Secure File sharing Mechanism with OTP Service in Big Data Environment

Sumanth V.
Assistant Professor,
Department of CSE,
R. R. Institute of Technology,
Bengaluru, Karnataka, India

Karthik
UG Student,
R. R. Institute of Technology,
Visvesvaraya Technological University,
Bangalore, India

ABSTRACT

File sharing has been an essential part of this century. Using various applications, files can be shared to large number of users. For the purpose of storage, the Hadoop Distributed File System (HDFS) can be used. HDFS is mainly used for the unstructured data analysis. The HDFS handles large size of files in a single server. Common sharing methods like removable media, servers or computer network, World Wide Web based hyperlink documents. In the proposed project, the files are merged using MapReduce programming model on Hadoop. This process improves the performance of Hadoop by rejecting the files which are larger than the size of Hadoop and reduces the memory size required by the NameNode

Keywords: Hadoop, HDFS, Map Reduce, Name Node, Data Node, Task Tracker, Job Tracker.

INTRODUCTION:

File sharing means sending and receiving of different types of file (audio, video, and picture) within the same network or different network. File sharing is done using techniques like file storage, distribution and transmission. File sharing is the act of circulating or giving access to carefully put away data, for example, PC programs, media (sound, pictures and video), reports, or electronic books. It might be actualized through an assortment of ways. Earlier file sharing applications have a drawback, once the file sharing link is discoverable, the file can be accessed by unauthorized user. The Hadoop system itself is for the most part written in the Java programming dialect, with some local code in C and charge line utilities composed as shell contents. Despite the fact that MapReduce Java code is normal, any programming dialect can be utilized with "Hadoop Streaming" to actualize the "guide" and "lessen" parts of the client's program. Hadoop is an open-source programming structure utilized for circulated capacity and handling of dataset of enormous information utilizing the MapReduce programming model. Managing big data is being a big problem to many organizations. Due to this reason, file sharing faces many problems. Every day 2.5 Quintillion bytes are created [12]. Hadoop can easily handle by Hadoop File Distributed System (HDFS). This proposed system has an advanced security i.e. OTP service. The sender will send the required file to receiver. While opening the file, the receiver has to enter an OTP number, which he will receive from the sender. This provides the system more security as the OTP is accessible to authorized user.

LITERATURE SURVEY:

In recent times organization process large quantity of data. This data is difficult to manage by any software as it is in large volumes. An easier solution to this problem is to buy many computers which will provide more storage option and increase the processor speed. But again this method becomes costly and not practical for small organizations. Another option is to build a big cluster, store the files in it and access them as and when needed by the organization. This concept of forming clusters is used in Hadoop which is an open source framework used for the purpose of data storing and distribution. [1]NHAR: Archive and Metadata Distribution! Why Not Both? By

Dipayan Dev, Ripon Patgiri: This paper proposes Hadoop ARchive Plus (1) using sha256 as the key, which is a modification of existing HAR. It is designed to provide more reliability which can also provide auto scaling of metadata. This paper focuses on reducing the access time for a file which is more in case of the NHAR over the HAR mechanism.[1] [2] An Improved HDFS for Small File by Liu Changtong China Small file problem of original HDFS is eliminated by judging them before uploading to HDFS clusters. If the file is a small file, it is merged and the index information of the small file is stored in the index file with the form of key-value pairs. This paper focuses on increasing the efficiency of the NameNode.[2] [3] Dealing with Small Files Problem in Hadoop Distributed File System by Sachin Bendea, Rajashree Shedgeb This paper focuses on the Comparative study of possible solutions for small file problem. The CombinedFileInput Format provides the best results.

PROBLEM STATEMENT AND OBJECTIVES:

A. Problem Statement:

To deal with the huge measure of little size records it requires greater investment (for name hub). The issue here is to enhance the execution of Hadoop in treatment of little records to accomplish the coveted yield by framework by utilizing productive consolidation.

B. Objectives

To Design and Implement Merging and Indexing plan for Improving access time required for handling of extensive number of little documents. To reduce the overhead on Name hub to use information hub for handling little documents.

ARCHITECTURE AND CHARACTERISTICS:

The main concept behind the Hadoop file system is the MapReduce. MapReduce is a concept which was developed by Google to sort the large data which is in petabytes by forming clusters of this data. MapReduce mainly has two parts i.e.-Map and Reduce. Map is a transformation step in which the singular records are handled in parallel. Reduce is a summarization step in which all the related records are handled in a single entity. The concept of MapReduce is that initially all the data is divided into small parts called as chunks. These chunks are then individually processed by a maptask. Later these chunks are physically partitioned and sorted accordingly. Each sorted chunk is send to a reduce task. The figure below shows the MapReduce concept in detail. Each record will be split into several parts (chunks). A map task can run on any node in the computer and there maybe multiple map tasks running simultaneously. The map task converts each record into a key/value pair. This key value pair will be partitioned and sorted. Each partition will then have a reduce task. Each partition's keys and values will have separate reduce tasks. There may be multiple reducetask running at a given time.

