Index

Welcome To Apache Giraph

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.

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 map-reduce computing model has been well adopted and the most deployed map-reduce infrastructure is Apache Hadoop. We have implemented a graph-processing framework that is launched as a typical Hadoop 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.

Giraph follows the bulk-synchronous parallel model relative to graphs where vertices can send messages to other vertices during a given superstep. Checkpoints are initiated by the Giraph infrastructure at user-defined intervals and are used for automatic application restarts when any worker in the application fails. Any worker in the application can act as the application coordinator and one will automatically take over if the current application coordinator fails.

To get started, visit: Quick Start Guide