International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 05 | May-2018 www.irjet.net p-ISSN: 2395-0072

An Overlay Database Management System

Anas Jamal[1], R. Suji Pramila[2]

[1]Student, Computer Science and Engineering, Noorul Islam Centre for Higher Education, kumaracoil, Kanyakumari, Tamilnadu
[2]Assistant Professor, Computer Science and Engineering, Noorul Islam Centre for Higher Education, kumaracoil, Kanyakumari, Tamilnadu

---

Abstract – *Management scheme for highly scalable big data mining has not been well studied in spite of the fact that big data mining provides many valuable and important information. An overlay- based parallel data mining architecture, which executes fully distributed data management and processing by employing the overlay network, can achieve high scalability. However, the overlay-based parallel mining architecture is not capable of providing data mining services in case of the physical network disruption that is caused by router/communication line breakdowns because numerous nodes are removed from the overlay network. To cope with this issue, an overlay network construction scheme is proposed based on node location in physical network and a distributed task allocation scheme using overlay network technology. The numerical analysis indicates that the proposed schemes considerably outperform the conventional schemes in terms of service availability against physical network disruption.*

Key Words: scalable; employing; capable; allocation; considerably

1. INTRODUCTION

Nowadays, increasing the amount of speed for information and communication technologies is becoming very sharp to the entire society. This phenomenon promotes the world to retrieve various types of data that is provided from the whole environment. The use of effective data mining is growing rapidly as an output of similar data which is referred to as the "big data mining" is growing tremendously. Since large comprises various types of data (such as document, pdf, and text) the nowadays big data mining has been increasing, therefore Hadoop and Map Reduce are the two technologies introduced in data mining architecture.

Data Mining is an analytical procedure intended to examine information (business sector related information).The main aim of data mining process is to find patterns, once these patterns are found then they can be used to make definite decisions for development of their further model. The purpose of data mining is predictive and expectation data mining is the most well defined form of data mining. Data mining combining research methods and computer technology by its nature should be considered as a research support system. It is the process of discovering interesting patterns and knowledge from large amount of data.

Data mining is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Data mining is the analysis step of the "knowledge discovery in databases" process.

In earlier architectures, the processing node performs the data processing but the master node performs the system organization task. The implementation and designing is centralized by management scheme which lacks scalability. If number of nodes increases performance of the system decreases due to overseen of master node. If any failure occur in master node, such that the service availability can spectacularly decrease. Parallel data mining architecture having some critical issues such as scalability and service availability.

An overlay network is a computer network that is built on top of another network. Nodes in the overlay network can be thought of as being connected by virtual or logical links, each of which corresponds to a path, perhaps through many physical links, in the underlying network. For example, distributed systems such as peer-to-peer networks and client-server applications are overlay networks because their nodes run on top of the Internet. The Internet was originally built as an overlay upon the telephone network, while today the telephone network is increasingly turning into an overlay network built on top of the Internet.

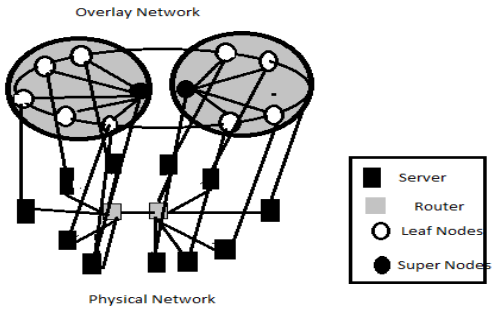

Fig .1 Network Based on Physical and Overlay

The network based on physical and overlay structure is shown in Fig.1. An Overlay based on Physical network that shows the relationship among physical and overlay networks. The overlay network consists of group of leaf nodes which is linked with the server .The router is connected to the server, and the routers are interconnected to each other. The overlay network means the interconnection of nodes therefore there is no collision occurs, here number of leaf nodes are connected with the number of server node. The main aim of the data mining task is to examine the previously unknown interesting models, group of data records and remarkable records, which also involves by using database techniques such as spatial index. This method can further be used for future studying such as in machine learning and logical analytics. For example this data mining process also identifies several groups in the given data, which can later be used for obtaining accurate end result by a resolution support system. Overlay networks are used in telecommunication because of the availability of digital circuit switching equipment and optical fiber. Telecommunication transport networks and IP networks (which combined make up the broader Internet) are all overlaid with at least an optical fiber layer, a transport layer and an IP or circuit switching layers.

