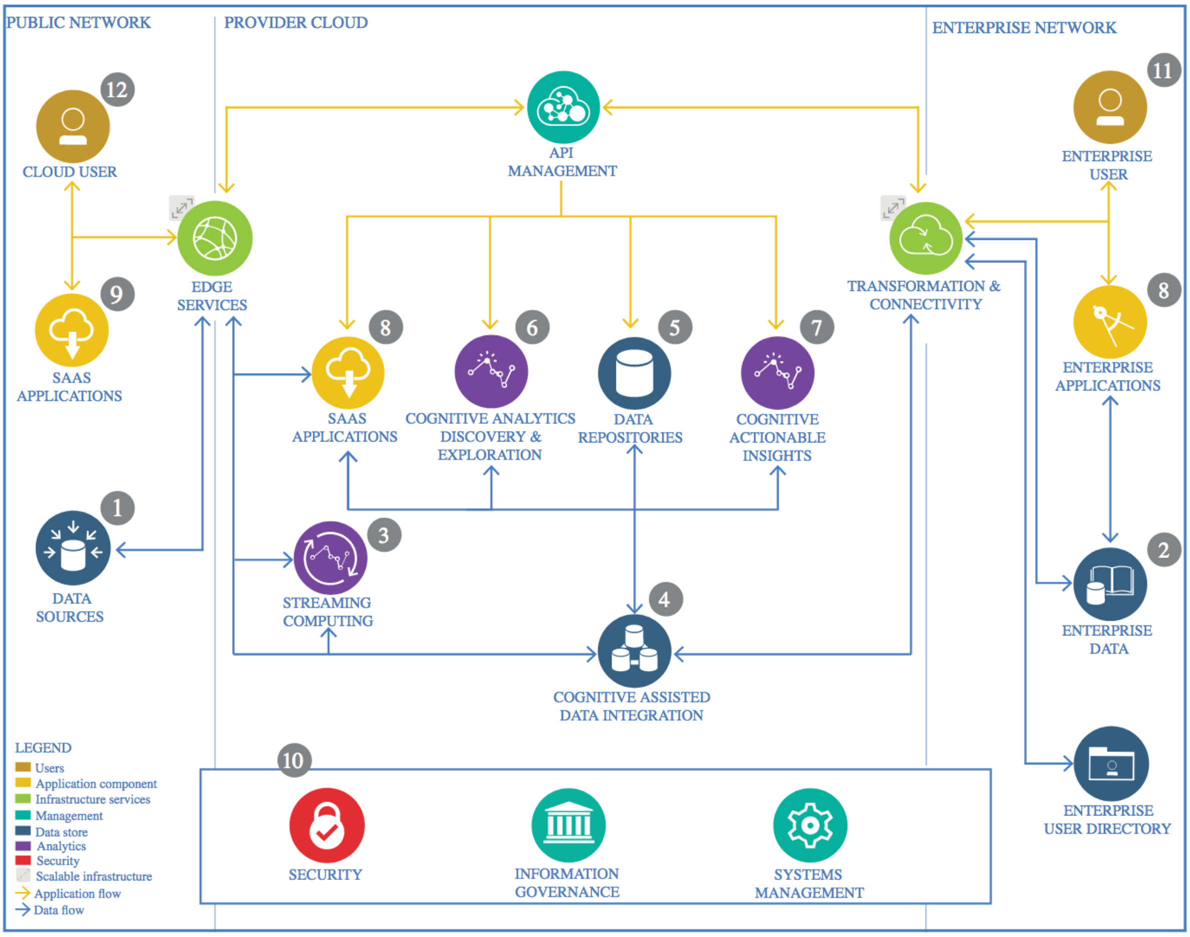
The Lightweight IBM Cloud Garage Method for Data Science

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

Understanding data is one of the most important part when designing any machine learning algorithm.The data was downloaded from Kaggle (https://www.kaggle.com/uciml/breast-cancer-wisconsin-data). CSV (coma separated values format).
123 KB of data.

1.1.2 Justification

The reason to download from Kaggle was availability and ease of use. The CSV file provided is a common format for table data, separator by ‘,’.

1.2 Enterprise Data

1.2.1 Technology Choice

GitHub repository

1.2.2 Justification

To available for every person every time on the repository

1.3 Streaming analytics

1.3.1 Technology Choice

NA

1.3.2 Justification

NA

1.4 Data Integration

1.4.1 Technology Choice

Not used.

1.4.2 Justification

Not used.

1.5 Data Repository

1.5.1 Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

1.5.2 Justification

Please justify your technology choices here.

1.6 Discovery and Exploration

1.6.1 Technology Choice

Jupyter Notebooks the following Python 3.6 libraries were used for Data Exploration and Visualization: - Pandas, Matplotlib and Seaborn.

1.6.2 Justification

Because I feel familiar with it and easy to use specifily with jupyter notebook you can know the parameter and read the documentation of it. 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, data visualization, machine learning, statistical modeling, and much more.

1.7 Actionable Insights

1.7.1 Technology Choice

The following Python 3.6 libraries: -
Pandas , numpy, sklearn and Tensoflow.

In Classifications, we will use following 2 Techniques to train our model and predict:

1. Random Forest
2. Support Vector Machine

1.7.2 Justification

We use sklearn library because is most common libraries that introduce the predicted model, We choose Random Forest because power to handle a large data set with higher dimensionality. for SVM because ususlly use for two classes.

We gone use F1 performance indicator because better measure of the incorrectly classified cases

1.8 Applications / Data Products

1.8.1 Technology Choice

A Jupyter notebook based report was generated.

1.8.2 Justification

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

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

NA.

1.9.2 Justification

NA.