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,
Yahoo!, Inc.

Tom Morse
11/21/13

Main idea

- Pig, high-level combination of SQL and Map-Reduce
- Map-Reduce is slow and clunky
- Controlling large amounts of data with simplicaty
- Bais of SQL spirit and Map-Reduce properties useful for users and systems
- Pig's Language called Pig Latin
- Focus of non-standard aspects of Pig
- Comparison of Pig's execution to Map-Reduce

Implementation

- Iterator Model
  - Simple single-threaded
- Push based implementation
- Extending the iterator model to avoid problems
- SPLIT and MULTIPLEX operators
- Nested programs
- Memory management

Analysis

- Data analysis easier for consumer and companies
- Yahoo backs it, potential there
- Well put together but improvements could be made
- Still very new and has yet to reach full potential

ADVANTAGES

- Bring best of both worlds together
- Impressive scalability
- New features being added
- Good performance

DISADVANTAGES

- Could use some improvements on storage structures
- Difficulties with memory management

Real-World Use Case

- Yahoo
- 60% of Ad-Hoc Hadoop jobs submitted via pig
- 40% of new Hadoop jobs coming through pig
- Futher adoption expected for newer users
- System projects have adopted pig for data processing pipelines