Big Data Analysis

Pregel: a System for Large-Scale Graph Processing

By: 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

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

Christopher Barnett

5/9/14

Main Idea of Pregel System

- Overcome various limitations associated with large graph databases
  - Ability to manage billions of nodes with trillions of edges
  - Scalability, add more data without sacrificing efficiency
  - Fault-tolerance, ensuring data is not lost in the event of hardware failure
  - Making the system easy to use and program for

Pregel Implementation

Design

- Designed for Google cluster architecture
- Divides graph into partitions

Process

- 1. Many copies of the user's Pregel program run on a cluster of machines
- 2. One of these copies is a master that determines the number of partitions and assigns partitions to worker machines.
- 3. The master assigns parts of user's input to each worker
- 4. Master has each worker perform a Superstep (sequence of iterations) which will loop through active vertices, in order to send messages

Fault-Tolerance

- Checkpointing
- At beginning of each Superstep, the master has workers save the state of their partitions to persistent storage

Implementation Analysis

- Provides a well structured graph for implementations such as social networks
- Improves efficiency by focusing on partitions of the graph
- Protects data from hardware failure
- Provided API can make programming simpler

Comparison to Large-Scale Data Analysis

- A parallel DBMS implementation uses the relational data model to ensure consistent, structured data with data constraints
- DBMS separates schema from the application and stores them in system catalogs which can be queried
- DBMS uses indexing by default to speed up data accessing
- Data can be accessed in a DBMS using SQL queries which are easy to write and understand

Advantages and Disadvantages

- Parallel DBMS implementations can be faster than Pregel due to indexing and other features
- Pregel/MR style implementations tend to have an easier fault-tolerance procedures at the cost of performance
- Pregel/MR style implementations also tend to have easier extensibility when compared to the older DBMS style