Gender Statistics Analysis

This project deals with identifying the gender factor in higher education aligned to the geographies that make an impact. The goal of this project is to identify special programs aimed at women across the globe.

Requirements

You must find the answers through the development MapReduce algorithms for the following Business Questions:

1. Identify the countries where % of female graduates is less than 30%.
2. List the average increase in female education in the U.S. from the year 2000.
3. List the % of change in male employment from the year 2000.
4. List the % of change in female employment from the year 2000.
5. Additionally, based on your data exploration and analysis, evaluate one business factor that you consider important, and make this your own requirement.

For each business question, there is the following architectural requirements, plus additional exceptions:

- Question 1 and 3 should be done with MapReduce.
- Question 2 and 4 should be done with Hive.
- Question 5, should be done with: Sqoop into MySQL for cleansing, Sqoop into HDFS, Hive for analysis, automated with Oozie.
- There should be an MR Unit test for each Map, each Reduce and each Map-Reduce combination.
- Data sets must be stored and processed within HDFS and a Hadoop Cluster.
- Performance optimization considerations are a critical factor.
- All Map Reduce solutions should have Javadoc documentation, specifically explaining what was the thought process, approach applied to the problem in question, *and any assumptions made*.
- Optionally, the configuration of an additional Data Node and showcase of jobs running in a multi-node cluster are encouraged. It is only mandatorily required to run it at least in a pseudo-cluster.

Mandatory Technologies

- Java
- HDFS
- Hadoop

- MapReduce
- MR Unit
- Hive
- Sqoop
- Oozie

Guidelines and Deadlines

- All requirements must be completed.
    - The data is available in the cloud in the following link in CSV format, multiple files.
    - Pre-transformation or data cleansing can be done to the data. Assume that in real-time you won't be able to transform Terabytes of data or more with just a grep command, for example.
    - You must show the analysis results in your presentation plus any findings, and you will also be asked to run any of these MapReduce jobs, live.
    - Project must be presented on Monday of Week 8.