KAPERNIKOV

A data pipeline for point cloud data

Romain Perret

Data engineering

INTRODUCTION

Data engineering consists in managing various data items during their entire processing, and sometimes lifetime, in what is referred to as a pipeline. In each step of a pipeline, the data might need to be validated and cleaned, then being processed or analyzed before potentially being transformed into another format and passed to subsequent steps. The goal of a pipeline is thus to automate a sequence of operations on some data, while being able to follow exactly what happens to that data at each step.

PRACTICAL PURPOSE

In many companies, data engineering has become an important tool to simplify the management of complex data infrastructures where many data sources have to be used to produce reports and metrics about what happens in the organization. But it also applies to the data accessed by clients of such companies if they provide digital services. In that case, data engineering reduces the efforts needed to maintain coherent data across the organization, and more importantly it facilitates the use and access of such data.

The internship

MISSION

This internship will allow the intern to work on a data pipeline implemented by Kapernikov for a famous Belgian public transport company.

Context

Originally, multiple c++ AI tools were developed by Kapernikov machine vision members to detect features based on a given point cloud input file, then generating different metrics, alerts and data files for the company. These binaries were then executed with a shell script, which was later on transformed into a Dockerfile allowing better portability through containerization.

However, even with docker it remains hard to track and manipulate data between the different c++ processes (steps), and allow for a flexible yet scalable configuration of these different steps. For this reason it was decided to create a pipeline that would automate all the processing done on a point cloud file, and allow easy configuration and error reporting. After experimenting with different solutions (python), we chose to rely on Dagster to achieve this and facilitate the use and extensibility of the pipeline.

Additionally, the project had to consider the use of the company's private cloud to leverage the power of computing clusters like increased fault tolerance (executing multiple instances of an app) and more parallelism for intensive workloads. Consequently, we also use kubernetes extensively as it is the predominant tool for deploying and managing containerized applications on clusters.

What we expect

The intern should be a master student motivated to work in the data engineering world. He or she will have to understand the current pipeline architecture and develop a new functionality on top of it. We expect a good understanding of computer science as a whole, but some competences are required for the task we propose:

- good experience with python
- good experience with shell scripting
- good experience with relational databases
- knowledge about networking
- can work on linux
- willing to learn kubernetes