MIDS W205

Course Project

Instructors:

Jari Koister, jari@ischool.berkeley.edu

Dan McClary, dan.mcclary@ischool.berkely.edu

Karthik Ramasamy, karthik@ischool.berkeley.edu

Arash Nourian , nourian@ischool.berkeley.edu

Manos Papagelis , papaggel@ischool.berkeley.edu

Introduction

This project is part of the final exam for this course. The intent is that students demonstrate that they meet the course objectives. The project will involves defining a problem, understanding the storage and processing needs (short term and long term) of a solution and select a technical approach. Finally the students are expected to create an initial demonstration implementation of the solution demonstrating their understanding if stitching and end-to-end solution together.

The students are allowed define their own problem, find the appropriate data, and define the solution. The sprit of the solution should be that of industrial-strength architecture that meets initial goals of quick prototyping, but also a path to a production strength system.

Guidelines

- You need to identify a business problem based on an exiting or new data set. There are no constraints here as long as all privacy or confidentiality constraints are meet.
- You need to implement process that computes the result in a repeatable fashion. Hence, it cannot just be a one-time computation.
- The result should be accessible for review through some kind of serving layer and presented in a form that would make sense in the real world scenario the problem is frame for
- You can pick any of the technical solutions discussed in the course as long as you can justify why you picked that solution, and your justification is grounded in a real world use case.

Evaluation and Acceptance Criteria

Deliverables and criteria

1. A proposal presentation in PPT or Google doc. You should think about the proposal as something that you are presenting to an executive. You are expected to justify solving the problem and motivating the solution you are proposing.
2. A final presentation. A presentation of the problem, the final product, and a roadmap for improving the solution with increase usage and data size.
3. All Code Submitted to Github per submission guidelines.
4. The instructor should be able to clone, build and run the project.
5. All required documents and presentations in Github. The repo should be completely self-contained.
6. All know limitations with respect to scale etc. should be documented in a README file.
7. There should be a runnable instance of the solution.

8. Analyze the complexity and storage needs for the application.

Problem

1. You should formulate specific problem and use case for the system/application.
2. It does not need to be a big data problem, but it should involve complexity along some dimensions such as the size of the data (Volume), the quality and variety of the data (Variety), the speed at which data arrives and need to be analyzed (Velocity).
3. It is permitted but not required to select a problem that requires advanced processing such as machine learning algorithms.
4. You can make certain assumptions such that data will be cleaned to a certain degree for example. But all such assumptions must be explicitly defined, and when appropriate reflected in the solution as tests, schemas etc.

Design and Architecture

Suggestions

Start by finding a data set. Based on the data set identify an interesting insight from the data using exploratory analysis. Implement a processing pipeline that can process and derive the insight repetitively. Determine what how frequently the result should be computed and how frequently you expect the data to be updated. Based on this determine what kind of architecture you need. Make sure that the architecture you choose can scale. If there are limitations on scale document them and check with an instructor that the limitation is acceptable.

Have clear deadlines so that you do not get stuck in a specific phase of the project.

Implement a steel thread of the solution quickly (a steel thread is a subset of the functionality implemented end-to-end)

Milestones

Week 4: Form groups & select problem area. Prepare slides for 10 minute presentation of your goals, challenges, how you will acquire your data. Also present what information organization challenges do you face as well as your initial plan to complete the project.

Week 6: A proposal (1-2-pages) must be sent to the instructor with sufficient detail of the problem being addressed and the supporting research that data can be acquired and organized.

Week 11: Progress Report I (2-3 -pages): description of first component of the project idea summary and justification, a partial description of data acquisition and organizing strategy and justification, tools/third party libraries description usages and initial performance evaluation on the adopted data acquisition strategy.

Week 15: Presentation of your project in class and final submission.

Submit any code and final report which includes (See above for details on acceptance criteria):

1. Overview of the problem being addressed.
2. Acquisition and organization of information for analytics.
3. The overall architecture of the solution and necessary implementation details.
4. The results of the project.
5. A retrospective on the project and suggestions for improvements.