Graph Analysis using Hive

Description

The purpose of this project is to develop a simple graph analysis program using Apache Hive.

This project must be done individually. No copying is permitted. **Note: We will use a system for detecting software plagiarism, called**[**Moss (Links to an external site.)**](http://theory.stanford.edu/~aiken/moss/)**, which is an automatic system for determining the similarity of programs.** That is, your program will be compared with the programs of the other students in class as well as with the programs submitted in previous years. This program will find similarities even if you rename variables, move code, change code structure, etc.

Note that, if you use a Search Engine to find similar programs on the web, we will find these programs too. So don't do it because you will get caught and you will get an F in the course (this is cheating). Don't look for code to use for your project on the web or from other students (current or past). Just do your project alone using the help given in this project description and from your instructor and GTA only.

Platform

As in the previous projects, you will develop your program on SDSC Comet.

Setting up your Project

Login into Comet and download and untar project7:

wget http://lambda.uta.edu/cse6331/project7.tgz

tar xfz project7.tgz

chmod -R g-wrx,o-wrx project7

You need to create an empty metastore database first (this must be done only once):

cd

rm -rf metastore\_db warehouse

export HIVE\_HOME=/oasis/projects/nsf/uot143/fegaras/apache-hive-2.1.0-bin

export JAVA\_HOME=/lib/jvm/java

export HADOOP\_CONF\_DIR=$HOME/cometcluster

module load hadoop/2.6.0

JAVA\_HOME=/lib/jvm/java $HIVE\_HOME/bin/schematool -dbType derby -initSchema

Go to project7/example and look at the join.hql example. You can run it in local mode using:

sbatch join.local.run

Optional: Use your laptop to develop your project

If you'd prefer, you may use your laptop to develop your program and then test it and run it on Comet.

To install Hive and the project:

cd

wget https://downloads.apache.org/hive/stable-2/apache-hive-2.3.7-bin.tar.gz

tar xfz apache-hive-2.3.7-bin.tar.gz

wget http://lambda.uta.edu/cse6331/project7.tgz

tar xfz project7.tgz

Every time you login, you need to execute these:

export HIVE\_HOME=$HOME/apache-hive-2.3.7-bin

export HADOOP\_HOME=$HOME/hadoop-2.6.5

export PATH=$HIVE\_HOME/bin:$PATH

export HIVE\_OPTS="--hiveconf mapreduce.framework.name=local --hiveconf fs.default.name=file://$HOME --hiveconf hive.metastore.warehouse.dir=file://$HOME/warehouse --hiveconf javax.jdo.option.ConnectionURL=jdbc:derby:;databaseName=/$HOME/metastore\_db;create=true"

You also need to create an empty metastore database first (this must be done only once):

cd

rm -rf metastore\_db warehouse

schematool -dbType derby -initSchema

Then, to evaluate Hive commands interactively, do:

hive

Go to project7/example and look at the join.hql example. You can run it in local mode (after you setup your PATH) using:

hive -f join.hql

To run your project in local mode, do:

hive -f graph.hql --hiveconf G=small-graph.txt

Project Description

You are asked to re-implement Project #6 (a simple graph algorithm) using Apache Hive. That is, your Hive program should calculate the number of incoming links for each graph vertex and should sort the nodes by the number of their incoming links in descending order, so that the first node is the one that has the most incoming links.

An empty graph.hql is provided as well as a script to run this code on Comet. The input graphs are the same as in Project6. Note: you can access the input graph in Hive (which are passed as a parameter) as '${hiveconf:G}'.

To run it in local mode over the two small matrices do:

sbatch graph.local.run

After you make sure that your program runs correctly in local mode (the output is the same as the solution), you run it in distributed mode using:

sbatch graph.distr.run

This will process the graph on the large dataset large-graph.txt. Your results should be similar to the results in the file large-solution.txt.

Documentation

You can learn more about Hive at:

* [Hive: Getting Started (Links to an external site.)](https://cwiki.apache.org/confluence/display/Hive/GettingStarted)
* [Hive Tutorial (Links to an external site.)](https://cwiki.apache.org/confluence/display/Hive/Tutorial)

What to Submit

Submit the zipped project7 directory, which must contain the files:

project7/graph.hql

project7/graph.local.out

project7/graph.distr.out