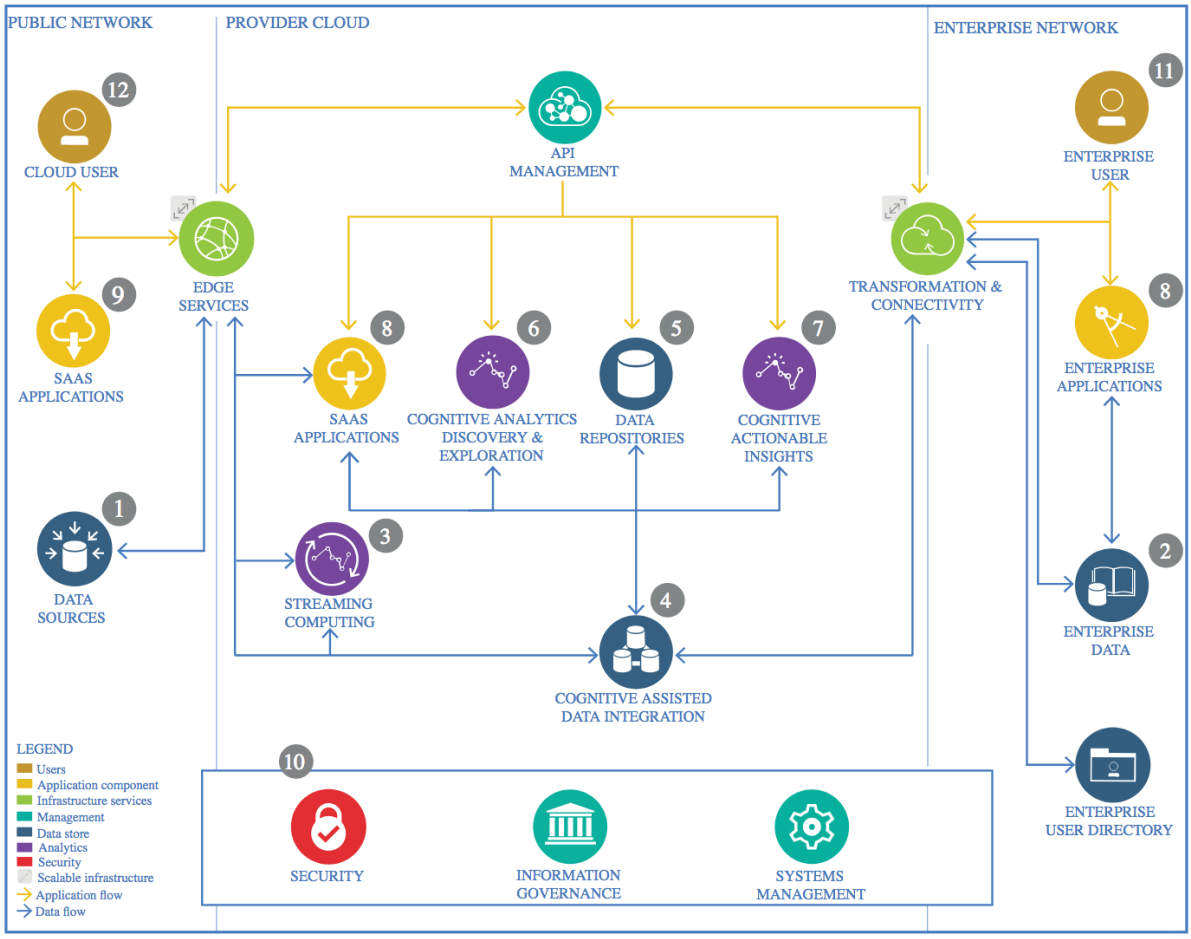
The Lightweight IBM Cloud Garage Method for Data Science

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

For training, the data source is a CSV file containing the information about the Portugal Bank clients. It's open and available on Kaggle.

1.1.2 Justification

The CSV is a format that can be easily open and transformed using Python, Pandas, PySpark, and many other libraries.

1.2 Enterprise Data

1.2.1 Technology Choice

The files will be stored at IBM Storage.

1.2.2 Justification

It can be easily accessed using IBM Watson's notebooks. As the file is very small, the space available for free is more than enough.

1.3 Streaming analytics

1.3.1 Technology Choice

Batch processing.

1.3.2 Justification

The client's information will be passed in a JSON format. This information will then be processed in batches.

1.4 Data Integration

1.4.1 Technology Choice

The data is cleansed and transformed to be ready to predictions using Pandas.

1.4.2 Justification

We have columns with high correlation (can be removed) and non-numerical features that need treatment.

1.5 Data Repository

1.5.1 Technology Choice

IBM Watson Storage.

1.5.2 Justification

High capacity, free and easy to integrate with IBM Watson Notebooks.

1.6 Discovery and Exploration

1.6.1 Technology Choice
Pandas and seaborn.

1.6.2 Justification
Pandas is an industry standard to work with not distributed data frames. It is efficient, clean, and easy to work. Seaborn is great to draw beautiful and meaningful visualizations. They both have a huge community and are widely adopted.

1.7 Actionable Insights

1.7.1 Technology Choice
Feature Selection for re-engineering and AdaBoosting, with 5 k-folds as a model.

1.7.2 Justification
We removed the following columns: 'education', 'emp.var.rate', 'cons.price.idx', 'day_of_week', 'month', 'marital', 'contact', 'campaign', and 'job'. During the training phase, we tested many algorithms, including ExtraTrees, GradientBoosting, Adaboosting, and MLP with different number of layers and neurons. Adaboosting was the best, and removing the columns also increased its performance from 42% to 52%.

1.8 Applications / Data Products

1.8.1 Technology Choice
Watson ML

1.8.2 Justification
We can deploy our model and predict using restful API. Using this model we don’t need to worry about availability, security, and scalability. All the details are covered by IBM.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice
All the security will be managed by the cloud.

1.9.2 Justification
The cloud has all the security protocols necessary. There is no need to implement anything.