Overview. The goal of the project is to allow you to more fully apply and go deeper into to the topics discussed throughout the semester. There are three different project options. The project you develop must have a big-data analytics theme or be in some tangible way related to big-data analytics. The project is divided into three steps (see below for details) where the first step asks for a project proposal, the second a progress status update, and the last step the final submission. The following are the three options (or general themes) for possible projects.

Option 1. Much of this course has focused on the underlying theory and system architectures that have made big-data analytics possible (and likely will continue to in the future). However, we only scratched the surface in terms of the papers we examined and the ideas covered. This project option allows you to find additional topics, via research papers, that you would like to learn more about. For this option, you must:

(a). Find two papers of interest. These can be on similar or different technology.

(b). Read the papers and take notes on each as in the reading assignments. Include observations as part of your notes (as in the reading assignments).

(c). In addition, as part of your notes (in a separate section), describe in detail how the work relates to what we have covered in class and through homework assignments (as applicable).

(d). Create a presentation (slides and a video) covering the main ideas of the papers. The target audience for the presentation would be the students in our class (along with myself), with the goal of making it clear what the main ideas are in each paper, their relation to big-data analytics, and their relation to the ideas we have covered in class. As part of your presentation, you must include for each paper the main technical ideas and contributions described in a way that is clear and easy to follow. You should aim for a presentation of around 20 minutes (at most) depending on the nature of the papers you select.

Google, Amazon, and Microsoft each provide access to published work they have done as part of their research and development programs. The following are links where you can find these papers (you will need to filter on topics relevant for the course).

- Amazon: https://www.amazon.science/publications/
- Google: https://research.google/pubs/

Option 2: Throughout the semester, you had the opportunity to learn about and use some of the basic services Google provides for enabling big-data analytics. However, this was not a
comprehensive survey, and there are a number of services offered by Google as well as by Amazon and Microsoft that we did not cover. For this project option, you can learn and use a service offered by Google related to big-data analytics that we did not cover, or learn and use a big-data service offered by Amazon (via AWS) or Microsoft (via Azure). For AWS or Azure, the service you select can either be similar to a service we looked at or can be different than those we covered (as long as the service is related to big-data analytics). For this option, you must:

(a). Select a service to learn and use that we have not or will not discuss. Note that in the rest of the semester, we will look at Google’s Dataproc, Dataflow, PubSub, Data Fusion, and VertexAI services.

(b). Similar to our homework assignments, identify materials (tutorials, labs, documentation) to learn about the technology.

(c). Create a “homework assignment” for the service. The assignment should be similar to our homework assignments in that it lists a series of steps, where each step involves a lab, tutorial, or documentation link for learning some aspect of the service. In addition, you should have at least two steps at the end that involve using the service for some purpose with a realistic dataset and/or data analytics scenario.

(d). Note that each step in your assignment does not need to list items to put into a notebook. However, you must turn in your “answers” to each step, i.e., create a document that demonstrates that you followed each step yourself.

(e). As part of (d) include a section that gives an overall description of the service you selected and its relationships to the services we covered in the class.

Option 3: Each homework assignment in the class covered how to get started and use the basics of a single service. For this project option, you can either put the services we’ve learned together to build an “end-to-end” data analytics demonstration project (e.g., from data ingestion through visualization) or select a specific service or technology (e.g., Spark, Dataflow, VertexAI, etc.) to dig deeper into also through a demonstration project. In either style of project, you must use datasets that have features or characteristics related to “big data”, e.g., large and/or unwieldy datasets, a large number of related datasets, data that requires significant cleaning and/or wrangling, streaming data, etc. For this part, you must:

(a). Describe whether you are developing an “end-to-end” analytics project or diving deeper into a particular technology as well as the services you plan to use. Note that if you decide to develop an “end-to-end” project, you should look at Google Cloud Composer as a tool to help automate the overall process.

(b). Identify and describe the datasets you are using including their location, their general contents, and challenges inherent in the data.
(c). Outline the goals of a corresponding demonstration project. The goals should be driven by the datasets you select, including what insights you hope to achieve and/or what data products you hope to produce. Note that the goals will also depend on the services you plan to use.

(d). Implement your project and submit all related code that you developed.

(e). Create a presentation (slides and video) that describes your project, explains your work including the service and/or services used and how, a demonstration of what you completed, an overview of what you learned, and a discussion of what your next steps would be if you had more time to work on the project. Your presentation should be around 10-15 minutes, but no longer than 20 minutes in total.

Step 1 (Proposal): Due Friday April 5. Turn in a hardcopy print out of your project proposal in class on Friday, April 5th. Your proposal must include the option you are choosing and details of what you will be doing as part of the option.

For option 1: provide the details of the two papers you will be reading (include authors, paper title, venue, and year published). If you are considering between a few papers, but haven’t narrowed down to two, include the ones you are picking between.

For option 2: provide the name of the service you will be examining and an initial list of resources (labs, tutorials, documentation) you have found to develop the project with.

For option 3: provide a description of the demonstration project you will be pursuing (including whether it is an end-to-end solution or a deep dive into a particular service), the services you plan to use, and an initial list of datasets you plan to use.

Step 2 (Check In): Due Monday April 22. For this part, you should submit a brief status update to the GitHub repo that will provided via Piazza (through GitHub classroom) for your project. Your status report should detail what progress you have made, any challenges you have faced, any modifications you are making to the project, and what you have remaining to complete. Note that you should have about 60% of your project completed at the check in.

Step 3 (Submission): Due on or before Friday May 10. For the final step, you must submit all of your work to your GitHub project repo. Include a readme (markdown) file for your project that includes a brief explanation of your project, a brief description of the files you submitted, and/or any additional information that will help in grading your project. For options 1 and 3, you must also include a link to your presentation video. Also be sure to submit your presentation (slides) as a single PDF file (for options 1 and 3).

Grading Rubric. The following describes how points will be allocated for your final project. Note that the project is worth a maximum of 60 points.
• **Steps 1 and 2 (10 points).** The proposal and check in are each worth 5 points towards your final project grade. You will receive full points for each by following the directions above. Points may be deducted if the instructions are not fully followed.

• **Project Completeness (35 points).** You will receive up to 35 points based on whether and to what extent the project was fully completed. Completion will be based on your original project scope and project goals. Completeness also includes an assessment of your documentation, where the specific documentation requirements depend on the type of project you select (see above).

• **Work Quality (15 points).** You will receive up to 15 points based on the overall quality of your project work. Assessment of quality will depend on the type of project you carry out and will be based on the project artifacts you submit (e.g., code, demo, notes, and so on).