**Harvard University Extension School**

**"Principles of Big Data Processing"**

**CSCI E-88, Fall 2019**

**Final Project Requirements**

## Final Project Requirements

* Select a publicly available Big Data source (logs, social feeds, demographic/health/financial feeds) - there are plenty of interesting public data available
  + Example streaming data sources <https://blog.k2datascience.com/real-time-data-sources-for-data-engineering-projects-9bcff65c8468>
  + you can also start your search for public datasets here: <https://www.google.com/publicdata/directory>
  + <https://datacatalog.worldbank.org/>
* Define a problem to solve, questions to answer, interesting use case to experiment with using this data
* Define processing pipeline for your implementation - your pipeline has to include at least 3 of the tiers from the full Big Data Processing pipeline we have learned about in the class:
  + Collection tier
  + Master Storage Tier
  + Messaging Tier
  + Batch Processing Tier
  + Stream Processing Tier
  + Real-Time Views Storage
  + Batch Views storage
  + Distributed Indexing/Search Tier
  + Visualization Tier
* For the pipeline implementation - you should pick at least one of the frameworks/technologies/languages/libraries that were **not exercised** in the class homeworks
* Your project cannot be a copy of/ be mostly comprised of some online tutorial for selected technologies/frameworks - it has to be your original work. You can, of course, refer to tutorials and reference some parts of them, with clear note of that in your solution document. For example, if you decide to use Twitter streaming and processing via Kafka consumers - you cannot just replicate this tutorial: <https://acadgild.com/blog/streaming-twitter-data-using-kafka/>
* Explain design of your pipeline, implementation details and all setup needed to be done for your pipeline
* Demonstrate your results
* All Final projects will be shared with the students of the class - so make sure that there is no confidential/proprietary information and/or data used in the project
* We will allow students to form teams of up to 2 people. This is totally optional and is your responsibility to contact/form groups if desired. For 2-people teams the requirements for the project are: you have to implement at least 4 tiers, and use at least 2 new technologies/frameworks

## 

## Final Project Milestones

1. In the next 1-2 weeks you will get an email with your assignment of the Final Project grader
2. **Final Project Proposal is Due: Sun, Dec 1, noon, EST - Points: 50 - No Late Days!**
   1. you have to submit your Proposal via Canvas and email to your assigned grader - to get it reviewed and approved
   2. we encourage you to submit your Proposal as early as possible to get it approved so that you could continue with your Final Project work
   3. also, as part of working on your proposal - try to do some proof-of-concept work to make sure what you are planning to do is doable - this will also speed up your work on the actual Final Project
   4. You CAN NOT start working on your Final Project until your Proposal is approved
3. **Final Project Due: Sat, Dec 14, midnight EST - Ponts: 250 - NO EXCEPTIONS** - we need to review all final Projects before the Final Project Presentations in class on Dec 17th!
4. **Final Projects Presentations: Dec 17-th**: the last class meeting will be dedicated to the Final Project presentations. Students will play the YouTube video of their project and answer questions from fellow students and teaching staff. We encourage all local students to join us for the last class in person in Boston to enjoy the results of your hard work over the semester!

## Submission Artifacts

1. **Final Project Proposal** - use the provided template, and refer to the example Proposal
2. **3-min YouTube video** - a summary and demo of your project. You can use any of the free video creation tool you can find, here is a list of some examples: <https://www.iskysoft.com/video-editing/video-creator-free.html> Once you create the video - you are required to upload it to YouTube - we cannot accept video uploads as media files (we just do not have the capacity). There are plenty of already posted Final project videos on YouTube from earlier semesters to peek at as an example - just search :) Required points to include into the video:
   1. Project goal/problem statement
   2. Architecture and technologies used
   3. Overview of the implementation
   4. Clear demo of the results
3. **Final Project document** - similar to your Hw solution document - please you the provided template for general structure. Main topics to include:
   1. Summary Page - 1st page: (similar to your proposal but with finalized details and real results)
      1. Project goals/ problem statement
      2. Big Data sources used (if applicable)
      3. Processing pipeline overview and selected technologies
      4. Brief summary of the results
      5. **URL of your YouTube video**
   2. Solution and implementation details
   3. Clearly demonstrated results
   4. Conclusions and Lessons Learned
4. Full source code, configuration files, input data (examples of events, data source), result data (if applicable) - has to be submitted as a separate zip archive

## Bonus Options:

* distributed deployment: deploy/demo your pipeline on AWS/ VMs/Docker : Bonus: +15
* add/implement extra Tiers to your pipeline (in addition to the 3 required ones): Bonus: +20
* use an additional new technology/framework (in addition to the one required one): Bonus: +20

## 

## Grading Rubrik

Total points: 300

|  |  |  |
| --- | --- | --- |
| Requirement | Details | Points |
| Final Project Proposal | Proposal is submitted on-time, in the specified format , includes required details, is approved | 50 |
| 3-min YouTube video | 3-min YouTube video with required details is provided | 100 |
| Final Project Document | Solution document with all required parts and details is provided | 100 |
| Source code | Full source code and all config artifacts are provided | 50 |
| Bonus options: | | |
| distributed deployment |  | +15 |
| Extra Tier[s] implementation |  | +20 |
| Extra new technology[s] |  | +20 |