Chosen Paper: "Building a High-Level Dataflow System on top of Map-Reduce: The Pig Experiment" *Alan F. Gates, Olga Natkovich, Shubham Chopra, Pradeep Kamath, Shravan M. Narayanamurthy, Christopher Olston, Benjamin Reed, Santhosh Srinivasan, Utkarsh Srivastava; VLDB '09 August 24-28, 2009, Lyon, France © 2009.*

Required Paper: "A Comparison of Approaches to Large-Scale Data Analysis" **Andrew Pavlo** *Brown University* [pavlo@cs.brown.edu], **Erik Paulson** *University of Wisconsin* [epaulson@cs.wisc.edu], **Alexander Rasin** *Brown University* [alexr@cs.brown.edu], **Daniel J. Abadi** *Yale University* [dna@cs.yale.edu], **David J. DeWitt** *Microsoft Inc.* [dewitt@microsoft.com], **Samuel Madden** *M.I.T. CSAIL* [madden@csail.mit.edu], **Michael Stonebraker** *M.I.T. CSAIL* [stonebraker@csail.mit.edu]; This work was supported in part by NSF Grant CluE – 0844013/0844480 VLDB Endowment VLDB '09 August 24-28, 2009, Lyon, France © 2009.

Required Video: **Michael Stonebraker on his 10-Year Most Influential Paper Award** at ICDE 2015.
http://kdb.snu.ac.kr/data/stonebraker_talk.mp4

Danny Puckett
Big Data Paper
03/07/17

Slide 0

"Building a High-Level Dataflow System on top of Map-Reduce: The Pig Experiment"

Main Idea:

The Main idea of this paper is the outlining of a multi-forming of languages between SQL and Map-Reduce. It practically aims at a sweet spot between the two and attempts to be a conformant of both within their strengths to differ their weaknesses. This is done with High Level Languages like DryadLINQ, Hive, Jaql, Scope, Hadoop.

Our recommended reading inferred that we either use SQL DBSM or we use Map-Request (a two operate system of pin input computation) for Large System analysis. Yet this paper, main reason of choosing this paper, is that there is an implementational middle ground if you will, with this concept of Pig Programming.

The great take away is that you can perform in both programming sequences and utilize either or each for their respective programming prosperities. If you are leaning toward a need for implementation of a Map-Reduce form and processing of data, that is feasible. Yet if you are leaning toward a SQL DBSM that is also quite possible with this same programming implementation.

So, Pig programming I'd like to surmise is an amalgam of the best you can do with SQL DBSM's and Reduce-Map. While having the capacity to perform data analysis in either fashion, you are a much more versatile programmer and are using a programming procedures that are in nature as versatile as they possibly can be for that environment of analysis.

Slide 1

"Building a High-Level Dataflow System on top of Map-Reduce: The Pig Experiment"

How the ideas are implemented:

The idea of Pig is merely a combination of known techniques that fulfills a practical need.
The PIG technique:

Parser → Logical Optimizer → Map-Reduce Complier → Map-Reduce Optimizer → Hadoop Job Manager

There are also three modes to Pig programming:
1.) Interactive mode
2.) Batch mode
3.) Embedded mode

Interactive mode: uses two commands, Describe and Illustrate. These are similar to Map-Reduce in essence. The Describe command displays the schema of a variable, While the Illustrate command displays a small amount of example data for a variable and the variables in its derivation tree.

Batch mode: In this mode, a user submits a prewritten script containg a series of Pig commands, typically ending in 'STORE'. The semantics are identical to interactive mode.

Embedded mode: Pig is also provided as a Java library allowing Pig Latin commands to be submitted via method invocations from a Java program. This option permits dynamic construction of Pig Latin programs, as well as dynamic control flow

This gives the Pig programming development a dual process capability when it relates to data management and analysis.

Slide 2

"Building a High-Level Dataflow System on top of Map-Reduce: The Pig Experiment"

My analysis of this idea implemented:

As a semi-proficient software engineer, elementarily constructed hardware developer and a fairly computer savvy individual I first found the idea of "multi-versed" programming quite intriguing! With that said Database Management is a brand new field of scope for my vision on the computational greatness computers truely are in their entirety.