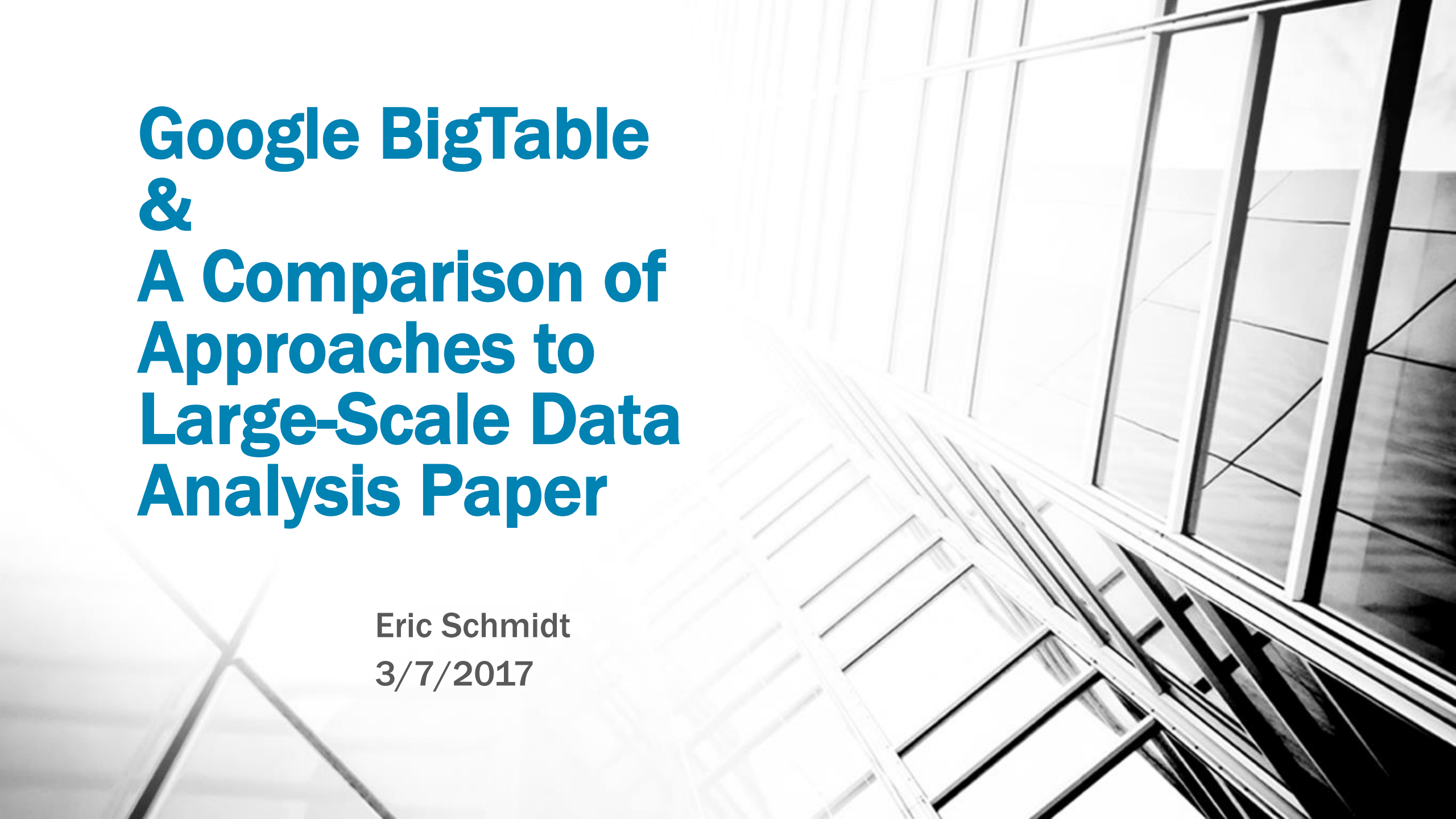
Google BigTable & A Comparison of Approaches to Large-Scale Data Analysis Paper

Eric Schmidt

3/7/2017



Google BigTable

- BigTable – Storage System that manages structured data, designed for huge amounts of data
- Focuses on how BigTable works and how google uses it do to handle their massive amount of continuous traffic
- Becomes easy to notice changes over a period of time (within a month, since last update, etc)
- Self managing system, scalable, and uses MapReduce
- Multidimensional sorted map
- Overarching columns (column families), columns, and rows, timestamps

