International Journal of Latest Trends in Engineering and Technology
Vol.(10)Issue(3), pp.236-238
DOI: http://dx.doi.org/10.21172/1.103.40
e-ISSN:2278-621X

EFFICIENT DEDUPLICATION USING HADOOP

Manjunath R. Hudagi[1], Sachin A. Urabinahatti[2]

Abstract :In cloud computing, we found that when user uploads the same file twice with same file name it doesn't allow saving the same file .Also doesn't allows to saving file with same file name with different content.Hadoop is high-performance distributed data storage and processing system. Hadoop doesn't provide effective Data Deduplication solution. Assuming a popular video or movie file is uploaded to HDFS by one million users and stored into three million files through Hadoop replication and thus it is wasting of disk space. Through proposed system, only single file spaces are occupied namely reaching the utility of completely removing plicate files. Before uploading data to HDFS we calculate Hash Value of File and store that Hash Value in Database for later use. Now same or other user wants to upload the same file name with same content. An SHA algorithm used to calculate Hash value and verify it to HBase (HBase is called the Hadoop database because it is a NoSQL database that runs on top of Hadoop). Now if Hash Value is matched with stored hash value then it will give message that "File is already exits".
Key words: Cloud storage, Deduplication, Hadoop, Hadoop distributed file system, Hadoop database.

1. INTRODUCTION:

Cloud computing is the most in demand advanced technology being utilized throughout the world. It is one of the most significant research ideas whose application is being researched recently. One of the prominent services offered in cloud computing is the cloud storage. With the cloud storage, data is stored on multiple third party servers, rather than on the dedicated server used in traditional networked data storage. All data stored in multiple third party servers are not concern by the user and no one knows where exactly data saved.[1]

Hadoop is high-performance distributed data storage and processing system. Two major subsystems of Hadoop are HDFS (for storage), and Map-Reduce (for parallel data processing). The Data Deduplication technology is widely used in Business File Server, Database, Backup Devices or lots more storage devices. Data Deduplication is the process of identifying the redundancy in data and removing it. It is found that Deduplication technique can save up to 90% storage, dependent on applications. Deduplication has proven a highly effective technology in eliminating redundancy in backup data. [1]

Hadoop doesn't provide effective Data Deduplication solution. Assuming a popular video or movie file is uploaded to HDFS by one million users and stored into three million files through Hadoop replication and thus it is wasting of disk space. Through proposed system, only single file spaces are occupied, namely reaching the utility of completely removing duplicate files. [1]

2. LITERATURE REVIEW

In Cloud computing one interesting thing we found that when user uploads the same file twice with same file name it doesn't allow saving the same file. But when user uploads the same file content with different file name Hadoop allows uploading that file. In general same files are uploaded by many users with different name with same contents so this leads to wastage of storage space. So we provided the solution of above problem and provide Data Deduplication in Hadoop. Before uploading data to HDFS we calculate Hash Value of File and stored that Hash Value in Database for later use. Now same or other user wants to upload the same file name with same file content, our DeDup module will calculate Hash value and verify it to HBase. Now if Hash Value is matched with stored hash value then it will give message that "File is already exits".

As we are aware Hadoop has two main components: HDFS and MapReduce. HDFS Client provides interface between user and Hadoop. So when user wants to upload data in Hadoop following steps are performed:

1. HDFS Client communicates with NameNode (via heartbeat messages)
2. NameNode finds appropriate DataNode.
3. NameNode provides details of DataNode.
4. HDFS Client upload file to DataNode.
5. DataNode divides files into blocks and stores it. It makes by default Three Replicas of that file.
6. DataNode provides blocks details to NameNode. So when user wants to download that file, HDFS Client communicates to NameNode and NameNode provides details of DataNode to HDFS Client. DataNode merge the blocks and it provide file for HDFS Client.

[1] Asst. Prof. Dept. of Computer Science and Engineering, TKIET, Warananagar
[2] Asst. Prof. Dept. of Computer Science and Engineering, TKIET, Warananagar

Hadoop gives message that file already exists when user upload the same file another time. (With same file name and same file content). But one interesting thing we found that when user uploads the same file content with different File Name Hadoop allows uploading that file. In general same files are uploaded by many users (Cross user) with different name with same contents so this leads to wastage of storage space. So we provided the solution of above problem. So we made changes in HDFS Client to provide Data Deduplication in Hadoop. Before uploading data to HDFS we calculate Hash Value of File and stored that Hash Value in Database for later use. As shown in Fig. 1 in the proposed approach, the main issue to be addressed is how to identify duplications and how to prevent duplicates from uploading to HDFS. For this issue, we use SHA algorithm to make a unique fingerprint for each file and set up a fast fingerprint index in HBase to identify the duplications. HBase is Hadoop database, which is an open-source, distributed, versioned, column-oriented database. It is good at real time queries. HDFS has been used in numerous large scale engineering applications. Based on these features, HDFS as a storage system and HBase as an indexing system are used in our work. So we made Deduplication module in HDFS client. When data will be uploaded for first time to HDFS, the Deduplication module will calculate hash value of the file and store it in HBase and store the file in HDFS. When new data will be uploaded by any users, the system will calculate its hash value and will check in HBase that if the hash value already exists or not. If hash value exists, then the system will give the message that file/content already exist in HDFS and will not allow uploading file for second time and will not store any entry of that file in HBase. If hash value does not exist then the system calculates Hash value of new file and then put the entry of new file and its Hash value is stored in HBase and file is uploaded to HDFS.

