CENTRAL EUROPEAN UNIVERSITY

CAPSTONE PROJECT SUMMARY

BALÁZS VADNAI

DEPARTMENT OF ECONOMICS AND BUSINESS

TECHNOLOGY MANAGEMENT AND INNOVATION

2021

Table of Contents

INTRODUCTION

As a software engineer, I adopted the “always be on the lookout for new technologies” approach that is required to keep my knowledge up-to-date and relevant. However, given my background I often notice myself wanting to explore something in-depth and in itself. I have always been fascinated with Big Data and data science, this is partly the reason why applied to the TMI program and took technical courses which were otherwise not mandatory for TMI

students. I chose my Capstone Project too in accordance with my needs of it being technical enough while also providing a solution for a relevant problem in today's world.

The Client

The client I have picked for this project does exactly that: they offer data-driven location intelligence and scenario planning by developing unique indices quantifying important measures that characterize the quality of living in urban areas and neighborhoods, where nowadays the majority of the world's population dwells.

The Problem

The proposed project contributes to this index-development phase by incorporating air quality, a major environmental indicator being of crucial importance in many aspects of a healthy urban life into the client's platform.

The Air Quality Index (AQI) provides citizens with the basic information needed to fully understand the health impacts of industrialised sectors of the economy and society. The AQI provides local authorities with a public service for the assessment of health risks and environmental contamination within their communities.

THE PROJECT

Parts of the project

First, we defined the subtasks which needed to be done in order to accomplish the initial goal. These subtasks were:

- Identifying relevant data sources for air quality measurements and describing their available features.
- Write crawling algorithms to collect them about the selected air-quality data sources
- Explore and analyze the quality of the different data sources, conduct comparative measurements
- Curate sets of urban features from the urban areas nearby the sensor locations

Simultaneously, given my extensive background in the fields of Information Technology, I took the role of a consultant, aiming to facilitate the data scientist team's transition to Microsoft's Azure cloud infrastructure to obtain a more effective workflow with a large amount of geospatial data - with cardinality in the hundreds of millions.

How the project went

Air quality measurement

Given the shortness of time, we narrowed my research area to Budapest. This made my task somewhat personal, as me falling under the sensitive group and the not-so-fortunate air quality conditions of my hometown made me vigilantly monitor the AQI for several years now.

I identified two data sources for further analysis:

- the sensor data from OMSZ, the official provider for Hungary. There are only eleven monitoring stations in Budapest, thus not providing granular enough data for street-level predictions
- a community-based air quality measurement project consisting of thousands of sensors worldwide, from which around 130 are located in Budapest. This can provide satisfactorily detailed data; however, outer city districts are still not covered

I wrote a crawling algorithm to scrape the OMSZ data from their website. The community project has an API for querying and filtering the current output of their sensors. The data collection went on for roughly a month, happening on an hourly basis.

As data was being gathered, we specified the final deliverables of the project, that is having the following parameters attached to each urban area:

1. absolute values of air polluters
2. relative level of different areas based on the level of air pollution and their individual characteristics
3. temporal trends of the air pollution levels (e.g. it has increased/decreased throughout a given period of time)

Cloud infrastructure

At the same time, I was working out solutions to move, store, access and consume immense data tables consisting of geospatial information, regularly discussing my findings with the data science and consulting with them about the next steps. The ultimate goal is to provide a technological environment for the client which facilitates Big Data analysis and thus the productivity of the company.

Results

As of this writing:

1. The gathered air quality data turned out to be not as clean as initially expected, further analysis is required

   As there are two types of data sources, a correlation must exist between these two in order to make predictions. Analyzing one month of data, unofficial measures do not correlate with the official ones, however, official sensors seem to have no correlation with other official sensors either, whereas unofficial data in itself does correlate. Binary prediction shows that the source of the data (official/unofficial) can be recognized just by looking at the measured values; this should not be the case.

2. The data team is not yet able to fully operate on the cloud platforms
   An Azure SQL database has been created; raw data has been migrated to data tables with heavy spatial indexing in order to filter millions of rows of geolocation information as quickly as possible. Yet there is a disparity between the original and the migrated data which needs to be solved before analytical work can begin on the cloud platform.

There were and still are challenges, but the project continues, as well as my involvement in it. The client and I have agreed on further cooperation.

PERSONAL REFLECTION

I had the chance to broaden my horizons by being in a work environment I was not previously familiar with. I learned about the tools and methods professional data scientists use, and also learned some of the limitations they encounter.