

Hive – A Petabyte Scale Data Warehouse Using Hadoop

Facebook Data Infrastructure Team. Hive – A Petabyte Scale Data Warehouse Using Hadoop. Facebook. Web. 7 May 2015

A Comparison of Approaches to Large-Scale Data Analysis.

Stonebraker, Michael, et al. "A comparison of Approaches to Large-Scale Data Analysis. 7 May 2015.

Morgan Baker

May, 7, 2015



Main Idea of Hive

- The size of business data is growing too rapidly, how will we contain it?
- Hadoop is a open source map reduce implementation
 - However, it's not easy to use
- Hive uses Hadoop but makes it easier to use and more reusable.



Where is it?

- This implementation uses HiveQL, a declarative language of SQL
- Allows for tables, arrays, maps, and the like
- Includes a system catalog for schemas and statistics
- Facebook uses it to contain over 700TB of user data.
- Also used by Yahoo!

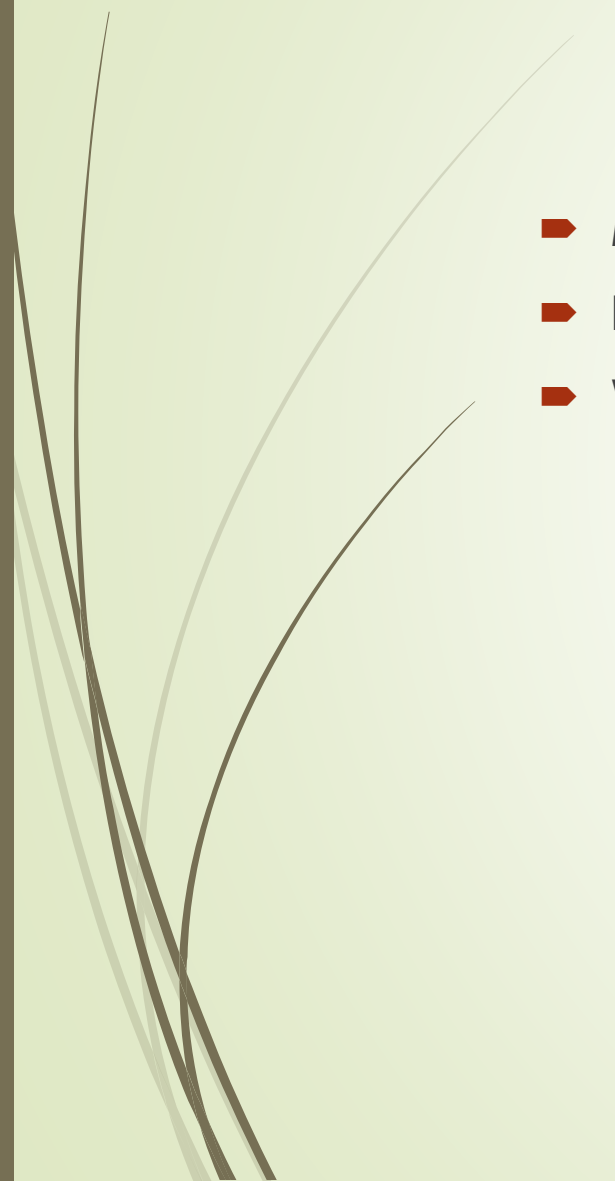


Analysis

- Seems to have more productivity
- Don't think the in unavailability of inserts has caused a problem yet, so that helps it for now, but will hurt later
- Seems to work well with Hadoop
 - Also makes less work for the DBA



Comparison Paper

- Map Reduce paradigm has gained popularity
 - But SQL DBMS has worked for so long
 - Which one works better?
- 

