IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

A Novel Approach to Enhance the Large Scale Computing System for Handling Distributed Data using Hadoop Framework

Ms. J. Monica Msc [1], Mr. P.Ponsekar Msc, MPhil., [2]

M.Phil Scholar, Department of Computer Science, Kovai kalaimagal College of Arts and Science, Coimbatore, India[1]

Assistant Professor, Dept. of Computer Science, Kovai kalaimagal College of Arts and Science , Coimbatore, India[2].

Abstract: Data analysis is an important functionality in Big Data processing and computing which allows a huge amount of data to be processed over very large clusters. Map Reduce is recognized as a popular way to handle data in the data intensive environment due to its excellent scalability and good fault tolerance features. Cost analysis shows that the user server cost still dominates the total cost of high scale data centers or cloud systems. Heterogeneous workloads are the problems in large scale data centers. Data analysis on huge datasets is processed by proposing an analysis framework for the task and resource provisioning which reduce the peak resource. Building indexes in data centres for analyzing the data. The Map reduce model uses effective processing to process the datasets in hybrid structure. It reduces the server cost on data centre. This process establishes hadoop distributed file system and parallel database for processing and indexing. An analysis framework is constructed through PSO, which incorporates parallel database and handles workloads.

Keywords: PSO, Map Reduce, Hadoop Framework, Clusters.

I. INTRODUCTION

Big Data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process them using traditional data processing applications. By evolving the current enterprise architecture, you can leverage the proven reliability, flexibility and performance of the Oracle systems to address the big data requirements. Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set.

Accuracy in big data may lead to more confident decision making. And better decisions can mean greater operational efficiency, cost reduction and reduced risk. The challenges include analysis, capture, search, sharing, storage, transfer, visualization, and privacy violations. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data.

Oracle is the first vendor to offer a complete and integrated solution to address the full spectrum of enterprise big data requirements. Oracle's big data strategy is centred on the idea that you can evolve the current enterprise data architecture to incorporate big data and deliver business value. Cloud computing is a technology that uses the internet and central remote servers to maintain data and applications.

II.CHARACTERISTICS OF BIG DATA

Big Data includes Call Detail Records, web logs, smart meters, manufacturing sensors, equipment logs, trading systems data. It includes customer feedback streams, micro-blogging sites like Twitter, and social media platforms like Face book. Machine-generated data is produced in much larger quantities than non-traditional data. For instance, a single jet engine can generate 10TB of data in 30 minutes. Social media data streams are not as massive as machine-generated data produce a large influx of opinions and relationships valuable to customer relationship management. Even at 140 characters per tweet, the high velocity of Twitter data ensures large volumes (over 8 TB per day**)**.Traditional data formats tend to be relatively all defined by a data schema and change slowly. In contrast, non-traditional data formats exhibit a dizzying rate of change. The economic value of different data varies significantly. Typically there is good information hidden amongst a larger body of non-traditional data, the challenge is identifying what is valuable and then transforming and extracting that data for analysis.

III.CHALLENGES AT LARGE SCALE

Performing large-scale computation is difficult. To work with this volume of data requires distributing parts of the problem to multiple machines to handle in parallel. Whenever multiple machines are used in cooperation with one another, the probability of failures rises. In a single-machine environment, failure is not something that program designers explicitly worry about very often: if the machine has crashed, then there is no way for the program to recover anyway.

DOI 10.17148/IJARCCE.2015.41221

In a distributed environment the partial failures are an expected and common occurrence. Networks can experience partial or total failure if switches and routers break down. Data may not arrive at a particular point in time due to unexpected network congestion. Individual compute nodes may overheat, crash, experience hard drive failures, or run out of memory or disk space. Data may be corrupted, or maliciously or improperly transmitted. Multiple implementations or versions of client software may speak slightly different protocols from one another. Clocks may become desynchronized, lock files may not be released, parties involved in distributed atomic transactions may lose their network connections part-way through, etc. In each of these cases, the rest of the distributed system should be able to recover from the component failure or transient error condition and continue to make progress. Of course, providing such resilience is a major software engineering challenge.

