Industrial Science
Vol.1,Issue.6/Aug. 2014 **ISSN** : 2347-5420

Research Papers

INNOVATIVE METHODOLOGY FOR EFFICIENT FILE HANDLING IN HDFS

More Vaishali P.[1] and Suhas D. Raut[2]

[1]Student, Nagesh KarajagiOrchid College of Engg.& Tech.,Solapur.
[2]Project Guide, Nagesh Karajagi Orchid College of Engg.& Tech.,Solapur.

Abstract

We live in the age of data. The huge amount of data is generated that can be structured and unstructured from the different sources due to the growth of technologies and services.Big data means the data which cannot be processed using traditional processes. Big data relies on a key foundation: having access to as much data as possible, retained over the longest period of time.To handle the Big Data, Hadoop is one of the most widely used technologies. Hadoop provides a reliable collective storage system - HDFS. In an educational organization, different file presents in different data format. This paper proposes a solution to store, retrieve and update the files in the educational organization using HDFS.

KEY WORDS:

Hadoop, HDFS.

I.INTRODUCTION

With the exponential growth of an educational organization,the size of the data is growing every day. In an educationalorganization, we were capture, share more data from different sources than ever before. Data is generating from the many sources in the form of structured as well as unstructured. [Sagiroglu et al]. Unstructured data is the fastest growing type of data, some example could be imagery, sensors, telemetry, video, documents, log files, and email data files.Foran educational organization, Hadoop is the latest technology. An educational organization requires storage of all file with a category and sub category to each and every file. After category for all files, it becomes very easy to retrieve them efficiently.

II. LITERATURE SURVEY

Hadoop [Apache.Hadoop] is an open source project, which develops software for reliable and scalable distributed computing. Hadoop Distributed File System [Schvachko et al] [White] is the flagship file system component of Hadoop. Inspired by the design of proprietary Google File System (GFS), HDFS follows the pattern of write-once and read-many-times [White]. HDFS has master/slave architecture [Apache.HDFS] as shown in fig. 1. The Hadoop Distributed File System provides global access to files in the cluster [Apache.HDFS]. HDFS consists of two services namely, the NameNode and DataNode. An HDFS[White] [King] cluster consists of a single NameNode, manages the file system namespace and controls access to files by clients. DataNodes, generally one per node in the cluster, handles storage attached to the nodes that they run on. HDFS[White][King]represents a file system

namespace and allows user data to be stored in files. Internally, a file is split into one or more blocks and these blocks are stored in a set of DataNodes[King]. File system namespace operations like opening, closing, and renaming files and directories are executed by the NameNode. It also determines the mapping of blocks to DataNodes. The DataNodes are responsible for serving read and write requests from the file system's clients. The DataNodes also perform block creation, deletion, and replication upon instruction from the NameNode. The existence of a single NameNode in a cluster greatly simplifies the architecture of the system. The NameNode is the arbitrator and repository for all HDFS metadata. The system is designed in such a way that user data never flows through the NameNode.

