Term Project & Report

USING SPARK FOR SCALABLE ANALYTICS
VERSION 1.1

DUE DATE: There are multiple deliverables associated with this assignment

As part of this assignment you will be doing a term project that involves using Apache Spark for performing analytics. You are free to use Spark for processing on-disk files or to use it for processing data streams.

This assignment may be modified to clarify any questions (and the version number incremented), but the crux of the assignment and the distribution of points will not change.

1 Requirements

As part of your term project you are required to implement a Spark application for analyzing one or more datasets. The requirements are the following:

1. Your Spark deployment should execute on a minimum of 10-15 machines. The problem should be data-intensive or [and optionally] compute-intensive.
2. Either the problem or the solution you propose has to be original. For example, if there is a well-known dataset and someone has used multiple linear regression to perform a set of analyses, you doing the same would be clearly unacceptable. But if you decided to use artificial neural networks, support vector machines, ensemble methods, or deep learning for the same thing that would be acceptable. However, your results must be better or at least in the same vicinity in terms of performance: accuracy/errors, latencies, and throughput.
3. All teams are required to coordinate using Slack. Your Slack group will be created by the course team, and will be used to determine whether all group members have contributed significantly to the project.
4. The minimum dataset size is 1 GB. You are not restricted to a single dataset. You are encouraged to combine your base dataset with other supplementary datasets to create a richer dataset. For larger datasets, it is recommended that you first stage it on HDFS and then from Spark.
5. Examples of unacceptable projects include:
    a. Assignments from Distributed Systems (x55 series at CSU) or Big Data (x35 series at CSU) courses at CSU, MOOCS, or other universities.
    b. Term projects from previous courses that you have taken (or are currently taking). The submission has to be an original submission that you have done specifically for CS455.
    c. Submitting projects that you are working on as part of your Research Assistantships or your day job. This is to ensure a level playing field for your peers.

Unlike the other assignments in CS455, this is a team project. You are required to work in groups of 2-3 for this assignment. One person in the group is required to send the class GTA an email about the composition of the group and the title of your proposed project. You can use the "Search for teammates" feature on Piazza as well. If you are not able to find team members, please let the GTA know so that team members can be assigned.

No Piazza discussions are allowed on this assignment. Spark setup is extremely seamless and there are a ton of resources available online; your team must resolve problems on their own.

2 Third-party Libraries and Restrictions

You are allowed to use 3rd party libraries ONLY AFTER you have received approval from the Professor or GTA. You are allowed to use ALL libraries from the Spark ecosystem. Once you have chosen the libraries that you will use, you are responsible for coping with issues that you encounter with them.

3 Project Proposal [TP-D1]

As the first deliverable, a 1-2 page proposal should be submitted. References or citations are not included in this page limit. Proposals that are not good enough will be subject to further refinements and a one-on-one discussion with the Professor/GTA. Please do not send e-mails about your topic till then.

The project proposal should include the following information.

- Title of the Project
- Full names of the team members
- Dataset (and supplementary datasets if there is any) – A brief introduction to the dataset including its origin, size, format, etc.
- Problem characterization
    - This is a technical description of the problem. Your audience is your peers so express it in a way that they can appreciate.
- Currently published related work done with the dataset – include references
- Analytic tasks you are planning to perform
- Evaluating the effectiveness of your solution
    - What are the metrics that you will use to assess how good your solution is? Examples of these include: accuracy, turnaround times, throughputs, number of false positive or false negatives, mean squared errors, etc.

4 Example Projects

Here is a list of some projects from previous years to kick-start your brainstorming sessions.

- Artificial song generation using the Lakh MIDI dataset
- Quantifying the effectiveness of Kiva Crowd micro-lending
- Customized restaurant recommendations based on Yelp reviews
- Analysis and visualization of your own Google Takeout data

There are many datasets out there and many interesting questions that can be asked. Consider topics that have a meaningful social or economic impact. Also, throwing machine learning libraries at a dataset and thinking that should do it is not wise.

4 Anatomy of the Term Project Report

CS455 is a capstone course and includes a writing component as well. The term paper report/paper *must include* several elements, each of which will be a separate section. These include:

- Introduction
- Problem characterization
- Dominant approaches to the problem
- Methodology
- Experimental Benchmarks
- Insights Gleaned
- How the problem space will look like in the future
- Conclusions
- Bibliography

There are several pitfalls that you must avoid when you are writing technical articles. Avoid cringe-inducing marketing lingo and hearsay *e.g.*, "My teammate Tony Stark thinks …". Quoting Professors and researchers in the University is not allowed. You are allowed to speculate, but these should be based on reasoned arguments. Avoid using words that are not part of your normal vocabulary – it is easy to know if someone had the thesaurus handy. Technical writing is meant to be clear while being accessible to those in the area.

Word Counts: The word counts set aside for each element of your term paper are specified below. Please do not try to skew the word limits for these sections so that you can reach the requisite word counts. Such skews are easy to spot and will be penalized. You are also not allowed to quote from cited papers just to pad the word counts.

4.1 Introduction

This section describes why the problem is important, where this research is being used, how this technology plays a role in our daily lives, etc. The introduction section is also a concise summary of your paper that outlines the rationale, organization, and key contributions of the term paper. It should be possible for a reader to know all the key aspects of your term paper just by reading your Introduction and Conclusions section.

You can also briefly inform this section with your past experience. Describe how you think your chosen area would be applicable to a project that you are working on or have worked on in the past.

Word count: 500 words

4.2 Problem characterization

This is a technical description of the problem. Your audience is your peers so express it in a way that they can understand and appreciate.