From a physical standpoint overlay networks are quite complex as they combine various logical layers that are operated and built by various entities (businesses, universities, government etc.) but they allow separation of concerns that over time permitted the buildup of a broad set of services that could not have been proposed by a single telecommunication operator, competitive telecom operators.

2. LITERATURE SURVEY

2.1. Designing P2P Networks Tolerant To Attacks and Faults Based On Bimodal Degree Distribution

In this paper, the authors [1] classified the existing distributed networks based on their degree distributions. Then, they demonstrated that they are not resilient to attacks and/or faults. For example, unstructured P2P networks, which have a power-law degree distribution, are vulnerable to attacks such as DOS. To address and resolve this issue, they proposed a method to construct a network following bimodal degree distribution, which is robust to deal with both attacks and faults. Performance evaluation is conducted through computer simulations, which show that the proposed method can achieve higher resilience compared with other existing networking approaches.

To achieve high robustness against both attacks and faults without increasing the average degree in a network, they employed a bimodal network, which has mixed features [1] from both regular and scale-free networks to exploit their benefits in the maximum way possible. A bimodal network can achieve high robustness against attacks since it has a lower degree of hub nodes which cause fewer network fragment attributed to attacks compared to scale-free networks. In addition, fault tolerance in bimodal networks is better than in regular networks since they have hub nodes which increase network connectivity. Thus, bimodal networks inherit attack resiliency from regular networks and fault tolerance from scale-free networks.

2.2. Map reduce: simplified data processing on large clusters

Map Reduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a map function that processes a key/ value pair to generate a set of intermediate key/value pairs, and a reduce [2] function that merges all intermediate values associated with the same intermediate key. Many real word tasks are expressed in this model. The major contributions of this work are a simple and powerful interface that enables automatic parallelization and distribution of large-scale computations, combined with an implementation of this interface that achieves high performance on large clusters of commodity PCs.

Massive-scale Big Data analytics is representative of a new class of workloads that justifies a rethinking of how computing systems should be optimized. This paper addresses the need for a set of benchmarks that system designers can use to measure the quality of their designs and that customers can use to evaluate competing systems offerings with respect to commonly performed text-oriented workflows in Hadoop™. Additions are needed to existing benchmarks such as HiBench in terms of both scale and relevance. We describe a methodology for creating a pet scale data-size text-oriented benchmark that includes representative Big Data workflows and can be used to test total system performance, with demands balanced across storage, network, and computation. Creating such a benchmark requires meeting unique challenges associated with the data size and its often unstructured nature. To be useful, the benchmark also needs to be sufficiently generic to be accepted by the community at large. Here, we focus on a text-oriented Hadoop workflow that consists of three common tasks: categorizing text documents, identifying significant documents within each category, and analyzing significant documents for new topic creation.

In this work, the map invocations are distributed across multiple machines by automatically partitioning the input data into a set of M splits. The input splits can be processed in parallel by different machines. Reduce invocations are distributed by partitioning the intermediate key space into R pieces using a partitioning function (e.g., hash (key) mod R). The number of partitions (R) and the partitioning function are specified by the user.

2.3. Exploiting dynamic resource allocation for efficient parallel data processing in the cloud

Ad-hoc parallel data processing has emerged to be one of the killer applications for infrastructure-as-a-service (IaaS) clouds [3]. Major cloud computing companies have been

started to integrate frameworks for parallel data processing in their product portfolio, making it easy for customers to access these services to deploy their programs. Therefore, the opportunities and challenges for efficient parallel data processing in clouds have been discussed in this paper and presented the research project Nephele. A parallel data processor centered on a programming model of so called Parallelization Contracts (PACTs) and the scalable parallel execution engine Nephele has been proposed. Nephele is the first data processing framework to explicitly exploit the dynamic resource allocation offered by today's IaaS clouds for both, task scheduling and execution. Based on this new framework, extended evaluations of Map Reduce-inspired processing jobs on an IaaS cloud system have been performed. The PACT programming model is a generalization of the well- known Map Reduce programming model that gives guarantees about the behavior of the function. The PACTs are second-order functions that define properties on the input and output data of their associated first-order functions.

