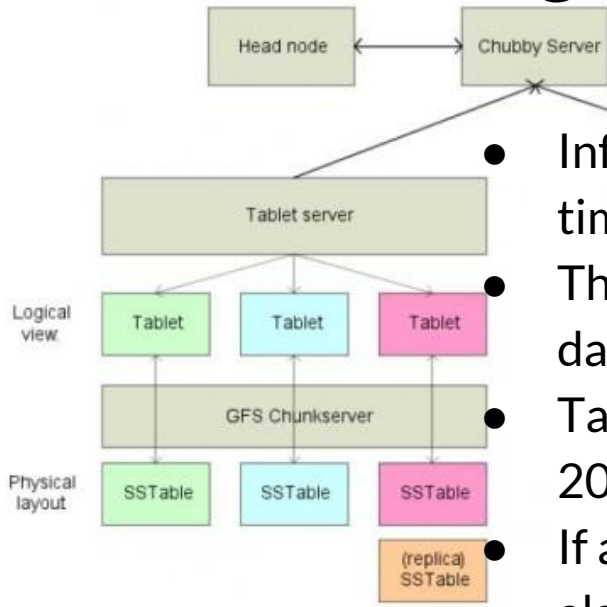
A Summary of
A Comparison of Approaches to Large-Scale Data Analysis
&
Bigtable: A Distributed Storage System for Structured Data

By Kevin Scharr
March 15th 2016

Bigtable: A Distributed Storage System for Structured Data

- Bigtable is a data storage system developed by googled to function of the cloud
- Client based data system, it can be adapted to fit the needs of clients
- Has “Timestamp” as another axis of the table, allows you to see values that used to occupy the cell as well as current
- Operates in a shared pool of machines that run a wide variety of other distributed applications

Bigtable Implementation



- Information is stored on 3 axis, row, column and timestamp
- This extra axis makes it very different from a relational database
- Tables are split into sections or “tablets” approximately 200 mb in size and stored in separate machines
- If any tablet later grows too big, Bigtable utilizes an algorithm for compressing data
- Uses columnstore to organize the universal descriptive values

Analysis

- I believe Bigtable is a very useful sounding data storage system
- The time stamp is an excellent way to retain outdated information which could still maintain usefulness
- However, if a cell or worse a column value is updated frequently this runs the risk of an explosion of storage space
- Otherwise i will always follow Google into its new world order

A Comparison of Approaches to Large-Scale Data Analysis

- Compares MapReduce and Parallel DBMS
- Schema-skeleton view of the logical structure of a database
  - MR does not require a schema, users must structure themselves or not at all
  - P-DBMS uses Hash or B-tree indexes, providing a preset structure for the database
- Data Analysis
  - Mapreduce uses the Map and Reduce functions that are written by a user process data
  - P-DBMS divies up tables between machines and each machine processes the data according to the user input

MR and P-DBMS Implementation

MapReduce

- Implementation in MapReduce largely depends on the user, since no schema is supplied, the user needs to also supply a method of processing data as well, they must write the Map and Reduce function.
- P-DBMS comes with a schema as well as a way to sort it. It utilizes multiple machines to process its data in pieces. One machine might take the main part of a query while another will handle a join of the same query. Handling the processing simultaneously rather than one after another

MR vs P-DBMS analysis

MR seems like a more versatile programming model that would allow users to customize their database to fit their needs more accurately. However it sounds easier to implement on smaller scale as any user created schemas need to be shared across other systems by the same group managed by different people. I would use it on smaller scales or in a case where customization is key.

P-DBMS seems much more streamlined. It seems user friendly and easy begin using more quickly than MR. However it seems more rigid, i would use it for simple and/or large amounts of data

Cross-Paper Analysis

All of these data systems seem useful in their own rite

- MR is for experienced users who are handling complex data, or users, allowing them to design their own schema
- P-DBMS is for more simple amounts of data but its streamlined system allows for more universal and multi-site usage
- Bigtable seems to bridge the too, but sits more closely to P-DBMS. It is a streamlined method of data management but tries to give more power to the user while also handling somewhat more complex levels of data with their time stamp index

Stonebraker Talk

Michael Stonebraker discussed the different types of fields that use Data Systems, mentioning some of the previously discussed data systems as well as others. The predominant fields for Data Systems he mentioned are

- Data Warehouse
- Transaction Processing
- Streaming Market
- Graph Analytics

He also discussed the NoSQL market, which uses a wide variety of Data models, some are copies or edits of existing models other are unique. Because of this he said this market has “No standards.”

Lastly he mentioned that Data Scientists will replace Business Analysts, as they will use databases to provide a much wider variety of information, that SQL cannot handle well. Such as regression, data clustering, predictive models, etc

Advantages and Disadvantages of Bigtable according to the talk

Advantages

- The timestamp value can help data analysts perform a wider variety of functions
- Is a column store data system so it functions better overall in the modern market
- Bigtable is a good client based data system

Disadvantages

- Is closely related to MySQL so it may not be able to handle as wide a variety as Data Analysts might want
- Does not support SQL-like queries, Bigtable does have its drawbacks in search variety as well as its advantages