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

A Comparison of Approaches to Large-Scale Data Analysis

Andrew Pavlo, Erik Paulson, Alexander Rasin, Daniel J. Abadi, David J. DeWitt, Samuel Madden, Michael Stonebraker

One Size Fits All – An Idea Whose Time Has Come and Gone

Michael Stonebraker

BY MITCHELL XANDERS ~ MARCH 15, 2016

THE PIG EXPERIENCE

- Goal to preserve simple properties of MapReduce systems while providing ability to manipulate data in the spirit of SQL.
- Allow developers to input user code at any point in the data pipeline.
  - Shy away from SQL modus operandi of importing all data into the database before manipulation.

IMPLEMENTATION

- Operates in Apache Hadoop framework.
- Extracts-Transforms-Loads data for processing.
- “Pig Latin”, the language of the platform, is influenced by Java and allows for MapReduce programming to reach the level of SQL, even utilizing User Defined Functions (UDF).
- Pig interpreter optimizes jobs inputted by user before execution.
- Data is stored in the Hadoop Date File System.

ANALYSIS

- Successful accomplishes the architect's goal of working with the best characteristics of MapReduce and SQL.
- Pig interpreter allows for flexibility of script usage in Pig Latin and execution of UDF.
- Use of streaming creates multiple pipelines for multiple outputs, which SQL is unable to do.

A COMPARISON OF APPROACHES TO LARGE-SCALE DATA ANALYSIS

- MapReduce systems are growing in popularity, but their effectiveness compared to Database Management Systems (DBMS) is up in the air.
- Benchmark testing to determine what makes each method more attractive for users to implement.

IMPLEMENTATION

- Tests were done to compare the effectiveness of the MapReduce system Hadoop with that of parallel database management systems DBMS-X and Vertica.
- Hadoop was found to be more user-friendly in setup and implementation and does an overall greater job at minimizing lost data due to hardware failure.
- The parallel DBMSs were found to be much faster at executing tasks than Hadoop, saving much more energy.

ANALYSIS

- MapReduce systems like the Hadoop framework appear to be more user-friendly than the DBMS.
- DBMS seem to be more effective at accomplishing desired tasks in less time, saving that potential time and, consequentially, energy lost.

PAPER COMPARISONS

- The MapReduce systems like Hadoop seem to operate keeping in mind how the user wants it to operate.
  - This is evidenced by the flexibility of Pig Latin in using different scripts and UDFs.
- DBMSs operate with the intention of providing the best performance in mind, accomplishing the task at hand in a straightforward manner at the cost of user-friendliness.

ONE SIZE FITS ALL

- In the earlier times of big data storage, it was commonplace to use the same relational DBMS to support all applications.
- As the market for data grows and develops, this “One Size Fits All” mentality applies to fewer and fewer markets.
  - DBMS are too heavy-weight and inflexible to deal with text search engine storage and application.
    - Google and Yahoo! developed their own MapReduce systems.
- Data warehouses are following a trend of including hundreds of attributes for every record, which makes old DBMS approaches and queries more complicated.

THE PIG EXPERIENCE: ADVANTAGES AND DISADVANTAGES

- Advantages
  - Much more user-oriented and flexible than typical DBMS
  - Better suited for keeping up with large data warehouses and growing data market
- Disadvantages
  - Not as efficient in performance as DBMS