GiraphProposal

Giraph : Large-scale graph processing on Hadoop

Abstract

Giraph is a large-scale, fault-tolerant, Bulk Synchronous Parallel (BSP)-based graph processing framework.

Proposal

Graph processing platforms to run large-scale algorithms (such as page rank, shared connections, personalization-based popularity, etc.) have become quite popular. Some recent examples include Pregel and HaLoop. For general-purpose big data computation, the MapReduce computation model is widely adopted and the most deployed MapReduce infrastructure is Apache Hadoop. We have implemented a graph-processing framework that is launched as a typical Hadoop MapReduce job to leverage existing Hadoop infrastructure, such as Amazon's EC2. Giraph builds upon the graph-oriented nature of Pregel but additionally adds fault-tolerance to the coordinator process with the use of ZooKeeper as its centralized coordination service. Additionally, Giraph will include a library of generic graph algorithms.

Background

Giraph was initially began development as a side project at Yahoo! at the end of 2010. It was made functional in a month and then started adding various features. Development has been focused on internal customers needs until this point.

Rationale

Web and online social graphs have been rapidly growing in size and scale during the past decade. In 2008, Google estimated that the number of web pages reached over a trillion. Online social networking and email sites, including Yahoo!, Google, Microsoft, Facebook, LinkedIn, and Twitter, have hundreds of millions of users and are expected to grow much more in the future. Processing these graphs plays a big role in relevant and personalized information for users, such as results from a search engine or news in an online social networking site.

Initial Goals

At this point, most of the functionality has been implemented and we are looking to get more adoption and contributions from users outside Yahoo!. We want to ensure that performance scales and that the code is robust and fault tolerant.

Current Status

Meritocracy

Giraph was initially developed by Avery Ching and Christian Kunz beginning in December 2010 at Yahoo!. There are other developers using Giraph at Yahoo! that are making suggestions and adding code. We are reaching out to other folks at social networking companies for additional usage and development.

Community

Several groups who are interested in either joining our project or using our code have contacted us. We certainly believe that there is a lot of interest and are actively looking to improve and expand the community.

Core Developers

- Avery Ching: Wrote a majority of the code
- Christian Kunz: Wrote most of the communication code and security integration with Hadoop

Alignment

Giraph uses several Apache projects as its underlying infrastructure (Hadoop and ZooKeeper). It also builds on Apache Maven.

Known Risks

Orphaned products

There are many social networking companies that would be interested in using this graph-processing framework and we have already received interest from some of them. Yahoo! is already using this code in production and will certainly continue to use it in the future as well.

Inexperience with Open Source

While the initial developers have limited experience on contributing to open-source projects, Yahoo! as a company has a strong commitment to open-source and we have several advisors that we can ask for help.

Homogenous Developers

At this time, the project is relatively young and the developers work at only two companies (Yahoo! and Jybe). However, given the interest we have seen in the project, we expect the diversity to improve in the near future.

Reliance on Salaried Developers

Currently Giraph is being developed by a combination of salaried and volunteer time. We expect that other corporations will take an interest in this project and likely contribute with salaried developers. Some individuals will likely spend volunteer time on it as well. It is still early in their project and we are hoping for a lot of growth.

Relationships with Other Apache Products

Giraph depends on many Apache projects: Hadoop, ZooKeeper, Log4j, Commons, etc. It is built using Apache Maven.

Giraph has some overlapping functionality with Apache Hama. However, there are some significant differences. Giraph focuses on graph-based bulk synchronous parallel (BSP) computing, while Apache Hama is more for general purposed BSP computing. Giraph runs on the Hadoop infrastructure, while Apache Hama uses its own computing framework.

An Excessive Fascination with the Apache Brand

The Apache brand is likely to help us find contributors, however, our interests in Apache are primarily because the other projects that we depend on are also Apache projects and it makes sense that all this software be available from the same place.

Documentation

Currently we have little documentation, but several examples. We are working on improving this situation.

Initial Source

The initial source of the code is from Yahoo! and began development in December 2010. It is already available on GitHub at https://github.com/aching/Giraph.

Source and Intellectual Property Submission Plan

We intend the entire code base to be licensed under the Apache License, Version 2.0.

External Dependencies

The required dependencies are all Apache compatible licenses. The following components with non-Apache licenses are enumerated:

- JSON – Public Domain

Cryptography

Giraph depends on secure Hadoop that can optionally use Kerberos.

Required Resources

Mailing lists

- giraph-private (with moderated subscriptions)
- giraph-dev
- giraph-commits
- giraph-users

Subversion Directory

https://svn.apache.org/repos/asf/incubator/giraph

Issue Tracking

JIRA Giraph (GIRAPH)

Other Resources

Giraph has integration tests that can be run with the LocalJobRunner. These same tests also designed to be run on a small (even single node) Hadoop cluster. While not required at this time, it would be nice if such a resource were available.

Initial Committers

- Avery Ching, aching at yahoo-inc dot com
- Christian Kunz, christian at jybe-inc dot com
- Owen O'Malley, owen at hortonworks dot com
- Phillip Rhodes, prhodes at apache dot org
- Hyunsik Choi, hyunsik at apache dot org
- Jakob Homan, jghoman at apache dot org
- Arun Suresh, asuresh at yahoo-inc dot com

Affiliations

- Avery Ching, Yahoo!
- Christian Kunz, Jybe
- Owen O'Malley, Hortonworks
- Phillip Rhodes, Fogbeam Labs
- Hyunsik Choi, Database Lab, Korea University
- Jakob Homan, LinkedIn
- Arun Suresh, Yahoo!

Sponsors

Champion

- Owen O' Malley

Nominated Mentors

- Owen O'Malley
- Chris A. Mattmann
- Alan Gates

Sponsoring Entity

Apache Incubator PMC