Data Engineering Career Track

Open-ended Capstone Overview

The goal of this capstone project is to encourage students to think like a data engineer. You are expected to select a (large) data set that you would want to work with for the rest of your project duration. You are expected to come up with a proposal of what you want to do with the data and how you can use it to derive value. Here, you need to think like a professional data engineer as to what tools you would use; how you would extract, transform and load the data; what scaling considerations would you have when you are designing your solution.

This project is divided into two phases. In **Phase 1 (step 1-5)** you will start with the ideation of the project you want to work at finding the dataset and writing a project proposal after which you will get your hands dirty with the data by doing exploratory data analysis and thereafter creating a working prototype of the data pipeline that you can run locally on your computer. In **Phase 2 (step 6-10)** you will focus on scaling up your solution and converting a solution that you were running locally into a solution that you can run in the distributed environment in the cloud.

In the end, we want you to be creative with your thought process and design your unique project.

Phase	Unit	Step	Estimated Hours
1	Unit 3. Intermediate Python	Step One: Project Ideas	4-6 Hours
	Unit 6. Git and GitHub	Step Two: Project Proposal	3-5 Hours
	Unit 8. Data Warehousing	Step Three: Data Collection	2-4 Hours
	Unit 11. Advanced SQL	Step Four: Data Exploration	6-9 Hours
	Unit 14. Data Structure and Algorithms	Step Five: Prototyping Your Data Pipeline	12-18 Hours
2	Unit 20. Apache Spark	Step Six: Scale Your Prototype	6-9 Hours
	More to come!		