3. PROPOSED SYSTEM

Hadoop is high-performance distributed data storage and processing system. Hadoop doesn't provide effective Data Deduplication solution. Assuming a popular video or movie file is uploaded to HDFS by one million users and stored into three million files through Hadoop replication and thus it is wasting of disk space. Through proposed system, only single file spaces are occupied namely reaching the utility of completely removing duplicate files. Before uploading data to HDFS we calculate Hash Value of File and store that Hash Value in Database for later use. Now same or other user wants to upload the same file name with same content. An SHA algorithm used to calculate Hash value and verify it to HBase (HBase is called the Hadoop database because it is a NoSQL database that runs on top of Hadoop). Now if Hash Value is matched with stored hash value then it will give message that "File is already exits".

4. IMPLEMENTATION OF PROPOSED SYSTEM

4.1Installation OS and Configuration

Installation of Ubuntu OS 14.4 Installation JDK 1.7
Adding a dedicated Hadoop system user hduser

4.2. Configuring SSH

Hadoop requires SSH access to manage its nodes, i.e. remote machines plus your local machine if you want to use Hadoop on it. We therefore need to configure SSH access to localhost for the hduser user we created in the previous section.

4.3 Installation of Hadoop

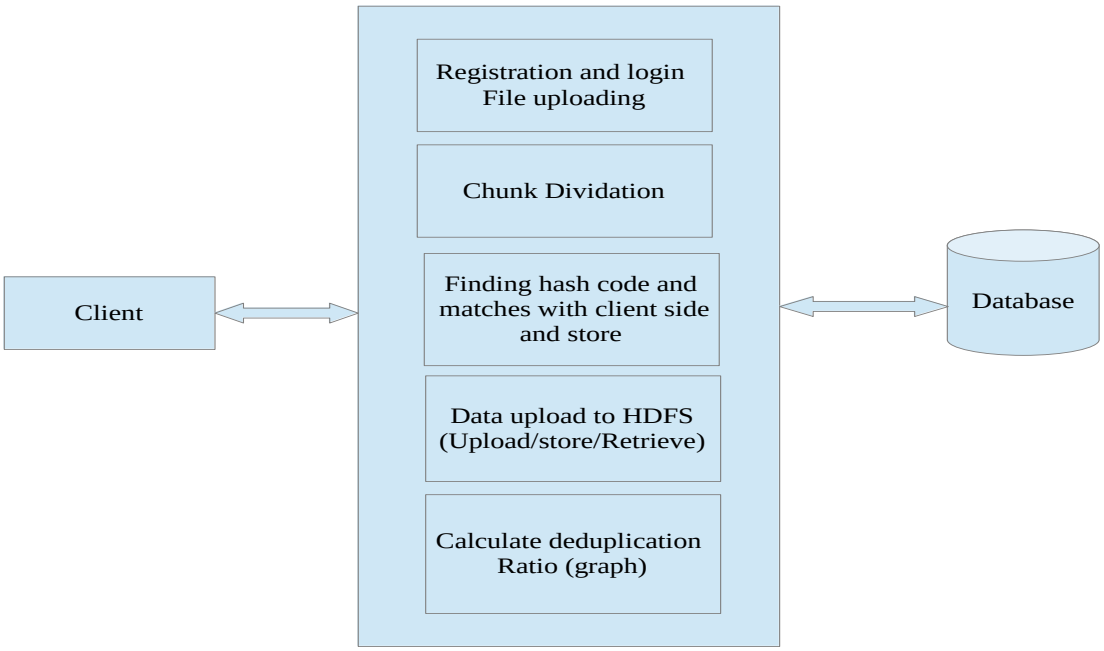
1.Download Hadoop from the Apache Download Mirrors and extract the contents of the Hadoop package to a location.
2.Update HOME/.bashrc: To configure the Hadoop.
3.Configurethefileshadoop-env.sh,conf/*-site.xml,onf/mapred-site.xml,conf/hdfs-site.xml.
4.Formatting the HDFS file system via the NameNode: The first step to starting up your Hadoop installation is formatting the Hadoop file system which is implemented on top of the local file system of your cluster.
5.Starting your single-node cluster

4.4 Modules and Architecture

4.4.1 Modules

1. User Registration and login And File uploading: In this module, we create user registration and login form to upload a file into HDFS (Hadoop Distributed File System).
Chunk Dividation: In this module, content of a file divided into small parts i.e. chunk.
Finding hash code and Matches hash code with client side & store: In this module, system find a hash code of a chunk and matches with the client side file and stored it into HDFS.
File uploads to HDFS: In this module, system uploads, store and retrieve data in HDFS.
Calculate Deduplication ratio graph: In this module, we find ratio of Deduplication files

5. ARCHITECTURE



6. CONCLUSION

The proposed system is designed to prevent duplicity of storage space in HDFS and to provide effective data storage solution. Data Deduplication is the process of identifying the redundancy in data and removing it. It is found that Deduplication technique can save up to 90% storage, dependent on applications. Deduplication has proven a highly effective technology in eliminating redundancy in backup data.

7. REFERENCES

[1] Cloud Computing: Principles, Systems and Applications Paperback –2012by Nick Antonopoulos, Lee Gilliam.
[2] Hadoop:The Definitive Guide by Tom white.