Architectural Decisions Document

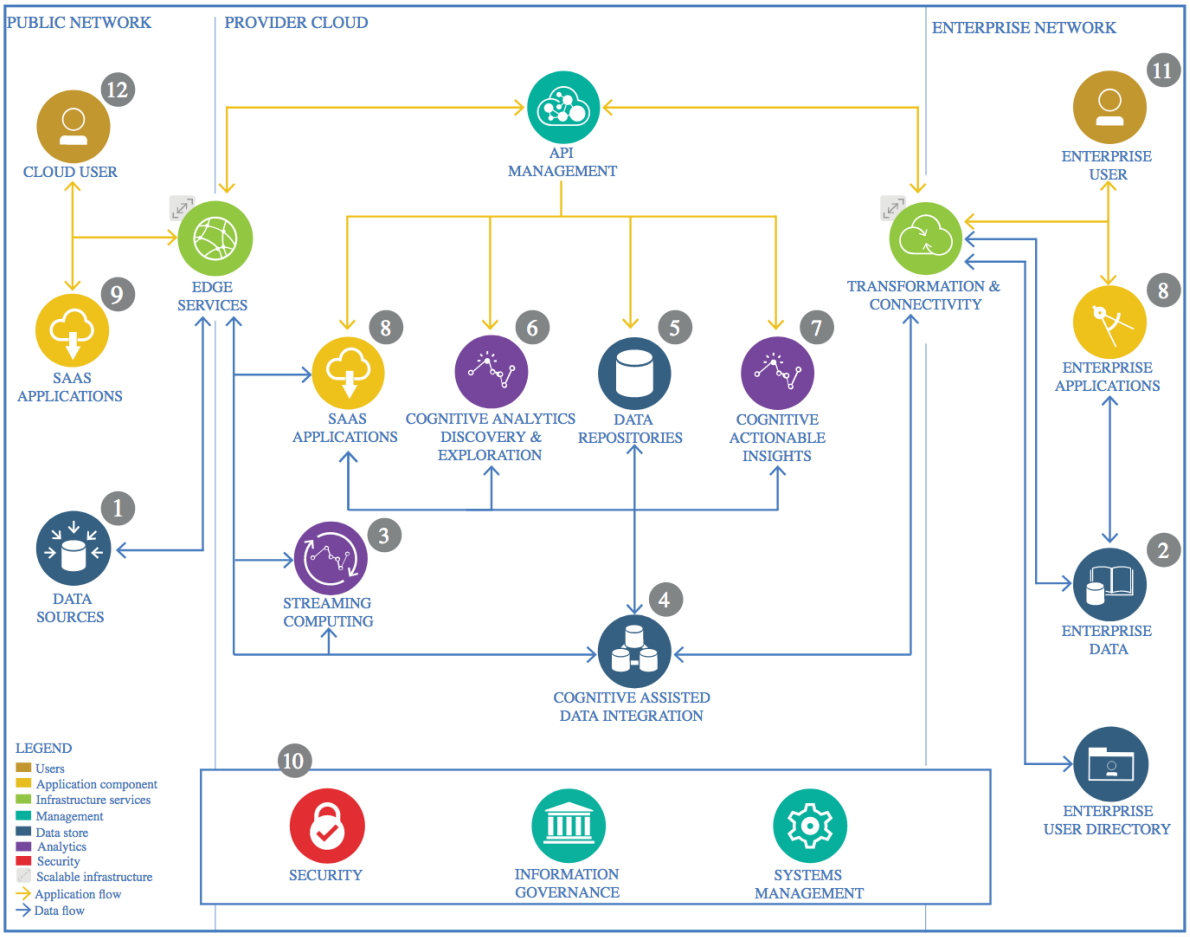
Project: Analyze, model and forecast extreme weather events in Basel (Switzerland, CH)

Edoardo Farina

Architectural Components Overview

The present work is based on the lightweight IBM Cloud Garage method process model. References to the approach method can be found at the following website: https://developer.ibm.com/articles/the-lightweight-ibm-cloud-garage-method-for-data-science/

In the following document, the principal components relevant in this project will be highlighted.



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

Data Source

Technology Choice
The main data source of the project relies on .csv formatted data provided by the meteoblue.com online service. It allows to download meteorological information, such as temperature, humidity, precipitation, wind and others for several locations on Earth, accessing several datasets spanning over a large period of time. In our case data related to Basel (CH) has been downloaded, from 1984 to 2021.

Justification
The data perfectly fits the project requirement: studying and modeling the meteorological events considered as “extreme”, over the years. Data are provided hourly; data cleaning and synthesis is therefore needed.

Enterprise Data

Technology Choice
Data are downloaded locally, despite some very convenient APIs are also provided. In the proposed study datasets are of the order of 100MB, easily fitting local computer disk. Cloud solutions are not strictly required for this project, despite the all analysis framework and analysis is designed to be executed in parallel and cloud environments

Justification

Streaming analytics

Technology Choice
No streaming analytics have been implemented.

Justification
New meteorological data are provided daily. A possible project improvement would be to automatize the new data fetching and addition to the analysis pipelines.

Data Integration

Technology Choice
Data are processed and managed as Apache Spark dataframes. This framework easily scale the workload on the available workers, benefiting of performance optimization, SQL-like queries, and user-friendly data-handling APIs. Since the project has been developed in python, in particular in Jupiter-notebook, Apache pyspark has been used. Data are cleaning, filtered for NULL values, ordered by date. Matplotlib plots are used for a quick data visualization. Data have been grouped by YEAR, allowing to appreciated trend in the temperature or precipitation records.
Finally, meteorological events considered extreme have been added to the data as additional column, and the trend over the years of such events have been computed.

A RNN based on LSTM has been developed to model the 34 years data available, in particular the daily maximum temperature. The framework used is keras, using Tensorflow as backend. The project is structured in several jupiter-notebooks: ETL, EDA, modelling and validation. Model performance has been assessed by means of RMSE indicator.

Justification

Apache Spark is extremely versatile, powerful and scalable. It is written in Scala, but the python interface allows to exploit the main features in Python Jupiter-notebook as well.
Matplotlib is among the references for data visualization for Python; Keras and Tensorflow among the leading frameworks in ML.
The RNN model has been developed to exploit correlations between the maximum temperature among days, and the layers based on LSTMs have been used as LSTM proved to be a solid and valid architecture for the purpose of the study.

Data Repository

Technology Choice

The data is maintained locally. Notebooks are available on GitHub, as well as the trained model and cleaned data. A second option to use ObjectStorage on Watson Studio is also envisaged.

Discovery and Exploration

Analysis shows that Basel is facing a temperature increase, both average and maximum since 1984. This is in line with the well-known global warming trend. Less obviously, Basel is experiencing less and less precipitations (intended as overall water over the year), but, on the other hand, characterized by more and more violent events, such as floods and storms.

Actionable Insights

Technology Choice

The LSTM model proved to efficiently model the past 34 years (provided the daily maximum temperature as parameter). The model works well on previously-unseen data, proving to be a possible tool also for data forecasting.
Forecast of the maximum monthly temperature have been also performed by the model for the next 6 months, from August 2021 to January 2022.

Applications / Data Products

Technology Choice

Forecasting the intensity and frequency of intense meteorological events is of utmost importance in the development and adaptation of our life and our cities to the climate change. These events are increasing and so, as shown in the project, also their violence.