Final Project

For your final project, you will investigate a research challenge in Big Data. We will allow two types of projects:

**Research Focused:** Your group must pick a problem under active research, read papers in the area, and implement a solution to the problem. Your solution can be a re-implementation of an existing research project or a combination/extension of current approaches. You will need to build and evaluate the performance of your system to compare it against prior research. If your focus is more algorithmic, you may want to build a simulator instead of a real platform so you have better control over the environment.

**Implementation Focused:** Your group must build a simplified version of an existing big data system. You still will need to read research papers about the system you are implementing, but your effort will be focused more on implementing the system and less on evaluating and comparing it against other proposed approaches.

You have about two months to complete the project so it should be more significant than a standard programming assignment, while still being feasible to complete within that time frame. The project will be broken into milestones.

**Due Dates:**

* Milestone 0: Form a Team - 09/19
* Milestone 1: Select a Topic - 10/02 (10 points)
* Milestone 2: Literature Survey - 10/09 (20 points)
* Milestone 3: Design Document - 10/27 (20 points)
* Milestone 4: Final Presentation - 12/15 (50 points)

Milestone 0: Form a team

* earlier if possible

You must form a team of 3-4 students. I suggest teams have 4 members if possible for the small teams. All members are expected to contribute equally to the project, although different students may have specialized roles. However, all students should at least partially contribute to both the coding and writing portions of the project. One student should be selected as the team leader who will act as the main contact point of the group.

When you form a team you do not yet need to have selected a project idea, however it is best if you decide an area (or areas) you want to work in and be sure all team members agree.

**Submission Instructions:**

* First, pick a team name. Your team name *must*  be that of a famous scientist, e.g., Galileo or MarieCurie. Pick someone interesting, inspiring, and unique.
* The team leader should  [create the team repository using GitHub Classrooms](https://classroom.github.com/a/WcN-Xf9T), entering the chosen team name. Make sure that nobody else has already picked your name! Other team members should then join the team. *It is hard for us to delete teams or move people between them, so be careful and don’t join the wrong team! We may deduct points if you cause us extra work.*
* Fill in the README.md file in the repository with your team members, team name, and optionally a brief description of the project idea(s) you are considering
* Create an Issue in your repository once your readme is complete. The title of the issue should be Milestone 0 - TEAM NAME (with your team name filled in). The body of the issue should list the team member names with leader first.

Milestone 1: Select a topic

* earlier if possible

For this milestone you will need to submit a 1-2 page report that answers the following questions:

* What is the topic area of the project? Why is it important and what specifically will you build or design?
* What are the main big data challenges you will face? (see our slides for the list of big data challenges)
* What are some sample papers (at least 4) that will guide your work?

All your milestone reports will be judged partially on how well they are written, so write complete sentences/paragraphs, not just a bulleted list!

We will not permit two groups to work on the same project idea. Precedence will be given to whichever group submits their proposal first. Your project proposal for Milestone 1 must be approved by the professors before you can submit the following assignments! If you submit early we will try to approve your project early.

**Submission Instructions:**

* Your writeup must be in [Markdown format](https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet) and saved in the file milestone-1/README.md
* Create an issue with title Milestone 1 - TEAM NAME and the body should have a link to your report file.
* Click the button on the top right side of the Issue view to *Assign the* issue to Professor Roozbeh Haghnazar.

Bad Project Ideas

Projects like the following will not be approved:

* Prediction of the home price based on time-series prices of homes.
* Taking code from an existing research project and measuring its performance. If code exists, you will be expected to modify / extend it in some way. If no code is available, then it is acceptable to try to directly mimic an existing project.

Milestone 2: Literature Review

“A literature review surveys books, scholarly articles, and other sources relevant to a particular issue, area of research, or theory, and by so doing, provides a description, summary, and critical evaluation of these works in relation to the research problem being investigated. Literature reviews are designed to provide an overview of sources you have explored while researching a particular topic and to demonstrate to your readers how your research fits within a larger field of study.”

– Fink, Arlene. Conducting Research Literature Reviews: From the Internet to Paper. Fourth edition. Thousand Oaks, CA: SAGE, 2014. See [this article](https://libguides.usc.edu/writingguide/literaturereview) for more information on writing a literature review.

