Big Data Paper Summary

Authored by: Chris Ognibene
Date: 8 May 2014

Paper 1: *Pregel - A System for Large-Scale Graph Processing*
Paper 2: *A Comparison of Approaches to Large-Scale Data Analysis*

Main Idea of Pregel

- Introduces the challenges of current graph algorithms in charting data
- Describes the computational model and implementation of Pregel graph processing
- Math and recursion-based algorithms used

Pregel Implementation

Model of Computation

- Directed graph of vertices
- Input, sequence of supersteps, output
- Parallel execution of vertices within each superstep
- Vertices “vote to halt” to end algorithm execution
- Vertices coordinate by passing messages, combiners, or aggregators
- C++ API for implementation using clustering framework or priority queue

Implementation

- Designed for Google Cluster Architecture
- Default partitioning of vertices using a hash function
- Many copies of user program on multiple machines in a clustered framework (one master copy)
- Fault tolerance handled through checkpoints, a “mean time to failure” model, and confined recovery

Analysis of Pregel Implementation

- Uses the idea of distributed computing, an effective method for systems that involve a lot of simultaneous data processing
- Perfect for graph applications (e.g. charting social network traffic) that track a lot of simultaneous behavior and also the frequencies of certain activities
- Each vertex has a unique identifier, a benefit if you want to pinpoint a specific node or behavior
- Confined recovery conserves resources by only redoing interrupted/aborted processes
- Yet, all nodes are connected together through a *master node*; this is the single point of failure.

Comparison with Data Analysis Paper

Similarities

- Cluster computing used: in this paper, with parallel database systems
- Implementation: Hash functions used
- Fault tolerance: strong
- Aggregation methods used

Differences

- Implementation language: C++ vs. MapReduce and SQL
- Computing: parallel execution of vertices vs. parallel query optimizer
- Ease of Use: parallel DBMS's is easier due to the intuitiveness and simpler syntax of SQL commands
- Functions of parallel DBMS's: HTML document processing, data loading

Advantages and Disadvantages of Pregel in Context of Data Analysis

Advantages

- Pregel seems more adaptable, scalable, and easy to modify, given that Pregel utilizes networks of nodes and it is easier to insert nodes than to create databases (more space required)
- Seems easier to debug, isolating issues at select nodes

Disadvantages

- Pregel implementation requires knowledge of object-oriented programming and data structures, while parallel DBMS methods deal mainly with SQL commands, which are easier for a wider number of users to understand.

Thanks for Viewing!