|  |  |  |
| --- | --- | --- |
|  |  |  |
| MIDS W205  Course Project |  | Instructors:  Jari Koister, [jari@ischool.berkeley.edu](mailto:jari@ischool.berkeley.edu)  Dan McClary, [dan.mcclary@ischool.berkely.edu](mailto:dan.mcclary@ischool.berekely.edu)  Karthik Ramasamy, [karthik@ischool.berkeley.edu](mailto:karthik@ischool.berkeley.edu)  Arash Nourian , [nourian@ischool.berkeley.edu](mailto:nourian@ischool.berkeley.edu)  Manos Papagelis , [papaggel@ischool.berkeley.edu](mailto:papaggel@ischool.berkeley.edu) |
|  |  |  |

|  |  |
| --- | --- |
| IntroductionStoring, managing, and processing datasets are foundational to both applied computer science and data science. Indeed, successful deployment of data science in any organization is closely tied to how data are stored and processed. This course introduces the fundamentals of data storage, retrieval, and processing systems. As these fundamentals are introduced, exemplary technologies will be used to illustrate how storage and processing architectures can be constructed.This course aims to provide a set of “building blocks” by which one can construct a complete architecture for storing and processing data. The course will examine how technical architectures vary depending on the problem to be solved and the reliability and freshness of the result. The problems are being considered in the context of data analytics. The course considers traditional architectures as well as so-called big-data architectures. Students should consider both small and large datasets because both are equally important, both justifying different trade-offs. Exercises and examples will consider both simple and complex data structures, as well as data ranges from clean and structured to dirty and unstructured. A project must select attempt to process data along some Guidelines  * You need to identify a business problem based on a data set. * You need to implement process that computes the result. * The result should be accessible for review through some kind of serving layer. * 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   1. A proposal presentation. You should think about the proposal as something that you are presenting to an executive where you will justify solving the problem and what solution you are proposing. 2. A final presentation. 3. All Code Submitted to Github 4. The instructor should be able to clone , build and run the project. 5. All required documents and presentations in Github. 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.   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. Milestones **Week 4**: Form groups & pick area and prepare slides for 10 minutes presentation about your goals, challenges, how will you acquire your data and what information organization challenges do you face and your initial plan to complete the project.  **Week 6**: A proposal (~ 2-page ) must be sent to the instructor with sufficient detail of the problem being addressed, analysis of related work, and the supporting research that data can be acquired and organized.  **Week 11** : Progress Report I (~ 5-page): 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 on 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. |  |