BigTable Implementation

- Major Components: Libraries, one Master Server, Tablet Server
- A library which is linked in every client
- Master Servers – assign each tablet to tablet servers, balance server loads, handles changes in schema
- Tablet Servers – A wide range of tablets for each serer which handles read and write requests that it has. It also splits a tablet once it has grown too large
- Tablet Servers become extremely self managing through this breaking up process. They are broken up at row boundaries to keep integrity

Google BigTable Analysis

- Implementation is efficient and well organized; typical of Google products
- Self-managing tablet servers allow for effective servers due to its ability to handle sizes
- Master Servers handles changes
- Data is split into many different levels within a tablet in order to make it easier to store and read. This creates an easy way to create information

Approaches to Large-Scale Data Analysis Main Ideas

Map Reduce

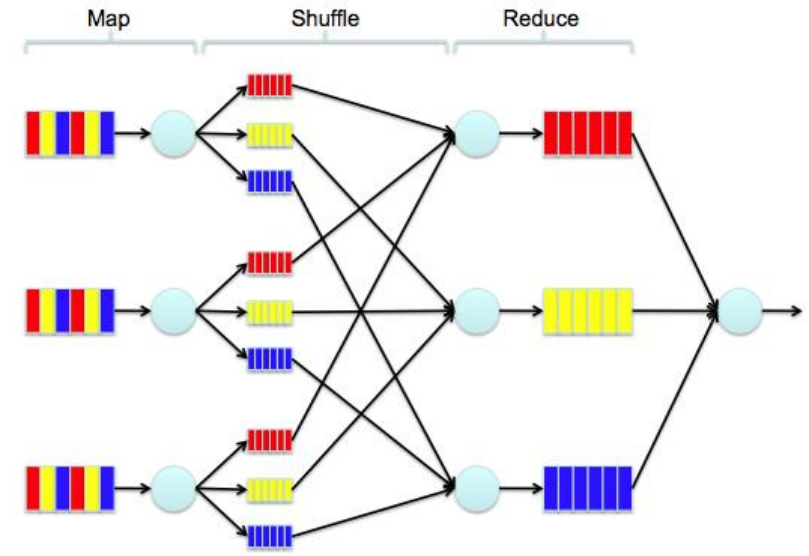
- Sorts Data into different sections/tables then filters
- Low Cost to implement, but difficult
- 2-3 Times slower in query execution than Parallel DBMS - X

Parallel DBMS

- Existed for 20 years -> more time to grow and adapt
- Very expensive to implement
- Easy to write SQL for complex analysis
- Slower to load data/information into

Approaches to Large-Scale Data Analysis Implementation

- MapReduce is typically implemented by importing the MapReduce function into your database and then applying it to your code.
- What this function does is sorts your information into different sections then searches each section individually.
- This is much slower then DBMS-X and Vertica