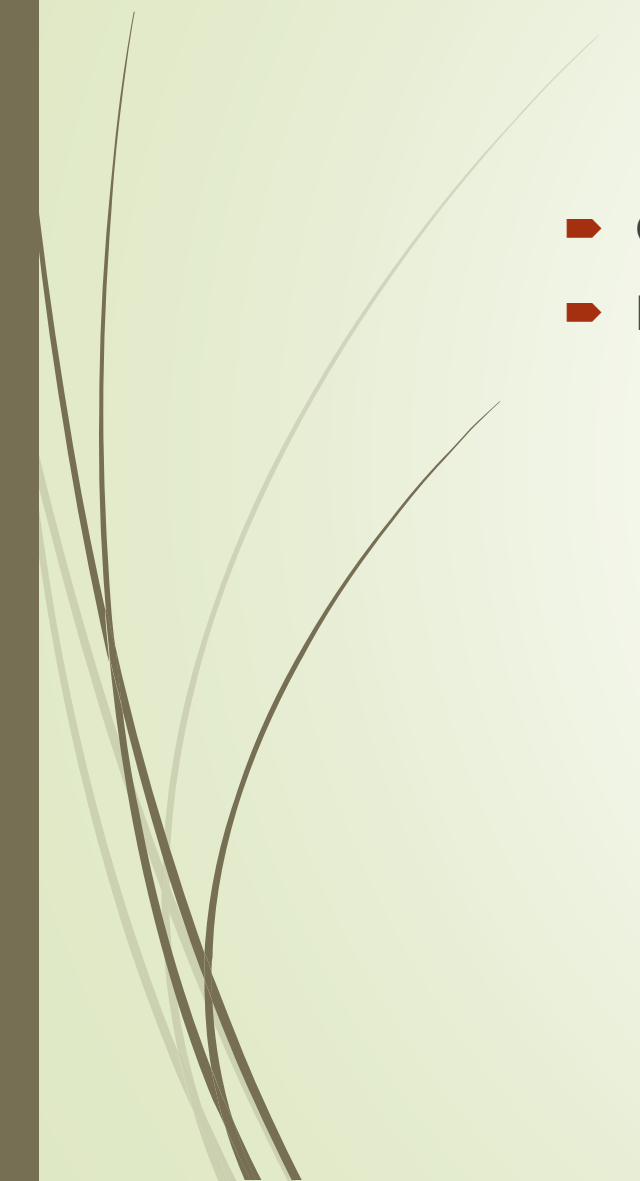


Implementation of Comparison

- The authors of the papers performed a test
- Given 100 nodes, test the performance of both methods
- After the tests, the authors found that DBMS's took longer, but also had better performance.

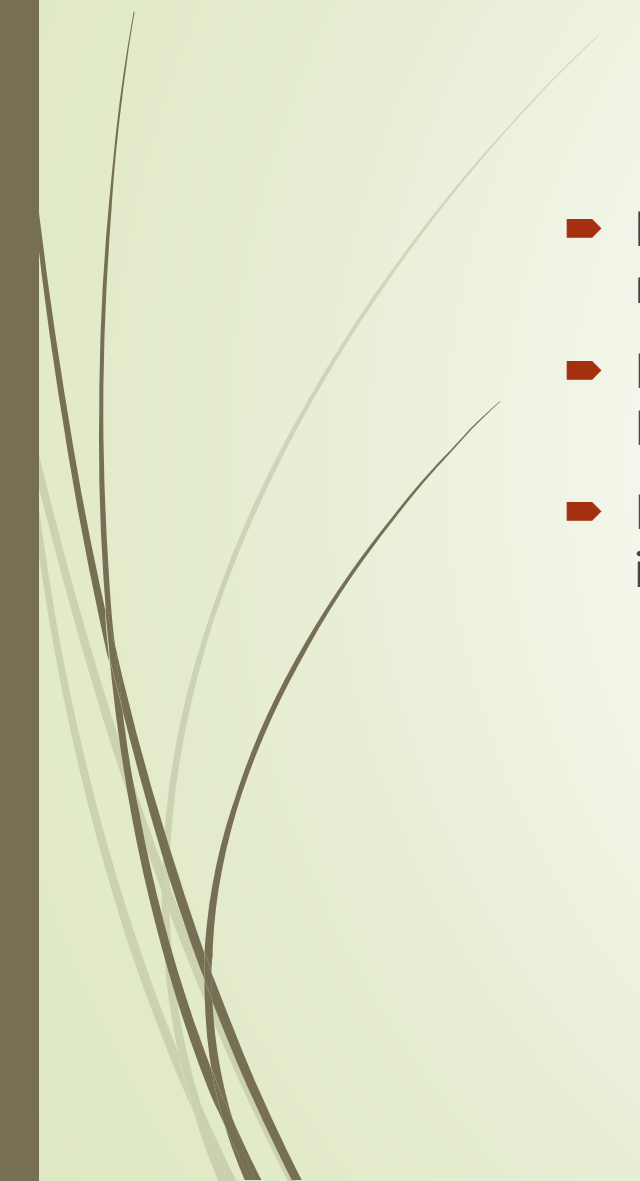


Analysis

- Considering the findings, it was definitely an interesting experiment.
 - However, the comparison seems like something is missing.
- 



Comparison of Both Papers

- I read the Hive paper first, so when reading the comparison paper, I noticed something.
 - Hive tries its darnedest to bring the best aspects of both SQL and the Hadoop implementation
 - Even though this is the case, Hive does not use a parallel DBMS system, but instead uses the MapReduce implementation.
- 



Main Ideas of Stonebraker's Lecture

- The whole time we've been trying to find the one size everyone will love.
- However, after ten years , Stonebraker claims that it doesn't exist.
 - There's a reason for why we have multiple implementations available



Analysis of Hive in the Scope of Comparison and Stonebraker

- It is possible that Stonebraker has the right idea that there is no right answer
 - Hive, a MapReduce System, could work for businesses like Facebook who deal in almost instantaneous periods of time
 - Or the Parallel DBMSes could work for organizations who need a high performance, like governments.
- It kind of feels like setting the graphics for a game.
 - If you want speed, and don't really need peak performance all the time, go with Hive
 - If you want performance and you're ok with a loss of speed, take the Parallel DBMS approach.