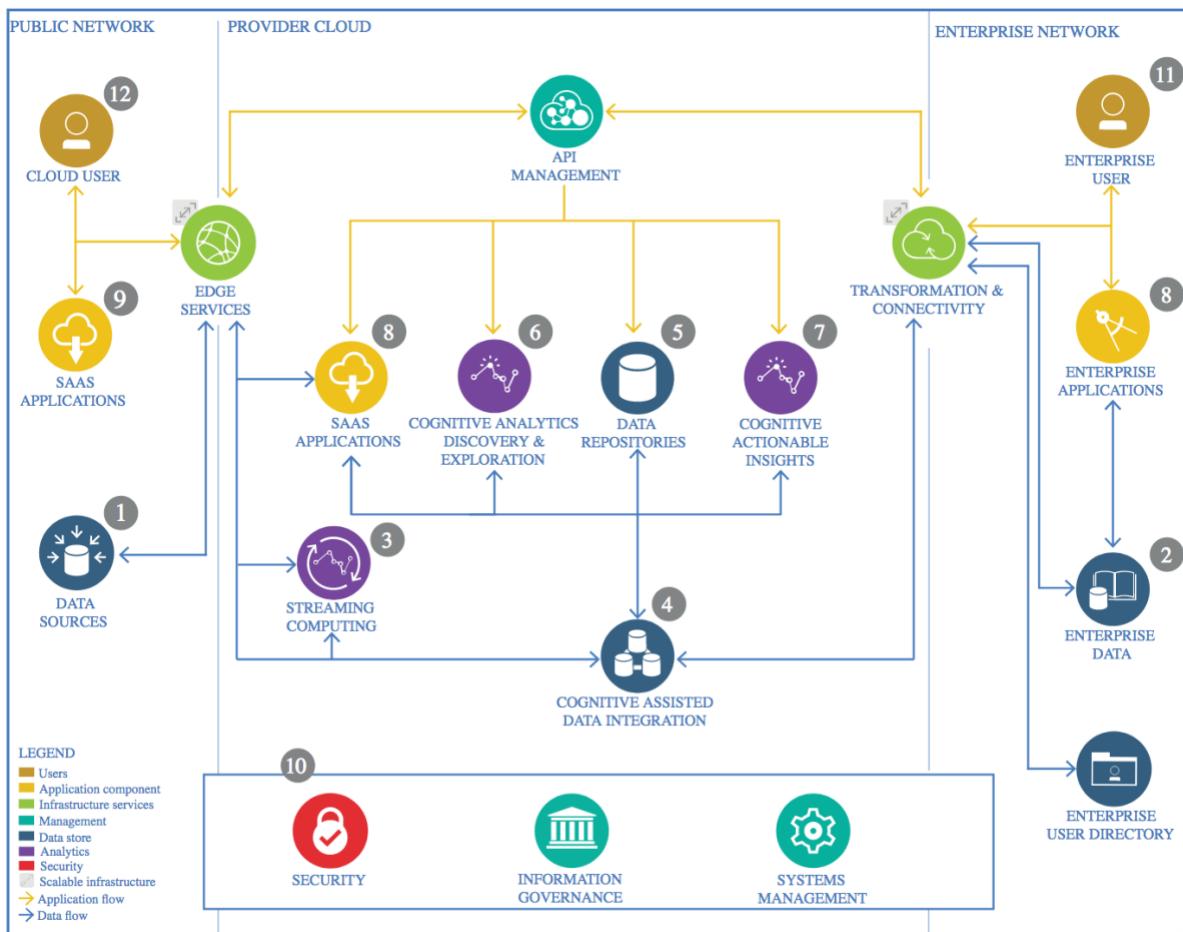


The Lightweight IBM Cloud Garage Method for Data Science

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

Dataset was identified and downloaded from Kaggle (<https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>).

1.1.2 Justification

Kaggle chosen as the data source hosting platform because of good accessibility and dataset overview. In addition, the datasets provided are well documented and verified.

The specific dataset was chosen because of its interesting ML applications, specifically to train a classifier.

1.2 Enterprise Data

1.2.1 Technology Choice

Github and IBM Watson Studio

1.2.2 Justification

Github selected for version control and cloud accessibility. IBM Watson Studio selected because of its free offering and feature-packed development environment with attached ObjectStorage.

1.3 Streaming analytics

1.3.1 Technology Choice

N/A

1.3.2 Justification

N/A

1.4 Data Integration

1.4.1 Technology Choice

N/A

1.4.2 Justification

N/A

1.5 Data Repository

1.5.1 Technology Choice

See 1.2

1.5.2 Justification

See 1.2

1.6 Discovery and Exploration

1.6.1 Technology Choice

For exploration, we used the NumPy, SciPy, Pandas and Seaborn python libraries in a Jupyter notebook.

1.6.2 Justification

These were chosen because of their ease-of-use and handy exploration tools. In addition, the size of the dataset allowed for non-parallelized libraries as 4 GB of memory was sufficient.

1.7 Actionable Insights

1.7.1 Technology Choice

Data is all numerical and nicely formatted. No missing or inconclusive values. Some features had outliers which were removed. After one iteration with the model, we went back and renormalized the features (z-score normalized) to aid in the accuracy, and we also revised the outlier removal.

For algorithms, we chose logistic regression and SVM algorithms. For model performance indicators, we used accuracy as our main metric, but also considered confusion matrices.

1.7.2 Justification

The classification algorithms we use are known to perform much better with normalized data. Z-score was chosen because of its simple but effective implementation. We chose the algorithms because of their applicability and relative ease-of-understanding/transparency.

1.8 Applications / Data Products

1.8.1 Technology Choice

As final deliverables, the model(s) were deployed as documented runnable Jupyter notebooks containing data exploration, EFL and modelling stages.

1.8.2 Justification

As the model itself is replicable and scalable independent of the data, a Jupyter notebook was determined to be sufficient material.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

N/A

1.9.2 Justification

N/A