Pig

Alan F. Gates, Olga Natkovich, Shubham Chopra, Pradeep Kamath, Sravan Narayanamurthy, Christopher Olston, Benjamin Reed, Santhosh Srinivasan, Utkarsh Srivastava

Joseph Rowell 5/7/2014

Main idea of Pig

Pig is for analyzing large datasets. It is a high-level framework built on Hadoop. It contains a high-level language for expressing data analysis applications.

3 features:

Easy to program: Complex tasks made up of multiple data transformations are explicitly encoded as dataflow sequences.

Optimization: The way in which tasks are encoded lets the system to optimize execution automatically

Extendable: Users can create custom functions

How the idea of Pig is Implemented

The initial implementation considered functionality and proof of concept more important than performance.

Designed publicly available benchmark called ‘Pig Mix’ to measure performance.

Started as a research project at Yahoo.

Pig is open source.

Pig is implemented in Java.

Analysis of Pig and its Implementation

- The syntax is excellent.
- The ability for users to incorporate User Defined Functions (UDFs) is helpful
- Java memory management is hard. Since Pig is built on top of Java, these issues would naturally arise.
- Although the authors admit more work needs to be done, having Pig open source means that it will definitely get better.

Comparison to the ideas of Pig v. Approaches to Large-Scale Data Analysis

- Pig aims at being the ‘sweet spot’ between SQL and MapReduce. MapReduce has a simple dataflow model, but this simplicity opens the door to low-level hacking to deal with the multi-step branching in dataflows.
- Parallel models are spread throughout systems. They offer better performance than MapReduce models like Pig.
- We are starting to see the convergence of both paradigmns with new technologies like Greenplum and Asterdata.

Advantages and Disadvantages to the Idea of Pig v. Approaches to Large-Scale Data Analysis

Pig Advantages: Easy syntax, things such as joins are easily written. Pig is relatively easy to set up compared to the parallel systems. Hadoop and Pig also cost less. Great fault tolerance.

Pig Disadvantages: Small support base and things like unstructured data require user defined functions.

Large-Scale data advantages: Many systems are established and have great documentation.

Large-Scale data disadvantages: Harder to learn, less extensible (poor support for UDFs) and costly.