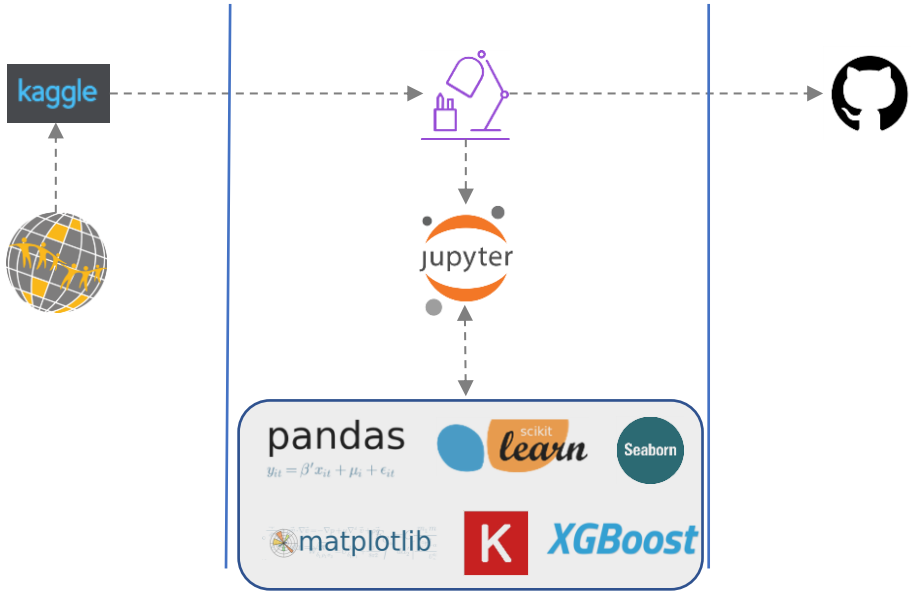
The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document

Advanced Data Science Capstone Project
By Luciano Guerra Domínguez

1 Architectural Components Overview



1.1 Data Source

1.1.1 Technology Choice

Data source comes from CSV files shared in a repository. Three different files with 24,8MB of data.

1.1.2 Justification

Data comes from a Kraggle competition: “DS1 Predictive Modeling Challenge”. The way the data is recollected by Taarifa and the Tanzanian Ministry of Water is manually, filling CSV files.

1.2 Data Integration

1.2.1 Technology Choice

IBM Watson Studio Jupyter Notebooks, scikit-learn, pandas, matplotlib, XGBoost, Keras, Seaborn.

1.2.2 Justification

1.2.2.1 IBM Watson Studio

IBM Watson Studio provides tools for data scientists, application developers and subject matter experts to collaboratively and easily work with data to build and train models at scale. It gives you the flexibility to build models where your data resides and deploy anywhere in a hybrid environment so you can operationalize data science faster.

1.2.2.2 Jupyter Noteboks

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning, and much more.

1.2.2.3 Scikit-learn

Scikit-learn (formerly scikits.learn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

1.2.2.4 Pandas.

In computer programming, pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals.

1.2.2.5 Matplotlib

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged. SciPy makes use of Matplotlib.

1.2.2.6 XGBoost

XGBoost is an open-source software library which provides a gradient boosting framework for C++, Java, Python, R, and Julia. It works on Linux, Windows and macOS. From the project description, it aims to provide a "Scalable, Portable and Distributed Gradient Boosting (GBM, GBRT, GBDT) Library". Other than running on a single machine, it also supports the distributed processing frameworks Apache Hadoop, Apache Spark, and Apache Flink. It has gained much popularity and attention recently as the algorithm of choice for many winning teams of machine learning competitions.

1.2.2.7 Keras

Keras is an open-source neural-network library written in Python. It is capable of running on top of TensorFlow, Microsoft Cognitive Toolkit, Theano, or PlaidML Designed to enable fast experimentation with deep neural networks, it focuses on being user-friendly, modular, and

extensible. It was developed as part of the research effort of project ONEIROS (Open-ended Neuro-Electronic Intelligent Robot Operating System), and its primary author and maintainer is François Chollet, a Google engineer. Chollet also is the author of the XCeption deep neural network model.

1.2.2.8 Seaborn

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

1.3 Data Repository

1.3.1 Technology Choice

Cloud object storage makes it possible to store practically limitless amounts of data.
It is commonly used for data archiving and backup, for web and mobile applications, and as scalable, persistent storage for analytics.

1.3.2 Justification

This project is more related to Machine Learning and simple Deep Learning models Proof of Concept on a small dataset so local machine storage looks to be relevant. We have no specific needs of specialized hardware such as GPUs, massive datasets storage, high throughput I/O to require a Cloud storage.

1.4 Discovery and Exploration

1.4.1 Technology Choice

Watson Studio, Jupyter Notebooks.
The data quality is assessed with EDA performed with pandas, pandas-profiler and matplotlib libraries.

1.4.2 Justification

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

1.5 Actionable Insights

1.5.1 Technology Choice

There exists an abundance of open and closed source technologies (IBM Watson Studio, Jupyter, Tensor Flow and Keras…).

1.5.2 Justification

Here, the most relevant are introduced. Although it holds for other sections as well, decisions made in this section are very prone to change due to the iterative nature of this process model. Therefore, changing or combining multiple technologies is no problem, although decisions let to those changes should be explained and documented.

1.6 Applications / Data Products

1.6.1 Technology Choice

IBM Watson Studio Jupyter Notebooks, scikit-learn, pandas, matplotlib

1.6.2 Justification

The components mentioned above are all open source and supported in the IBM Cloud. Some of them have overlapping features, some of them have complementary features.