CSE4/587 Data-intensive Computing Spring 2019
Due Date: 4/19/2019 by midnight.

LAB2: DATA AGGREGATION, BIG DATA ANALYSIS AND VISUALIZATION: B. RAMAMURTHY

OVERVIEW:

In this lab, we will expand our skills in data exploration developed in Lab1 and enhance them by adding big data analytics and visualization skills. This document describes Lab2: Data Aggregation, Big Data Analysis and Visualization, involves (i) data aggregation from more than one source using the APIs (Application programming interface) exposed by data sources, (ii) Applying classical big data analytic method of MapReduce to the unstructured data collected, (iii) store the data collected on WORM infrastructure Hadoop and (iii) building a visualization data product.

We will leverage the data collection and exploratory data analysis skills developed in Lab1 to accomplish the goals of Lab2.

LEARNING OUTCOMES:

- ✓ Automate data collection from multiple sources using the APIs offered by the businesses
- ✓ Explain the importance of evaluating the reliability of data (for example: social media vs news media)
- ✓ Apply classical big data analytical methods: MapReduce for word count and related family of algorithms such as word occurrence and n-grams
- ✓ Work on Hadoop 2.x, and HDFS and process the data using big data algorithms
- ✓ Learn a high level language-based data analysis by exploring Python as data processing language
- ✓ Apply modern visualization methods and disseminate results using the web/mobile interface

OBJECTIVES:

The lab goals will be accomplished through these specific objectives:

1. **Explore MapReduce (MR)** model and programming using this model. Understand the model.
2. **Read** this award winning report on data analysis and prediction [6]. Don’t skip this step. You read to get ideas about they use data analysis in real world. You will use as a guideline for this lab and the next. It describes has all the steps in the data engineering pipeline: from architecture to viz.
3. **Choose** a topic of interest to you. It could be “sports”, “snow”, “facebook” or anything of current interest. Make sure you will get enough data from your data sources on the topic you have chosen.
4. **Aggregate** data from multiple sources to corroborate any findings and outcomes of data analysis.
5. **Install** a Hadoop infrastructure: (i) cloudera docker image from the document we provided, or (ii) virtual machine (VM) image for data storage in HDFS and Hadoop infrastructure from Hortonworks, for example, (iii) amazon aws or another one you are familiar with.
6. **Code** solutions in Java or Python to process data in <key,value> format using MR model.
7. **Visualize** the outcome of the MR analysis using “wordcloud” and charts using visualization tool.

8. **Compare** the outcomes of the same analysis for at least two different sources: first an opinion social media source such a twitter and two others, a reliable researched source such as NYTimes [2] and the common crawl data [7].
9. **Create** a responsive web interface (web tool) for visualizing the outcome of your analysis.
10. **Document** the complete development process as a README in your submission. Use the report in step 2 as a guideline. It does not have to be that long, that report has lot more operations than you are doing in this lab.
11. **(optionally) Extend** the work to collect "real" big data on a topic and apply sophisticated methods such as Latent Dirichlet Allocation (LDA) and a generate conference poster or paper.

LAB DESCRIPTION:

Introduction: An important and critical phase of the data-science process is data collection. Several organizations including the federal government (data.gov) have their data available to the public for various purposes. Social network applications such as Twitter and Facebook collect enormous amount of data contributed by their numerous and prolific user. For other businesses such as Amazon and NYTimes data is a significant and valuable byproduct of their main business. Nowadays everybody has data. Most of these data generator businesses make subset of their data available for use by registered users for free. Some of them as downloadable data files (.csv, .xlsx) as a database (.db, .db3). Sometimes the data that needs to be collected is not in a specific format but is available as a web page content. In this case, typically a web crawler is used to crawl the web (pages) and scrap the data from these web pages and extract the information needed. Data generating organizations have realized the need to share at least a subset of their data with users interested in developing applications. Entire data sets are sold as products.

Recall from Lab 1, that an API or application programming interface is a standard, secure and programmatic access to data by an organization that owns the data. An API offers a method for one or two way communication among software (as well as hardware) components as long as they carry the right credentials. These credentials for authentication for programmatic access is defined by another standard OAuth (Open Authentication) delegation protocol [1] or API key in some case as in NYTimes data access [2].

We will collect data about from at least three sources, one opinion-based social media in twitter, research data in New York Times, and the third is the common crawl data for the same topic or key phrase, and similar time periods. Process the three data sets collected individually using classical big data methods. Compare the outcomes using popular visualization methods.

LAB 2: WHAT TO DO?

