Name: Yufeng Yuan

UML ID: 01506240

E-mail address: [Yufeng\_Yuan@student.uml.edu](mailto:Yufeng_Yuan@student.uml.edu)

1. **Computing Selection by MapReduce:**
2. **Question Description**

*Find cities whose population is larger than 300,000.*

1. **Data Source**

City (ID, Name, CountryCode, District, population)

1. **Solution**

The java source code for this question is in [*SelectionByMapReduce.java*](SourceCode/SelectionByMapReduce.java)*.*

In map, first we need check if the population of this city is greater than 300,000, because we don’t need to count, so we just build a map function like: *<Name, “”>*,

Where Name is the name of city whose population is larger than 300,000, and “” is just an empty string.

In Reducer function, we just output each key.

1. **Result**

The result file was showing in directory [result1](result1/part-r-00000).

1. **Computing Projection by MapReduce**
2. **Question Description**

Find all the name of the cities and corresponding district.

1. **Data Source**

City (ID, Name, CountryCode, District, population)

1. **Solution**

The java source code for this question is in [*ProjectionByMapReduce.java*](SourceCode/ProjectionByMapReduce.java)*.*

In map, because there might be some cities belongs to same distinct, we the district should not be the key, so we create the key/value pair as <Name, District>.

In Reduce Function, we just simply output the key and value.

1. **Result**

The result file was showing in directory [result2](result2/part-r-00000)

1. **Computing Natural Join by MapReduce**
2. **Question Description**

Find all countries whose official language is English.

1. **Data Source**

Country (Code, Name, Continent, Region, SurfaceArea, IndepYear,

Population, LifeExpectancy, GNP, GNPold, LocalName,

GovernmentForm, HeadOfState, Capital, Code2)

CountryLanguage (CountryCode, Language, IsOfficial, Percentage)

1. **Solution**

The java source code for this question is in [NaturalJoinByMapReduce.java.](SourceCode/NaturalJoinByMapReduce.java)

This question we need to use two table, so I created two map function, first one is *MapCountryLanguage*, in this class, we need use the data from *CountryLanguage.txt*, first we need to check each country has a Language value equals to “English”, and IsOfficial value equals to “T”, if both situation is true, this means this country’s official language is English, then emit *<ConutryCode, “FindCountryCode”>*. Here the string “*FindCountryCode*” just means this *CountryCode* is selected.

The other Map function is *MapCountry*, in this class, we just simply output all Code and Name pair, where the Code in file Country.txt is the *CountryCode* in CountryLanguage.txt. Thus, in this mapper, we got <Code, Name>.

In reduce, we can find if the key has two values (“*FindCountryCode”, Name*), the Name value is what we want to output, if there is only one values (*Name*), that means this country’s official language is not English. So just simply check if how many values in each key, if it equals to 2, then output the name.

1. **Result**

The result file was showing in directory [result3](result3/part-r-00000).

1. **Aggregation by MapReduce**
2. **Question Description**

Find how many cities each district has.

1. **Data Source**

City (ID, Name, CountryCode, District, population)

1. **Solution**

The java source code for this question is in [AggregationByMapReduce.java](SourceCode/AggregationByMapReduce.java).

For this question, this was very similar with the official sample WordCount. In this case, the district will be the key, for each city in this district emit *<district, one>.*

In reducer, we just add those ones together and then we can calculate the number of cities in each district.

1. **Result**

The result file was showing in directory [result4](result4/part-r-00000).

**Problem:**

There is a problem makes me confused a long time is that the Text encoding problem.

I found that the original file text encoding format is Western (Mac OS Roman), but the output format is utf-8, I have tried many ways to fix that, one way is to decode the output to ISO-Latin-1 or ISO-Latin-15, but the result was just transfer those unknown characters into ‘?’. The output shows in those result directories are not decode. I hope I can solve this issues in the future.