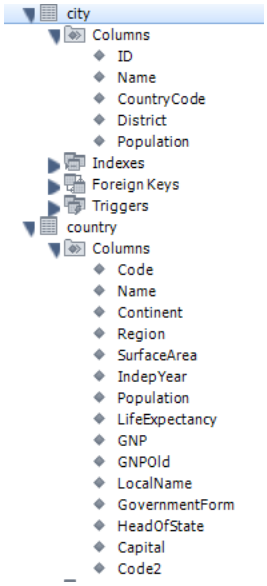
Homework #1: SQL via Hadoop MapReduce & Apache Spark

TA handling this homework: Qianmu Yu qianmuyu@usc.edu

Due: February 4, Monday

100 points

In this homework, we consider using Hadoop MapReduce and Apache Spark to answer SQL queries on a set of tables. In particular, we consider two tables: city and country, whose attributes are shown below. Note that CountryCode of city table refers to Code of country table. The sample contents of two tables are provided to you in text files (each row is record with column value separated by tabs). You can assume that files are stored in separated directory as shown in the example execution format below. Note that your program should run on additional samples (used by TA for testing and grading, not posted) of table contents.



Consider the following SQL query:

```
SELECT country.Name, count(*)
FROM city, country
WHERE city.CountryCode = country.code and city.Population >= 1000000
GROUP by country.Name
HAVING count(*) >= 3
```

1. [70 points] Write a Hadoop MapReduce program "SQLCount.java" to implement the above query.
   Execution format:
   hadoop jar SQLCount.jar SQLCount <input-dir-to-city> <input-dir-country> <ouput-dir>

   Note that each input directory contains only one text file, storing the sample content of table as described above.

2. [30 points] Write a Spark program “SQLCount.py” to implement the same query on the data set.
   Execution format:
   spark-submit SQLCount.py <input-dir-to-city> <input-dir-country> <ouput-dir>

Note that for both programs, the <output-dir> should store the results in a text file, with each line contains a country name and count, separated by a tab.