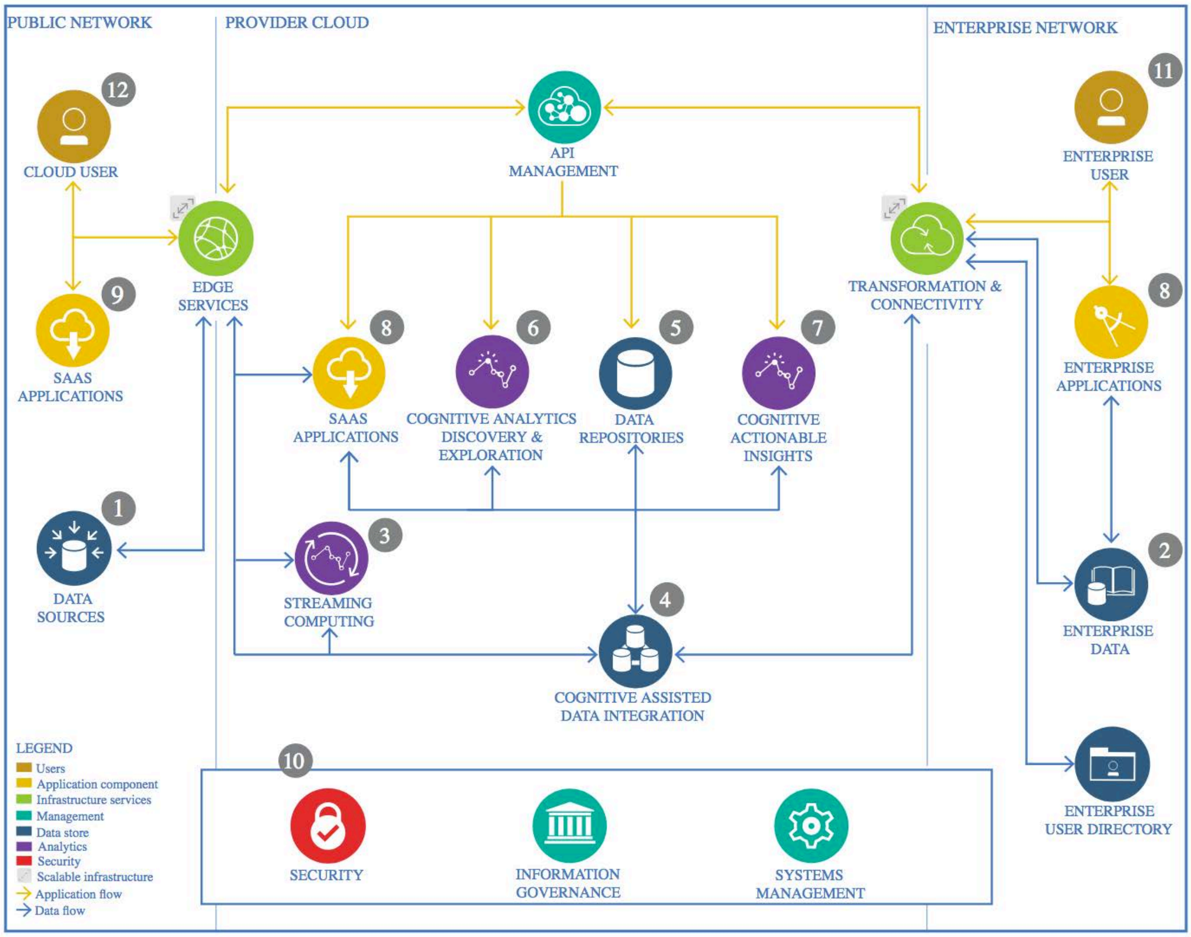
The Lightweight IBM Cloud Garage Method for Data Science

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 source of the data is from kaggle website in the following link.
https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009
The data is downloaded directly into the IBM Watson cloud.

1.1.2 Justification

As the file size is less, I have decided to directly download the file into IBM Watson cloud.

1.2 Enterprise Data

1.2.1 Technology Choice

The data is directly called into the IBM Watson notebook.

1.2.2 Justification

It is easier to deal with the data as the file size of the data is not as much as it would have a computational problem while dealing with the data.

1.3 Streaming analytics

1.3.1 Technology Choice

There is no streaming of real time data. So there isn't much use of it in this case.

1.3.2 Justification

As the data is a collection of different readings, there isn't any live data which has to be fed into the model in real time.

1.4 Data Integration

1.4.1 Technology Choice

IBM data stage and IBM Watson on cloud is used.

1.4.2 Justification

As the dataset is downloaded into the IBM cloud, it is cleaned, analyzed and models are prepared.

1.5 Data Repository

1.5.1 Technology Choice

IBM DB2 is used as the data repository.

1.5.2 Justification

IBM DB2 provides a very easy interface to deal with the data and it can be easily called in Watson Notebook through SQL commands.

1.6 Discovery and Exploration

1.6.1 Technology Choice

Jupyter Notebook, Pandas, Matplotlib, Seaborn.

1.6.2 Justification

Jupyter, Python, scikit-learn, pandas, Matplotlib are all open source and supported in IBM Cloud.Pandas is used for statistical analysis and seaborn is mainly used as the visual library.

1.7 Actionable Insights

1.7.1 Technology Choice

Python, pandas and scikit-learn and Keras

1.7.2 Justification

Python is a simple programming language with many different libraries. Scikit learn is used to build Classifiers such as Random Forest and XGBoost models. Keras is used to build a dense neural network. Keras provides an abstraction layer on top of TensorFlow. Hence this frameworks are seamlessly supported in IBM Cloud through Watson Studio and Watson Machine Learning.

1.8 Applications / Data Products

1.8.1 Technology Choice

D3 is selected as the technological choice.

1.8.2 Justification

D3 is one of the most prominent and most widely used visualization widget frameworks with a large open source ecosystem contributing a lot of widgets for every desirable use case.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

IBM Cloud Internet services is the selected as Security & Information Governance

1.9.2 Justification

IBM Cloud Internet Services provides global points of presence (PoPs). It includes domain name service (DNS), global load balancer (GLB), distributed denial of service (DDoS) protection, web application firewall (WAF), transport layer security (TLS), and caching.