


Apache HBase Homework

Realtime and Big Data Analytics

Fall 2018



Homework

Analytics Project

1. Complete the HBase practice commands pdf to try out HBase commands using the HBase shell.

2. Develop code to ingest your project data, or copy the data into HDFS on Dumbo.

Submit this code, and/or commands, and/or a description of how you moved the data to NYU Classes - this is an individual assignment - only upload your own ingest steps.

3. Cleaning and Profiling Code (Complete this on Dumbo with your data in HDFS)

Profile: Develop code using Hadoop MapReduce to characterize (profile) the data in each column you plan to use from the data source you are responsible for. Use MapReduce for data profiling- **all team members must use MapReduce** to profile their respective data sources. (In future assignments where you'll develop the actual analytic code, you will use non-MapReduce tools.)

Clean: Develop code using Hadoop MapReduce to ETL (clean/format) your data sources as needed (there should be at least one data source per team member). Submit this code. The profiling steps above should be helpful in deciding how to clean/format your data. For example, you might want to drop some columns, you might want to normalize data in a column (for example you may want to change all references to NYC in a 'City' column to 'New York City' instead of NYC/nyc/NYCity/NYC), you might want to detect badly formatted rows that might be missing important data.

****Any problems with access to data sources need to be resolved quickly or you will fall behind.**

Submit this code in NYU Classes - this is an individual assignment - only upload your own code.

4. Submit the Data Schema for your dataset(s)

Submit the information from profiling the columns of data - this will be output from your profiling program that shows the range of values for each of your columns and the maximum string lengths you expect (where applicable). You might also include the schema (column/field names and data types). This should be just a text file generated by your code, you can annotate it as needed.

Submit this in NYU Classes - this is an individual assignment - only upload results for your data source.

Readings

5. Please read Chapter 20 on HBase in TDG - pages 582-587.