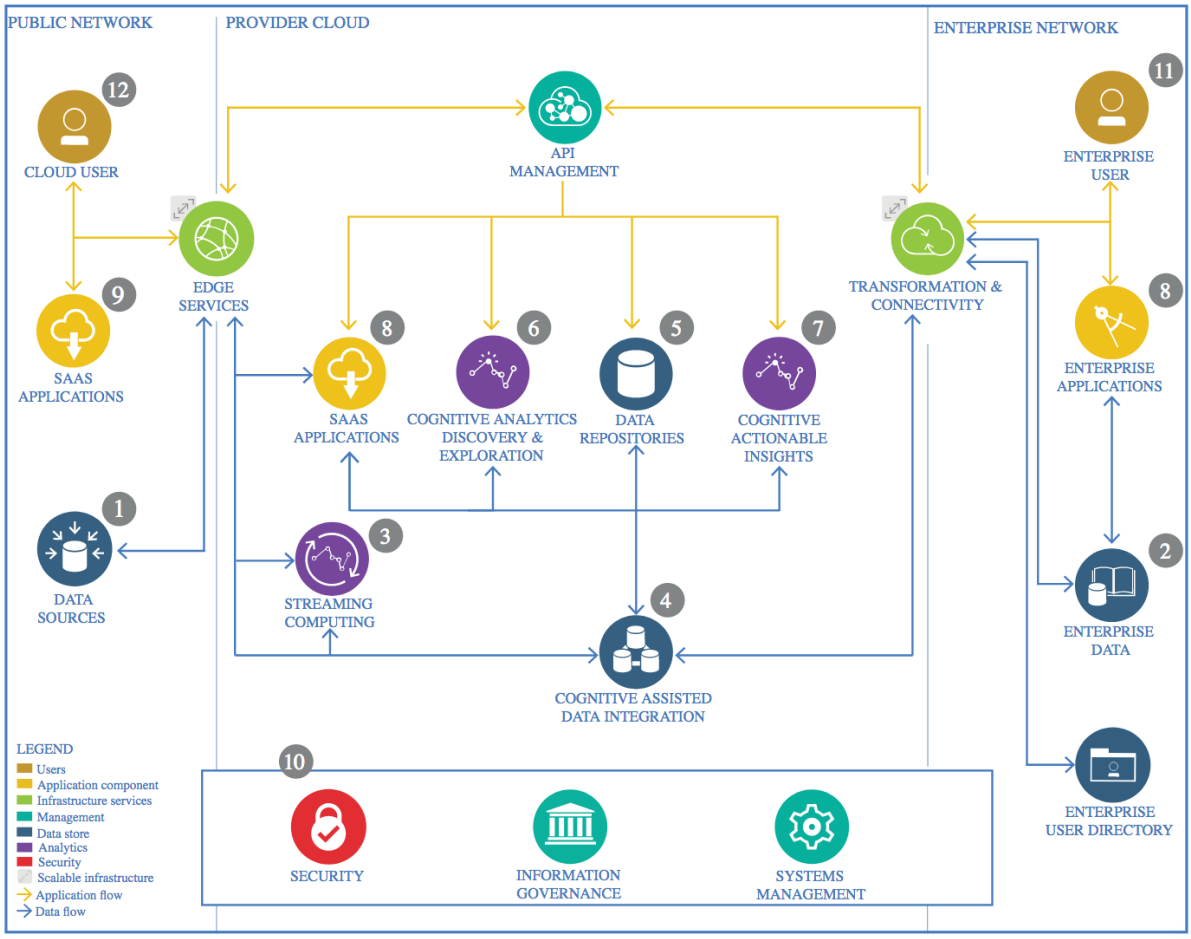
The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document Template

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

- Why have I chosen a specific method for data quality assessment?
- Why have I chosen a specific method for feature engineering?
- Why have I chosen a specific algorithm?
- Why have I chosen a specific framework?
- Why have I chosen a specific model performance indicator?

1.1 Data Source

1.1.1 Technology Choice
The Data was taken from a CSV file.

1.1.2 Justification
CSV files are one of the most common way to store the data.

1.2 Enterprise Data

1.2.1 Technology Choice
Not applicable

1.2.2 Justification
Not applicable

1.3 Streaming analytics

1.3.1 Technology Choice
Not applicable

1.3.2 Justification
Not applicable

1.4 Data Integration

1.4.1 Technology Choice
The Data is processed with Anaconda Jupyter Notebook.

1.4.2 Justification
Anaconda is one of the best tools to work with data science projects. It contain the majority of necessary libraries that we need it during our work.

1.5 Data Repository

1.5.1 Technology Choice
IBM Watson Cloud object storage

1.5.2 Justification
It provides a secured way to put on your files

1.6 Discovery and Exploration

1.6.1 Technology Choice
We used Label encoding library for the Feature engineering. We made our deep learning model with the help of TensorFlow and keras .

1.6.2 Justification

Keras is an open-source neural-network library written in Python. It is capable of running on top of TensorFlow. It's often in the majority of Deep Learning projects

1.7 Actionable Insights

1.7.1 Technology Choice

I did used Random forest, xgboost, light gbm, catboost, stacking between the tree-based models and keras.

1.7.2 Justification

Those models are the best models for binary classification problems based on researches, especially tree-based model (xgboost, light gbm, catboost) which are the best machine learning models for now. Also stacking is a great technique which combine multiple different models so that the resulting one will have a better score.

1.8 Applications / Data Products

1.8.1 Technology Choice

Not applicable

1.8.2 Justification

Not applicable

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

Not applicable

1.9.2 Justification

Not applicable