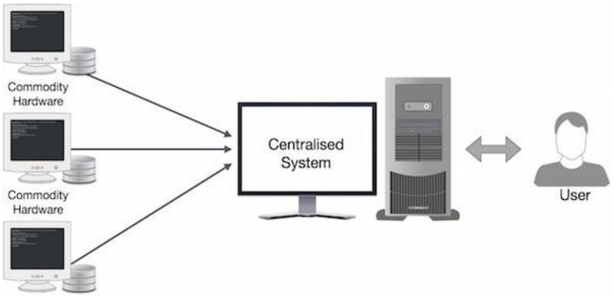

Fig 1:Distributed file system data flow

Hadoop is designed to efficiently process large volumes of information by connecting many commodity computers together to work in parallel.

IV.CLUSTER BASED SCALABLE NETWORK SERVICES

In the literature, three fundamental requirements for scalable network services: incremental scalability and overflow growth provisioning, 24x7 availability through fault masking, and cost effectiveness. A argue that clusters of commodity workstations interconnected by a high-speed SAN are exceptionally all-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, a 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 compassable workers that perform transformation, aggregation, caching, and customization (TACC) of Internet content. For both performance and implementation simplicity, the architecture and TACC programming model exploit BASE, a weaker than ACID data semantics that results from trading consistency for availability and relying on soft state for robustness in failure management. Our architecture can be used as an “off the shelf” infrastructural platform for creating new network services, allowing authors to focus on the “content” of the service (by composing TACC building blocks) rather than its implementation. A discuss two real implementations of services based on this architecture: TranSend, a Ab distillation proxy deployed to the HotBot, the commercial implementation of the search engine. A present detailed measurements of TranSend’s performance based on substantial client traces, as all as anecdotal evidence from the TranSend and HotBot experience, to support the claims made for the architecture.

V.CLUSTERING RELATED SOLUTION

Clustering and Classification is important issue for big data which relates to Volume .Scaling and Scheduling has lead to Velocity issue in Big data. Concept Drift and Feature evolution has lead to variety issues to big data. Resource Provisioning dominates the total cost on the high Scale data centers.

Clustering the large data is one of the essential tasks in mining unstructured data, posts significant challenges on both modeling- similarity between structured objects and unstructured objects through developing efficient computational methods. The previous methods in literatures using traditional partitioning clustering methods like k-means and density-based clustering.

Such methods cannot handle unstructured objects that are geometrically indistinguishable, such as products or database from healthcare, Finance company etc and very different variances in customer ratings is also undetermined. Surprisingly, probability distributions, which are essential characteristics of unstructured data objects, a systematically model uncertain objects in both continuous and discrete domains, where an uncertain object is modeled as a continuous and discrete random variable, respectively. The Job Tracker knows which node contains the data, and which other machines are nearby. If the work cannot be hosted on the actual node where the data resides, priority is given to nodes in the same rack.

An Objective of proposed System is to establish an efficient data analysis framework for handling the large data Drift in the workloads from enterprise through the exploration of data handling mechanism like parallel database. Recent cost analysis shows that the server cost still dominates the total cost of high-scale data centers or cloud systems. Proposal has a hybrid technique with map reduce and parallel database for resource provisioning problem heterogeneous workloads are a fact of life in large scale data centers, and current resource provisioning solutions do not act upon this heterogeneity.

The contributions are threefold: first, to propose a Efficient framework for task and resource provisioning solution, and take advantage of differences of heterogeneous workloads so as to decrease their peak resources consumption under competitive conditions; second, for four typical- heterogeneous workloads: parallel database is utilized for parallel operation like evaluating the jobs, loading job and building the indexes for statistical analysis in the data centers, third a use map reduce programming model for effective processing and utilizing the hybrid structure to process the dataset with less utilization time and high throughput.

DOI 10.17148/IJARCCE.2015.41221

VI.DISTRIBUTED RESOURCE CLUSTERS THROUGH HADOOP FILE SYSTEM

HDFS is designed to store a very large amount of information This requires spreading the data across a large number of machines. It also supports much larger file sizes than NFS.HDFS should store data reliably. If individual machines in the cluster malfunction, data should still be available. HDFS should provide fast, scalable access to this information. In nearly all cases, a Map Reduce job will either encounter a bottleneck reading data from disk or from the network (known as an IO-bound job) or in processing data (CPU-bound). An example of an IO-bound job is sorting, which requires very little processing and a lot of reading and writing to disk.

