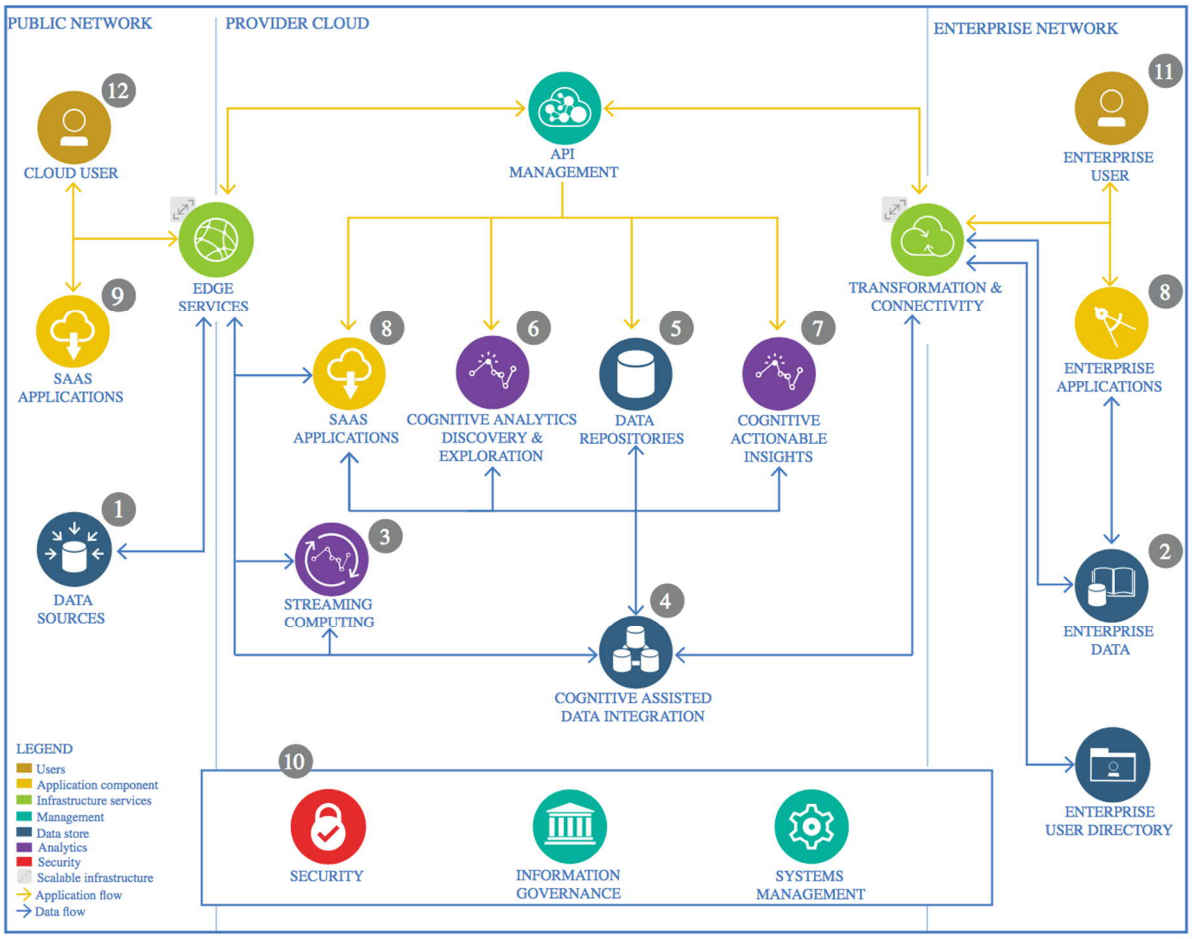
The Lightweight IBM Cloud Garage Method for Data Science

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 source is a CSV file and is uploaded to the IBM Watson Studio.

1.1.2 Justification

Since the dataset is small (<10k datapoints) there is no need for more complex architectures.

1.2 Enterprise Data

1.2.1 Technology Choice

The data is uploaded to the IBM Watson Studio.

1.2.2 Justification

Since the dataset is small (<10k datapoints) there is no need for more complex architectures. In case of a more sophisticated application (e.g. several embedded devices uploading data on a regular basis) a more robust approach should be considered, evaluating availability, redundancy, data security and integrity checks.

1.3 Streaming analytics

1.3.1 Technology Choice

In the scope of the capstone there is streaming capability.

1.3.2 Justification

Since the dataset is static there is no need for streaming analytics. In a real world application streaming would need to be taken into account.

1.4 Data Integration

1.4.1 Technology Choice

The data integration platform of choice is jupyter notebook.

1.4.2 Justification

Jupyter notebook is an adequate tool for data cleaning and exploration. However, for a real world application the choice should be revised.

1.5 Data Repository

1.5.1 Technology Choice

Since the dataset is static it was simply uploaded to the IBM Watson Studio.

1.5.2 Justification
Given that the data is static and easily available at the UCI repository, simply uploading to the IBM Watson Studio is appropriate. For a real world application a more robust choice of data repository is needed.

1.6 Discovery and Exploration

1.6.1 Technology Choice
Jupyter notebook is an adequate tool for data cleaning and exploration.

1.6.2 Justification
Jupyter notebook is an simple enough and can be used to explore data and comment design choices.

1.7 Actionable Insights

1.7.1 Technology Choice
Main insights were the need for data validations (% of missing values) and normalization.

1.7.2 Justification
The use case considered is rather simple and the database provided is of reasonable good quality.

1.8 Applications / Data Products

1.8.1 Technology Choice
LSTM can perform a good reconstruction of the desired signal. However in a real world application issues of model recalibration (e.g. sensor loss, aging) of fault detection should be addressed.

1.8.2 Justification
LSTM are a simple but effective approach. GRU should be investigated as well.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice
This was not approached in the capstone project.

1.9.2 Justification
For a capstone project this would be out of scope. In a real world application issues of model recalibration (e.g. sensor loss, aging) of fault detection should be addressed. Furthermore issues of system security and availability should be considered.