The Nephele's [3] architecture follows a classic master-worker pattern. Before submitting a Nephele compute job, a user must start a VM in the cloud which runs the so called job manager (JM). The JM receives the client's job, is responsible for scheduling them, and coordinates their execution. It is capable of communicating with the interface the cloud operator provides to control the instantiation of VMs. This interface is called Cloud Controller. By means of the Cloud Controller the JM can allocate or deallocate VMs according to the current job execution phase.

2.4. PrIter: a distributed framework for prioritizing iterative computations

In this paper, the authors [4] Yanfeng Zhang et al explored the opportunity for accelerating iterative computations by prioritization. Instead of performing computations on all data points without discrimination, they prioritized the computations that help convergence the most, so that the convergence speed of iterative process is significantly improved. They developed a distributed computing framework, PrIter, which supports the prioritized execution of iterative computations. PrIter either stores intermediate data in memory for fast convergence or stores intermediate data in files for scaling to larger data sets. They evaluated PrIter on a local cluster of machines as well as on Amazon EC2 Cloud. In addition, PrIter is shown better performance for iterative computations than other state-of-the-art distributed frameworks such as Spark and Piccolo.

In this paper,[4] they demonstrated the potential of prioritized execution for iterative computations with a broad set of algorithms. This motivates the desire of a general priority-based distributed computing framework. They designed and implement PrIter, [4] a distributed framework that supports the prioritized execution of iterative computations. To realize prioritized execution, PrIter allows users to explicitly specify the priority value of each processing data point. PrIter allows either to store data in memory for better performance or to store data in files for better scalability. In addition, PrIter is designed to support load balancing and fault tolerance so as to accommodate diverse distributed environments.

2.5. Cayleyccc: a robust P2P overlay network with simple routing and small-world features

In this paper, the authors [5] introduced a general group theoretic method and define a new Cayley graph. They then use these constructs to derive a novel P2P overlay structure. CayleyCCC, their proposed overlay network, has many desirable features, including short query paths, compact routing tables, high clustering, and robustness. Because of its symmetry, their design offers simpler routing (searching) and several other desirable properties compared with many previously proposed overlay networks, such as Chord and Ulysses. Performance evaluation results, reported in this paper, quantify the advantages of CayleyCCC in terms of query path length, routing table size, and robustness relative to some recent proposals.

CayleyCCC is particularly useful in distributed computing under relatively stable conditions. Their method's generality makes it applicable to the design of P2P systems possessing diverse sets of features. The method is particularly useful in distributed computing, under relatively stable conditions. In addition to providing an efficient resource searching mechanism, CayleyCCC supports explicit grouping of peers, thus facilitating effective resource browsing. Theoretical analysis and experimental evaluation show that CayleyCCC can reach the lower bounds of routing table size and query path length at the same time.

3. PROPOSED WORK

An efficient overlay based data mining architecture is introduced for improving scalability, and it is more secure data mining. This architecture is planned based on the grouping of two networks they are physical networks and overlay networks. It improves the service availability against server breakdowns. In the proposed work, number of users can execute processing function by using overlay network. The overlay network defined as computer networks that are constructed on top of another networks, it is connected by virtual or logical links each of which corresponds to a path possibly through many physical layer in the network. Additionally proposed work introduces a secured process that allows each user to access the information. If the user wants to store any secret information, then security is provided using SHA-1 algorithm by computing SHA-1 hash of the file.

In this section, introduces the parallel data mining architecture based on the overlay network. Then it describes the existing works that aim to improve the service availability. The data will be more secured, and using this splitting method it will be more memory efficient this process is shown below in .

