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

CLAUDIA ROJAS

OCTOBER 18, 2016

Wilbur

The paper:

- Introduce Pig as a compromise between SQL and Map-Reduce.
- Highlights the uniqueness of Pig compared to other SQL databases
- Explains how well Pig performs
- Gives details to the development and deployment of Pig

Peter Porker

Pig has many forms of user interface that helps the user understand it

- Interactive mode provides the user with an interactive shell that uses Pig commands
- Batch mode provides a template script of Pig commands
- Embedded mode allows Pig Latin to be submitted through Java

Pig stages:

1. The parser verifies that the program is syntactically correct.
2. Logical optimization pushes the plan created by the parser into Map-Reduce jobs.

- Stages 3- 6 carries out the program through the stages of Map-Reduce

Porky Pig

I think Pig Latin is understandable and well organized.

By providing different interactions(interactive, batch, and embedded) beginner users are able to easily understand Pig.

It is attractive because of its fast development, extensibility, protection against Hadoop, and ease of debugging.

Pig is also easy to learn through its readable code, has fast iteration because of different algorithms, and easy to use for collaboration projects.

A Comparison of Approaches to Large-Scale Data Analysis

The paper:

- Conducts experiments between MapReduce and parallel database models
- Compares the function of MapReduce to DBMSs
- Provides a background on MapReduce
- Compares how the effectiveness between MapReduce and DBMSs

Petunia Pig

Map instances use the same hash function which stores them all in the same output file.

Reduce gathers the records assigned to it and records them in an output file.

MR itself conducts how many Map instances should run and which nodes to place them in.

Parallel DBMs has two aspects: most tables are sectioned and the system uses an optimizer to translate SQL commands into a query

Parallel DBMS has three stages as opposed to MapReduce.

Babe

For the experiments conducted DBMS-X was used, which was the latest version of DBMS at the time.

Hadoop, a popular example of MapReduce, outperformed DBMS by having the shortest loading times.

During the task execution experiment Hadoop ran the slowest when it was given 1TB of Data while DBMS was a fraction faster.

The selection task results indicate that each system loading times increase when introduced to an increasing amount of nodes.

After looking though the results of the experiments it can be inferred that DBMS ran faster compared to Hadoop.

Ace

Compared to DBMS, Pig Latin is more compact which allows it to be more readable and have faster iteration.

Since Pig Latin is easy to understand it also allows for easy collaboration.

Pig seems easier to understand than DBMS because it provides templates to help users understand it.

Although in order to fully determine if Pig is faster than DBMS experiments like those in the comparison paper must be conducted.

Berkshire

Stonebraker and his team tried to make RDBMSs universal but failed.

Ten years later, Stonebraker realized that DB2, Oracle, and SQL server where obsolete and good for nothing.

Column stores in Data warehouses are faster than row stores

Data scientists will be trained to run regressions and do data clustering which will be defined on arrays rather than tables.

The elephants are in danger of losing their market share unless they change to a new engine.

Hamm

Advantages of Pig:

- Easy to understand
- Fast development speed
- Faster implementation

Disadvantages of Pig:

- Fairly new dataflow system
- Not universal
- Only Pig Latin may be used