A Survey on Big Data Processing in Large Scale Computing System

N.Boopal,S.Gunasekaran,S.Karthiban

Abstract—Big data defines to data that cannot be handled in traditional system and the data capacity is too large to handle. Data analysis which has huge set of data in the cluster nodes of cloud system and the data can be processed while analysing. Map Reduce is the key to access the big data environment with good scalability. The cost of the server is leading to be huge while compared with the total cost during analysis of data. Heterogeneous workloads are major problems in data centres, the hybrid structure of cloud system that establishes HDFS and parallel database process for processing and indexing. PSO are constructed to proceed parallel data processing and workloads on each node. In this paper, we introduce data classification mechanism to task partitioning, a system that efficiently process complex data analysis tasks by improving Map Reduce runtime framework on large clusters.

Index Terms—Big Data, Cloud computing, MapReduce, Hadoop, Parallel Processing, Data Scheduling, HDFS..

1 INTRODUCTION

The data that is too large and cannot be accessed on small system are considered to be big data. It refers to traditional enterprise data, machine generated data, sensor data and social data. Traditional enterprise data includes customer information from CRM and web transactions. Machine generated/sensor data include smart meters and manufacturing sensors. Social data includes data on social platform like twitter, facebook, etc., Big data is defined by four key characteristics and they are volume, velocity, value and variety. Volume is based on generating larger data sets which are larger than non-traditional data. Velocity is based on social media streams. Variety is based defined data formats with data schema and change. Value of each data varies significantly

1.1 Cloud Introduction

Cloud Computing is a technology that maintains data and applications by using the internet and central remote servers[10]. More inefficient computing is given by this technology by the centralization of storage, memory, processing and bandwidth. Cloud computing allows consumers and businesses to access their personal files around anywhere with the help of internet access and also provides applications without installation.

1.2 Types of Cloud

Public cloud: If the computing infrastructure is being hosted by the cloud vendor at his own premises, then that cloud becomes a public cloud. This public cloud can be shared between many organizations, where the customer will have no visibility and control over the computing infrastructure being hosted.

Private cloud: Private Cloud are more expensive and they seem to be more secure when compared to public clouds, since because the computing infrastructure is dedicated to a single particular organization and its not being shared among many. Some experts consider that private clouds are not real examples of cloud computing, because it gives less access to people when it is a private cloud.

Hybrid cloud: Hybrid cloud is relatively called Cloud Bursting. Organizations hosts certain critical applications which needs more security on the private cloud and the applications which can be shared on the public cloud. This type of usage of both the public and private cloud is called hybrid cloud.

1.3 Hadoop Introduction

Hadoop is a large scale distributed batch processing infrastructure and it can be used on a single machine. Large amount of work can be efficiently distributed across few set of machines using Hadoop. Hadoop is developed to process Web Scale on the data which ranges in Terabytes and Pentabytes. HDFS processes each data inputs by breaking up and sending those data to several clusters for processing.

2 LITERATURE SURVEY

2.1 Survey on data analysis in a distributed environment

According to [1] [2] the scheduling of data for distributed system. In [1] address the problem of scheduling on cluster with the tasks and in each node the application data is stored. Scheduling tasks to the nodes on a cluster is crucial for performance because the data in the nodes are close to the tasks. Data intensive computing benefits the resource sharing model with semi static resource allocation. Sharing of each resource in semi static allocation and allocation on cluster implements data on the nodes. Scheduling problem is mapped to a graph structure. The process evaluates the Quincy framework for the implementation on hundreds of computer system by implementing Quincy against queue based algorithm. [2] Scheduling tasks on a parallel system is challenging on heterogeneous cluster. The problem of scheduling multiple applications made of collections of independent and identical tasks, on a heterogeneous master-worker platform. The applications are submitted through internet access which means the knowledge of workload distribution in unknown during execution. Optimal algorithm is used for the offline version of the problem.

2.2 Cluster based scalable network

The survey paper [3] on fundamental requirements for scalable network services: incremental scalability and overflow growth provisioning, availability through fault masking, and cost effectiveness. We argue that clusters of commodity workstations interconnected by a high-speed SAN are exceptionally well-suited to meeting these challenges for Internet-server workloads, provided the software infrastructure for managing partial failures and administering a large cluster does not have to be reinvented for each new service. To this end, we propose a general, layered architecture for building cluster-based scalable network services that encapsulates the above requirements for reuse, and a service-programming model based on compatible workers that perform transformation, aggregation, caching, and customization (TACC) of Internet content.

2.3 Virtualization for Dynamic Computational Domains

The literature [4], a large organization such as a university, commonly supplies computational power through multiple independently administrating computational domains. Each computational domain faces the conflict between dynamic workload and static capacity. This is clearly inefficient at times when some clusters have idle nodes while others experience excessive workload. An opportunity arises to resolve this conflict by dynamically adapting the capacity of clusters by borrowing idle machines of peer domains. In this literature, we present the design, implementation, and evaluation of Vio-Cluster, a virtualization based computational resource sharing platform. Through machine and network virtualization, Vio-Cluster enables virtual computational domains that safely trade machines between them without infringing on the autonomy of either domain. Our performance evaluation results show that dynamic machine trading between virtual domains increases their resource utilization and decreases their job wait times.

2.4 Cost aware resource provisioning

According to [5], Heterogeneous workloads are considered to be the biggest problem that is not acting with the resource provisioning solutions. Resource provisioning decreases the peak resource workloads in cooperative process. Parallel processing on each job enables cooperative solution and the server costs are saved. Phoenix cloud enables resource provisioning of cooperative process that works on MapReduce jobs. The cloud services like EC2, Google Apps, Amazon Azure posses high cost of computing and storage while transferring data from client to data centres. System capacity, power and cooling infrastructure are the problems in data centres. Cooperative resource provisioning develops the resource provider on heterogeneous workloads by enabling phoenix cloud.