PAIR PROGRAMMING: We are going to allow pair programming for this lab. You will work in groups of one or two. **(No groups >2)**. You will get an F for the course if your group plagiarizes or copies somebody else's work or some other group's work. You can discuss anything ONLY with your pair team member. **Each team will have to submit all the material in their submit directory.**

Preparation: Here are the preliminary requirements for the lab.

1. Development language: We plan to use python. If you do not know the language, you can use Java. These are the two languages we will support.
2. For twitter, NYTimes and Common Crawl data you will need to get the appropriate Oauth and API keys. You do have the Oauth keys from Lab1 and you can reuse it for twitter search API. For NYTimes or any other API, you will have to apply and get the API keys ready. Now you know how to get access to many other data sources using the standard APIs the data organizations provide. Common crawl may not need API keys.
3. For data analytics, you will need to either use the Hadoop VMimage or docker image or native hadoop on your Linux machine. We have provided the details of a Hadoop installation. You may also install it from scratch if you have prior experience with this. Many organizations such as Cloudera provide their bundle. You can also use cloud offerings by aws (amazon), and Google cloud, if you are familiar with them. You have too many choice. Cannot decide? Just use the docker instructions we have provided.
4. Now for the visualization of the results. We want you to use the d3.js, a very popular javascript library for visualization. We have chosen to introduce d3.js for you understand origins of d3.js and how it came about from NYTimes need for complex visualizations [3]. You can also use Tableau, a very popular but proprietary viz tool. But you get an academic edition free. Why not use it? It can directly access your hadoop output (result) files.

Part 1: **(15%) Prototype data collection:** PLEASE DON'T WAIT till last minute to collect data. If this step is not completed on time, you will not be able to obtain reasonable results for visualization. Lab2→data→cc→sample(for common crawl), Lab2→data→tw→sample (twitter), Lab2→data→nyt→sample (for NY times). Let the data be big data in size, of the same topic(s) and similar time periods. The API provides only 1 week data and we won't accept last year tweets later in evaluation (**Students should collect data on current semester**). The way you should collect data is using topics and subtopics. For example, if your topic is sports, your sub topics can be cricket, football, etc. Big Data is really big (peta and terabytes of data) but for this project, the students should aim to collect at least 500 articles for each topic (so maybe 100 articles for each sub-topic) from NY Times, at least 500 articles from Common Crawl and at least 20000 tweets to perform analysis. These 20000 tweets should be unique tweets i.e. all duplicate tweets and retweets should be removed.

Part 2: (5%) Set up big data infrastructure: Set up a Hadoop infrastructure for storing and your big data: Don't delay this part. Get help in the first week. No excuses, we have already demoed this in lecture and given you detailed written instructions in the lecture notes. Run the sample programs for word count that we have provided. Save the demo code and data for word count Lab2→demo

Part 3: **(80%) Analyze and visualize**: Now that you have the data (three sources: cc, nyt, and tw) and the infrastructure ready, analyze these using big data methods. More specifically, we will use MapReduce algorithm.

Choose a topic of current interest to people in the USA. Something that is in the news. Use the topic as the key word or phrase to aggregate tweets, news articles, and common crawl data about the topic for the same period. You may have to tweak the phrase to get a good yield of tweets and news articles.

Load the data aggregated in Part 1 from your local file system into the big data infrastructure, three directories: twitterData, newsData, crawlData. Each directory can have many files of data. Start with small data set (say for a day), to prototype your data pipeline.

