The Google File System

Christopher Maher

11/25/2013

Ghemawat, S., Gobioff, H., & Shun-Tak, S. (2003). The google file system. In Bolton Landing: ACM.

Idea of Google File System

- A scalable distributed file system
- Using large numbers of available inexpensive hardware
- Supporting Large amount of data processing workloads

Implementation

- Slightly larger 64 MB chunk sizes
- Minimum Data stored on Master
- Minimum Client to single Master interaction
- Benefits of GFS chunkserver on linux file system
- Mutations
- Rebalancing, Replication, and Recovery
- Meta Data
  - File and chunk namespaces
  - Mapping from files to chunks
  - Locations of each chunks replicas

Analysis

- Appears to be Cost Effective
- Almost Acid compatible
- Durability due to multiple Replicas
- Consistent due to Stale Replica Detection
- Atomic due to each name being handled by Master
- Fails Isolation

Advantages and Disadvantages

- Increases Fault Tolerance.
- Consumes More Raw Storage
- Relies on Extensive Logs to fix isolation issue
- Expectations of Hardware Failure
- Garbage Collection Delay

Real World Use

- Group access to a project
- Can be used to store polling information
- Storage of collected information from users
- Can be used to deal with Marist Registration