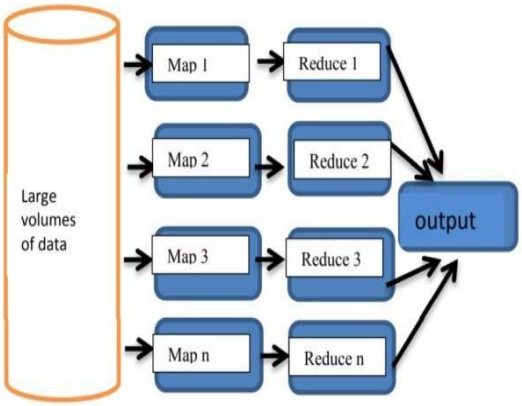


Fig.1: Map and Reduce Concept

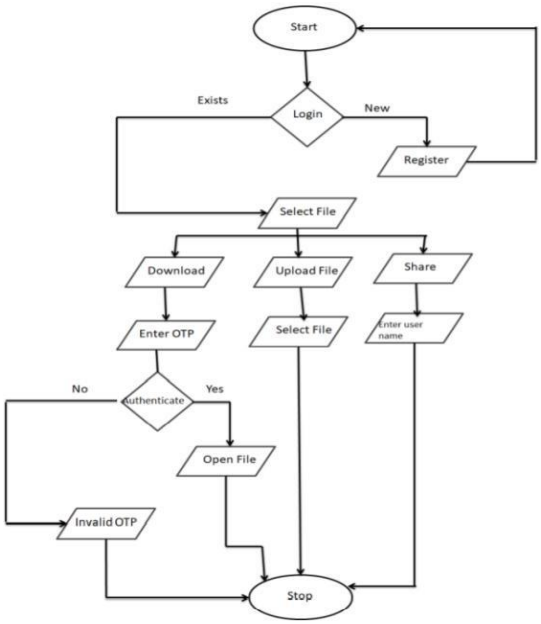
Any developer needs to provide to the Hadoop Framework 4 items i.e. a class which will read the input record and transform it into key/value pair, a map class, a reduce class and a class which will transform the key/value pair into output record. MapReduce requires a shared file system. Shared file system does not mean any file level system but it needs a distributed system with a plug in available to the framework. When HDFS will be used as a shared file system,Hadoop has an advantage that it will recognize which node has a physical copy of the input data and will check to read the data that is on the machine. If the user doesn't want the reduce task, the user need not specify the reduce class. The framework will partition the input and schedule the map task. If needed the

output of the map task will be sorted and given to the reduce task and the final output will be given to the user. The framework has two processes that handle themanagement of the MapReduce, they areTaskTracker and JobTracker. The TaskTracker provides the individual execution of map and reduce tasks. The JobTracker provides job submission, monitoring and control. The Hadoop File Distributed System: Is the one which is designed for the MapReduce concept. This HDFS system uses the MapReduce concept to sort large chunks of data, sort them and provide them as large chunks of output files. The HDFS services are provided using two main processes i.e. NameNode and DataNode. The NameNode manages the file system data byproviding the management and control services. The DataNode provides data storage and retrieval services.

ALGORITHM:

Step1. Open the windows application
Step2.If new user then register else login with user id and passwordStep3.Select the files to be uploaded
Step4.Select the file to be shared Step5.Enter the user to who file will be sent
Step6.The receiver will have to enter the OTP to open the fileStep7.Repeat the steps 4 to 6 any number of time.
Step8.Stop

FLOWCHART:



MODULES:
Admin Module: Manages the user managementUser Module:

- User will register
- User will login
- User will share the file

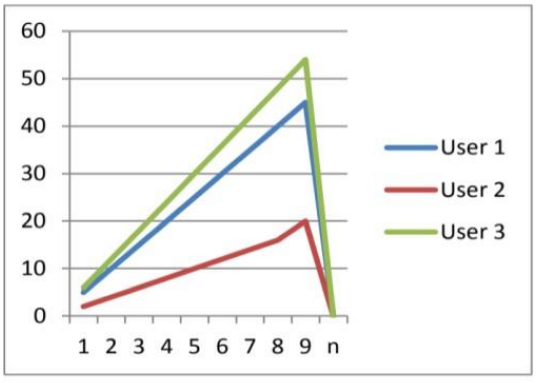


Fig.2. Graph to generate OTP for 3 users

The above graph shows the generation of OTP for 3 users. Each time a user requests for an OTP, the system uses an algorithm to generate a random OTP to a user. This OTP is time bound as it expires after the time specified in the algorithm. For example the User 1 will be given the 1 st OTP by multiplying the values of 1,10,20,30,40,50,60,2,3, n. and so on for n users.

ADVANTAGES:

- Simple to share any kind of record like picture, sound, video, application and so forth
- Does not require any USB link or Bluetooth association
- Numerous clients can get to the kind of record in the event of a little workstation where there are issues of vastplate space
- Simple organization& Gives better security

FUTURE SCOPE AND CONCLUSION:

While sharing a file the main drawbacks include storage problem and file security. These drawbacks make an application weak and ineffective over the long run. These drawbacks are solved in this project. Hadoop File system is able to store large amounts of data. This data is secured as the concept of OTP is used while accessing the file. The OTP service is put forth in such manner that the receiver has to enter the OTP which is send to him from the sender after verifying the receiver. Once the OTP is entered the file is accessible else the file cannot be opened. This paper deals with the file sharing aspect with help of HDFS using the concept of MapReduce. The Hadoop system is used to store large number of large files; it cannot handle the large number of small files.

REFERENCES:

An approach to solve a Small File problem in Hadoop by using Dynamic Merging and Indexing Scheme
An Improved HDFS for Small File by Liu Changtong China
Dealing with Small Files Problem in Hadoop 1Distributed File System
Dealing with Small Files Problem in Hadoop Distributed File System by Sachin Bendea, Rajashree Shedgeb
Dean, Jeffrey; Ghemawat, Sanjay. "MapReduce: Simplified Data Processing on Large Clusters" [5] "Google Research Publication: The Google File System".
Defining Hadoop". Wiki. Apache. Org
Google Research Publication: MapReduce".
https://www.ibm.com
June 2012
NHAR: Archive and Metadata Distribution! Why Not Both? By Dipayan Dev, Ripon Patgiri
What is the Hadoop Distributed File System (HDFS)?"
Palmer, B., “Hadoop: Strengths and Limitations in National Security Missions”, SAP National Security Services,

---