Big Data Processing using Apache Hadoop in Cloud System

Mr. Yogesh Pingle, Vaibhav Kohli, Shruti Kamat, Nimesh Poladia
Vidyavardhini`s College of Engineering & Technology
Mumbai University
E-mail: vaibhavkhl@yahoo.co.in

Abstract. **The ever growing technology has resulted in the need for storing and processing excessively large amounts of data on cloud. The current volume of data is enormous and is expected to replicate over 650 times by the year 2014, out of which, 85% would be unstructured. This is known as the 'Big Data' problem. The techniques of Hadoop, an efficient resource scheduling method and a probabilistic redundant scheduling, are presented for the system to efficiently organize "free" computer storage resources existing within enterprises to provide low-cost high-quality storage services. The proposed methods and system provide valuable reference for the implementation of cloud storage system. The proposed method includes a Linux based cloud.**

Keywords: cloud Storage, hadoop, resource scheduling.

I. INTRODUCTION

With the rapid development of Internet, we are experiencing an information explosion era. Large amounts of data are being stored and managed. Cloud storage has the potential of providing geographically distributed storage services since cloud can integrate servers and clusters that are distributed all over the world and offered by different service providers into one virtualized environment.

Traditional storage systems always cost much on software and hardware which is unaffordable to many all and medium enterprises. Compared with traditional storage system since the change from storage device to storage service is achieved by using application software, cloud storage significantly reduces the number of server that is required for storage, which furthermore can reduce construction cost of storage system. Cloud storage has become a trend in the future development of storage. On the other hand, currently , most devices owned by enterprises are idle or not full used, the cloud storage proposed in our work could efficiently make use of such free and idle resource and increase the utilization ratio of resource.

However, there are still lots of works remaining in current cloud storage technique. For instance, Amazon's Simple Storage Service (S3) once brokedown and been off-line for 3 hours. Occasional troubles that happen to Google GMAIL as well cause users to pay attention on security technique such as data backup. Therefore, in future development of cloud storage, problems on performance, security, reliability and scalability will still need to be solved. We have done some work in these problems and proposed a Hadoop-based high performance and economical cloud storage system in this paper.

Using open-source framework Hadoop, efficient resource scheduling, probabilistic redundant scheduling, the proposed storage system organizes and utilizes currently idle computers and storage resource to provide cheap and good quality storage service.

II. RELATED WORK

Cloud storage is an expansion and development of cloud computing. Similar to cloud computing, cloud storage connects different storage devices and make them work together via application and applying functions such as cluster application, grid or distributed file system [1]. Once cloud storage was proposed, this concept attracted much attention and support from enterprises. Academia has also started lots of research [3-8] on the new technique. Reference [2] suggests a cloud computing system, GridBatch that addresses the problem of large-scale data-intensive batch processing. It implements the data-intensive-application oriented parallel program to satisfy enterprises' growing demand for data integration and analysis. Reference [3] proposes a cloud based infrastructure that is optimized for high performance, wide area networks and designed to support the ingestion, data management, analysis, and distribution of large terabyte size data sets. This type of the infrastructure consists of a high performance storage cloud called Sector and a compute cloud called Sphere that are designed to store large distributed data sets and to support the parallel analysis of these data sets. Experiments show that the performance of distributed data computing of this system on wide area network is the same as computing with local data. In order to reduce the construction cost of storage system, reference[4] proposes an economical and efficient cloud storage service model using disks connected by network to replace disks being directly connected, which makes it feasible to add original PCs to cloud storage system and use current PCs to achieve economical storage service. Reference[5] points out that the major issue existing in current cloud storage is how to convince users and make them submit their data storage and data processing to external virtual cloud storage system. This requires cloud storage to carefully address issues of security, efficiency and service quality. Reference[6] points out that cloud storage infrastructure has to store redundant data to ensure usability of data. Nevertheless, redundancy may unnecessarily increase storage and communication cost. Therefore, for the purpose of implementing scalability, works are needed to decrease redundancy. Reference[7] indicates that all current shared access protocols need to be linearized when dealing concurrent execution, which means the time needed to respond read operation should be in accordance with the order of all read operations. It is mentioned in the paper that

the integration of SSS(short secret sharing) and DHT (distributed hash table) into cloud storage implements secure, reliable, efficient cloud storage system. DHT algorithm is used to implement near-by share lookup and SSS technique is applied to secure sharing access of data.