For Milestone 2 you must perform a Literature Review which *summarizes and synthesizes* information from key sources related to your project. You will present the importance of the project area, motivate the difficulty of the key challenges, and describe the state of the art in resolving them.

**Article Selection**: You must present information from *at least 6 articles for 3 person teams* or *at least 8 articles for 4 person teams*.

* Use a tool like [Google Scholar](https://scholar.google.com/) to find articles related to your topic area. Note that you may need to be on the [GW VPN](https://guides.himmelfarb.gwu.edu/offcampus/vpn) in order to access research article PDFs.
* In general, we prefer you to use research articles such as those from conferences or journals published by ACM, IEEE, or Usenix, since they have more depth and are peer-reviewed (and thus more likely to be accurate) than random webpages.
* At most half of your articles (preferably less) may be technical “blogs” or other articles found online. However, these should be thorough, in-depth articles, not just how-to documents or news articles designed for everyday readers. For example, this article on [game network coding and performance](https://technology.riotgames.com/news/peeking-valorants-netcode) is a good choice, but this article on [cloud gaming and 5G](https://www.forbes.com/sites/forbestechcouncil/2020/04/21/cloud-gaming-is-booming-but-is-the-network-ready/#200f40303d97) would be a bad choice since it lacks technical depth. If you think someone *without* a CS background could easily understand the article, then it is probably a bad choice since it will not have sufficient technical detail to teach you anything.
* Your articles should be recent. It is fine to have some “fundamental” papers from prior decades, but be sure to also include some recent work that accounts for the state of the art in the area.
* You may find it useful to look at [ACM Computing Surveys](https://dl.acm.org/journal/csur), a journal that publishes literature reviews in many CS areas. You might find one related to your topic and it will give you an idea of what an in-depth literature review looks like (yours will be much shorter).

**Report Structure**: Below is a suggested structure for your paper. You may use something different but must cover all these points.

* Introduction: Give the background for your proposed project and the importance of resolving it. Refer to your articles to provide motivational data or use cases that show why this is something useful and important. (~0.75 page)
* Challenges: Describe the key big data challenges related to your project. Use the terminology defined in our class (scalability, heterogeneity, etc) and show how they can impact your project. Focus on the problems presented by the related work or any challenges which you feel the articles fail to address sufficiently. (~1 page)
* Approaches: Summarize the designs and algorithms proposed by related work. Try to categorize them so that you group different types of approaches together in a way that adds meaning. (~1.5 pages)
* Project Proposal: Very briefly outline how your project relates to the problems and solutions described above. For example, you could explain how you are focusing on a specific sub-problem or extending an existing piece of work. This section can be short: 1-2 paragraphs at most.

**Format and Submission:** Your report should be *2-4 pages long* and follow these guidelines.

* As with all writing, *it is****never acceptable****to copy/paste from another author*. All of your writing must be your own! If we find any sentences copy/pasted from another source (unless it is a direct quote), we may deduct 10-50 points.
* Your document must be submitted as PDF and should be in the form of a technical article. You may want to use a document template, [such as those provided by ACM](https://www.acm.org/publications/authors/submissions), in Word or Latex format.
  + We suggest single column, justified text, size 10 font, 1.25x line spacing, or similar.
  + If you use Latex (not required), we encourage you to use a service like [Overleaf](http://overleaf.com/) to simplify collaboration and generation of the document.
* Your writing must include citations where appropriate using a style such as [Wood18] to cite a paper by Wood et al in 2018.
* Save your document as milestone-2/literature.pdf and create an issue with a link to the document assigned to the professors when you are ready to submit. While not required, we suggest storing copies of all the related articles you surveyed into your repository.
* In consultation with the group members, the leader should prepare a [Phase Time Sheet](https://gwdistsys2023.github.io/DistributedSystem/project/phase-n-time-sheet.xlsx) and save it as milestone-2/phase-2-time-sheet.xlsx

The output of this report will be the fundamental basis of the next Milestone. For your next Design report, you need to make a decision about the solution you will develop for your problem according to the pros and cons of the studies cited in your literature review. Your solution can be chosen among the cited studies or it can be a combination of them, or you might propose your own solution and method to address the issues in your problem space. But the important point is that your proposed method should be based on an understanding of what others have done in the field. For the Literature Review Milestone you only need to give a basic outline of your plan, which you will then fully develop in Milestone 3.

