ITMD 521 – Week-02 – Chapter-02 Comparative Assignment

Objectives:

- Write large scale queries using UNIX tools and SQL tools
- Compare the relative time each job takes to execute and relate that to the amount of data being used
- Generate theories to optimize the processing time of these operations

Outcomes

At the completion of this lab you will have built from scratch your own database, database schema, import script, and application code that will find the highest temperature per year using a Java based SQL application and using the UNIX tool AWK. I will run your code on my Vagrantbox and it needs to run for you to receive credit.

Procedure

Part 0

You are to transfer your picture and introduction information you posted in the discussion board into the ReadMe.md in your provided Github Account

Part 1

Retrieve the gzipped file data from the link provided in Blackboard (note there is a large amount of data here so you need to start downloading this early. If you are in a place with bad internet you need to find a place with good internet – no excuses at the master's level.) Note – the data is too large to open in Notepad, you have to use the Linux commandline.

Once data is downloaded place the unextracted data into a directory named **all** that resides on your Vagrant box. You need to clone the sample code git repository as well for you will need the max_temperature.sh code from ch02-mr-intro. https://github.com/tomwhite/hadoop-book/

You need to run the max_temperature script 3 times – noting the execution time of each job. (Hint – make use of the UNIX `time` command)

- Run against the 1990 data set
- Second time against 1990 and 1992
- Third time against 1990, 1991, 1992, 1993

List each output in a chart and graph – give a brief explanation of the amount of time each step took. Also note the amount of RAM and your CPU speed.

Part 2

Using the same datasets and the schema provided in the beginning of chapter and in the slides for chapter 02, you are to develop a java program that will parse the datasets and insert them into **3 tables** in a mysql database, in the same way as listed in Part 1. (that you install into your Xenial64 Ubuntu Vagrantbox)
(Make the root password combo: safestsystemever)

You will need to:

- Include a script that will create the database and 3 tables with schema
- Include code to insert all the data into your tables
- You can assume that the data has already been decompressed from the prior example.
- You Java code will need to make a connection to the database and query the table of data you entered. Your result is the highest temperature per year.
- Note the time each query took on the three datasets and chart and graph these results along with the amount of RAM and CPU you have.

Deliverable: Note that you need to submit all of this source code to your github repo in a folder named **week-02** under a root folder named **itmd521**. (Assume the java mysql library is pre-installed on my system). Include all necessary code and a *DETAILED* Readme.md file giving me instructions how to run your project. Nothing should be pre-compiled just the *.java file.

Submit your detailed analysis in a **file in the repo with the code named: lastname-firstname-week-02-analysis.pdf** ← this is important, wrong format means no credit – the Blackboard submission is just your Github repo URL.

Finally there should be a commit history in your Github repo as you develop this project, I should see multiple commits, project will not be accepted if there is no history.

Sources – if you find code samples on the internet, please place a comment with a URL of the original site of retrieval or reference. When in doubt just comment.