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

Written by: Alan F. Gates, Olga Natkovich, Shubham Chopra, Pradeep Kamath, Shravan M. Narayanamurthy, Christopher Olston, Benjamin Reed, Santhosh Srinivasan, Utkarsh Srivastava

For: Yahoo! Inc.

Anthony Cali

25 November, 2013

What is Pig?

- Pig is a non-relational dataflow system.
- It eases the overhead in large datasets.
- User extendible to perform specific tasks.

How does it work?

- Pig is split into two parts
  - Pig Latin – the SQL styled language
    - ex: a = LOAD ‘data’ USING BinStorage AS (user);
  - Hadoop – the map-reduce system
    - Split into multiple phases
    - Each phase does one operation
    - Phases include: map, sort, combine, shuffle, merge/combine, and reduce

Comments:

- The system is designed for large datasets
- Map-reduce is useful for the types of datasets the system is made for
- Easy to understand syntax is a good thing
- Some datasets are better in a relational model than this system

Advantages and not:

- If the dataset is small or highly organized
  - Only small advantages over relational systems
- If the dataset is large or unorganized
  - Has a lower performance impact than a generic map-reduce system
  - Can find relations in unorganized datasets

Real implementations:

- By June 2009 60% of all ad-hoc Hadoop requests were using Pig.
- 40% of Production pipelines using Pig.
- Yahoo! search uses Pig.
- Ipreo extensively uses Pig in their SAAS platforms (presented by Alan from Ipreo).