2.5 Prioritizing iterative computation

Iterative computation possess mining algorithm which like K-means algorithm adopted for large amount of data. MapReduce algorithm is the way to handle data on the cloud system. Hadoop, Hive, Pig, Pregel are proposed for handling big data. In [6] design of PrIter is for handling distributed framework and prioritizing execution. PrIter increases the speed of implementation of MapReduce on hadoop, PrIter supports iterative process on MapReduce and in iMapReduce process takes the output of reduce and map input to the next iteration. Prioritizing iteration defines the value of each node and maintained for iteration with selected subset based priority value.

2.6 MapReduce model for data analysis

According to [7], MapReduce model process the data with Map() and Reduce() function for analysis. Filtering aggregation on homogeneous data is not convenient so that the filtering- join-aggregation model is an extension of MapReduce model and HDFS possess Join() function on filtering aggregation model.

Map-Join-Reduce shuffle input data set to reducer and the function applies all the data sets in a parallel manner. Each data sets are combined and produce aggregation result for frequent checkpoints. Map-Join-Reduce job to join the data with hybrid processing strategy in parallel database.

On basis of [8] , the scheduling increases the cluster utilization and minimize the completion time and resources are scheduled to Map and Reduce tasks. The abstraction on each tasks which schedule the sub optimal job to improve the make span by processing the jobs in single order. Based on Johnson scheduling five strategies for accuracy and performance.

- Min strategy
- Max strategy
- Min sim
- Max sim
- Balanced pool

Balanced pool technique partitions cluster into two pools, each pool jobs are scheduled and both pools executes the job based on sub cluster jobs and report the overall make span.

3 EXPERIMENTAL RESULTS

The process executes the hadoop cluster of the heterogeneous resources based on the metadata information of the workloads. The workloads based on the different strategies and conditions based on optimization principles of the system. Each execution on the MapReduce model simulates the heterogeneous workloads that saves the server cost[5] with phoenix cloud. In [7] Map-Join-Reduce model improves the MapReduce runtime system for data analytical tasks on large cluster. Join() combines multiple data set to aggregate computation. Phoenix cloud analysis the data with the rate of 568 seconds and which is comparatively high than non cooperative resources of 310 seconds.

PrIter on each massive data set which consumes time that reduces the execution time of the selected subset data. Prioritizing execution of iterative computation accelerates iterative algorithm by consuming less time by enabling iteration

- *S.Karthiban is currently pursuing masters degree program in Software Engineering at CIET,India, PH-7708579973. E-mail:karthee.karthee@gmail.com*
- *N.Boopal is currently working as an Assistant Professor in Computer Science Engineering at CIET, PH-7402107222. E-mail:c.sathyachandran92@gmail.com*
(*This information is optional; change it according to your need.*)

for selected subset..

4 Conclusion

This paper based on hadoop framework by adopting MapReduce model in a heterogeneous environment by executing each data by performing data repartitioning and virtual machine mapping to reduce the completion time. Establishing the distributed resource clusters through HDFS. Parallel database in HDFS for processing and indexing, data analysis structure is constructed through PSO which to incorporate parallel database for resource utilization and performance can analysis through make span and response time of each node.

Acknowledgment

The authors wish to thank the institution for their support and encouragement.

References

[1] M. Isard et al., “Quincy: Fair Scheduling for Distributed Computing Clusters,” Proc. ACM SIGOPS 22nd Symp. Operating Systems Principles (SOSP ’09), pp. 261-276, 2009.

[2] A. Benoit et al., “Scheduling Concurrent Bag-of-Tasks Applications on Heterogeneous Platforms,” IEEE Trans. Computers, vol. 59, no. 2, pp. 202-217, Feb. 2010.

[3] S. Chase et al., “Managing Energy and Server Resources in Hosting Centers,” Proc. 18th ACM Symp. Operating Systems Principles (SOSP ’01), pp. 103-116, 2001.

[4] P. Ruth et al., “VioCluster: Virtualization for Dynamic Computational Domains,” Proc. IEEE Int’l Conf. Cluster Computing (Cluster ’05), pp. 1-10, 2005.

[5] Jianfeng Zhan et al., “Cost-Aware Cooperative Resource Provisioning for Heterogeneous Workloads in Data Centers,” Proc. IEEE Transaction on computers (Volume 62), Nov 2013.

[6] Yanfeng Zhang et al., “PrIter: A Distributed Framework for Prioritizing Iterative Computations,” Proc. IEEE Transaction on parallel and distributed system (Volume 24), Sep 2013.

[7] Dawei Jiang et al., “MAP-JOIN-REDUCE: Toward Scalable and Efficient Data Analysis on Large Clusters,” Proc. IEEE Transaction on knowledge and data engineering (Volume 23), Sep 2011.

[8] Abhishek Verma et al., “Orchestrating an Ensemble of MapReduce Jobs for Minimizing Their Makespan,” Proc. IEEE Transaction on dependable and secure computing (Volume 10), Sep/Oct 2013.

[9] An oracle white paper., “ Big Data for the Enterprise,”, June 2013.

[10] Mohiuddin Ahmed et al., “An Advanced Survey on Cloud Computing and State of the art Research Issues,” IJCSI (Volume 9), Jan 2012.

[11] Wikipedia about Apache Hadoop.

[12] Wikipedia about MapReduce model.