The Lightweight IBM Cloud Garage Method for Data Science

Project Description:

This project serves as the capstone project of the IBM Advanced Data Scientist certification and its purpose is to show an E2E pipeline where data is used to make data-driven decisions. My specific project bases on the work of Moosavi et al (2019) and aims to analyze the traffic data in the USA between February 2016 and June 2020 in order to predict the severity of traffic accidents depending on various features like geographical, time, environment, etc.

The project is divided in three scripts (Jupyter Notebooks):

1. *USAccidents_data_exp.ipynb*: First data exploration and high-level data preprocessing. Many plots are displayed in order to get as many insights as possible. This script outputs an new csv file (*US_Accidents_June20_CLEAN_Week1.csv*) that will be used in the next script. It also outputs a table that summarizes the features of the dataset with feature description, data types and an example.
2. *USAccidents_etl_and_feature_eng.ipynb*: This script prepares the data to be fed to the model. In a first step it typecasts all features to the right category. Then it creates and transforms features (i.e. frequency domain) that are more useful for the model. In a last step, categorical features are encoded. This script outputs a new csv file (*US_Accidents_June20_CLEAN_Week2.csv*).
3. *USAccidents_model_def_train_evaluate.ipynb*: This model researches the prediction power of different machine learining models. In a first step, a basic model is deployed in order to define a baseline performed. Afterwards, two more state of the art machine learning models are implemented and finally a basic deep learning model is also compared to the other models.

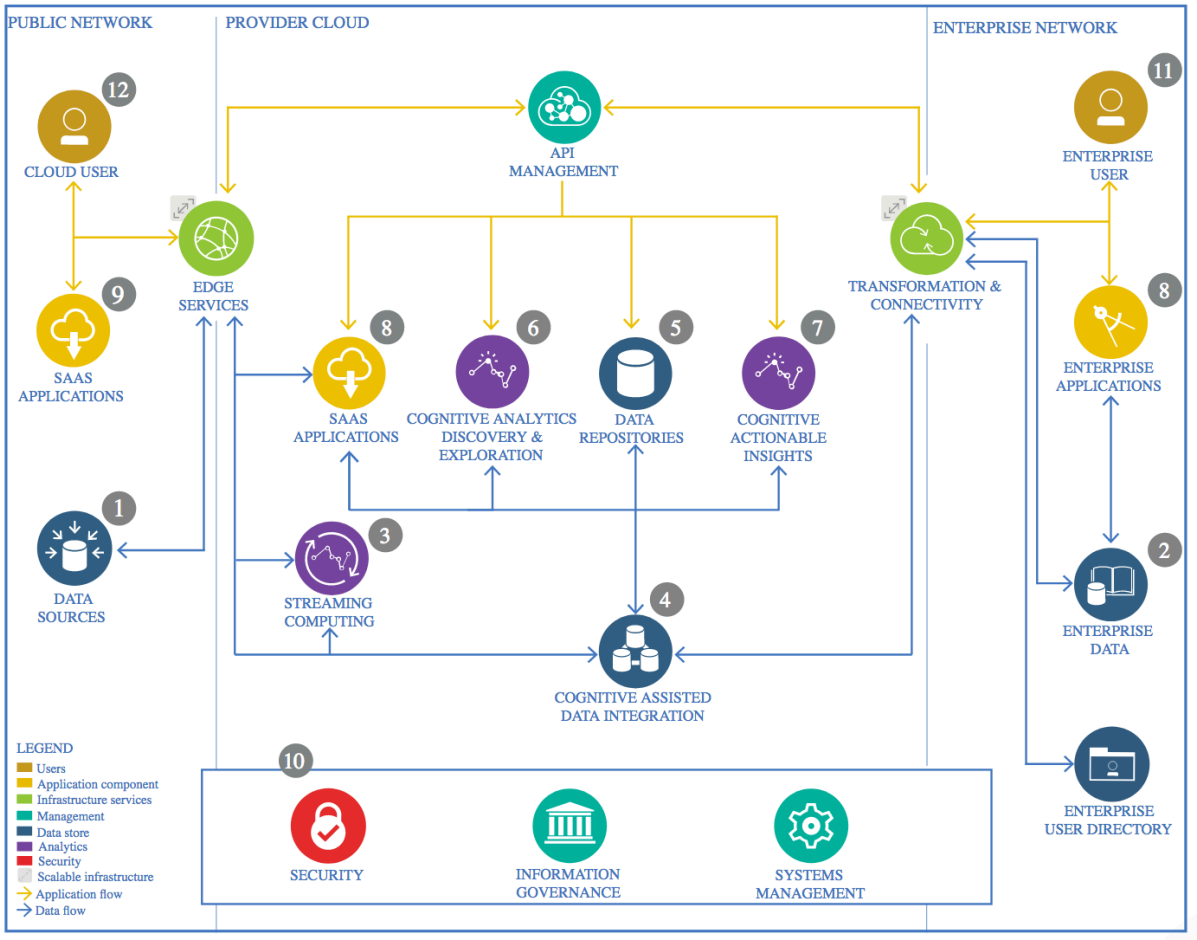
All scripts were run on a local machine and all data required is in the csv provided by the publication of Moosavi et al.

