Distributed File Systems and Their Big Data

Ethan Moore

September 17, 2021

1 Introduction

Big Data is a phrase that doesn't have a clear definition but is used frequently in today's world. The U.S. Department of Commerce defined Big Data in their special publication NIST Big Data Interoperability Framework: Volume 1, Definitions that "Big Data consists of extensive datasets-primarily in the characteristics of volume, variety, velocity, and/or variability-that require a scalable architecture for efficient storage, manipulation, and analysis." [U.S.DepartmentofCommerce, 2015]. This definition shows that their is a need to store this data while keeping in mind the different characteristics. The servers that solve this issue do it by utilizing a special file system called a distributed file system which aims to solve the issues of points of failure, files and metadata storage, and system of distributing or centralizing storage.

2 Distributed File Systems and Big Data

The textbook OS Concepts defines DFS as "a file system whose clients, servers, and storage devices are dispersed among the machines of a distributed system. Accordingly, service activity has to be carried out across the network. Instead of a single centralized data repository, the system frequently has multiple and independent storage devices." [Silberschatz et al.,]. DFS's two most popular architectures is the client-server model and the cluster-based model. The client-server model uses one or more servers to deliver files through the network to clients. The cluster-based model uses a

metadata server that is communicated to by the client and then the metadata server communicates with the chunk servers to deliver the clients their files. These different architectures make storing Big Data different. Component failure safety, metadata and file storage, and whether the data is centralized or distributed is handled differently between the two DFS architectures.

3 Strategies and Methods to store Big Data

The storage of Big Data on a client-based model doesn't copy the data multiple times. The cluster-based model however does create multiple copies. The google file system paper says "For reliability, each chunk is replicated on multiple chunkservers." [Ghemawat et al., 2003]. This strategy is better than the client-based model since a server could be inactive yet the data would not be lost. Another strategy is how files and metadata are stored and used. The textbook OS concepts says "The server stores both files and metadata on attached storage." [Silberschatz et al.,]. While the GFS paper explains that "Clients interact with the master for metadata operations, but all data-bearing communication goes directly to the chunkservers." [Ghemawat et al., 2003]. These contrasting methods of having a master server be used for finding files instead of a the server that holds the file allows for "faster access to the data" since their are more copies that can be accessed [Silberschatz et al.,]. The storage of Big Data using DFS is decentralized. OS concepts states "Instead of a single centralized data repository, the system frequently has multiple and independent storage devices." [Silberschatz et al.,]. The Big Data can require multiple hard disk drives or solid state drives across different servers.

4 Conclusion

Big Data's characteristics of "volume, variety, velocity, and/or variability" rely on strategies for the efficient storage and retrieval of Big Data [U.S.DepartmentofCommerce, 2015]. Component failure safety, faster file access via multiple copies, and the distribution of drives strategies contribute to the characteristics efficiency. The Big Data phrase's ubiquitous use is due to the ingenious strategies and methods developed for the secure and speedy storage of Big Data.

References

[Ghemawat et al., 2003] Ghemawat, S., Gobioff, H., and Leung, S.-T. (2003). The google file system. In *Proceedings of the 19th ACM Symposium on Operating Systems Principles*, pages 20–43, Bolton Landing, NY.

[Silberschatz et al.,] Silberschatz, A., Galvin, P. B., and Gagne, G. Os concepts. pages 757–761.

[U.S.DepartmentofCommerce, 2015] U.S.DepartmentofCommerce (2015). Nist big data interoperability framework: Volume 1, definitions. page 13.