Hadoop[9] is an open-source implementation of Google's distributed file system and the MapReduce framework for distributed data processing. Hadoop primarily consists of HDFS (Hadoop Distributed File System), MapReduce (distributed computing system) and HBase (distributed query system). Hadoop has currently been widely used in the Internet, as well attracted general attention from research community. Hadoop Distributed File System (HDFS) is a distributed file system running on common hardware and its implementation follows the framework of Google GFS. HDFS is highly fault-tolerant and is designed to run on low-cost hardware. The master/slave architecture [10] is applied in HDFS. A HDFS cluster is constituted by one Namenode and a certain number of Datanode. Namenode is a central server which is the brain of the entire file system. It is responsible for managing namespace of file system and access of files. Datanode is generally a node in cluster that manages the storage on Datanode itself. Datanode performs creation, deletion and copy of block under the Namenode's command. Both Namenode and Datanode are designed to run on common low-cost Linux machine. In Hadoop MapReduce framework, JobTracker is in charge of arranging the order of task execution and scheduling computing resource. Thus it directly influences the overall performance of Hadoop platform and utilization of resource. FIFO algorithms is default in Hadoop cluster, it neglects the different needs of tasks. Such algorithm is too simple to meet the need of some specific task. So far, Fair Scheduling method and Capacity Scheduling method provided by third parties have taken into account the equitable use of resource (providing certain amount of resource to every task) and the priority of tasks that, to some extend, improves utilization of resource. But it goes against the execution of short tasks (tasks with high priority will be performed first) and doesn't help optimize the scheduling performance of interactive tasks, neither support preemptive execution. Based on analysis above, we can reach the conclusion that the following problems still exist in current cloud storage technique.

(1) Input Cost: current providers of cloud computing service focus only on decreasing management cost, ignoring significant pre-investment cost. Cloud storage service is usually built on costly and highly expandable commercial hardware. Therefore, better techniques and methods are needed to establish a more economical cloud storage service.

(2) Reliability: Before users are willing to put their data in cloud, they must be convinced that cloud storage service is reliable and secure enough that data on cloud will never be lost. Redundancy technique is generally used to ensure security of data. However the use of redundancy technique may increase the overhead of storage and communication. Hence, it is important to balance redundancy and performance.

(3)Although the current popular Hadoop framework could run on low-cost common computer, there are flaws in Hadoop: 1)The construction and management of Hadoop is complicated; 2) Currently task scheduling problems: the default FIFO method is simple but doesn't satisfy different user requests; 3)Resource scheduling is based on the assumption that the resources(computers) are homogeneous. Datanode is randomly selected to store data, which means there is no guarantee on system performance.

Let's see some of the practical examples for big data processing.

A. Linkedln:
- For discovering People You May Know and other fun facts.
- Item-Item Recommendations
- Member and Company Derived Data
- User's network statistics
- Who Viewed My Profile?
- Abuse Detection
- User's History Service
- Relevance Data
- Crawler Detection

B. Mobile Analytic TV:
- Natural Language Processing
- Mobile Social Networking Hacking
- Web Crawlers/Page Scrapping
- Text to Speech
- Machine generated Audio & Video with remixing
- Automatic PDF creation & IR

C. Datagraph:
- Executing long-running offline SPARQL queries

D. GumGum-lin-image ad network:
- GumGum is an analytics and monetization platform for online content.
- Image and advertising analytics

E. Lineberger Comprehensive Cancer Center - Bioinformatics group

F. Pharm2Phork Project – Agricultural Traceability

- Processing of observation messages generated by RFID/Barcode readers as items move through supply chain.

