Google File System and Big Data

Frank Siderio

3/15/2016

Idea

- System for data-intensive applications
  - Runs on inexpensive commodity hardware
    - Fails often
    - Constantly self monitors
- Deployed as storage platform for:
  - Generation and processing of data
  - Research and development
- The design is created by observations from:
  - Application workloads
  - Technology environment

How it’s Implemented

- GFS cluster contains single master and multiple chunk servers and is accessed by multiple clients
- Files get divided into fixed-size chunks
  - Each chunk is identified by a 64-bit chunk handle
- Chunk servers store chunks as Linux files
- Linux machine running
  - Runs both chunk server and a client

Personal Analysis

- Extremely efficient
  - Chunks are a good way to organize data
  - Diagnostic tools: helps identify problems immediately
- Inexpensive components
  - Failing is the norm instead of the exception
- High availability
  - Fast recover
  - Efficient replication

Comparison of Approaches to Large-Scale Data Analysis Idea

- Evaluate and compares MapReduce paradigm and Parallel Database Management System
  - In terms of performance and complexity
- The performances of DBMS performed better than MapReduce

How the Idea is Implemented

- Parallel DBMS uses a relational paradigm while MapReduce structures their data in any way or even have no structure at all
- All current DBMSs use hash or B-tree indexes to improve access to data while MR does not provide built-in indexes (the programmer must implement any indexes)
- Both systems provide runtime support
- MR is mainly objected oriented programming and DBMS is mainly SQL

Personal Analysis

- DBMS performs better and provides structure so it might be a better choice
- MR provides less structure and leaves things up to the programmer to do which gives the programmer more customization
- DBMS seems better suited for big business since it provides structure and indexing
- MR seems better suited for smaller business or just personal use since it allows for more customization

Comparison of the Two Papers

- DBMS implements by using a relation paradigm and MR in any way or none at all
- GFS uses chunks that are given a unique 64-bit identifier and are stored on a chunk server running on a linux machine

Main Ideas of Stonebraker Talk

- There is a huge diversity “one size fits none”
  - Data warehouse market: Column stores
  - OLTP: Lightweight transaction systems
  - NoSQL Market: No standards
  - Complex Analytics: Business intelligence products
  - Streaming Market: Has some market share
  - Graph Analytics: Simulation in c store or array engine
- There are a lot of new ideas and implementations
- Up against “innovators dilemma”
  - Vendors selling old technology have a hard time transitioning to the new tech
  - SAP vs. Oracle

Advantages and Disadvantages of GFS

- Advantages
  - Use of chunks on a chunk server to modularize data
  - Duplication of the data is good for preventing data corruption
- Disadvantages
  - Inexpensive components creates a lot of overhead and component failures
- GFS does not use a traditional DBMS or a MR paradigm
- GFS is Stonebrakers example of a noSQL market
  - Google created its own data model and architecture based off its applications