Milestone 3: Design Document

The design document should be composed of sections as follows:

Abstract

Start with a very brief introduction to the problem you are trying to solve, followed by a high level overview of your proposed project. You may be able to partially reuse some text from your Milestone 2 Introduction and Project Proposal sections for this purpose, but this section should be only 1-2 paragraphs long.

Project Design

This section should describe the key components and algorithms that will make up your project. Describe your architecture and how the different components will interact. Where possible, you should use the terms and design patterns that we have discussed in class, Data Pipeline Architecture Style. You also should specifically mention how your design allows the system to handle some of the core big data challenges: capture, curation, storage, search, sharing, transfer, analysis, and visualization.

When describing your components and how they communicate with each other, you must include several diagrams as described below to illustrate your design. Not including figures, this section should be 2-4 pages long. You do not need to go into too much detail about what each component will do, but you give enough information that we will be confident you know how to proceed with your implementation.

Design Diagrams

There are many ways to show and explain the structure of your system. One of the best tools, to trace your architecture and algorithms, is UML diagrams. These diagrams enable the designers to explain their mindset and thoughts, to the developers, and also the QAs can execute test scenarios and test cases to find the possible flaws.

Since we would like to follow a standard strategy to produce a big data system, we are going to use UML diagrams to illustrate the system’s components and interactions. One of your diagrams *must be a Use Case style diagram*  to show the overview of your system functionality and different functions in your system. According to the nature and characteristics of the project, you are free to choose *two or more additional diagrams* among all UML types. For example, your Use Case diagram might show the overall system, then you can use the Deployment and Component diagrams to show the software and hardware components of your system. A Communication diagram also can illustrate the connections between the components, Sequence diagram explains the sequence of actions, events, and processes between the components, objects, and actors in your system. You can also use Activity and State diagrams to explain more detail about a process or object in your system. We expect to see at least 2 or 3 diagrams in level 0 and 1 to illustrate a proper overview of your system.

We have made you [two examples to illustrate a small part of a hypothetical system](https://gwdistsys2023.github.io/DistributedSystem/project/project-uml.pdf) to understand how can we use these diagrams to design our system.

Also please check the following articles to have a better understanding of UML diagrams and also modeling big data with UML.

Please notice that you might see different annotations of UML in different resources but we strongly encourage you to check the https://www.uml-diagrams.org/ annotations, definitions, and examples.

* <https://www.uml-diagrams.org/>
* <https://link.springer.com/chapter/10.1007%2F3-540-40011-7_13>
* <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.102.2388&rep=rep1&type=pdf>
* <https://www.researchgate.net/profile/Dimosthenis_Anagnostopoulos/publication/220923234_Using_UML_to_Model_Distributed_System_Architectures/links/54d8bbad0cf25013d03ee68d/Using-UML-to-Model-Distributed-System-Architectures.pdf>
* <http://agilemodeling.com/artifacts/componentDiagram.htm>
* <https://www.sciencedirect.com/topics/computer-science/deployment-diagram>
* <https://www.visual-paradigm.com/guide/uml-unified-modeling-language/what-is-communication-diagram/>
* <https://www.visual-paradigm.com/guide/uml-unified-modeling-language/what-is-sequence-diagram/>
* <https://www.visual-paradigm.com/guide/uml-unified-modeling-language/what-is-state-machine-diagram/>

Project Schedule

The final section of the report must include your plan for completing the project by the end of the semester. We suggest breaking down your project into a series of prototypes, from the simplest possible demo up to your complete system. Each new prototype should add more functionality or complexity to your system. Finally, you should include some kind of evaluation plan describing how you intend to test the system’s correctness and performance. Be realistic; there is not too much time and we will expect you to complete what you propose on schedule.

You must include a timeline for when you will complete each of your prototypes, ideally in the form of a Gantt chart or similar. The work breakdown should also specify which team member will be responsible for each component or feature.