III. SOLUTIONS TO SOME KEY PROBLEMS

Complex Construction and Management of Cluster. To establish a Hadoop-based efficient and economical cloud storage system, the first work would be effectively organizing distributed computer resource. The construction of Hadoop cluster needs complicated manual management and the deploying and installing process are also complex and could not be automated.

To address the complication of construction and management, we propose the methods of automatic cluster construction and self-organized management. For the purpose of implementing automatic cluster construction, the technique of virtual machine migration is adopted to migrate the mirror of Hadoop virtual machine to some specific node, and then let it automatically be deployed. To address the complication of management, we propose a new resource management method, a resource self-organization model. This model achieves self-organization management of cluster by dynamically optimizing resource organization according to change of resource.

Efficiency Resource Scheduling. Only physical position of nodes is taken into account when HDFS schedules resource (DataNode). Heterogeneity of nodes and utilization of resource is not in consideration. Therefore, when there is data to be stored, the system needs to optimize data storage and resource allocation according to nodes' network location and utilization ratio of resource.

Efficiency Resource Scheduling. Only physical position of nodes is taken into account when HDFS schedules resource (DataNode). Heterogeneity of nodes and utilization of resource is not in consideration. Therefore, when there is data to be stored, the system needs to optimize data storage and resource allocation according to nodes' network location and utilization ratio of resource.

In order to achieve efficient data storage and resource scheduling, we propose a resource scheduling model based on network performance and resource utilization ratio of data nodes. The model can implement data balancing between data nodes, as well supports a good performance of data read/write by optimizing the selection (scheduling) of data nodes.

Conflict between Data Storage and Performance. Because of the heterogeneity and unreliability of nodes inside cloud storage system (computers owned by enterprise internally could be very different), data replication is designed to reliably store data across heterogeneous machines within the enterprise. Nevertheless, data replication results in extra overhead and waste of resource. Thus, it is necessary to find a point between reliability and performance at which the reliability of data is assured and the performance is as good as possible.

Therefore, we propose a reliable cloud storage model based on probabilistic redundant scheduling. By using probability, the proposed model calculates and gets optimized replication factor to avoid unnecessary data copy, which furthermore resolves the conflict between storage reliability and system performance.

IV. HADOOP BASED PRODUCTS IN CLOUD

In Cloud Computing, we have few options available for Hadoop implementation. 1) Amazon I aaS 2) Amazon MapReduce 3) Cloudera. Amazon Elastic Compute Cloud (Amazon EC2-IaaS) is a web service that provides resizable compute capacity in the cloud . It is designed to make web-scale computing easier for developers. If you run Hadoop on Amazon EC2as shown in Fig.1 you might consider using AmazonS3 for accessing job data (data transfer to and from S3 from EC2 instances is free). Initial input can be read from S3 when a cluster is launched. The final output can be written back to S3 before the cluster is decommissioned. Intermediate, temporary data only needed between MapReduce passes, is more efficiently stored in Hadoop's DFS. It became a popular way for Big Data processing and that led to the emergence of another service called Amazon Elastic MapReduce.

