Architectural Decisions Document

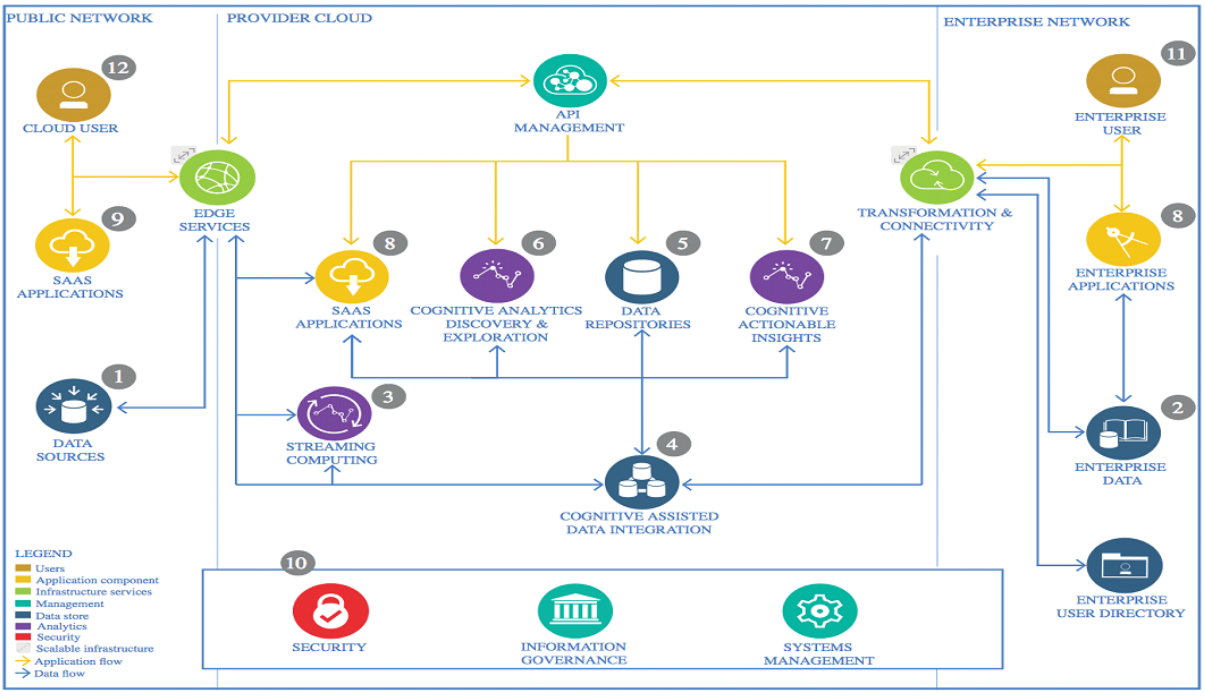
Project Title: Water Quality : Drinking water potability

Author: Hrishikesh Kini

Architectural Decisions Document

1. Architectural Components Overview

The project uses the lightweight IBM Cloud Garage Method process model. The lightweight IBM Cloud Garage Method for data science includes a process model to map individual technology components to the reference architecture. This method does not include any requirement engineering or design thinking tasks. Because it can be hard to initially define the architecture of a project, this method supports architectural changes during the process model.



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

2. Data Source

Technology Choice :

Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges. This dataset is downloaded from Kaggle (https://www.kaggle.com/adityakadiwal/water-potability)

Justification :

Primary reason to download from Kaggle was availability and ease of use.

3. Enterprise Data

Technology Choice :

Data can be downloaded in pdf, xml, csv or excel formats with different time intervals from different locations and weather using government websites. In this case, we use csv files.

Justification :

These datasets are not huge. Therefore, they can be easily downloaded and stored in csv format, which is very easy to read and to work with in python.

4. Streaming analytics

Technology Choice :

Government and some NGO websites provide data on water quality readings nearly in real-time. That means, the prediction model can be fed in almost real-time and we can have streaming analytics and forecasting. In this project, streaming analytics is not used for simplicity. But it can be implemented at any time. For example, using IBM Streaming Analytics service.

Justification :

IBM Streaming Analytics provides fast streaming application delivery using Python. Data Scientists and Developers can use existing Python code for building Streams applications without starting from scratch.

5. Data Integration

Technology Choice :

All the datasets have been downloaded to a local machine. In the preprocess jupyter notebook can be seen that data is cleaned, merged and get ready for building a model. (In case of a real project or huge data, they can be loaded into a data warehouse for example IBM Object Storage)

Justification :

Jupyter notebooks and python are now mostly used by data scientists and they are easy technologies to work with. That’s why everything is done using python.

6. Data Repository

Technology Choice :

Part of the job is done locally. So, there's a directory with all the data on the local machine. Moreover, they are pushed regularly to a GitHub repository as backup. The other part of the job, which includes training ANN, is done on cloud, specifically on IBM Watson studio. The models are then stored to IBM Object Storage, and finally downloaded to a local machine.

Justification :

IBM Watson Studio provides an environment with 4 vCPU and 16GB RAM for free. Specially, for the ANN, which takes a lot of time for training, this works much better than a local machine. Also, storing the models in Object Storage can be used later for building an interactive product using a REST API for example.

7. Discovery and Exploration

Technology Choice :

There is a Jupyter notebook especially for EDA. In these notebooks, data is explored.

The following Python 3.8 libraries were used for Data Exploration and Visualization: - Pandas,

Matplotlib,

Seaborn

Justification :

The size of the dataset was the key factor in deciding data exploration tools. The current data small enough to be processed on a single computer ruling out the need for distributed processing (Spark, pyspark)

8. Actionable Insights

Technology Choice :

There is a Jupyter notebook especially for EDA. In these notebooks, data is explored.

The following Python 3.8 libraries were used for Data Exploration and Visualization: -

Model used for predicting the potability of water are:

- Random Forest
- Support vector Machine
- Tensorflow keras

Justification :

To understand the Correlating features a white-box model was required. Tree based algorithms were identified as a good match. ThusRabdom Forest was used. Neural network based algorithm was used as a reference for the Tree based model. Easiest and Fastest implementation is possible in keras. Tensorflow is the backend.

9. Applications / Data Products

Technology Choice :

A Jupyter notebook based report was generated. The models are then stored in IBM Object Storage. And using the IBM machine learning service model has been deployed on the ibm cloud and scoring endpoint has been generated. Then predictions are being made using a scoring endpoint.

Justification :

As only the correlating factors needed to be identified, a Jupyter notebook based report was considered sufficient. As only the correlating factors needed to be identified, a Jupyter notebook based report was considered sufficient.

10. Security, Information Governance and Systems Management

Technology Choice :
Since the project is deployed on cloud the https and ssl network is used.

Justification :
NA