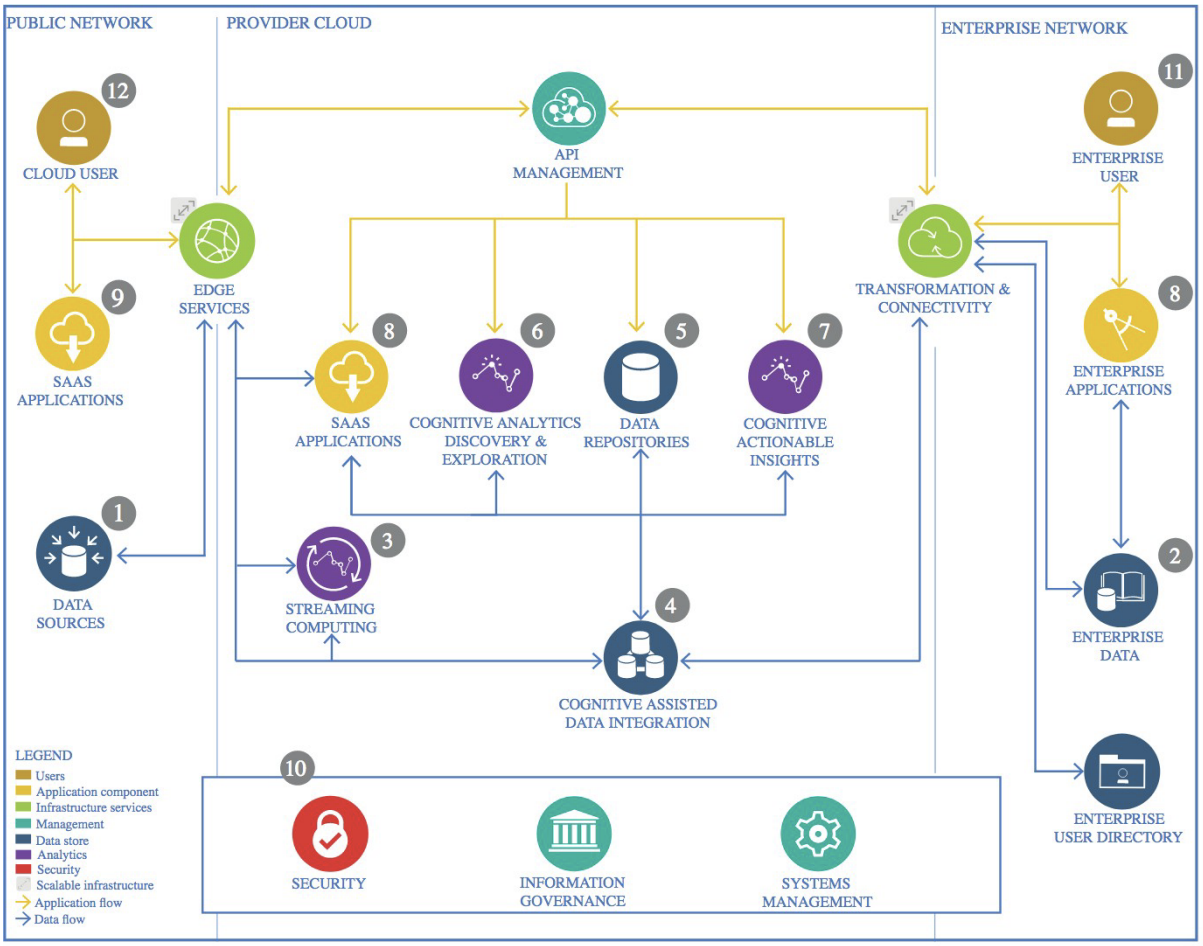
The Lightweight IBM Cloud Garage Method for Data Science

1. Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

- Data Source
  - Technology Choice
    - IBM Watson Studio to download the data into the system.
  - Justification
    - CSV file with the results of web scrapping. It is a convenient way to import data.
- Enterprise Data
  - Technology Choice & Justification: NA
- Streaming analytics
  - Technology Choice & Justification: NA
- Data Integration
  - Technology Choice: Pandas, sklearn.
  - Justification
    - Pandas were selected as it had a simple interface and a big community where you can find the help if you need. It is a flexible and easy to use data analysis and manipulation tool. Also, the dataset size is not big, so

we don’t need parallelization. Sklearn provides easy tools to impute features and do further analytics.

- Data Repository
    - Technology Choice: IBM Watson
    - Justification: It’s easy to integrate and call from the notebook. It allows you to can save and load data between the notebooks. It allows you to divide the process of development into structured steps and save the data on each of them.
- Discovery and Exploration
    - Technology Choice: Pandas, Matplotlib, Seaborn
    - Justification: Matplotlib and Seaborn are common, and handy tools for data visualization. They got a lot of built-in functions to plot correlation matrices, histograms, scatter plots, and other useful things.
- Actionable Insights
    - Technology Choice: Keras
    - Justification
        - User-Friendly and Fast Deployment.
        - Quality Documentation and Large Community Support.
        - Pretrained models,
- Applications / Data Products
    - The realization of our research as data product lies beyond the project, but the possible application, that can be based on it is the website or analytical system, that could provide a coach help in deciding the chances of an athlete winning medals. All of the above can be done on the base of the primary analysis, as we proved that even the result could be predicted based on the physique data. Possible technology is any standard web stack, for example, MySQL + Django application.
- Security, Information Governance and Systems Management
    - Security should be maintained on the analytics side to avoid un-anonymizing of patients against their will. In our project, we are using already anonymized data, where each athlete goes with just an ID, and we don’t need to perform any actions on the security side.

2. Development process

The development process is better described and documented in the notebooks.

2.1. Why have I chosen a specific method for data quality assessment?

I extracted only a particular part from the data set, and conducted standard preprocessing, of course, to make models work.

2.2.Why have I chosen a specific method for feature engineering?

I tested various methods of feature engineering. First of all, I followed the standard guidelines on the question. Scaling and One-Hot encoding are de-facto nowadays. The imputation of missing values was used.

2.3.Why have I chosen a specific algorithm?

I have chosen Logistic Regression as this is one of the simplest machine learning algorithms and is easy to implement yet provides great training efficiency.

2.4. Why have I chosen a specific framework

I have chosen Keras for the user-Friendliness and fast Deployment. It also has quality documentation and large Community Support with pre-trained models.

2.5.Why have I chosen a specific model performance indicator?

I used two performance indicators - accuracy and the AUROC. I have chosen them as standard quality metrics for binary classification. In the end, I decided to concentrate on maximizing the AUROC, as it leaves the opportunity to decide which threshold you want to set in favor of minimizing desired error (type 1 or type 2) but still improves the model. AUC is, I think, is a more comprehensive measure in my case than simple metrics.