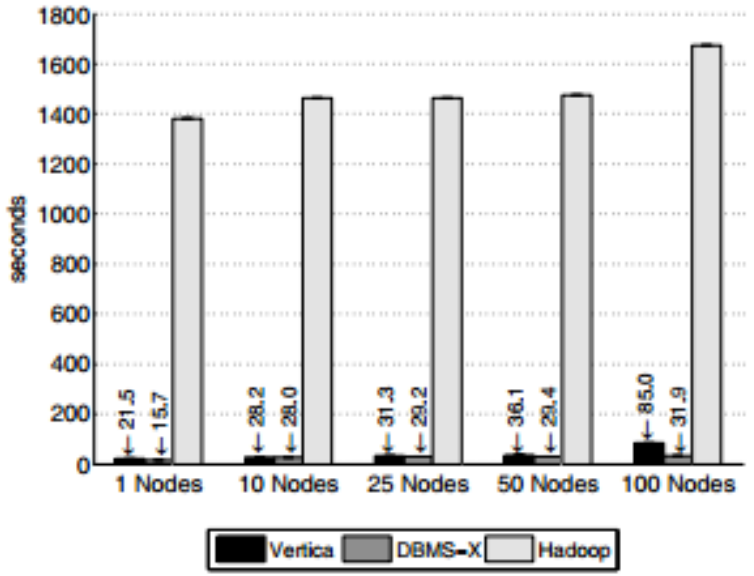
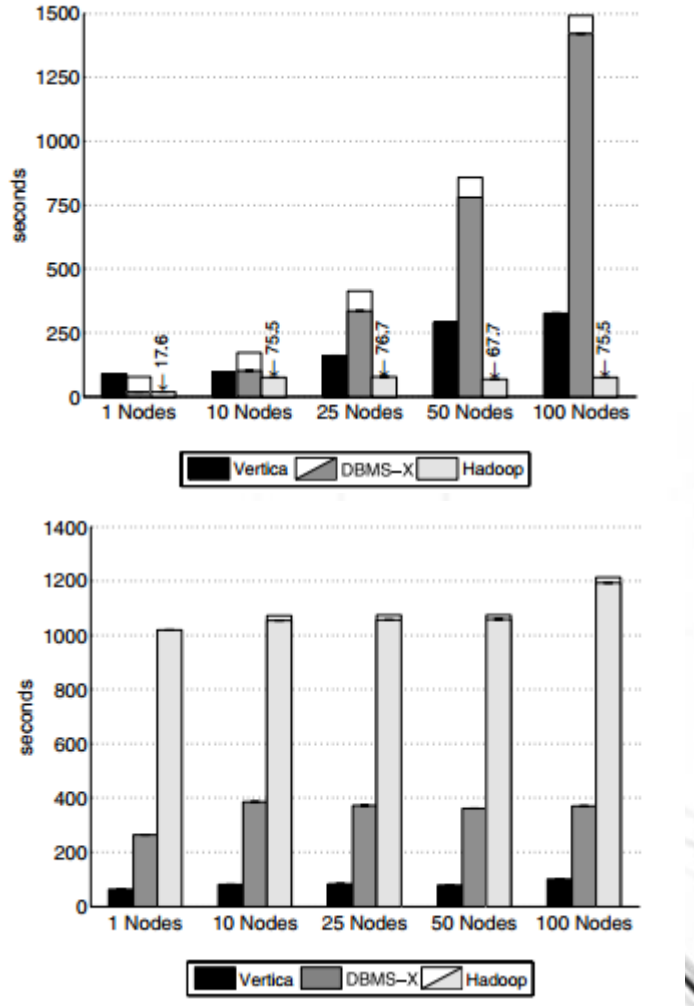


Figure 9: Join Task Results

Approaches to Large-Scale Data Analysis – Analysis

- Based on the time it took for MapReduce to complete the tasks of
  - Load Times
  - Cluster Data
  - Selections
  - Aggregations
- MapReduce took more time to complete each task compared to Vertica and DBMS-X but is significantly cheaper and therefore more popular.



Comparison of BigTable and Large-Scale Data Analysis

BigTable

- System that Google uses for many of its projects to store information
- Implementation is cost effective and going through data is simple
- Its use of column families, row keys and timestamps allow for a fast data retrieval system

Large-Scale Data Analysis

- Compares MapReduce and Parallel DBMS
  - Both are doing similar functions to BigTable but MapReduce does so at a much slower rate of data retrieval
- Implementation of MapReduce is cost efficient when compared to Parallel DBMS but not as time efficient as the others

Main Ideas of Stonebraker Talk

- One type of Database engine does not meet the needs of all different kinds of databases
- There are a large number of available database engines
- Traditional Row stores are not effective in Data Warehouse, OLTP, NoSQL, Streaming and Graph Analytics Markets
- During the 80's and 90's people believed in a one size fits all methodology
  - DOES NOT WORK ANYMORE! -says Stonebraker
- It may become difficult to introduce new engines into the market as powerhouse ones like Oracle may just adapt quick enough

Stonebraker on Google BigTable with MapReduce

Advantages

- MapReduce is going for cost efficient data sorting and retrieval like Google BigTable
- Dynamic and not trying to fit into one size for all data engine
- Adaptable to changing sizes through the tablet splitting self-managing technique

Disadvantages

- Google BigTable's massive size creates a much smaller space for innovation in database engines
- Can pigeon hole DBMS researchers into only developing BigTable further instead of potentially creating more diverse engines.