Distributed systems assignment_02

Users can access the data and files just like they can locally thanks to the Sun Network file system. In 1984, Sun Microsystems created it. Like two computers that can communicate with one other across a local network using TCP/IP, this system typically functions on all IP-based networks. The file system's data sharing, authorization, and authentication are managed through NFS. Once they have been given permission, clients can access the data on their internal drives and local systems alike. It is also a stateless protocol. The server only provides the information that a client requests. It is uninformed of the actions of the clients, the transfers, etc. The client can easily obtain the data in the event of a server crash by resubmitting the request. Its widespread use can be attributed in part to this. Many of the contemporary operating systems, including Linux, Microsoft Windows, etc., have it implemented. Additionally, it utilizes current IP infrastructure and is simple to set up. Since there is less needed to save software or other files locally, it reduces the amount of disk space needed by individual users. Essentially, it is a low-cost network file sharing solution.

Andrew File System: It is a distributed file system which provides a global namespace to all its client work stations through a group of trustworthy servers. It is developed at Carnegie Mellon University. allows the server to inform all clients with open files about any updates made to that file by another client, through what is known as a call back. allows the server to inform all clients with open files about any updates made to that file by another client, through what is known as a call back. Call- backs to all clients with a copy of that file are ensured as a call back promise is issued by the server to a client when it requests a copy of the file. There are two key software components in AFS called as Vice and Venus. Vice is the server side process that resides on top of the Unix kernel, providing shared file services to each client and Venus is the client side cache manager which acts as an interface between the application program and the vice.

Compare the architectures of HDFS Vs GFS :

Both the file systems are designed to support large data sets and intensive computing. GFS is exclusive to google only whereas HDFS is an open-source project which can be used by any company which is performing processing of big data. The GFS works on the Linux platform, whereas HDFS works on cross-platforms. Both systems use a master-slave architecture. GFS has two servers-master node and chunk servers; and After logging onto workstations that communicate inside the Distributed Computing Infrastructure, users exchange data and programs. The goal is to facilitate large-scale information exchange by reducing client-server communication. This is accomplished by moving whole files between server and client computers and caching them until the servers get a more recent version. AFS has stateful servers. The server keeps track of which files are opened by which clients unlike NFS. It has many benefits such as File security, File backups, Reliability and availability, Authentication etc. Name node and data node servers are present in HDFS. Large files containing data are

divided into manageable pieces in HDFS. There will be a default size for each block, and it can be increased as needed. Each block may be distributed stored with several copies on various nodes. There will be a single master and several chunk servers in GFS. Files will be split into chunks of a defined size, and each chunk will be duplicated on various chunk servers. Therefore, even if one of the replicated chunk servers crashes, the data will still be available in the other servers.

Common features and Differences

Some of the similarities among these file systems are: data replication, master and slaves, and in their structures. Both the systems use single master and multiple slave modes and maintain multiple copies of information for better performance.
The structure of these systems is also similar, which is a tree-structured file system consisting of copy, rename, move , delete, etc. operations.

Developed in Java.	Developed in C, C++
It has Name node and Data node	It has Master node and chunk servers
Open source Framework	Owned by Google
Default block size is 128 Mb	Default block size is 64Mb
Deleted files can be renamed into particular folder and will be removed via garbage.	Deleted files will be renamed in a hidden space and will be deleted if not in use.

Compare and analyse the databases of GFS vs. HDFS

HDFS is the main storage system used by Hadoop applications. It cannot be called a database, because usually databases store information in the form of tables with a particular defined schema. Traditional databases store only structured data, whereas HDFS can store unstructured, semi-structured, and structured data. HDFS is more of a distributed file system which stores and processes huge volumes of data sets. Such huge data files are stored across multiple machines, making them available for parallel processing.

Google created the distributed file system known as the Google File System to meet its needs for data processing. It is composed of various storage systems constructed from hardware

parts. There is a single master and several chunk servers. On local disks, chunk servers keep data in the form of Linux files. The data is split up into 64 MB-sized parts and copied throughout the network. Data will still be accessible and very dependable even if certain nodes in the system fail.

References:

- https://www.sciencedirect.com/topics/computer-science/google-file-system

- https://www.techtarget.com/searchdatamanagement/definition/Hadoop-Distributed-File-System-HDFS

- https://www.geeksforgeeks.org/andrew-file-system/

- https://www.tutorialspoint.com/hadoop/hadoop_hdfs_overview.htm