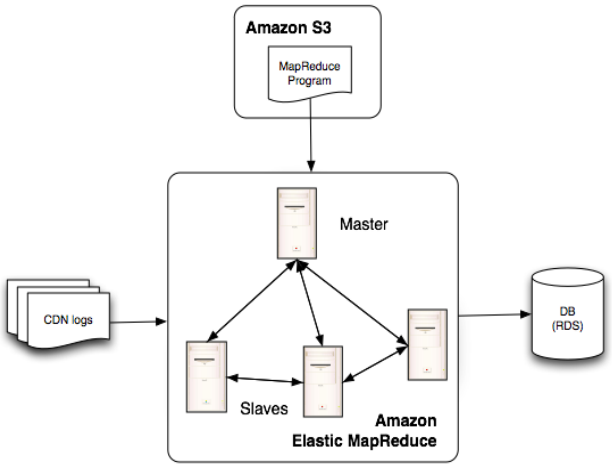

Fig.1Amazon MapReduce Workflow

It utilizes a hosted Hadoop framework running on the web-scale infrastructure of Amazon Elastic Compute Cloud (Amazon EC2) and Amazon S3. In a nutshell, the Elastic MapReduce service runs a hosted Hadoop instance on an EC2 instance (master). It's able to instantly provision other pre-configured EC2 instances (slave nodes) to distribute the MapReduce process. All nodes are terminated once the MapReduce tasks complete running. *Cloudera* has two products: Cloudera's Distribution for Hadoop (CDH) and Cloudera Enterprise. CDH is a data management platform (incorporates HDFS, Hadoop MapReduce, Hive, Pig, HBase, Sqoop, Flume, Oozie, Zookeeper and Hue). It is available free under an Apache license.

V. ARCHITECTURE OF HADOOP-BASED EFFICIENT AND ECONOMICAL CLOUD STORAGE SYSTEM

Based on the solutions to key problems described above, we designed the architecture of hadoop-based efficient and economical cloud storage system. It is a four-layer architecture: System Resource Layer, Cluster Management Layer, Storage Service Layer and Web Access Layer.

(1)System Resource Layer: This is the most fundamental part of cloud storage system. This layer provides storage resources for the cloud storage system. Such resources could be server or PC located in many different departments of one company and connected by the Internet.

(2)Cluster Management Layer: An important issue to be addressed is how to effectively organize those "idle" computers distributed in one company. In our work, Hadoop cluster technique is used to organize idle computers in internal network to constitute a private cloud that provides platform for company to construct a cloud storage system. Such technique could fully utilize existing IT resource possessed by company.

(3)Storage Service Layer: Storage Service Layer is the core part and also the most difficult one for implementation of the cloud storage system. Storage Service Layer uses different techniques including dynamical resource management, probabilistic redundant scheduling, management of Master node and scheduling separate to implement efficient and reliable cloud storage. This layer provides better storage and access of data via a unified storage service built on effective organization of computer resource owned by company.

(4)Web Access Layer: Web Access Layer is the gateway for users to use the cloud storage system. We apply the popular SSH framework to implement a flexible and easy-to-use WEB access interface. Any authorized user could login the cloud storage system through Web Access Layer and access different services provided by the cloud storage system.

VI. IMPLEMENTATION OF HADOOP-BASED EFFICIENT AND ECONOMICAL CLOUD STORAGE SYSTEM

The proposed cloud storage system consists of one master server, multiple data storage node, storage client, and WEB access. Master server and multiple storage nodes constitute a distributed file storage clusteras shown in Fig.2. The primary function of the storage platform is using Hadoop cluster to organize and coordinate different types of computer storage devices on internal network, and jointly provide data storage service using HDFS.

