1. HIVE – A PETABYTE SCALE DATA WAREHOUSE USING HADOOP
2. A COMPARISON OF APPROACHES TO LARGE-SCALE DATA ANALYSIS
3. MICHAEL STONEBREAKER ON HIS 10-YEAR MOST INFLUENTIAL PAPER AWARD (ICDE 2015)

BY: DIETRICH MOSEL | NOVEMBER 3RD, 2017

Pavlo, Andrew, et al. "A comparison of approaches to large-Scale data analysis." *Proceedings of the 35th SIGMOD international conference on Management of data - SIGMOD 09*, 2009, doi:10.1145/1559845.1559865.

Stonebreaker, Michael , and Ugur Cetintemel. *One Size Fits All - An Idea Whose Time Has Come and Gone (2005)*. Database Group MIT/Brown, kdb.snu.ac.kr/data/stonebraker_talk.mp4.

Thusoo, Ashish, et al. "Hive - a petabyte scale data warehouse using Hadoop." *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, 2010, doi:10.1109/icde.2010.5447738.

MAIN IDEA BEHIND HIVE

- Built in January 2007, open source in August 2008
  - Size of data sets is growing and requires new and powerful warehousing solutions (old solutions are becoming more expensive)
  - MapReduce programming is low level and demands developers who can write custom programs (time consuming and hard to maintain)
  - As a result, Hive was created which can support large and increasing data while providing users the opportunity to plug in custom scripts
  - Bring familiar concepts of tables, columns, partitions and other features of SQL to unstructured world of Hadoop
    - Maintain the flexibility and extensibility of Hadoop

SEAMLESS IMPLEMENTATION

- Popular within Facebook so implementation was direct and popular
  - Applicable to a multitude of jobs from simple summarization to business intelligence and machine learning
  - Hive is meant to cater to diverse applications and users which can scale to different situations
  - Easy implementation as most people are familiar with SQL syntax and concepts like tables, columns, rows and partitions
    - Supports basic primitive types in addition to more complex ones such as maps, lists and structs
  - Jobs that would normally take more than a day to complete can now be executed within a few hours because of the partnership between Hadoop and Hive (community hardware)

ANALYSIS OF HIVE & ITS IMPLEMENTATION

- A useful and scalable program that addresses a major need in the marketplace
- Hive provides users with the flexibility to incorporate data into a table without having to transform it
  - Saves a substantial amount of time
  - Users can edit and create complex scripts, making it appealing to multiple groups
  - Hive query language (HiveQL) encompasses part of the SQL language making it a familiar and seamless transition for those who wish to use it
- System can be used by the most novice or advanced user with new users being able to use Hive after only an hour of training

MAIN IDEA OF COMPARISON PAPER

- There is a considerable amount of enthusiasm for MapReduce platforms which rival parallel DBMS
  - Stonebreaker and his team seek to understand and analyze the differences between MapReduce and parallel database systems in reference to how they perform large-scale data analysis
  - MapReduce and parallel DBMS are compared in terms of performance and development complexity
  - Paper includes tests run on an open source version of MR as well as on two parallel DBMS
  - They speculate causes of performance differences and debate future implementation concepts

IMPLEMENTATION (OR THE DEBATE OF IT)

MapReduce

- Attractive…
  - Simplistic model (Map and Reduce)
  - Users can express sophisticated programs
  - Deep interest from educational community
- Undesirable…
  - Few data sets require the 1,000 nodes that MR provides
  - Lack of constraints may not be conducive for long-term and large projects
  - MR programmer must perform data distribution tasks manually

Parallel DBMS

- Attractive…
  - Robust, high performance computing platforms (Microsoft SQL, DB2 and Oracle)
  - SQL DBMSs were significantly faster and required less code for each task
  - At 100 nodes, the two parallel DBMS' run 3.1 to 6.5 faster on a variety of tasks
- Undesirable…
  - Took longer to tune and load data
  - All DBMSs require data to conform to a well-defined schema
  - Does not significantly minimize the amount of work that is lost in the event of a hardware failure

ANALYSIS OF THE COMPARISON & THEIR IMPLEMENTATIONS

- MapReduce and Parallel DBMS both have their pros and cons
  - MapReduce is new in the minds of some and is simple enough for those who are new to it, while users who wish to utilize it for more complex tasks can
  - Parallel DBMS are backed by major companies on reliable platforms which are significantly faster than any MR and require less code
- Overall, people are working on improving the functionality of MapReduce and commercial as well as open-source systems which will see developments in terms of the parallelization of user-defined functions
- Both systems are moving toward each other (Greenplum and Asterdata)

