**Final Project**

For your final project you will need to create an end to end data pipeline and perform some analysis on the data. Some examples of public data set can be found [here](https://github.com/awesomedata/awesome-public-datasets" \l "agriculture).

You will identify a data source that is an API, simulator, or data store and create a data pipeline for ingesting the data using one of the tools we will cover in class or a tool of your choice (you may use Azure Stream Analytics, Kafka, Spark Streaming, Python scripts, etc.).

You will then have to include some analysis about the data that may include statistical tests, machine learning algorithms or relevant visualizations.

You may do the work individually or in groups of up to 3 members.

The final project will consist of three parts:

**Setting up the pipeline**

In the first part you will describe your data pipeline using code, screenshots or any information that you think is necessary. You will describe the technologies used to retrieve the data. You will mention the frequency of data retrieval (only once, once a day for a month, constant stream, etc.). You will also describe all the steps taken to process or clean the data.

**Analysis**

In the second part you will describe the analysis you performed and the questions your analysis answered. For example, you would like to know whether the amount of precipitation in California varies greatly from Florida during the month of July. Another possible question is if there are more postings for housing on Craigslist on the weekends. This analysis must include at least one visualization that can be produced using the tool of your choice.

**Portfolio**

The code and a markdown writeup and code will be in github or similar site.

The project write up is due July 24th at 9pm as a new project in Github and will include your choice of data source, your choice of tools and what questions you intend to answer with your analysis.

The final project will be presented on the last two weeks of class.

**Examples of projects**

If you have trouble thinking of a project. Here are some examples of the type of data you will be analyzing for the project.

Social Network: You can take [Facebook data](https://archive.org/download/oxford-2005-facebook-matrix) and using spark or a graph database review how people are related. You could look at the average number of friends someone has in other schools, or how students of different demographics or major relate to the amount of friends they have.

Gun Violence: You can take [gun violence data](https://github.com/jamesqo/gun-violence-data). Then you would clean the data, do factor analyses, review trends, and predict the future of gun violence.

Bike Share: You could analyze [bike share data](https://github.com/BetaNYC/Bike-Share-Data-Best-Practices/wiki/Bike-Share-Data-Systems) to determine the optimal number of bikes needed in a city. You could compare different city’s data with external data such as weather or terrain.