

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 Map Reduce algorithms for the following Business Questions:

* Identify the countries where % of female graduates is less than 30%. (assuming tertiary graduation, 28448, assume average of years, for processing consideration start at 28890 to exclude regional data and look for indicator se.ter.cmpl.fe, partition by country?, combine to get %?, reduce for final list?) (gonna try to do all of this at once but a better approach would be to modularize the process, mapper will average numbers, reducer will remove values less than 30% or 0%, key would be country for reducer)
* List the average increase in female education in the U.S. from the year 2000.
* List the % of change in male employment from the year 2000. (emp rates, pay, etc.?)
* List the % of change in female employment from the year 2000.
* Additionally, based on your data exploration and analysis, evaluate one business factor that you consider important, and make this your own requirement. (how much data isn’t being used, what future data can we get?)

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

* 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 single-node cluster.

**Mandatory Technologies**

* HDFS
* Hadoop
* MapReduce
* MR Unit

**Guidelines and Deadlines**

* All requirements must be completed.
  + The data is available in the cloud in the following [link](https://s3.amazonaws.com/hadoop-dataset/genter-stats-data.zip) in CSV format, multiple files.
  + Pre-transformation or data cleansing can be done to the data, but if done, it must also be done with Map Reduce. Assume that in real-time you won’t be able to transform Terabytes of data or more with just a grep command, for example.
  + It’s important to share thoughts and debate with your colleagues, but keep this as a very individual project, because you will only get the most exposure by trying to fix issues yourself in the first place.
  + You must show the analysis results in your presentation plus any findings, and you will also be asked to run any of these Map Reduce jobs, live.
  + Project must be presented on Friday of Week 5.

Write python script for running the hadoop stuff