IBM DB2

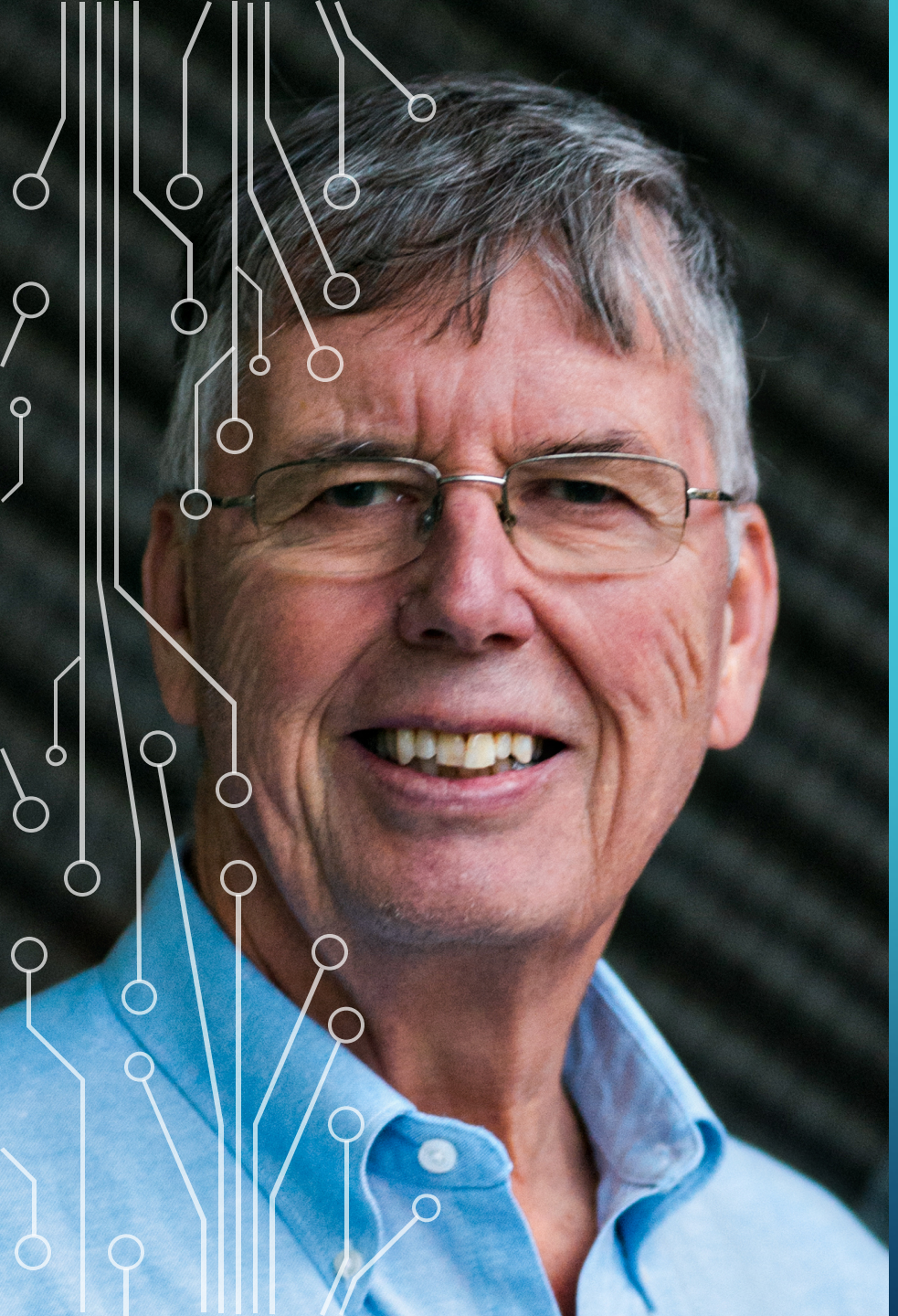
Microsoft® SQL Server®

ORACLE DATABASE

COMPARING THE TWO PAPERS

- Both papers present good arguments with reputable opinions and supporting data

1. The Hive paper argues that companies data capacities are expanding and they need new data warehousing options that can grow and cater to their constantly changing needs
2. The comparison paper admits that MR systems are important and developing rapidly but at the same time are slower as compared to parallel DBMS which are robust and high-performing systems backed by reputable firms



STONEBREAKER'S MAIN IDEAS

- Data warehouses will all soon have column stores; they are faster
- OLTP (transaction processing) databases are not that large
  - Put data into main memory as it is very cheap
- NoSQL market is almost non-existent with 100 or so vendors and no standards
- Data scientists will replace business analysts and row stores will fade in market share
- The streaming market is alive and well, not based on traditional row stores as OLTP engines have greater market share
- Graph analysts simulate in column stores & array engines but not in row stores
- There is diversity among engines which are all oriented toward specific verticals or applications (One size fits none)
- Industry elephants will have a hard time morphing into selling new systems without losing market share
- Now is a great time to be a database researcher as there will be a lot of new implementations in coming years
- We were dead on our feet from belief that one size fits all in 80s and 90s but we have since evolved

ADVANTAGES & DISADVANTAGES OF HIVE IN CONTEXT OF COMPARISON PAPER & STONEBREAKER TALK

- Hives advantages are that it scales in a cost effective manner, maintains simplistic SQL syntax while giving users the option to input more complex scripts
- In terms of the comparison paper, Stonebreaker and his team believe Hive is useful and up-incoming in the industry, however it needs work
  - Slower than parallel DBMS systems
  - Customers may not need the 1000 node capacity
  - Programmer must perform some manual tasks which waste time
- In his talk, Stonebreaker initially starts to tear parallel DBMS apart but at the end gives us a glimmer of hope
  - In relation to Hive, Stonebreaker believes there is a transition going on from new to old that will suit new players
  - There is room to improve the industry and programs like Hive may do it