BIG DATA PROCESSING

Robert Lynch
Dr. Labouseur

CMPT 308 – Database Management, Fall 2016
20 October 2016

GOOGLE FILE SYSTEM – MAIN IDEAS

- Largescale distributed Filesystem
- Designed for scalability, reliability, and availability
  - Anticipate component failure
  - Work with multi-gigabyte files
  - Applications and file systems designed for flexibility
  - Optimized for multiple users
  - Process Data Fast
- Based on a Hierarchical Model
  - Master and chunkservers

GOOGLE FILE SYSTEM - IMPLEMENTATION

- Data is broken into 64 MB chunks stored on a chunkserver
- Chunkservers are clustered with other chunkservers and one Master
- Chunkservers hold chunks and chunk replicas (3 times)
  - Edits (amends) are made to a chunk replica, then copied to the other replicas
- Data flow
  - Client requests from the master, master gives the client the chunkserver
  - Client then only communicates with the chunkserver

ANALYSIS OF THE GOOGLE FILE SYSTEM

- An effective file system for largescale data sites
- Data is precise and consistent
  - Using checksums
- Data is not rewritten (written over), but amended
  - Helps with edit history (Google Drive)
  - Could be a waste of space

COMPARISON – MAIN IDEA

- While MapReduce databases are not as fast as DBMS, they are much easier to set up
- DBMS are faster the MapReduce by up to 2x
- MapReduce databases are much easier to cope with hardware failure
- DBMS are easier to use functions on
    - Use much less code

ANALYSIS & IMPLEMENTATION OF DBMS vs. MR

- While DBMS are very fast systems, the ultimate choice comes is decided by who the system will be serving
  - Large data farm
  - Small cluster
- In a large data farm, fast (and accurate) results are necessary
  - Downside – Not as easy to replace failed hardware
- In a small cluster where time is not an problem, MRs are much better
- Implementation
  - 100 clustered DBMS and MR servers
  - 100 clients

GFS vs COMPARISON

- GFS is a modified MR DB
  - Allows for all the advantages of MR, with more speed
- DBMS and GFS are not specifically compared
  - Not able to see which is faster
- GFS is a real world example of MR and shows that it is a very functional DB for big data
  - Doesn’t prove its better or worse than DBMS, but shows that it as an extremely effective MR
- GFS is better than the basic MR DB used in the testing

STONEBRAKER TALK – MAIN IDEAS

- RDBMS is not the answer to modern DB needs
- The standard row/column based models do not work with large data
  - One size does not fit all
- Need DBMS with database management and statistics
  - Possible column store
  - Possible graph analytics
- Different instances require different structures

GFS in relation to DBMS vs MR & STONEBRAKER

ADVANTAGES

- Not a basic row store
  - MapReduce
- MapReduce allows for quick set up and failure response
- Modern database system

DISADVANTAGES

- Not as fast as DBMS
  - MR is not as fast, but GFS is modified
- Molded to fit Google
  - Might not work in other instances