[2], Ankita Nirmal[3], Diksha Wadkar[4]

Assistant Professor, Dept. of Computer Engineering, JSCOE, Hadapsar, Pune, India[1]

UG Student, Dept. of Computer Engineering, JSCOE, Hadapsar, Pune, India[2,3,4]

ABSTRACT: Big data is a collection of large amount of data which cannot be processed using traditional computing techniques. Big data contains data produced from Black box data, social media data, transport data. It also includes huge volume, high velocity and extensible data of the following types as structured, semi structured and unstructured data. Big data faces some challenges as storage, sharing, analysis, presentation, etc. Hadoop is proposed to scale up from single server to thousands of servers, each does computations and storage .HDFS stores very large data sets and frequent access. HDFS also develops applications for parallel processing. Map-Reduce is a programing model for generating and processing large amount of data. Users uses map function to process key/value pair to generate intermediate key/value and reduce function is used to merge all values related to same intermediate key.This paper focuses on the different algorithms and techniques to handle the big data. This paper is very useful for new researchers for study various techniques which are covered in this paper.

KEYWORDS: Big Data, Hadoop, HDFS, MapReduce, Hadoop acceleration, Cloud Computing**.**

I. INTRODUCTION

Big data is a word used to describe a massive volume of both structured, semi structured and unstructured data that is so large. It is difficult to process using traditional database and software techniques. Hadoop is an Apache open source framework written in java that is used to store and process big data in distributed environment across clusters of computers. HDFS(storage) and MapReduce(processing) are the two component of Apache Hadoop. HDFS is used tostore very large amount of data and provides easier access. Hadoop provides a command interface to interact with HDFS. It provides file permissions and authentication. Following figure shows the architecture of HDFS.

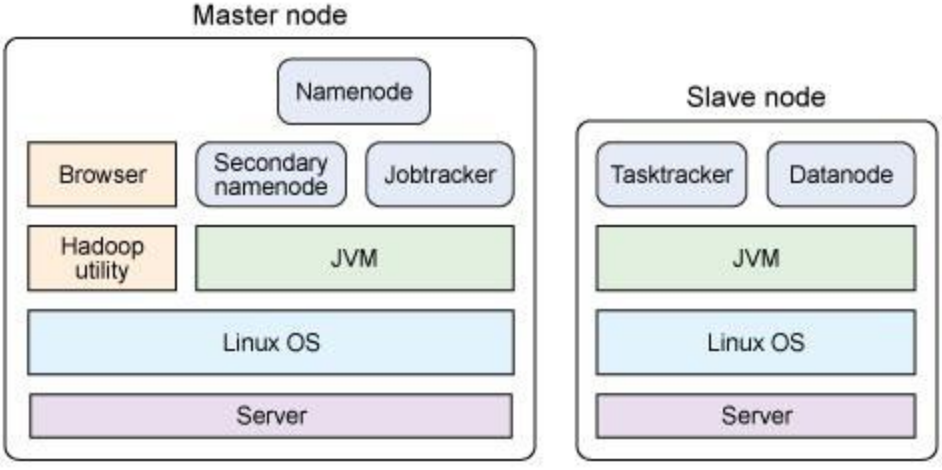

Figure1:- Architecture Of HDFS

HDFS follows the master-slave architecture and it has the following elements.

Master node:
1.Namenode

Slave node:
1.Datanode

ISSN(Online) :2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

2.Jobtracker 2.Tasktracker

Namonode: It is Commodity hardware that contains the Linux operating system and namenode software. It executes file system namespace operations like opening, closing, and renaming files and directories. Namenode regulates the clients access to files. Datanode: The data node is a commodity hardware having Linux operating system and data node software. It is used to store the data in the Hadoop File Systems. It performs read-write operations on the file systems, as per client request. They also perform operations such as block creation, deletion, and replication as per instruction of name node. Tasktracker: It accepts the tasks like Map, Reduce, Shuffle from jobtracker. After finishing the tasks it notifies to jobtracker [11].