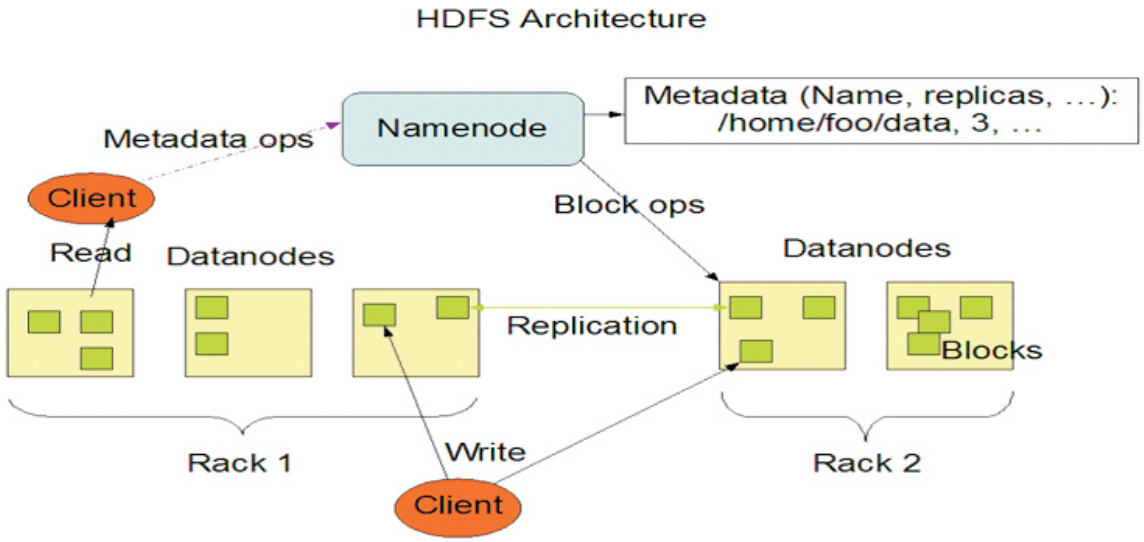

Fig.2. HDFS Architecture

The NameNode [HDFS] keeps an image of the entire file system namespace and file Block map in memory. This key metadata item is designed to be compact, such that a NameNode with 4 GB of RAM is plenty to support a huge number of files and directories [Bakshi].

III.PROJECT UNDER CONSIDERATION

In educational Organization, there are multiple types of files collected from the different sources. Thisfilesthat needs to be accessible immediately; filesthat needs to be accessed within a few seconds or minutes; and filesthat is accessed infrequently. While these types of filesplay different roles within anEducation organization, each is valuable. These different types of filesrequire different kinds of storage solutions. For handling of such heterogeneous file format we use Hadoop framework. In Hadoop, storage of different documentsis provided by HDFS (Hadoop Distributed File System). Also storage of documents according to category is one of the most important tasks. Availability of a document and need of providing a category to a document motivated to implement this work.

IV. PROPOSED ARCHITECTURE

Our proposed project contains single NameNode and DataNodes. Our project provides Graphical User Interface for storing, retrieving and updating files of educational organization having a category and subcategory. Project allows user heterogeneous files be store or retrieve in or fromHDFSaccording to their category and subcategory.

V.EXPECTED RESULT

Proposed system contains master/slave architecture with a single mastercalled the NameNode and multiple slaves called DataNodes. The files stored into HDFS arereplicated onto any number of DataNodes as per configuration,to ensure data availability.Only authorized user stores the file in HDFS,

retrieves and update the stored files.Proposed system provides reliable storage for very large and small files, on top of HDFS, which run on commodity hardware.

VI. CONCLUSION

We are in the growth area of big data. So we must understand a latesttechnology to handle Big Data and Hadoop is one of the most widely used technology. HDFS provides reliable storage file system.This paper proposes a project which consists of user friendly GUI forefficient way to store, retrieve and update file in HDFS.

REFERENCES

1.[Apache.HDFS] "HDFS Architecture Guide - Apache Hadoop"
http://hadoop.apache.org/docs/r1.0.4/hdfs_design.html
2.[Bakshi] Bakshi, K.; Cisco Syst. Inc., Herndon, VA, USA "Considerations for big data Architecture and approach" Aerospace Conference, 2012 IEEE, IEEE Conference Publications
3.[Sagiroglu et al] Sagiroglu, S., Sinanc, D. (20-24 May 2013),"Big Data: A Review" Collaboration Technology& System (CTS), International Conference
4.[Schvachko et al] K. Schvachko, H. Kuang, S. Radia, R. Chansler. "The Hadoop Distributed File System". In Proceedings of IEEE 26th symposium on Mass Storage Systems and Technologies (MSST), Incline Village, Nevada, USA, May 2010.
5.[King] "HDFS | Hadoop King" http://hadoopking.com/hdfs/
6.[White] Tom White, "Hadoop: The Definitive Guide", 2nd ed. O'Reilly Media, Yahoo! Press, Jun. 2009, pp. 41 45.