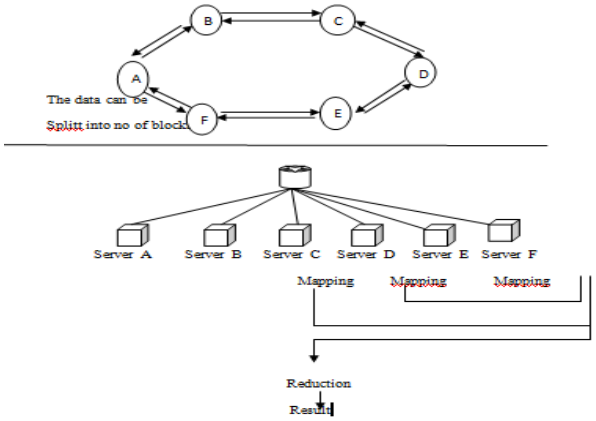

Fig.2 Mapping and reduction processes in overlay-based parallel data mining architecture

This diagram depicts the data splitting process into number of blocks. The number of servers are used for storing the data by introducing mapping process while mapping, each blocks of data is stored into each servers. Figure shows an example of mapping and reduction processes in the overlay-based parallel data mining architecture. When a data processing request is injected, a block that received the request executes a reception function by using the overlay network. Then, a mapper that initially finished the mapping process becomes a reducer, and it requests to other mappers to transmit the processed data to itself, where the request message can be forwarded by using flooding scheme. After receiving the processed data from mappers, the reducer executes the reduction process and outputs the analyzed result.

3.1. Uploading Process

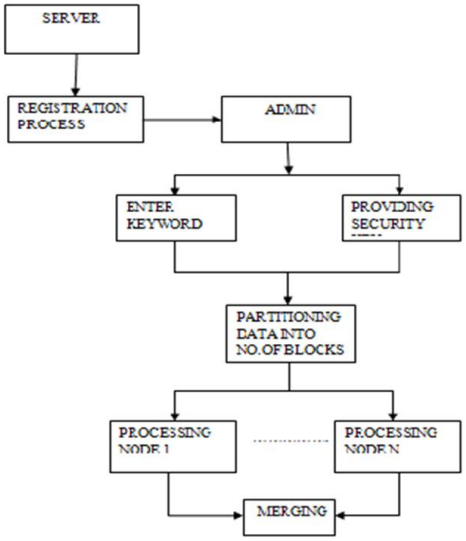

Fig.3 Work Flow Diagram of Uploading Process

Fig.3 represents the uploading process of secured overlay data mining architecture here number of server and users are introduced. The first process is to move on to the server registration process, here server page shows server id, name and URL, also no of servers can be created using this procedure. After completing this process then move on to the admin page, here document, file, pdf can be uploaded by the admin. The keyword is given according to the uploaded file, admin will provides a security key by SHA1 algorithm. After the completion of uploading process the data will be split into number of blocks according to the amount of data then the merging process will also be done by the admin.

3.2. Downloaded Process

Fig.4 depicts the downloading process of the secured overlay data mining architecture, this process will be done by a user here number of users can be created, the required file will be searched from the server list and corresponding keyword is given according to the required file, while downloading the file, the security key is given which has been provided by the admin after this procedure the file can be downloaded and then viewing that file.

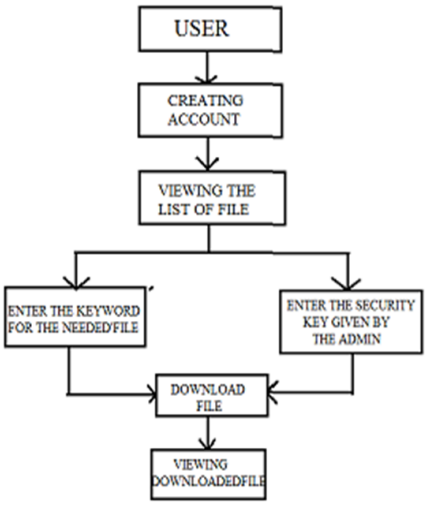

Fig.4 Workflow diagram of downloaded process

4. PERFORMANCE ANALYSIS

4.1. Service Availability

Service Availability refers to service that is available irrespective of software and hardware or user fault and consequences. Many of the customers describe availability that reflect the complex in which real IT systems actually fail allowing for partial failures that impacts on just some functions or just some users performance to be unacceptably slow. In real life scenario it is not practical to include all of these factors in a calculation since the complicity will be great.

4.2. Traffic Load

It refers to the total traffic carried on communication links during a specified interval of time. Load refers to the amount of data (traffic) being carried by the network. It

focuses on optimizing resource utilize, exploit throughput, reduce response time and avoiding the overlay of single resource. In a network, the traffic between servers must be divided, so that the data can be sent and received without major delay.

