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

Alan F. Gates, Olga Natkovich, Shubham Chopra, Pradeep Kamath,
Shravan M. Narayanamurthy, Christopher Olston, Benjamin Reed,
Santhosh Srinivasan, Utkarsh Srivastava
Yahoo!, Inc.

Mason Crane
11/25/2013

Main Idea

- This paper's main idea is to describe the processes involved in creating the High-Level Dataflow System on top of Map-Reduce, called Pig.
- The paper explains how Pig offers "composable high-level data manipulation constructs in the spirit of SQL, while at the same time retaining the properties of Map-Reduce systems that make them attractive for certain users, data types, and workloads (1)."
- The Paper also explains that Pig was initially a research project at Yahoo, but due to its success as a prototype, management decided to convert Pig into a production system and build a team dedicated to developing Pig.

Implementation

- The development team was able to create Pig by using the Pig Latin program as input, compiling that program into one or more Map-Reduce jobs, and then executing those jobs on a given Hadoop cluster.

- Four important steps of this process are 1. Step-by-Step dataflow language 2. High-level transformations 3. specifying schemas as part of issuing a program 4. The use of user-defined functions

- Having three modes of user interaction, Interactive, Batch, and Embedded Mode, provides all levels of users to be able to use Pig efficiently.

My analysis

- I think the idea of making something as complicated as high-level dataflow easier to manage was a great idea.

- The developers of Pig realized that for the product to be beneficial, it would have to be user friendly. By using three different levels of users, as well as giving users the ability to edit things their way, through user defined functions was a very smart idea. Even better was building the program around something people already understood, Map-Reduce.

- I also found it impressive that Pig went through both a research engineering team, and a development engineering team. Trying to work with other people's ideas can be strenuous, but they were able to work together and finish Pig.

Advantages and Disadvantages

- The advantages to using Pig
  - User Friendly
  - Allows different functionalities dedicated to increasing a user’s experience
  - Impressive scalability and fault-tolerance properties
  - Optimizes Map-Reduce
  - Allows user code at any point in the pipeline
  - Easy to learn
- The Disadvantages to Pig
  - Pig is procedural, so if you want to use a declarative language, SQL would better suit your needs

Real World Uses

- Pig is being used for many things such as:
  - Data Warehousing
  - Building text indexes
  - Training collaborative filtering models for image and video recommendation systems
  - Research
  - Anyone in need of a fast iterations through algorithms.