This section should describe the theoretical, physical, social, and/or engineering aspects that make the problem particularly challenging. A clear discussion of the challenge also makes the reader look forward to reading the remainder of the paper.

Word count: 500 words

4.3 Dominant approaches to the problem

The section must contrast and identify possible approaches and also identify inefficiencies in each of these schemes. For each work that you cite you need to describe the advantages, disadvantages, and the scope of the work. Your objective is not to defend any work, rather you should let the facts speak for themselves. Finally, in your write-up you need to describe each reference in the context of the overall narrative.

If you compare features across two systems and say that one system outperforms the other include a citation for this. List what the comparison point is. This could be latency, throughput, scaling, efficiency, accuracy, price, etc.

Citations have a specific purpose. They: (1) relate to work that you are surveying, (2) substantiate your claims, and (3) could be used by readers to delve deeper. Remember to number your references and list them in your bibliography in the order that you referenced it. If an article is in your bibliography it must cited in the main text. Citing at the right location indicates what your source is for a particular piece of information, and also demonstrates that you have read the article. Make sure that you cite all your references including Wikipedia and online lecture notes that you may have perused. References that are not cited should not be in your bibliography.

Word count: 300 words

4.4 Your Methodology

Describe your methodology. This includes the analytic tasks you have performed, the approach that you have taken, and the justification for your methodology.

E.g. if your approach involves fitting models to the data, describe the rationale for your choice of the model fitting algorithm.

Word count: 1000 words

4.5 Experimental Benchmarks

Include a description of the benchmarks that you performed. Your performance metrics must be amenable to quantitatively assessing the quality of your solution. Examples of these include: accuracy, turnaround times, throughputs, number of false positive or false negatives, mean squared errors, Area under the curve for the receiver operating characteristic, etc.

Word count: 500 words

4.6 Insights Gleaned

These are things that you *did not* know before you started this project The best solutions are the ones that you may have not thought of, but seem incredibly obvious once you have come up with them.

Word count: 400 words

4.7 Transformation of the problem space in the future

This is a thought experiment. You will be looking ahead and visualizing a future where there could be proliferation of certain types of devices, new types of services, changes in usage patterns, etc. You must describe the forces that you think will drive this change. Once you have these forces in place, identifying how the problem space will evolve in the future should be easier. Ultimately, you are describing what technology advancements and the way we interact with services will affect the problem space of your research area.

Word count: 400 words

4.8 Conclusions

A conclusion is not a summary. You must make a set of assertions about your work.

Word count: 400 words

4.9 Bibliography

The final term paper must have at least 8-10 references. All references must be cited in the paper. Citations must be numbered and sorted in the order that they appear. The Bibliography is NOT included in your word count.

4.10 Overall Word Count

Excluding the bibliography, the total word count for your term paper is 4000 words.

5 Deliverables

The term paper is split into three deliverables. We will have a session after the first deliverable which will include a critical analysis of the deliverables including the mistakes found in this deliverable.

5.1 Deliverable Zero

This deliverable requires submission of the composition of your group. You must send an e-mail to cs455@cs.colostate.edu with this information by 3/30 @ 5:00 pm MT. The composition of the group can involve a mix of on-campus and distance students.

Any problems that you are having with your non-performing teammate should be reported 2-3 weeks after Deliverable Zero. Timely intervention will allow us to resolve problems before they fester.

5.2 Deliverable One

You should submit a 1-2 page project proposal. One member of the team must submit a PDF document of the project proposal using checkin. The folder set aside for this submission is **TP-D1**.

5.3 Programming Component

Please submit the source codes for your term project along with a README file. numbers for this assignment. Group Members should be listed as authors in each of the deliverables. The folder set aside for the final submission using checkin is **TERMPROJECT**.

5.4 Term Project Report

The third deliverable is the completed term paper. This will be about 4000 words and must include at least 20 references. This should include ALL the sections (described in section 2 of this document) while avoiding deviations of more than 10% from the prescribed word limits. numbers for this assignment. Group Members should be listed as authors in each of the deliverables. The folder set aside for the final submission using checkin is **TERMPROJECT**. This should be a **PDF** file.

6 Grading

This assignment will contribute a maximum of **15 points** towards your final grade. The grading will also be done on a 15-point scale. The score distribution for the assignment is listed below:

There is a **1 point deduction** for not submitting Deliverable-0 and Deliverable-1 on time.

1 point for the Term Project Proposal

9 points for Term Project Completion including source codes and demonstration. The demos including one-on-one interview sessions will be scheduled using SignupGenius as we get closer to the deadline.

5 points for Term Project Report

7 What to Submit

Deliverable 0 is an e-mail to cs455@cs.colostate.edu with the composition of your team. The deadline for this is 4/1/2020 @ 5:00 pm MT. If you are unable to find a team member you should let us know by then. There is a 1 point deduction towards your cumulative course grade if you miss this deadline.

Deliverable 1: Term project proposal due 4/10/2020 @ 5:00 pm MT. The folder set aside for the final submission using checkin is **TP-D1**. There is a 1 point deduction towards your cumulative course grade if you miss this deadline.

Final Deliverables: Source codes (4/29/2020) and report (5/1/2020) are due @ 5:00 pm MT.

The source codes including instructions to compile and execute your programs should be included.
The term paper should be submitted as a PDF file. You will be assigned Group numbers for this assignment. Group Members should be listed as authors in each of the deliverables. The folder set aside for the final submission using checkin is **TERMPROJECT**.

8 Change history

Version	Date	Comments
1.1.	3/24/2020	First public release of the assignment.