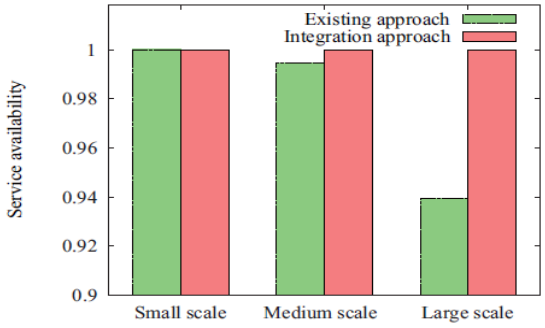

Fig. 5 Service Availability

Fig.5 depicts the result of increase on service availability on three scales of physical network failures. From the given result it is clearly proved that integration approach gains higher service availability and ensure almost 100% availability in spite of the scale of physical network failures. In the small scale and medium scale the service availability is similar to the existing approach. The difference appears in the large scale service providers where the services are greatly available. The proposed approach constructs the overlay network by considering the neighborhood of the physical network failures. Moreover, the proposed work allocates tasks to sufficient nodes according to some of the feature of the physical network failures.

Fig. 6 describe about the predictable grouping approach on the traffic load of each link in the physical network according to the three scales. It is confirmed that the proposed approach reduces the traffic load of the network. In particular, the existing approach consumes more

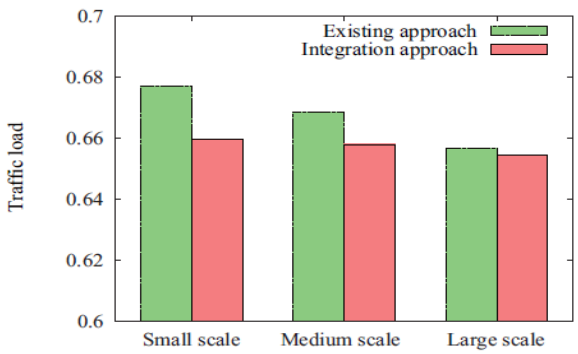

Fig.6 Average Traffic Load on Communication Links

network resources in the case of small scale failures. This happens because the existing approach makes constant replicas without taking into consideration about the characteristics of physical network failures, which result in unessential data transmissions. On the other hand, the proposed data allocation principle replicates an adequate number of data automatically. Thus, the proposed approach is able to reduce the traffic load in contrast with the existing data allocation scheme.

5. CONCLUSION AND FUTURE ENHANCEMENT

Efficient overlay-based parallel data mining architecture, which fully distributes processing function and management by using, overlay network methods. By using this method users can store and retrieve any amount of data or document, and important data can be protected by using SHA-1 algorithm. This method also proposes a task allocation and neighbor selection scheme based on integration of the overlay and physical networks. In order to improve the opportunity of data mining beside physical network interference, overlay network is constructed using neighbor selection based on node position in physical network and the task allocation scheme selects nodes from different diagonally-cornered groups in the overlay network as mappers, to protect the data from the hackers a secure key is also used.

REFERENCES

[1] K. Suto, H. Nishiyama, X. S. Shen, and N. Kato, "Designing P2P networks tolerant to attacks and faults based on bimodal degree dis- tribution," Journal of Communications, vol. 7, no. 8, pp. 587–595, Aug. 2012.

[2] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," in Proc. of 6th Symposium on Operating Systems Design and Implementation, San Francisco, USA, Dec. 2004, pp. 137–150.

[3] D. Warneke and O. Kao, "Exploiting dynamic resource allocation for efficient parallel data processing in the cloud," IEEE Transactions on Parallel and Distributed Systems, vol. 22, no. 6, pp. 985–997, Jun. 2011.

[4] Y. Zhang, Q. Gao, L. Gao, and C. Wang, "Priter: A distributed framework for prioritizing iterative computations," IEEE Transactions on Parallel and Distributed Systems, vol. 24, no. 9, pp. 1884–1893, Sept. 2013.

[5] Wenjun Xiaoa, Yunyan Xionga, and Huomin Lianga," Cayley: A Robust Overlay with Simple Routing and SmallWorld Features for Wireless Sensor Networks" Procedia Engineering 15 (2011) 3008 – 3016,