The Google File System

SOSP'03, October 19–22, 2003, Bolton Landing, New York, USA. Copyright 2003 ACM 1-58113-757-5/03/0010

A Comparison of Approaches to Large-Scale Data Analysis

SIGMOD'09, June 29–July 2, 2009, Providence, Rhode Island, USA. Copyright 2009 ACM 978-1-60558-551-2/09/06

fortnow. "Michael Stonebraker IEEE ICDE 2015 Ten-Year Influential Talk." *YouTube*, YouTube, 20 June 2015, www.youtube.com/watch?v=9K0SWs1mOD0.

Delroy Mathieson

10/27/17

the main idea of the paper you chose

- The Google File system is a distributed file system designed to be reliable and support millions of clients. GFS organizes and manipulate large data files used to provide users with infrastructure they need to create and access their data. The system must be efficient and able to handle millions of files available to clients when needed. The GFS meets the demands of the rapidly growing demand of Google data processes to support performance, scalability, reliability and availability. The Google File System implements efficiency and provides fault tolerance within the system running on inexpensive hardware able to deliver performance coherent to the design. It processes large data while being able to solve problems it faces without human interference, overcoming its challenges and creating more opportunities for successfully meeting our requirements.

How that idea is implemented

- The Google File System is implemented by using divided 64MB chunk size data blocks which are stored on a chunk servers on a Linux file. The files are able to be accessed by multiple clients and are operated using the create, delete, snapshot, open, close, read, write and record append functions. The master also stores three types of meta data, the file and chunk namespaces, the mapping from file to chunks, and the location of each chunk’s replicas, which is kept in the master key memory.

Your analysis of that idea and its implementation

- The Google File System serves as a reliable system storage technique that secures data, preventing it from being lost and is able to meet the needs of millions of clients all while preventing system failures. The amplitude of space creates a large environment that can be used endlessly with less maintenance and efficient storage.

The main idea of the comparison paper

- The comparison of approaches to Large-Scale Data Analysis compares and contrasts the MapReduce and Parallel SQL database management system. The paper observes the process of each data analysis noting its performance and how it is implemented.

How those ideas are implemented

- MapReduce functions, map and reduce, reads a set of records from an input file, filters the data and then outputs a set of records in the form of new key pairs. Then producing a split function that records into R disjoint buckets, applying a function to the key output record.
- The Parallel DBMSs are partitioned over nodes in a cluster using a system optimizer that translate SQL commands into a query plan whose execution is divided amongst multiple nodes.

Your analysis of those ideas and their implementations

- The file system MapReuce implements the file data system as it is applied similarly to that of the Google File System. The System inputs store data in cluster. Parallel DBMs translates SQL commands into a query plan, both file systems can be optimal depending on your needs and requirements.

comparison of the ideas and implementations of the two papers

- The Google File System is an efficient data storage option for data analytics. The GFS implementation is similar to that of the MapReduce system involving a master to control its queries. However, Parallel SQL even though may run queries faster handles failures not as efficiently as the GFS and requires more human interference than the other two file systems.

the main ideas of the Stonebraker talk

- Michael Stonebraker argues that relational databases aren't the solution to future database management problems. Stonebraker demonstrates how many markets such as data analyst and relational data systems are not the best option for a file system.

advantages and disadvantages of the main idea of the chosen paper in the context of the comparison paper and the Stonebraker talk

- The Google File System is a optimal file system for performance and reliability. GFS uses inexpensive hardware while self diagnosing its failures within its data system. The parallel SQL can be a reliable file system. Stonebraker brings about great points however, with rapidly improving technology and processor diversity, relational systems may be more efficient than he claims.