a. (20%) Code and execute MapReduce word count [5] on each of the data sets. Map will clean and parse the data sets into words, remove stop words, stem the words (ex: running to run) and reduce will count the useful words. twitterdata→twitterWords and newsData→newsWords, crawlData→crawlWords. Review and visually compare the output for representative words about the topic. You may have to change the search word, obtain new sets of data that may comparable sets of output words. You can use Python or java for your coding language, and appropriate libraries for stemming and removing stop words.
b. (10%) Visualize each of the outputs using d3.js, or Tableau and on a simple web page that you create for this lab. A simple display will be a drop down of search word of phrases you have been using, and output is the top 10 words corresponding to that phrase, in cc, nyt, and tw data set.
c. (10%) Now repeat the steps a) to c) for larger data set collected over week. May be you will see some convergence in your output among the outputs from the three sets. You may need to do a lot a lot of trial and errors since the data is indeed from three diverse sources. Allocate time for these trials, and be prepared for this. That is the reason for long duration we have given you for this lab.
d. (10%) Now design a web page (can piggy back on the one in step b) and feed the results by embedding d3.js code (with replaceable worldclouds) in it, finalizing the display of results. In fact, you should be able to create interactive data product! Input a search topic, we will return the word cloud associated with that topic! You can also do this using Tableau and its server. You do not need to do it in real time. You just need to design the webpage in such a way so that you to show us your results by navigating along the website (maybe set up some buttons so that by clicking on those, the wordclouds that you already collected will appear on the page). The website can be locally hosted.
e. (20%) We want to drill deeper into our analysis. Using the smallest data sets you collected in step a), analyze each set (Crawl, Twitter and News) word co-occurrence for only the top ten words from your word count exercise in step a. Assume context for co-occurrence is the "tweet" in the case of twitterData, and the paragraph of the news article in the crawlData and newsData. Your "map" function emits <word, co-occurring word> and your "reduce" function should collate the co-occurrences for the top ten words and output them in a suitable format. This is like the bigram in the reading material [6] we have assigned for this lab.
f. (5%) Document all the activities and how we can use your explorations and repeat them with some other data. Use block diagrams where needed. A well-organized directory structure is a requirement. Show this directory structure in the documentation and the demo below.
g. (5%) A short video that explains your data analysis and visualization process. Both teams should be involved in the presentation.

Infrastructure: We will provide a VM image. You can also install Hadoop from the scratch if you are good at installing and managing software.

Submission:

1. You will create a folder in timberlake named lab2. (Timberlake is a cse server).
2. Every file should have your name only at the top of the notebook and your team member's name in the second line. Both teams mates should have created the files on timberlake and submit invidually.
3. Store or transfer all the file to lab2 folder on timberlake: yourLastNamePart1.ipynb, yourLastNamePart2.ipyn, **all the data used including curated tweet**s;
4. On timberlake tar the lab2 files into yourLastNameLab2.tar
5. Submit using submit_cse487 filename.tar or submit_cse587 filename.tar

DUE DATE: 4/19/2019 BY MIDNIGHT. ONLINE SUBMISSION AND DEMO REQUIRED.

HOW CAN DO WELL IN THIS LAB?

- Start working on it today. If you have any difficulty, don't wait till the minute. I don't believe that problems simply appear on the due date!
- Please install the infrastructure. Make sure it works with the sample data we have provided with in it.
- Leverage your data acquisition knowledge from Lab1. Start collecting data about the topic of your choice. This lab2 could be your passport to getting a good job. Data pipeline expertise is a sought after skill.
- You may not get the data you want in the last minute. You cannot copy data from others.
- Plan, design, prototype, test and iterate through these steps.
- Choose a partner so that you can complement each other in skills and learn from each other.
- Attend TA office hours and recitations every week. Attend any number of office hours by any TA until your questions are answered.
- Enroll in Piazza (CSS4/587) and ask questions. Don't post code. Be civil. This is a public forum.
- Login into timberlake.cse.buffalo.edu and make sure you have an account on cse servers. If not send mail to cse-consult@cse.buffalo.edu to get an account.
- Create a lab2 folder with dummy files, tar/zip the file, submit the zip file and check it out it goes without any problem.
- Finally, no cheating. Do not copy or get the code from somebody. By this you are building a disadvantage for yourself. You are missing a golden opportunity to learn. The lab, the languages and tools may be hard for non-programmers, but that is no substitute for hard work. Of course, we will make sure people who cheat are appropriately penalized.

REFERENCES:

1. Twitter API. Twitter Developer https://dev.twitter.com/, last viewed 2017.
2. New York Time Developer Network. https://developer.nytimes.com/, last viewed 2018.
3. D3.js, https://en.wikipedia.org/wiki/Mike_Bostock, last viewed 2018.
4. J. Lin and C. Dyer. Data-Intensive Text Processing with MapReduce, Synthesis Lectures on Human Language Technologies, 2010, Vol. 3, No. 1, Pages 1-177, (doi: 10.2200/S00274ED1V01Y201006HLT007). Morgan & Claypool Publishers. An online version of

this text is also available through UB Libraries since UB subscribes to Morgan and Claypool Publishers. Online version available. Use it for coding word count and word co-occurrence.

5. M.Knoll, MR in Python. http://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/
6. R. Singh et al., Using Open Data to predict market movements., DELL EMC, 2017, https://education.emc.com/content/dam/dell-emc/documents/en-us/2017KS_Ravinder-Using_Open_Data_to_Predict_Market_Movements.pdf
7. Common crawl (open data). http://commoncrawl.org/, last viewed 2019.
8. Tableau, https://www.tableau.com/, last viewed 2019.