Format and Submission

Your report should be approximately 3-6 pages long. You should follow the same formatting template instructions given for the Milestone 2 report.

* Save your document as milestone-3/design.pdf, then create an issue with a link to the document and assign it to the professors when you are ready to submit.
* We suggest storing copies of your design diagrams and the source files used to generate the PDF (i.e., the word or latex files) in the repository too
* In consultation with the group members, the leader should prepare a [Phase Time Sheet](https://gwdistsys2023.github.io/DistributedSystem/project/phase-n-time-sheet.xlsx) and save it as milestone-3/phase-3-time-sheet.xlsx

Milestone 4: Final Report

Your final project submission must include a ~15 minute video and a final report.

Final Report

Your final report will primarily be a combination of the reports you wrote previously. It should be a “stand alone” piece of writing, meaning that a reader should be able to understand your full project without needing to look at the other documents. You should use the following layout which re-uses your prior reports and adds two new sections:

1. **Introduction**: Background on the problem and your proposed solution (Reports 1, 2, and 3)
2. **Related Work**: Descriptions of key related systems or papers (Report 2)
3. **Project Design**: Explaining and illustrate your overall project structure and design decisions (Report 3)
4. **Big data Challenges** (partially new section): This should be an extended version of what you wrote in Reports 1 and 3. You must list *all of* our key big data design challenges and discuss how they relate to your project. Even if you have not solved one of the challenges, you should still explain how it relates to your project and what would need to be solved for a real system deployment.
5. **Project Outcomes and Reflection** (new section): In this section you should describe the overall result of your project and reflect on your design decisions. You should include some sort of measurements to understand how it performs or works (ideally in comparison to another system). For example, you could measure the performance of your system in terms of throughput and latency. Finally you should discuss how well your design choices worked and what you think the potential of your ideas are. You should be honest about your results – we don’t really care how well your approach does, but we do want to see that you have thought about its potential benefits and drawbacks.

Final Video

You should create a ~15 minute long video explaining your project and demoing how it works. The goal of your video is to “sell” your project to us and get us excited about what you have done. Your presentation should be understandable to anyone who has taken our big data course.

We suggest the following structure:

1. Introduce your project and the problem that it solves.
2. Briefly present related work/systems.
3. Describe the key design choices of your system, with a discussion of how they relate to some of our big data design challenges.
4. Show us some (NOT all) of your code and explain how the code relates to your design. For example, pick one or two key components and show us the main functions, classes, or data structures.
5. Demo a portion of your project running. You do not need to show every part of your project, just a key aspect. For example, if your project focuses on reliability, show what happens when you cause a crash.
6. Discuss the overall impact of your design and present some evaluation of how well it works.
7. Conclude your presentation with a summary of your project’s key points.

Note that this is a lot of content to cover in 15 minutes, so you will need to be brief and only focus on the main points. If your video is more than 20 minutes we may deduct points.

The simplest way to create your video is to have your team members join a Zoom/Skype call and record as your present. While we expect your presentation to look good (i.e., using clearly readable, nicely structured slides), we do not expect you to do any fancy video editing.

Submission and Deadline

Your final submission **must be made by December 15, 5PM**. However, we will award **1 extra bonus point** to your project for each day early that you submit, up to 8 points extra if you submit by the original deadline of December 15th 11:59PM.

To submit, complete the following steps:

* All of your code and written reports should be in your GitHub Repository. The final report should be in the milestone-4/ folder.
* Your video should be uploaded using [this form](https://gwu.app.box.com/f/fdea26b01a164cf8af0843f886173d93). You **should not** include your video in your github repository since it is a large file.
* In consultation with the group members, the leader should prepare a final [Phase Time Sheet](https://gwdistsys2023.github.io/DistributedSystem/project/phase-n-time-sheet.xlsx) and save it as milestone-4/phase-4-time-sheet.xlsx
* When the above steps are complete, create an issue in your repository and **Assign** it to the instructor to notify that you are ready for grading. Include a link to your final report in your issue.

After that, have a great break and a successful future building big data! We hope you have enjoyed the semester and learned a lot!