Pregel: A System for Large-Scale Graph Rescaling

Grzegorz Malewicz, Matthew H. Austern, Aart J. C. Bik, James C. Dehnert, Ilan Horn, Naty Leiser, and Grzegorz Czajkowski

A Comparison of Approaches to Large-Scale Data Analysis

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

By: Alex Stigliano
Date: 10/25/17

The Main Idea of Pregel

- Being able to process large graphs with ease.
- Fault-tolerant and scalable platform.
- Flexibly able to express arbitrary graph algorithms.
- Takes inspiration from the Bulk Synchronous Parallel model.

Implementation of Pregel

- Designed for Google Cluster architecture.
- Fault-tolerant through the use of checkpointing.
- Master operations to keep track of everything.

Analysis of Implementation

- Allows for the making of large-scale graphs to be much easier.
- Strongly prevents mistakes from being made through fault-tolerance.
- Increases ease of work for both the master and workers.

The Main Idea of "A Comparison of Approaches to Large-Scale Data Analysis"

- MapReduce's easy use with only Map and Reduce.
- Parallel DBMS' use of nodes and optimizer for SQL to query commands.

Implementation of MapReduce and Parallel DBMS

- Files that call for MapReduce must simply be placed into Map and Reduce programs. If there are already MapReduce implementations, then the code will simply call for more complex structures of data.

- Parallel DBMS' query optimizer reduces the traffic of data over a network, while keeping a balance of its workloads.

Analysis of "A Comparison of Approaches to Large-Scale Data Analysis"

- In the case of MapReduce, its methods do make for the simplification of large scale data to be broken down.
  - Even more so with just it's two programs of Map and Reduce that are used in tangent to reduce the load.

- For Parallel DMBSs, it allows coders to achieve their goal with it's effective use of balancing and replication.
  - Along with the handling of the nodes the data is divided into.

Comparison of the Ideas and Implementations of the two papers

- Pregel: A System for Large-Scale Graph Rescaling
  - Only seems to be able to handle large-scale graphs.
  - Fault-tolerance helps keep tracks of mistakes being made in the project.
  - Very flexible.

- *A Comparison of Approaches to Large-Scale Data Analysis*
  - MapReduce and Parallel DMBS seem to be able to handle data of any kind.
  - Parallel DMBS balances, whereas MapReduce is simple and to the point.

Main Ideas of the Stonebraker Talk

- One Size Fits None
  - Systems from vendors ten years ago was not good for what it did back then, in the present time, it is completely obsolete and does not work for anything.
- New Ideas Are Still Possible
  - With non-volatile RAM, big main memory, vectorization, processor diversity, etc, it will be possible to create markets and implementations in the future that will succeed.

Advantages and Disadvantages of Pregel in context of comparison paper and Stonebraker talk

- Advantages
  - API is flexible and easy to use.
  - Good for the use of large graphs.

- Disadvantages
  - Is outdated.
  - Can *only* work for graphs.
  - In the future, will most likely be obsolete and replaced with another system that is much more efficient.