Chance Pickett (011004180)

DASC 1223 – Intro to Data Science (Dr. Schubert)

Data Description and Data Management

What would you say are the benefits and pitfalls of centralization or de-centralized DataOps?

First, the difference between centralized and decentralized DataOps is control. With Centralized DataOps, “the organization can standardize metrics, control data quality, enforce security and governance, and eliminate islands of data. The issue is that too much centralization chokes creativity” (Medium article). One benefit of centralized development is that it can more efficiently implement worthy ideas to scale. So the ideal development process would have a balanced use of both central and independent processes to reap the benefits of both.

If you were building a company from nothing, what would be the first three steps you’d take in creating a data-driven culture/business?

The first step is to decide the type of data development pipeline to use. An example of a pipeline is sandbox management, develop, Orchestrate, test, deploy, orchestrate, and monitor. This can be altered to be a more waterfall type of pipeline and the amount of centralization and decentralization of the pipeline can also be altered. After planning the pipeline, I would start working on the data sandbox. The data sandbox is a dev environment that is a copy of the data factory. When creating the copy, I will need to be careful of security, governance, or licensing restrictions. The next step is to develop models and algorithms for predictive analytics in the sandbox environment. This step will be the bulk of the coding, version control, and integration with the sandbox.