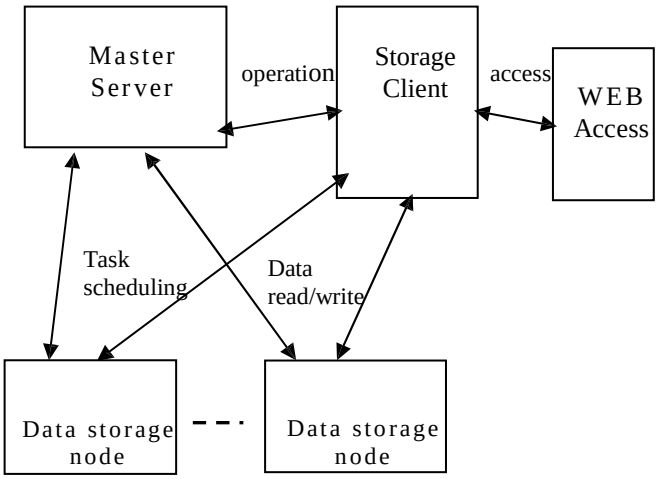

Fig.2 Prototype System of Cloud Storage

Web Access Layer is the gateway for users to use cloud storage system. Storage Client is both the client of distributed file storage platform and the server of WEB application in cloud storage system. It is responsible for receiving data storage requests from WEB access and transfers them to Master server and read/write data according to the result returned from server using HDFS API. Master server is responsible for managing metadata and task scheduling of data storage. Data storage node is responsible for handling read/write requests from storage client and creating, deleting, copying data blocks under the unified scheduling of Master server.

Below we have demonstrated in Fig.3 the actual working of hadoop on single cloud server system.

```
**********************************************************/
hduser@ubuntu:~$ /usr/local/hadoop/bin/start-all.sh
starting namenode, logging to /usr/local/hadoop/bin/../logs/hadoop-hduser-nameno
de-ubuntu.out
localhost: starting datanode, logging to /usr/local/hadoop/bin/../logs/hadoop-hd
user-datanode-ubuntu.out
localhost: starting secondarynamenode, logging to /usr/local/hadoop/bin/../logs/
hadoop-hduser-secondarynamenode-ubuntu.out
starting jobtracker, logging to /usr/local/hadoop/bin/../logs/hadoop-hduser-jobt
racker-ubuntu.out
localhost: starting tasktracker, logging to /usr/local/hadoop/bin/../logs/hadoop
-hduser-tasktracker-ubuntu.out
hduser@ubuntu:~$ cd /usr/local/hadoop
hduser@ubuntu:/usr/local/hadoop$ bin/hadoop dfs -copyFromLocal /tmp/gutenberg /u
ser/hduser/output
copyFromLocal: File /tmp/gutenberg does not exist.
hduser@ubuntu:/usr/local/hadoop$ bin/hadoop dfs -copyFromLocal /tmp/input /user/
hduser/output
copyFromLocal: File /tmp/input does not exist.
hduser@ubuntu:/usr/local/hadoop$ bin/hadoop dfs -copyFromLocal /tmp/Input /user/
hduser/output
hduser@ubuntu:/usr/local/hadoop$ bin/hadoop dfs -ls /user/hduser/output
Found 3 items
drwxr-xr-x   - hduser supergroup          0 2011-09-25 10:02 /user/hduser/output
/Input
drwxr-xr-x   - hduser supergroup          0 2011-09-23 11:40 /user/hduser/output
/_logs
-rw-r--r--   1 hduser supergroup     774172 2011-09-23 11:48 /user/hduser/output
/part-r-00000
hduser@ubuntu:/usr/local/hadoop$ bin/hadoop jar hadoop*examples*.jar wordcount /
user/hduser/output /user/hduser/output1
11/09/25 10:04:28 INFO input.FileInputFormat: Total input paths to process : 2
11/09/25 10:04:29 INFO mapred.JobClient: Running job: job_201109250957_0001
11/09/25 10:04:30 INFO mapred.JobClient:  map 0% reduce 0%
11/09/25 10:04:59 INFO mapred.JobClient:  map 50% reduce 0%
```

Fig.3 Practical working of hadoop on single server cloud system

For Management, Processing of enormous unstructured data above on single sever cloud system we have used Mapreduce Technique.

MapReduce is one of the solutions, for big data processing. MapReduce as shown in Fig.4 is a programming model and an associated implementation for processing and generating big 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 function that merges all intermediate values associated with the same intermediate key. Computational processing can take place on data stored either in a file system (unstructured) or within a data base (structured). Programs written in this functional style are automatically parallelized and executed on a big cluster of commodity machines. This allows programmers without any experience with parallel and distributed systems to effortlessly utilize the resources of a large distributed system.

