Big Data Analysis: PIG

By:
Rosemary
Bulgarelli

12/9/2014

Paper Titles

- 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.

- A comparison of approaches to Large-Scale Data Analysis
  - BY: Andrew Pavlo, Erik Paulson, Alexander Rasin, Daniel J. Abadi, David J. Dewitt, Samuel Madden, Michael Stonebraker

Main idea of PIG

- The Pig system overcomes various limitations as a large graph database.
  - Dataflow language (Pig-Latin) for Map-Reduce
  - Data
    - Unstructured Elements
      - Web page text, images
    - Structured Elements
      - WebPage Click Records
      - Extracted Entity Relationship Models
  - Map- reduce offers
    - Scalability
    - Lower Cost
    - Easability
  - Schemas are optional
  - The great median between SQL & Map Reduce

PIG Implementation

- Three modes
  - Interactive Mode
    - User presented with interactive Shell (Called Grunt) which accepts PIG commands
  - Batch Mode
    - Submits a prewritten script with PIG commands
  - Embedded Mode
    - A java like library which contains PIG Latin commands
- PIG Mix
  - Joins
  - Grouping and Co-grouping
  - Distinct Aggregation
- Parser verifies if the program is syntactically correct. The next step is logical optimizer then map-reducer compiler which is the logical to physical compilation.

PIG Analysis: My Thoughts

- I think the name is fabulous. The sub-pig names are also very entertaining.
- After reading the paper I realized that programming is very easy.
  - The code is easy to write, understand and maintain.
- PIG has easy to create functions
  - Users can create there own functions that to do whatever they like.
- The work of SQL and PIG together make it easier for traversal between the two languages
- Overall PIG was created for the User

Comparison to Large-Scale Data Analysis

- PIG uses a “command line” just like in Hadoop which yields the best results in both loading process and task execution.
- Parallel DBMS are able to execute faster and for less money because there tables are already sorted where in Hadoop they arent so they have more nodes to go through and it has an increased start-up cost as more nodes are added.
- 2 DBMS performs tasks faster then Hadoop alone. When hadoop map-reduces it has to do both map and reduce where DBMS uses a column-store system .
- DBMS also uses SQL queries instead of command line (grunt)

Advantages and Disadvantages

- Map-Reduce in PIG doesn’t support complex N-step dataflow.
- PIG creates a “command line” for there program which creates a custom data loader program.
- PIG doesn’t use simple to read queries like DBMS which could become confusing.