Apache Giraph

*Functional Area:* Parallel Horizontally Scalable Data Processing:Graph

*Overview:* Apache Giraph is an iterative graph processing system built for high scalability. In other words, Apache Giraph is an [Apache](http://en.wikipedia.org/wiki/Apache_Software_Foundation) project to perform [graph processing](http://en.wikipedia.org/wiki/Graph_(computer_science)) on big data. It utilizes [Apache Hadoop](http://en.wikipedia.org/wiki/Apache_Hadoop)'s MapReduce implementation to process graphs. Giraph provides graph processing for Hadoop V2. It is used for tasks that don’t fit Map Reduce as it provides better performance for those tasks. For example, it is currently used at Facebook to analyze the social graph formed by users and their connections.

Giraph originated as the open-source counterpart to Pregel, the graph processing architecture developed at Google and described in a 2010 [paper](http://dl.acm.org/citation.cfm?id=1807184). Both systems are inspired by the [Bulk Synchronous Parallel](http://en.wikipedia.org/wiki/Bulk_synchronous_parallel) model of distributed computation introduced by Leslie Valiant. Giraph adds several features beyond the basic Pregel model, including master computation, sharded aggregators, edge-oriented input, out-of-core computation, and more. With a steady development cycle and a growing community of users worldwide, Giraph is a natural choice for unleashing the potential of structured datasets at a massive scale.

*Requirements:*  
Java 1.6

Maven 3 or higher  
ZooKeeper  
Hadoop

Yarn (2.03-alpha) or  
 Version 0.20.x

*Installation:*  
To install Hadoop go through the installation steps for Hadoop. After that execute the following steps:

su – hdadmin

sudo apt-get install git

sudo apt-get install maven

mvn –version  
  
cd /usr/local/

sudo got clone https://github.com/apache/giraph.git

sudo chown –R hduser:hadoop giraph

su – hduser

After that edit $HOME/.bashrc for user account hduser with the following line:

Export GIRAPH\_HOME=/usr/local/giraph  
  
Save and close the file. And then package GIRAPH into JAR files by running the following commands.  
source $HOME/.bashrc

Cd $GIRAPH\_HOME  
mvn package –DskipTests

The argument –DskipTests will skip the testing phase. On the first run this may take a while as Maven is downloading the most recent artifacts into the repository. The command may need to be executed a couple of times for it to succeed as the remote server may time out before the downloads are complete. On suceessful packagaing the Giraph core JAR $GIRAPH\_HOME/giraph-core/target/giraph-1.1.0-SNAPSHOT-for-hadoop-1.2.1-jar-with-dependencies.jar  
  
  
*Few useful links:*

<http://giraph.apache.org/quick_start.html>

<https://www.youtube.com/watch?v=LnwNr7MaQVA>