Big Data Paper Summary

Pregel: A System for Large-Scale Graph Processing

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

Google, Inc.

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

Brown University, University of Wisconsin, Yale University, Microsoft Inc., M.I.T. CSAIL

Summary by Tyler Cavera
5/9/14

Main Ideas of Pregel

- There are many computing problems dealing with large graphs, such as social networks or the Webgraph, with some having billions of vertices and trillions of edges.
- It is hard to efficiently process and gain information from these very large graphs using common methods alone, such as established graphing algorithms or machine distribution, and current alternatives have some faults.
- This team from Google created Pregel, a flexible programming framework that allows users to create graph algorithms.
- Pregel works by dividing the input graph into partitions, with each partition being a set of vertices and their outgoing edges, and then the partitions are distributed to multiple machines to be processed.
- Pregel is successful in making graph computations and algorithm creation simple and friendly to processing power, faults, and scalability.

Pregel Implementation

- Pregel works by establishing a master and worker machines, by which the graph will be partitioned.
- The program executes on a group of machines, and one machine will act as the master. Instead of analyzing a partition of the graph, the master coordinates and manages which partitions will go to which workers.
- Workers are assigned identifiers, addressing information, and partition assignments. The master keeps a record of this information.
- Workers constantly talk to the master, and vice versa, by using pings. If a worker doesn't receive a ping from the master after a certain amount of time, it terminates processing. If the master stops receiving pings from a worker, the master marks the process as failed. Failed processes are simply reassigned to open workers by the master.
- Users write Pregel programs in C++ .

My Analysis of Pregel

- Pregel seems to be a great system to solve the problems of analyzing large graphs. The paper acknowledges getting information about transportation routes or disease outbreak has been done for plenty of years. However, analyzing and getting insightful data about these graphs quicker and easier by using a system like Pregel can be incredibly beneficial to finding efficient or quick traffic routes or containment plans for diseases faster than before.
- With Pregel being flexible and scalable, there are many possible uses for the system. Users can, and some have already, create incredibly useful and efficient programs to successfully analyze graphs.
- Ultimately, it is up to the user to determine how to go about graph analysis, which has proven to be satisfactory among early users of Pregel.

A Comparison of Approaches to Large-Scale Data Analysis ideas

- A Comparison of Approaches to Large-Scale Data Analysis discusses and compares the architectures of two types of data analysis systems, MapReduce systems and parallel database management systems.
- Cluster computing has become incredibly popular recently, and MapReduce is one of the earliest and most well-known tool using cluster computing.
- This paper compares various system architectures and how they measure to one another with data processing. The paper concludes that the parallel systems tested performed better that MapReduce systems, using less processing power and energy.
- MapReduce isn't as rich in functions as some parallel database systems, and more effort must be put into MapReduce performance.
- However, both types of systems are important to learn from and both can always be improved.

Advantages and Disadvantages of Pregel in context of A Comparison of Approaches to Large-Scale Data Analysis

Advantages:

- Pregel is said to be more flexible and scalable than some DBMSs.
- Pregel's ability to detect and fix faults in data processing is better than parallel DBMSs.
- Arguably easier for users to set up and use Pregel than other DBMSs.

Disadvantages:

- More processing resources/time necessary for Pregel to match performance of DBMSs using less resources/time.
- Pregel program maintenance can be tedious and annoying.
- DBMSs have been around for longer, so support and user communities may be bigger and more helpful.