An example of a CPU-bound job is classification, where some input data is processed in very complex ways to determine ontology. Several more examples of IO-bound workloads are Indexing, Grouping, Data importing and exporting and Data movement and transformation. Several more examples of CPU-bound workloads are Clustering/Classification, Complex text mining, Natural-language processing and Feature extraction.

Because Cloudera's customers need to thoroughly understand their workloads in order to fully optimize Hadoop hardware, a classic chicken-and-egg problem ensues. Most teams looking to build a Hadoop cluster don't yet know the eventual profile of their workload, and often the first jobs that an organization- runs with Hadoop are far different than the jobs that Hadoop is ultimately used for as proficiency increases.

Furthermore, some workloads might be bound in unforeseen ways. For example, some theoretical IO-bound workloads might actually be CPU-bound because of a user's choice of compression, or different implementations of an algorithm might change how the Map Reduce job is constrained. For these reasons, when the team is unfamiliar with the types of jobs it is going to run, as an initial approach it makes sense to invest in a balanced Hadoop cluster. The Name Node role is responsible for coordinating data storage on the cluster, and the Job Tracker for coordinating data processing. (The Standby Name Node should not be co-located on the Name Node machine for clusters and will run on hardware identical to that of the Name Node.)

VII.CONCLUSION

As leveraged the differences of heterogeneous workloads and proposed a cooperative resource provisioning solution to save the server cost, which is the largest share of hosting data center costs. To the end have to built an innovative system Phoenix Cloud to enable cooperative resource provisioning for heterogeneous workloads of parallel batch jobs. Proposed an innovative experimental methodology that combines real and emulated system As Enhancement, to propose a classification technique in the map reduce Models for data evolution (new Data analysis) through the Ensemble classifier in the efficient data analysis framework to cater the workloads and processing of it in commodity clusters in the datacenters which yields better performance in terms of fault handling.

REFERENCES

1. 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.
2. B. Rochwerger et al., "The Reservoir Model and Architecture for open Federated Cloud Computing," IBM J. Research Development, vol. 53, no. 4, pp. 4:1-4:11, 2009.
3. Hoelzle .U et al, 2009 The Data Center as a Computer: An Introduction to the Design of Warehouse-Scale Machines. Morgan and Claypool Publishers.
4. Irwin. D et al, 2006 "Sharing Networked Resources with Brokered Leases," Proc. USENIX '06 Ann. Technical Conf. p. 18.
5. Jiang .D, Antony K.H. Tung, and Gang Chen et al, 2011 "MAP-JOIN-REDUCE: Toward Scalable and Efficient Data Analysis on Large Cluster" IEEE Transaction.
6. J. Moreira et al., "The Case for Full-Throttle Computing: An Alternative Datacenter Design Strategy," IEEE Micro, vol. 30, no. 4, pp. 25-28, July/Aug. 2010.
7. J. Zhan et al., "Phoenix Cloud: Consolidating Different Computing Loads on Shared Cluster System for Large Organization," Proc. First Workshop Cloud Computing and Its Application (CCA' 08), http://arxiv.org/abs/0906.1346, 2008.
8. Lin .B et al, 2005 "VSched: Mixing Batch and Interactive Virtual Machines Using Periodic Real-Time Scheduling," Proc. ACM/IEEE Conf. Supercomputing (SC '05), p. 8.
9. P. Ruth et al., "VioCluster: Virtualization for Dynamic Computational Domains," Proc. IEEE Int'l Conf. Cluster Computing (Cluster '05), pp. 1-10, 2005.
10. P. Wang et al., "Transformer: A New Paradigm for Building Data-Parallel Programming Models," IEEE Micro, vol. 30, no. 4, pp. 55-64, July 2010.
11. W. Gao et al., "BigDataBench: a Big Data Benchmark Suite from Web Search Engines," Proc. Third Workshop on Architectures and Systems for Big Data (ASBD 2013) in conjunction with ISCA'13013, 2013.
12. Zhan. J et al, 2011 "Cost-Aware Cooperative resource Provisioning for heterogeneous workloads in data centers" Pro IEEE Transaction on computer".

DOI 10.17148/IJARCCE.2015.41221