Secondary Namenode: When the primary Namenode (i.eNamenode) fails then secondary namenode comes and perform the action. Secondary Namenode is continuously connected with primary Namenode and take snapshots of Namenode memory structures. These are used to recover the failed Namenode and recent memory structure. Job Tracker: It is used to keep the track of which Map-Reduce jobs are executing, schedules Maps, Reduces operation tasks to specific machine, monitors the success and failures of these tasks[12].

MapReduce: MapReduce is parallel programming model and processing technique for extracting and analysing the data from big data sets. It contains two tasks, Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs).Input data is in taken form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data. Reduce tasks takes the output from a map as input and combines those data tuples into smaller set of tuples. Reduce stage is the combination of the Shuffle stage and the Reduce stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.

Algorithms and Techniques

I. Camdoop

To reduce network traffic and for better performance of map reduce programming model new technique is proposed called Camdoop[1].

II. Online Aggregation, HOP(Hadoop Online Prototype) and Pipeline

To simplify fault tolerance data needs to be pipeline between operators where intermediate keys are pipelined with the help of three techniques are 1.Online Aggregation 2.HOP(Hadoop Online Prototype) and 3.Pipeline[3].

III. Automation Setting of Tuning Parameter

Automation setting of tuning parameter of map-reduce programming is used to provide ad-hoc map-reduce program on large amount of data[7].

IV. Phoenix Technology

To reduce the delay in query processing one pass analytic technique is used[5].Phoenix technology is used to increase the performance of multicore systems and for error recovery in parallel programming. This technology automatically manages thread creation, data partitioning and fault tolerance[8].

V. MapReduce is a Programming Model

MapReduce is a programming model used to handle a distributed data and to handle different computations on large data sets. Where map function is used for processing set of key/value pairs and to generate set intermediate key/value pairs and reduce function will then merge all the values associated with the same key. To achieve scalability flexibility storage of data like structured and unstructured is made independent[10].

II. LITERATURE SURVEY

In today's technical era handling the big data is major issue. The many Authors are working on the different techniques and algorithm for handling the big data. The author Paolo Costa, et al. built Camdoop technique. Social media like Facebook ,Google, Microsoft etc. generates the huge amount of data which is responsible for high network

ISSN(Online) :2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

traffic and difficult to support using traditional system. Thus, Camdoop technique is used to reduce network traffic and it also provides the high performance[1].

To simplify fault tolerance in map-reduce programming model to actualize the output of each map and reduce the task before it used. The authors Tyson Condie, et .al proposed modified map-reduce architecture data to be pipeline between operator where intermediate data is pipelined. There are three techniques introduced by author Online Aggregation, HOP(Hadoop Online Prototype),Pipeline.1.Online Aggregation:-As soon as mapper produces the data which has been immediately processed by reducer. Thus task can be completed in efficient time.2.HOP(Hadoop Online Prototype):-It is used to support continuous queries that map-reduce job run continuously, accept the new data when it arrives and analyzes it. 3.Pipeline:-It increases the chances of parallelism, it improves utilization and reduces the response time[3].

Current map-reduce system requires dataset to be loaded into cluster before running queries because of that there is issue arise like high delay to start query processing. Thus authors Edward Mazur, Boduo Li introduced one pass analytic technique. This technique is used to reduce the delay in query processing[5]. Map-reduce based system is slower than Parallel Database system in performing variety of tasks. Because there is performance gap between the Map-reduce and Parallel Database system. For achieving scalability and flexibility in Map-reduce system authors Dawei Jiang, et .al designed a system such that it is independent of storage system and it analyzes various kind of data like structured and unstructured. And it improves the performance of map-reduce system[6].

In traditional system to run single program in map-reduce framework the system administrator need to set number of running parameter. The author Shivnath Babu, make a case for technique to automate setting of tuning parameter of map-reduce programming which used to provide ad-hoc map-reduce program on large amount of data[7].Parallel programming techniques like message passing, shared memory, it is difficult to write correct and scalable parallel code .The author Colby Ranger, et .al implemented phoenix technology it automatically manages thread creation, data partitioning, and fault tolerance across processor nodes. Phoenix technology used to increase the performance of multicore system and recover the errors[8].

