The Google File System by Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung.

A comparison of Approaches to Large-Scale Data Analysis by
A.Pavlo, E.Paulson, A.Rasin, D.Abadi, D.DeWitt, S.Madden, M.StoneBraker

“10 year test of time” award talk
by Michael Stonebraker

Presentation by Bradley Lamitie

Due on 10/20/2016

The Main Idea of “The Google Filing System”

- Their main goal was designing a database that was:
  - Cheap: They used commodity hardware to run the system on.
  - Efficient: The system needed to be able to handle hundreds of Terabytes quickly and effortlessly.
  - Flexible: The filing system had to have the ability of appending data. Not only do clients have to be able to append data, but they also have to be able to do it concurrently.

Implementation of “The Google Filing System”

- Some of the Techniques they used to implement the system:
  - Garbage collection – After a delete a file is hidden and marked as deleted, but is not deleted yet. The master disposes of the file after a set time interval past the deletion timestamp.
  - Snapshots – Allows users to branch off of the master document, so they can work on it simultaneously. Using copy on write allows the master to ensure that all changes to the document are known and copied to the master.
  - Chunk Servers – The data clusters are divided into one master node and many chunk servers. Each chunk server stores chunks of the broken down files.

Analysis of “The Google Filing System” and It’s Implementation

- It seems that Google tried to make their system as simple as possible, by having just one master with multiple replicas of a given file.
  - However, this could be an issue due to one master handling all the updates.

- Overall, the System seems to be efficient in handling large masses of data, despite the above trade-off.

Main ideas of “A comparison of Approaches to Large-Scale Data Analysis”

- In their paper A.Pavlo et al. compare Hadoop, DBMS-X and Vertica.
- The purpose of the paper is to see if using Hadoop, which uses MapReduce, is better than using traditional databases.

Implementations used in “A comparison of Approaches to Large-Scale Data Analysis”

- They found that while Vertica and DBMS-X are better for running queries quicker, while Hadoop, which uses MapReduce, is much easier to set up.
- At times Hadoop would take in data faster due to the fact that it doesn’t have to enter the information into schemas.
- Due to the way that Hadoop and other systems use MapReduce to ship data to each of the nodes, write and reread, Performance costs increase as nodes increase.

Analysis of “A comparison of Approaches to Large-Scale Data Analysis”

- It seemed to me that the Hadoop database seemed vastly overpowered by the capability of traditional databases like Vertica and DBMS-X.
  - Hadoop runs queries much slower, despite being able to import data faster.
  - RDBMS’s seem to offer a lot more at much faster rates than Hadoop, despite importing data slower.

Comparison of “A comparison of Approaches to Large-Scale Data Analysis” and “The Google Filing System”

- It is difficult to compare such a new database system like the GFS(Google Filing System) to Hadoop and traditional DBMSs.
- If you need to do just importing of data without queries or have to deal with lots of unstructured data, Hadoop seems like a good system to use.
- If you need to run queries fast and often or index files quickly I would say traditional DBMSs are you best bet.
- GFS is great because of the ability to update data from multiple clients in one place, at the same time. Also, it’s usually cheaper than traditional DBMSs

Main ideas of the Stonebraker Talk

- DBMS have a lot of room for growth, Stonebraker encourages DBMS researchers to try to explore new implementations.
- In order to make RDBMSs universal they added many things and thousands of pages of SQL spec.
- In 2005 Stonebraker determined that streaming applications cant be wedged into traditional RDBMS.
- NoSQL is only supporting 100 or so clients.
- The Streaming Market is growing, alongside other engines.
- Traditional row stores are becoming obsolete slowly.

Advantages and disadvantages of the chosen paper in context of comparison paper and Stonebraker talk

- The Google Filing System is efficient, and can deal with simultaneous editing of files.
- Google Filing System is cheap, flexible, and efficient, able to deal with hundreds of terabytes effortlessly.
- However, Google Filing System lacks what a lot of traditional DBMS are capable of, like relational views of data and replication.
- Overall, The Google Filing System is gearing up for something different from traditional DBMSs. While DBMSs are really just geared towards the past and present, The Google Filing System is preparing us for the future.