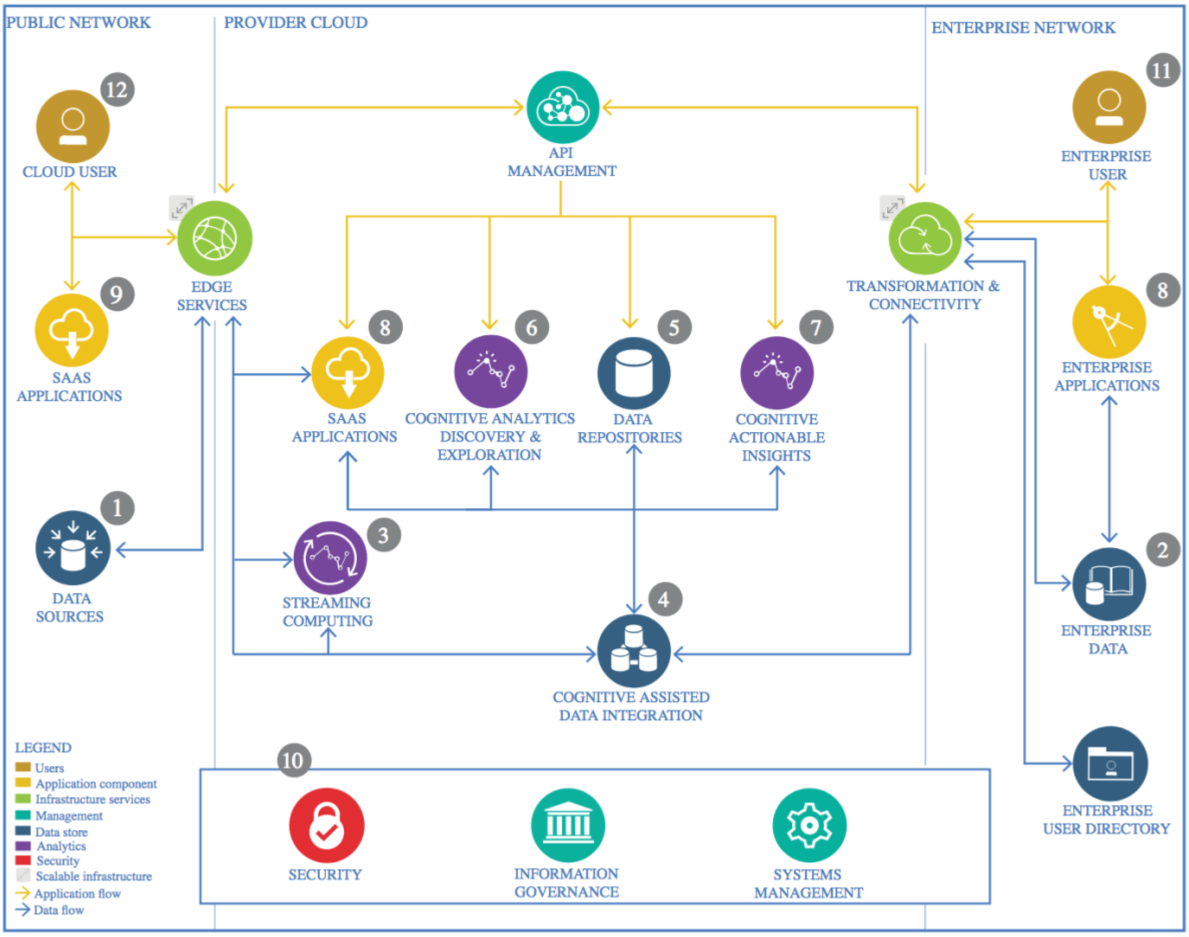
The Lightweight IBM Cloud Garage Method for Data Science

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 data was downloaded from Kaggle (https://www.kaggle.com/uciml/breast-cancer-wisconsin-data)

1.1.2 Justification

Kaggle contains data on a wide variety of areas of knowledge, being my main interest biomedicine. Also, Kaggle offered a friendly interface for data preview.

1.2 Enterprise Data

1.2.1 Technology Choice

NA

1.2.2 Justification

The data I used corresponded to a publicly available biomedical research carried out in Wisconsin by Dr. Street et al.

1.3 Streaming analytics

1.3.1 Technology Choice

NA

1.3.2 Justification

In this particular cases, analyses were static, so no streaming analytics tool or technologies were taken into consideration.

1.4 Data Integration

1.4.1 Technology Choice

NA

1.4.2 Justification

There is only one data source which already contained all the instances and features for the goal of my study.

1.5 Data Repository

1.5.1 Technology Choice

In this case a GitHub repository was used, alongside with IBM Cloud Storage and local disk.

1.5.2 Justification

This option would enable keeping data up-to-date with multiple modifications and storage for the different dataframes and graphics created.

1.6 Discovery and Exploration

1.6.1 Technology Choice

In this case, data exploration consisted in the following steps. (i) Dataset import and exploration of data types and features. (ii) *Ad hoc* data documentation and information retrieval regarding the case-study, in this case mammary biopsies. (iii) Explorational visualization (boxplot) (iv) Analysis of correlated features. The following Python 3.7 modules were used for Data Exploration and visualization:

- Numpy
- Pandas
- Seaborn
- Matplotlib
- Sklearn

1.6.2 Justification

In this case, the size of the dataset did not require to use distributed processing (Spark). The usage of the previously-mentioned modules is considered a standard in the Data Science workflow. Such modules enable a global understanding of the most relevant features in the dataset.

1.7 Actionable Insights

1.7.1 Technology Choice

In this case, data analysis and model generation consisted in the following steps. (i) Data cleansing (ii) Feature Engineering (iii) Model generation. The following Python 3.7 modules were used for Data Analysis and model generation:

- Numpy
- Pandas
- Sklearn
- Scipy

1.7.2 Justification

In order to solve a classification problem, we built two baseline ML models that were reviewed in the previous courses in this specialization diploma: Random Forest and Support Vector Machines (SVM). The chosen metric for model evaluation was the accuracy.

1.8 Applications / Data Products

1.8.1 Technology Choice

Jupyter notebook-based report and google slides presentation, thus only informative purposes.

1.8.2 Justification

Simple and easy to use, both techniques expose the technical aspects aspects of the work in a way that is approachable for data scientists and biomedical researchers in this case.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

NA

1.9.2 Justification

I did not focus on the security aspects nor the systems management of this project.