Authors Jiuxin, et .alproposed design of MPI over infini-band which acquires the benefit of RDMA to large messages as well as small and control messages which improves the better scalability by combining RDMA operations like send or receive. RDMA based design used to reduce latency by 24% and increases the bandwidth by over 104% also reduce the overhead up-to 22%[9].In past decades author implemented some special purpose computation that process large amount of raw data like crawled documents, web request logs etc. for computing various kind of derived data but this computations are straightforward which cannot handle issues like how to parallelize the computation ,distribute the data and handle computations with simple computation on large amount of data. Jeffery Dean, et .al who propose a new approach which used for map reduce programming model where map function process key value pairs and generate intermediate key value pairs and reduce function merges all values associated with intermediate key[10]

III. CONCLUSION AND FUTURE WORK

Big data is a collection of large amount of data which cannot be processed by using existing techniques. Hadoop is a framework which is used to process the big data. It provides the storage and processing with the help of HDFS(Hadoop Distributed File System) and Map-Reduce programming model. Many researchers implement the different algorithms and techniques for processing large amount of data. In this paper we covered number of techniques and algorithms implemented by researchers for processing the big data. The efficiency and throughput is achieved by using different algorithms and techniques. We come to the conclusion that this paper is very beneficial for new researchers for the innovation of new techniques and algorithms for processing of big data. In future work we invent the algorithms or techniques which is very useful to process the big data using hadoop and it also improves the efficiency existing hadoop.

REFERENCES

[1]P. Costa, A. Donnelly, A. Rowstron, and G. O'Shea, "Camdoop: Exploiting in-Network Aggregation for Big Data Applications," Proc. Ninth USENIX Conf. Networked Systems Design and Implementation (NSDI '12), p. 3, 2012.

[2]X. Que, Y. Wang, C. Xu, and W. Yu, "Hierarchical Merge for Scalable MapReduce," Proc. Workshop Management of Big Data Systems (MBDS '12), pp. 1-6, 2012.

ISSN(Online) :2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

[3] T. Condie, N. Conway, P. Alvaro, J.M. Hellerstein, K. Elmeleegy,and R. Sears, "MapReduce Online," Proc. Seventh USENIX Symp.Networked Systems Design and Implementation (NSDI), pp. 312-328,Apr. 2010.
[4] Vaibhav Dhore and Sonali R. Jagtap, A Survey on Hierarchical Merge for Hadoop-A, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438
[5] B. Li, E. Mazur, Y. Diao, A. McGregor, and P. Shenoy, "A Platform for Scalable One-Pass Analytics Using MapReduce," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '11), pp. 985- 996, 2011.
[6] D. Jiang, B.C. Ooi, L. Shi, and S. Wu, "The Performance of MapReduce: An In-Depth Study," Proc. VLDB Endowment, vol. 3,no. 1, pp. 472-483, 2010.
[7] S. Babu, "Towards Automatic Optimization of MapReduce Programs," Proc. First ACM Symp.Cloud Computing (SoCC '10), pp. 137-142, 2010.
[8]C. Ranger, R. Raghuraman, A. Penmetsa, G.R. Bradski, and C. Kozyrakis, "Evaluating MapReduce for Multi-Core and Multiprocessor Systems," Proc. IEEE 13th Int'l Symp. High Performance Computer Architecture (HPCA '07), pp. 13-24, 2007.
[9] J. Liu, J. Wu, and D.K. Panda, "High Performance RDMA-Based MPI Implementation over InfiniBand," Int'l J. Parallel Programming, vol. 32, pp. 167-198, 2004.
[10] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Proc. Sixth Symp. Operating System Design and Implementation (OSDI '04), pp. 137-150, Dec. 2004.
[11] http://www.tutorialspoint.com/hadoop/index.htm
[12] Apache Hadoop Project, http://hadoop.apache.org/, 2013.