CSE487 Data-intensive Computing Spring 2016

Due date: 4/16/2016 by 11.59PM or earlier, online submission on cse machines.

Goal: Learn Parallel Processing of Big Data using Hadoop MapReduce and a Build a Dashboard(s) for Analysis and Visualization of the Results

Context: Class room scheduling for courses is complex problem. It is all the more difficult in a department where the enrollments are increasing and number of courses and class sizes are increasing. Consider the case of this course (CSE4/587): I requested a larger room at the beginning of the semester. We have to send in a formal request through a departmental secretary and the reply comes a week later and it is always negative. For example they could not give a room larger than NSC 215 (150 cap) for this course. I requested a larger room for the midterm exam. They answer was negative. I found out that all the information about courses and classrooms is in a database and it is publicly available through a web site: http://www.buffalo.edu/class-schedule?semester=spring for example gives the Spring semester's courses. A web crawler can get this information by scraping the web site resulting in very large unstructured text data. Any authorized person can also get this information directly from the database. That's what we have. I will send you the link in a message to the class. DO NOT SHARE it with anybody beyond this course. Download and save this data in csv file as CourseRoom.csv. This is the main data set you will work with. You can get other data sets based on the needs of your analysis.

Problem Statement: Understand the data by doing an exploratory data analysis using RStudio. The data is for courses and class rooms from 1931 to 2017. Data analysis you perform will be driven by problems you want to solve and questions you want answered. Users for this "data product" are (i) course schedulers (ii) teachers teaching a course (iii) university planners and (iv) university development department for fund raising related activities.

Understand the domain and design a list of relevant questions. You are required to design and implement MR algorithms to extract useful intelligence to answer the questions. This intelligence will be offered as answers to questions through a web interface (dashboard). The answers to these questions could be through visualization, textual output or numerical information. The dashboard will feature ability for the user to interact by choosing from a set of questions and also by configuring the parameters for the questions and the visualizations. Sample questions to get you started are given below one for each type of user: (i) List all the class rooms of capacity greater than 200 for Mondays and Wednesdays between 5 and 6.20PM for Spring 2016 (ii) List a class room with capacity greater than 200 for May 9, 2016, 7-11pm (iii) track the trend in growth of seats per semester per building in the last 10 years in an interactive visualizations and (iv) compare enrollment increase over last 10 years to the building space increase in a visualization.

More specifically we want to analyze the data and provide a presentation of the charts (in the form of the dashboards) explaining the usage of the spaces (classrooms). How it has changed, what is trending, who is using the classrooms. As one of the Assistant Dean stated (I am paraphrasing here), ".. the classrooms ought to be continuously scheduled.". Are the rooms being efficiently used? Why was

cse4/587 scheduled in a 150 capacity NSC 215 when there was a classroom with modern facilities available at Knox 109 and Knox 110?

Data Characteristics: We will use structured and semi-structured data stored in csv files. The entire data could be stored in several files in tables, and spread over multiple directories (depending on the data). The project will deal with data from multiple sources. The core data set is available at ftp source that was sent you earlier. The data does not have any heading row. Understand the data structure and add appropriate heading: this is an important step for EDA [1]. You can pre-process the data.

Implementation details:

The general application architecture for this project is as follows. The data given and any other data you collect will be stored in single input directory. Understand the data by doing an EDA on RStudio. Suggested application architecture is shown below. **While MR is a requirement for data analysis you are free to use any other visualization for the charts. Tableau is recommended**.

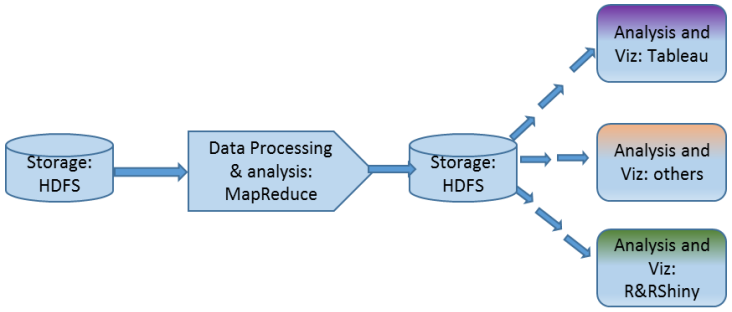

Figure 1 Application Architecture

Application architecture above shows the overall approach to be taken. Understand the problem domain and develop a set of relevant questions you want answered. Develop general MR solutions to solve the problems. The data analysis and visualization should present a coherent interactive user interface for users to interactively specify/select the questions and parameters and get a visual response. For example, (i) we want to know how efficiently is the classroom space in Knox Hall used in the Spring semester of 2016? (ii) Create a dashboard of charts to convince the administrators that certain classrooms cannot be used for single day/one time reservations. (Knox 109 and Knox 110 are reserved for one time use reservation for Spring 2016.)

Problem 1: Parallelizing data processing using MR (30%): Study the details of basic MapReduce provided in the Apache Hadoop documentation [4]. You can run the word count tutorial directly on your unix environment (machine/partition on your machine) or a VM that your TA has created. Then you will use the class room data provided and use MR to determine the rooms used in each hall for each semester. Example: outputs: <key, value> pair: (Key=<"Knox_Spring 2016">, Value=6000). **Output for this problem is all the rooms with years and capacity served.** This could be done by reading a line, selecting the three needed tokens (room, year and capacity), setting the key with (room concatenated with year) and setting the value (capacity) and output the <key,value> pair. **Firm deadline for this problem 1 is 4/2.** Please do not submit data: Just the MR source code and sample input and real output.

Problem 2: Design of questions (Experiments) (40%): Design a set of 20 questions that will provide insights into the situation of classroom scheduling at UB North Campus. This could be trends over the years, how the space needs have grown and how efficiently are the classrooms used, what is most popular time and day etc. Design MR code to solve these problems/questions and extract the answers. Submit the questions, output/results and also the MR source code. **Firm date of completion for this part: 4/9.**

Problem 3: Building an analysis and visualization user interface (30%): Ingest the results of the MR analysis in the above parts to create visual insights into the data. Any user should be able to browse through a dashboard of charts that explain the overall status of classroom usage in UB North Campus. I am looking for Tableau like story or dashboards using Tableau, RShiny or other visual software you are familiar with. **Firm deadline for this part is 4/16**.

Guidelines:

1. **You work on with groups of one or two, no more. Your documentation should clearly indicate the responsibilities of the team members.**
2. You can source your data from multiple sources while this is NOT a requirement. You may have to clean the data. For example there another standard source available for every university called Common Data Sets that is available here [5].
3. You are required to prepare a detailed report and documentation that will allow us to repeat the experiment/analysis you have carried out and also provide the provenance for the results you have generated.
4. Use elegant directory structure and naming conventions for directories and files to capture all the work for project 2 and then tar or zip each problem into its own compressed file of self-explaining names and problem# and username.
5. Submit the solutions as soon as you complete them. Do not wait till the last minute to work on all the parts.
6. **Do not publicize the code on github or any other similar public forum until the semester is over**.

References:

1. C. ONeil and R. Schutt. Doing Data Science. Orielly, 2013.
2. MapReduce documentation. http://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html, last viewed 2016.
3. Common Data Sets for UB. **http://www.buffalo.edu/provost/oia/reports-documents/common-data-sets.html , last viewed 2016**