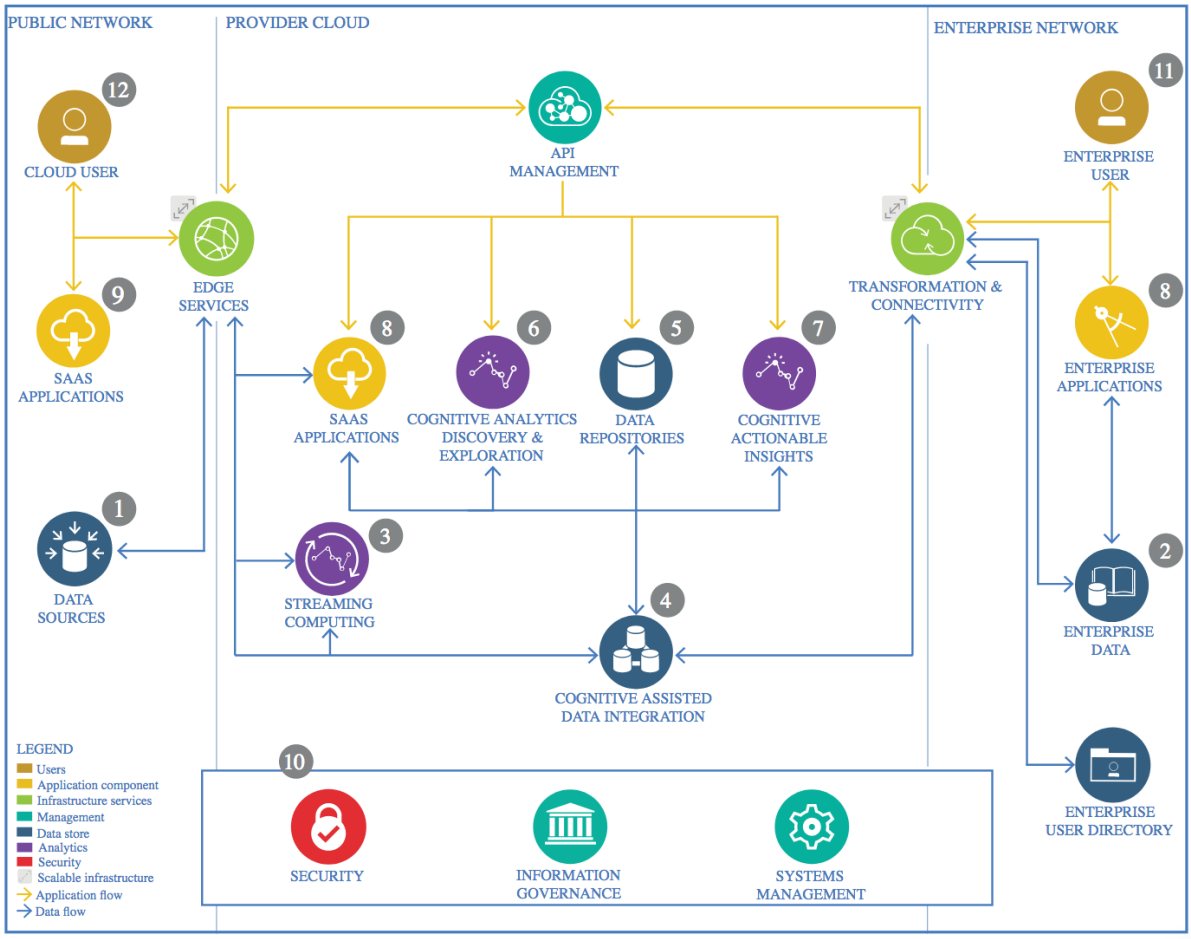
The Lightweight IBM Cloud Garage Method for Data Science

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

Data is sourced from the UCI Machine Learning Repository as a csv file:
http://archive.ics.uci.edu/ml/machine-learning-databases/00222/bank-additional.zip

1.1.2 Justification

Publicly available dataset. Documentation included

1.2 Enterprise Data

1.2.1 Technology Choice

N/A

1.2.2 Justification

N/A

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

Dataset is near-ready for use

1.5 Data Repository

1.5.1 Technology Choice

Local Drive Storage

1.5.2 Justification

Low cost, ease of access and ease of use

1.6 Discovery and Exploration

1.6.1 Technology Choice

Jupyter Notebook, Apache Spark, Pyspark, SparkML

1.6.2 Justification

Open source. Active developer community. Parallel Computing

Apache Spark Engine

Apache Spark is a parallel cluster computing engine that is widely adopted for scaling data science projects. It is famed for its speed, support for multiple languages (Scala, Java, R, and Python), and advanced analytics offering one of which is the dedicated machine learning library, SparkML. The SparkML module is engaged to train the predictive model using the Pyspark API, which is Python's syntax interface with Apache Spark computing engine.

1.7 Actionable Insights

Feature Selection
Model Selection
Model Evaluation

1.7.1 Technology Choice

Jupyter Notebook, Apache Spark, Pyspark, Pandas.

1.7.2 Justification

Open source. Active developer community.

Feature Selection

Subsequent to exploratory analysis, the 'default' feature was found to have low information content and was deemed redundant. All rows having 'unknown' entries were dropped.

Model Selection

The Random Forest and Gradient-Boosted Tree models gave comparable performance. However, the Random Forest model is preferred for its lower computational cost and amenability to parallel computing.

Model Evaluation

The Random Forest and Gradient-Boosted Tree models have comparable performance. However, the Random Forest model is preferred for its lower computational cost and amenability to parallel computing.

1.8 Applications / Data Products

1.8.1 Technology Choice

Interactive Jupyter Notebook

1.8.2 Justification

Client requires the interactive Jupyter Notebook in order to have a feel of the models and evaluation metrics adopted

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

N/A

1.9.2 Justification

N/A