During phases 2 and 3, (‘*ETL and Feature Creation’* and ‘*Model Definition, Training and Evaluation’* respectively) several iterations took place, following the **Lightweight IBM Cloud Garage Method for Data Science**.

Below, one can find the Architectural Decisions Document (ADD), where each of the components of the architecture of the pipeline are described, as well as the specific justifications for each of them.

Architectural Decisions Document Template

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

1.1.1 Technology Choice

The only data source is the *US-Accidents: A Countrywide Traffic Accident Dataset*, published by Moosavi et al (2019) *(https://smoosavi.org/datasets/us_accidents).*

1.1.2 Justification

The dataset covers all the countrywide traffic accidents between February 2016 and June 2020. It covers the accidents captured by a variety of entities in the US and includes data from traffic cameras and traffic sensors. There are over 3.5 million accident records. Since the dataset already consolidates traffic data from several different sources, no other data sources were used.

1.2 Enterprise Data

1.2.1 Technology Choice

None (see justification below).

1.2.2 Justification

My project is just a PoC with the data recorded until June 2020. In case a real-time solution should be built for an enterprise, a cloud-based solution would make sense in order to fit scalability and responsiveness, but in this case, it does not make sense to include this module.

1.3 Streaming analytics

1.3.1 Technology Choice

None (see justification below).

1.3.2 Justification

Same justification as for 'Enterprise Data'.

1.4 Data Integration

1.4.1 Technology Choice

Jupyter Notebooks

1.4.2 Justification

Jupyter Notebooks allows running the code cell by cell, which is really useful for debugging. Also, it has a lot of libraries that ease handling the data.

1.5 Data Repository

1.5.1 Technology Choice

The cleansed data are stored **locally** in a separate csv-file (*US_Accidents_June20_CLEAN_Week2.csv*), that is ready to be fed to the predictive models.

1.5.2 Justification

Due to the small size of this project, it doesn't make sense to store it on the cloud.

1.6 Discovery and Exploration

1.6.1 Technology Choice

Jupyter Notebooks. Name of the file: *USAccidents_data_exp.ipynb*

1.6.2 Justification

Jupyter Notebooks allows running the code cell by cell, which is really useful for debugging. Also, it has a lot of libraries that ease plotting the data: *matplotlib* and *seaborn* were used in this project.

1.7 Actionable Insights

1.7.1 Technology Choice

Jupyter Notebooks. Name of the file: *USAccidents_model_def_train_evaluate.ipynb*

1.7.2 Justification

With the library scikit-learn one gets to develop machine learning models very quickly.

1.8 Applications / Data Products

1.8.1 Technology Choice

None.

1.8.2 Justification

Not sure yet I'll do something with the insights I get from this project. Maybe an interactive dashboard would be an option in the future.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

None.

1.9.2 Justification

For a Proof of Concept (PoC) of this size, I am not going to cover security matters.