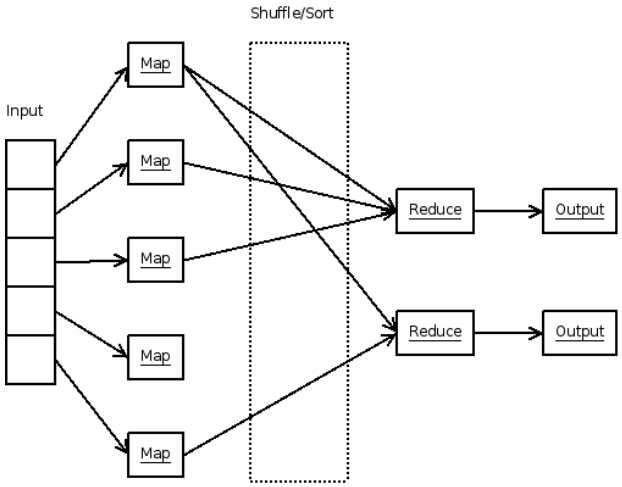

Fig.4 MapReduce Workflow

VII. SUMMARY

This paper analyzed some imperfections in current cloud storage technique and then proposed a Hadoop-based efficient and economical cloud storage system. The solutions to some key problems in the implementation of the cloud storage system are given in this paper. Then we provide some methods to implement Hadoop-based efficient and economical cloud storage system. So far, we have already developed the prototype system of Hadoop-based efficient and economical cloud storage system which implements functions such as file storage, user management, resource monitoring, and we have adopted the resource scheduling model that based on network performance and resource utilization ratio of data nodes in the prototype system. Yet it is certain that there is still work to do on data security, performance optimization of the system.

REFERENCES

[1] F. Dan: Research and Progress in Key Technologies of Network Storage, MOBILE COMMUNICATIONS,2009(11):35-39.

[2] H. Liu and D.Orban: GridBatch: Cloud Computing for Large-Scale Data-Intensive Batch Applications. Proc. of 8th IEEE International Symposium on Cluster Computing and the Grid,May 2008:295-305.

[3] Robert L. Grossman, Yunhong Gu, Michael Sabala and Wanzhi Zhang. Compute and Storage Clouds Using Wide Area High Performance Networks.Future Generation Computer Systems archive,Vol. 25(2),February 2009,pp:179-183.

[4] Shin-gyu Kim, Hyuck Han, Hyeonsang Eorn, Heon Y. Yeorn. Toward a Cost-effective Cloud Storage Service,The 12th International Conference on Advanced Communication technique (ICACT), 2010:99-102.

[5] Gabriel Antoniu. Autonomic Cloud Storage: Challenges at Stak, International Conference on Complex, Intelligent and Software Intensive Systems,2010:481-481.

[6] Lluis Pamies-Juarez, Pedro García-López, Marc Sánchez-Artigas, Blas Herrera. Towards the design of optimal data redundancy schemes for heterogeneous cloud storage infrastructures,Computer Networks,2010.

[7] Yunqi Ye, Liangliang Xiao, I-Ling Yen, Farokh Bastani, Secure, Dependable, and High Performance Cloud Storage, 29th IEEE Symposium on Reliable Distributed Systems, 2010, srds, pp.194-203.

[8] Sara Bouchenak. Automated control for SLA-aware elastic clouds,In FeBiD '10: Proceedings of the Fifth International Workshop on Feedback Control Implementation and Design in Computing Systems and Networks (2010), pp. 27-28.

[9] Tom White. Hadoop.The Definitive Guide[M].1st ed.USA:O'Reilly Media, 2009.

[10] Apache. Hadoop user Guide[EB/OL]. http://hadoop.apache.org(2010)

[11] Amazon EC2, http://aws.amazon.com/ec2/

[12] Andrew W. McNabb, Christopher K.Monson, and Kevin D.Seppi,"Parallel PSO Using MapReduce", IEEE congress on evolutionary computation, Singapore, 23-27 Sept 2007, pp. 7Cloudera, Apache Hadoop for Enterprise, http://www.cloudera.