International Journal of Sciences & Applied Research

www.ijsar.in

Big data processing using Apache Hadoop in Cloud system

Snehal D. Dahake*, Amol S. Dudhe

Department of Computer Science and Engg., Babasaheb Naik College of Engg., Pusad (MS).

Correspondence Address: *Snehal D. Dahake, Department of Computer Science and Engg., Babasaheb Naik College of Engg., Pusad (MS).

Abstract

The ever growing technology has resulted in the need for storing and processing excessively large amounts of data on cloud.

To solve the current data Security problem for cloud disk in distributed network, for example transmission, storage security problems, access control and data verification, a network cloud disk security storage system based on Hadoop is proposed. The techniques of Hadoop, an efficient resource scheduling method is 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 and a network cloud security storage system based on Hadoop is proposed.

Keywords: Cloud Storage, Hadoop Cluster, HDFS

Introduction

Cloud computing is a powerful technology to perform massive-scale and complex computing. It eliminates the need to maintain expensive computing hardware, dedicated space, and software. Massive growth in the scale of data or big data generated through cloud computing has been observed. Addressing big data is a challenging and time demanding task that requires a large computational infrastructure to ensure successful data processing and analysis. The rise of big data in cloud computing is reviewed in this study. The definition, characteristics, and classification of big data along with some discussions on cloud computing are introduced. The relationship between big data and cloud computing, big data storage systems, and Hadoop technology are also discussed [1]. Apache Hadoop project develops open-source software for reliable, scalable as well as distributed computing, and its library is a framework which allows for the distributed processing of large data sets across clusters of computers using simple programming models. Apache Hadoop is designed to scale up from single servers to thousands of machines as well as each offering local computation and storage[5].

SYSTEM ANALYSIS

Cloud Disk's weak Security System:

Transmission Security: Data in communication process may be intercept, but the data communication is not working

with the strong encryption security measures.

Access Control: Access control power is weak, the client data stored in the clouds without setting access power, the client lost absolute right to monitor.

Data Storage: Client upload data after the clouds, it is likely to be distributive stored, Client's do not know the exact position where the data is stored. And the private data and non-private data stored are not classified, which may cause the outflow of data.

Data Verification: The cloud makes no confirmation and examination on the data uploaded. It can't guarantee that the uploaded data is corresponding to the right client's data or the unique data from the client[2].

Hadoop distributed file system (HDFS):
Hadoop Distributed File System is extended version of the Google's Google File System (GFS). The wirk of HDFS is responsible for storing the data on cluster of machines. The Hadoop runtime system coupled with HDFS manages the details of parallelism and concurrency to provide ease of parallel programming with reinforced reliability. In hadoop cluster, a master node controls a group of slave nodes on which the Map and Reduce functions run in parallel. The master node assigns a task to a slave node that has any empty task slot. There is single master node and multiple slave nodes possible in HDFS. Master node contains meta information of file. HDFS divide the data into 64MB block and divide among the nodes in the cluster.

Typically, computing storage node and processing node in a Hadoop cluster are identical from the hardware's point of view. In other words, set of homogeneous nodes manages processing as well as storing the data. HDFS operates on top of native UNIX file system. HDFS provides replicated storage for data using cheap commodity hardware. By default replication factor is 3. HDFS is read only, random writes are not allowed. It will give you best performance when small numbers of large file are there to process[3].

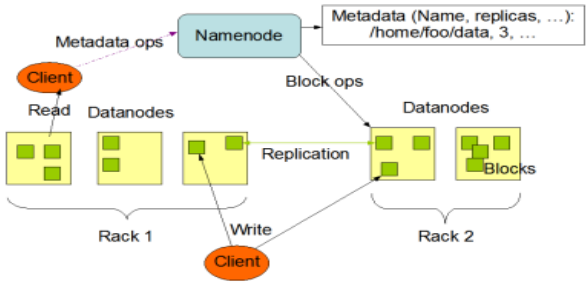

Fig. 1: HDFS Architecture.

BASIC CLOUD SYSTEM
The common design for cloud disk structure comprising of the different end devices, such as smart phone, laptop, tablet PC, etc. The client could use a browser to login and right to use the data, which lowers the requirement by the terminal. The operation is easy, suitable and fast. Inspecting from its physical structure, the system is divided into three components, the client, server and cloud group, as shown in Fig 2.

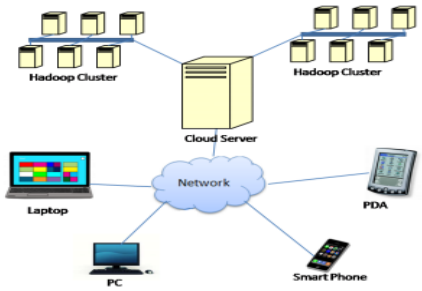

Fig. 2: Cloud Structure.

According to the software component, the system can be separated into client component, server call component, secret key production and distribution component, data encryption component, data signature

component, the data transmission component, data authentication component, and data storage component, in which the client component includes the data encryption component, data transmission component ,data signature component while the server call component includes the secret key production and distribution component, data authentication component and data storage component[2].

Disadvantages of cloud storage technique

Based on analysis above, we can reach the conclusion that the following problems still exist in current cloud storage technique.

[1] Input Cost

[2] Reliability

[3]The construction and management of Hadoop is complicated.

SOLUTIONS TO SOME KEY PROBLEMS

Complex Construction and Management of Cluster: To establish a Hadoop-based 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.

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.

PROPOSED SYSTEM

Architecture of Hadoop-based Cloud Storage System:

Based on the solutions to key problems described above, we designed Proposed System based on architecture of hadoop-based cloud storage system. It is a four-layer architecture:

(1)System Resource Layer: This is the most fundamental Master Server Task scheduling operation Data read/write Storage Client access WEB Access part of cloud storage system. This layer provides storage resources for the cloud storage system.

(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.

(3)Storage Service Layer: Storage Service Layer is the core part and also the most difficult one for implementation of the cloud storage system. 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. Any authorized user could login the cloud storage system through Web Access Layer and access different services provided by the cloud storage system.

Conclusions

This paper analyzed some imperfections in current cloud storage technique and then proposed a Hadoop-based cloud storage system. The solution to some key problems in the implementation of the cloud storage system are given in this paper. Then we provide some methods for Hadoop-based cloud storage system. Also distributed encryption system that could reduce the load on the server, and lastly achieve security, stability, efficient and successful storage.

References

[1]Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, Samee Ullah Khan "The rise of big data on cloud computing: Review and open research issues" Information Systems 47 (2015) 98–115 .

[2]Karthik D, Manjunath T N, Srinivas K " A View on Data Security System for Cloud on Hadoop Framework" International Journal of Computer Applications (0975 – 8887) National Conference on Knowledge, Innovation in Technology and Engineering (NCKITE 2015) page no. 3-4.

[3]Mikin K. Dagli, Brijesh B. Mehta "Big Data and Hadoop: A Review" IJARES Vol. 2, ISSUE 2, Feb. 2014 page no. 193.

[4]Mr. Yogesh Pingle, Vaibhav Kohli, Shruti Kamat, Nimesh Poladia "Big Data Processing using Apache Hadoop in Cloud System" International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622" National Conference on Emerging Trends in Engineering & Technology (VNCET-30 Mar'12) page no.475-477.

[5]Rakesh Kumar, Bhanu Bhushan Parashar, Sakshi Gupta, Yougeshwary Sharma, Neha Gupta " Apache Hadoop, NoSQL and NewSQL Solutions of Big Data" Volume 1, Issue 6, October 2014. Impact Factor: 1.036, Science Central Value: 10.33 page no.29.