91.673 Fall 2015 Project

Due 11:59pm Saturday December 12

I **strongly suggest that you start as early as possible** on this project. It may take longer than you expect!

1. Programming Environment

You are required to use Hadoop and Java to do this project.

Before you can get started installing Hadoop, you should have a Linux environment configured and ready to use. If you are using windows or mac, you can install a virtual machine first, i.e., VMware workstation, then install Linux under the virtual machine.

Since Hadoop developers usually test their scripts and code on a pseudo-distributed environment, which is a virtual machine that runs all of the Hadoop daemons simultaneously on a single machine, I suggest that you make sure your project can run on such a pseudo-distributed environment.

After installing Hadoop, you can install Hadoop-eclipse-plugin to compile and run your MapReduce program.

2. Database Schema

The database we use is a world database. Following is the schema for the whole database. Primary keys of tables are underlined.

City (ID, Name, CountryCode, District, population)
Country (Code, Name, Continent, Region, SurfaceArea, IndepYear, Population, LifeExpectancy, GNP, GNPold, LocalName, GovernmentForm, HeadOfState, Capital, Code2)
CountryLanguage (CountryCode, Language, IsOfficial, Percentage)

3. What you need to do

There are three data files provided, corresponding to the data of three tables, respectively. Each line of a data file corresponds to a record of the table, where fields are separated by ",". Your Java program needs to parse the text files to get the data. You should have a number of mapper processes initially (e.g., 10), each parsing a portion of the records (e.g., 1/10). Your program needs to return the results of the following four queries.

3.1 Computing Selection by MapReduce:

Find cities whose population is larger than 300,000.

3.2 Computing Projection by MapReduce:

Find all the name of the cities and corresponding district

3.3 Computing Natural Join by MapReduce

Find all countries whose official language is English.

3.4 Aggregation by MapReduce

Find how many cities each district has.

4. Logistics

4.1 Collaboration

You have plenty of time and this project is manageable by a single person. Therefore, to the benefit of your learning experience, no collaboration is allowed in this project. You must write your own program. If needed, you could choose to discuss with another student on high-level ideas, but you must indicate whom you have discussed with in your write-up. Even so, you still must understand the project and implement it by yourself. You will get no credits for this project if we find you cheating.

4.2 Submission

You need to turn in the following:

- **Write-up:** It should contain the following
  - Your full name
  - UML ID

    - E-mail address
    - Results from the printout of your program
    - A brief README to explain the program and results, what you have done and what you have not done.
- Your complete java program

Pack all the above files into a zip or tar file and email it to the professor at ge@cs.uml.edu by the due date (11:59pm Saturday December 12). Make sure you receive an email reply acknowledgement confirming the receipt of your submission and keep that email; you are responsible for ensuring the receipt of your submission by the due date. If you have specific questions about the project, you may also get some help from my student assistant Lijian Wan at: xiaowan91@gmail.com.