2.1 Sprint Project: Building an Analytics Ingestion System

The purpose of this project is to build a system which mimics the properties of a data ingestion system. In particular, the modern analytics stack works by taking data generated by a client and then sent to a server for processing.

This process contains a few moving parts and your goal is to build them all in a series of sprints, which are described in the associated assignments. This document will broadly describe what needs to be built. Each specific sprint will address a part that needs to be completed. Before starting there are three key of the system that need to be defined:

1. **Client:** The client is the system which is sending data. This could be any data generating device, such as a Mobile Phone, an IOT device or even another computer. Importantly, for our purposes, the data is being generated on a device different than the device processing it.

2. **Server:** The server is the system which accepts and processes the data from the client. This piece of the data pipeline handles data ingestion and is frequently the starting point for an "ETL" (extraction, transformation and loading) process.

When thinking about the coding process, you should not only keep in mind the system that you are working with, but also the environment that you are working on. While this particular project is simple, most production data science systems will use some form of the below:

- **Local Environment:** The local environment is where code is developed. Generally this will be the computer the developer is working on.

- **Development:** The "Dev" environment is the playground or sandbox where people work.

- **Test:** This is where development code is pushed in order to be tested.

- **Staging:** A staging environment is generally an exact copy of the production environment. It is where code goes to be "staged" for production. Generally a bit of testing is done in this environment, but very little. By the time gets to here, the code should be solid.

The development process usually involves some combination of the above. For our simple exercises, you will probably only need to have a local environment and a production environment.

2.1.1 Sprint 0: Team Building

The purpose of this project is to lay out, organizationally, how you and your team are going to work. You should consider this document a contract signed by all of you.

Part of the difficulties around any project (including those projects involved with data science or software development) is organizational – how do you make sure that everyone on the team is being used efficiently. For most projects (especially when you are still a student), very little work and energy is put into project planning.

If we had infinite amounts of time we could, for each task, define *who* is responsible, *what* the exact tasks is, *when* it is due and *what* degree of quality is required. Unfortunately, we do not have infinite time and instead rely on incomplete contracting – assuming that we can negotiate issues that arises at the time that it arises.

As part of this project, there are two things that are going to be required – a "Sprint 0" contract (what is being done here) and a final reflection assignment which will require you to evaluate both the outcome of the assignment as well as how efficiently you worked as a team.

In order to complete this assignment, please answer the following questions:

1. Who are the members of your team?
    - Focus on relative strengths and weaknesses for each of you.
2. What is your goal for the entire Sprint project?
3. What is your process for making sure that work is distributed evenly?
4. What is your conflict resolution process?
5. What does your work process (as a team) look like? In particular, please explain:
    - When you plan on working on the project. (days / time)
    - Where, physically, do you plan on working on the project.
6. For each of the (computational) tasks, who is responsible and what is your estimate of the amount of work that it will take?
    (a) Setting up (as well as managing and documenting) a test environment (A server that the rest of the team can work on)?
    (b) Setting up (as well as managing and documenting) a code repository?
    (c) Manage credentials?
7. How are write-ups (such as this) going to completed? What does your non-coding process look like?
    - Whose job is it to make sure that a hard-copy is turned in?
    - How will you handle grammar, spell-checking?
    - How will you handle making sure that everyone is comfortable with the result?

The final piece of the Sprint assignment will require revisiting the above assignment. Taking time now to think through this will mean less work in the future. Also, during the assignment you will also be required to do some tracking around the time that you take on different assignments.

Requirements

In order to complete this assignment, please answer the questions above and turn them in (hard copy) at the start of the class in which it is due. The entire document should be around two to three pages.

The assignment will be graded along a few dimensions: effort and grammar being the two most important.