THE GOOGLE FILE SYSTEM

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung

{sanjay, hgobioff, shuntak}@google.com

Google, Inc.

James Sior

11/25/2013

What is Google File System?

- The Google File System (GFS for short) is a scalable distributed file system for large distributed data-intensive applications that provides fault tolerance while running on inexpensive commodity hardware.
- In other words, it's a file system used for large applications that is reliable (fault-tolerant) and runs on cheap hardware.

Google File System Implementation

- Large clusters of hundreds or thousands of computers access the file system.
- Files are stored in chunks on chunkservers. Chunk size is large (64 mb)
- There is one master that controls the chunkservers and stores important information like metadata and namespaces and oversees operations like read, write, create, etc.

Analysis of Google File System

- The Google File System is very powerful because it allows for easy storage of large chunks of data.
- It also enables multiple client machines to view and manipulate the data at the same time.
- The system is reliable because it stores all important operations in the operation log on the master. This log is backed up frequently and can be restored if an error occurs.

Advantages and Disadvantages

- **Advantages**
- Can support a very large number of machines and clients.
- Reliable storage of files, access is controlled by the master.
- Data very rarely gets corrupted.
- Data is distributed very quickly.

- **Disadvantages**
- Although data is rarely corrupted, it can become temporarily unavailable.

Real World Uses

- Google File System could be used for many real world applications that require a robust file system.
- For example, large financial institutions (Morgan Stanley, Goldman Sachs) could use GFS to store large quantities of data easily and still know it is secure.
- Overall, I think Google File System could be used by many different companies in the real world because it's so robust.