Paper Summary: Building a High-Level Dataflow system on top of Map-Reduce: The Pig Experience.

Jonathan Higgins

Alan Labouseur

CMPT 308

November 24, 2013

Gates, Alan F., Olga Natkovich, Shubham Chopra, Pradeep Kamath, Shravan M. Narayanamurthy, Benjamin Reed, Christopher Olston, Santhosh Srinivasan, and Utkarsh Srivastava. "Building a High-Level Dataflow System on Top of Map-Reduce: The Pig Experience." n.d.: 1-12.

Main Idea of Pig

-Map-Reduce Alone: is simple and scalable, but cannot perform data manipulation, perform more complex data flows, also, cannot process more than one data set at a time.

-Pig allows users to process large sets of data quickly and efficiently.

-"Pig is a high-level dataflow system that aims at a sweet spot between SQL and Map-Reduce."(Page 1)

How Was it Implemented?

-Programs are written in pig latin.

-Compile the pig latin into Map-Reduce jobs.

-Execute Map-Reduce jobs on a Hadoop Cluster.

-Four Major steps:

- -step-by-step dataflow language
- -high level transformations
- -specify schemas
- -user defined functions

Analysis of Pig

-I believe that Pig is a great idea and it is very useful because it makes high-level dataflow easier by making it more user friendly.

-Having three different types of user modes make using Pig easier for users with different skill types.

- I Think that it is great tool for a world that is reliant on the connection and interpretation of unrelated data and information.

Advantages & Disadvantages

Advantages-

-User friendly
-easy to learn
-great scalability
-optimizes Map-Reduce

Disadvantages-

-can't make use of optimized storage structures
-memory overflow issues
-lack of memory management
-restricted user defined functions

Real World Uses for Pig:

-Research

-Data warehousing

-text indexes

-Yahoo

-web searches