Big Table: A Distributed Storage System for Structured Data
Google Inc.

&

A Comparison of Approaches to Large-Scale Data Analysis
Sigmod 09’

Scott DiBisceglio
CMPT 308 – Spring 2015
“Big Data Paper”

Paper 1: BigTable Main Points

- Distributed storage system providing a sample data model to clients, which would allow them to manage data control over layout, formatting, memory allocation.
- The use of Google File System to store data, which is similar to things like GoogleEarth and GoogleFinance.

Implementation of BigTable

- Data is stored over several different tables.
- All tables have a different makeup, ranging in size, data percentage, and complexity
- Three main functions;
  - Google Analytics
    - Goes over traffic patterns of websites
    - Uses two tables for data storage – raw click table and summary table
  - Google Earth
    - Provides high-resolution satellite imagery and navigation of the earth
  - Personalized Search
    - Records user searches and clicks across web searches etc.

Analysis

- Big Table has the ability to scale across thousands of servers that store data.
- Branches away from the traditional ideals behind data storage.
- Use of Tablet servers makes ease-of-access of adding and removing servers from a cluster to accept changes.

Main Idea of Other Paper

- "Shared Nothing" Collection of computers.
  - Divides data into partitions
  - Relational DBS require a structured schema and programming, while Hadoop's allows flexibility in regards to the schema.
  - Allows for creation of original and independent indexes.
  - Parallel DBMS uses knowledge of data distribution and location to their advantage

Implemented

System Install

- Parallel Databases are much harder to manage than Hadoop's
- Point of tuning differs depending on which system

- Data Loading
  - In DBMS the data is loaded on each node sequentially.
  - Load time increases as the amount of data increases
  - In MR nodes copy all data files from the disk

  - Execution
    - DBMS performs full scans of tables on data
    - MR, large use of control messages

Analysis

- Hadoop's MR has a lower start up cost
- Upon further research many users stated that SQL was difficult to use and stayed away from it when given the choice.
- Even if one was installed easier, the performance of the parallel database system was far more impressive and well liked throughout the industry.

Compare the ideas and implementation of the two

- The two papers had majorly different ideals, the first paper focused on Google's storage system while the second, focused on systems and traditional database systems

- MapReduce function was used in both papers.

Stonebraker Talk

- Pretty much everything this man said, went against what we spent the semester talking about
- Stated that everything we use today will one day be obsolete
- When doing database work you must focus on NON VIOLATE RAM, memory, high speed networks etc.
- Databases must become more durable than they are now

Advantage and Disadvantage

Advantages	Disadvantages
Supports access control	Does not use joins
High scalability	Does not use constraint checks
More efficient	Not available on Open Source
Doesn't use SQL	Costs lots of money
	Not as accurate