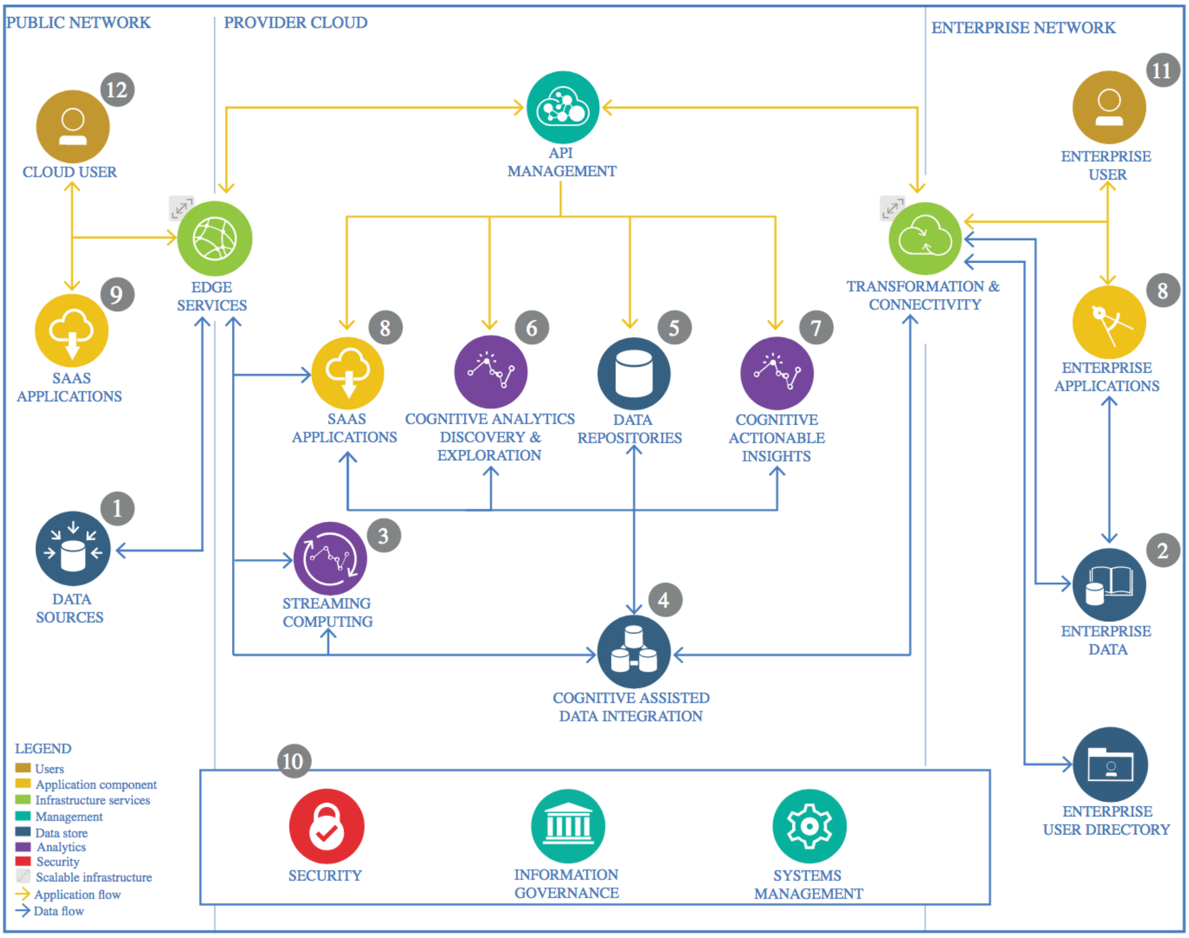
# The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document

Credit Card Default Prediction

## 1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

### 1.1 Data Source

### 1.1.1 Technology Choice

The dataset is available at https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

### 1.1.2 Justification

Ease of use and readymade available

## 1.2 Enterprise Data

### 1.2.1 Technology Choice

GitHub Repository

### 1.2.2 Justification

The data will be available up to date in the repository

## 1.3 Streaming analytics

### 1.3.1 Technology Choice

Not applicable for this project

### 1.3.2 Justification

Not applicable for this project

## 1.4 Data Integration

### 1.4.1 Technology Choice

Not used for this project

### 1.4.2 Justification

Not used for this project

## 1.5 Data Repository

### 1.5.1 Technology Choice

Local hard disk and Jupyter notebook used for storage the data.

### 1.5.2 Justification

Used local disk since the data set is not much big.

## 1.6 Discovery and Exploration

### 1.6.1 Technology Choice

The following technology tools are used for the data exploration and visualization:

- Python Pandas
- Seaborn
- Matplotlib

### 1.6.2 Justification

Size of the current data set is not that much big, so the above tools are sufficient to run the data exploration and visuals in a reasonable configured system

## 1.7 Actionable Insights

### 1.7.1 Technology Choice

Methods used for the data quality assessment:

a. Descriptive and Exploratory Data Analysis

Methods used for feature engineering:

a. Removed unnecessary features
b. Rename attributes
c. Extract new features

Algorithms used in the project:

a. Random Forest
b. Logistic Regression
c. Artificial Neural network (Deep Learning)

Framework used in the project:

a. Scikit-Learn
b. Keras and Tensorflow

Model performance indicators used in the project:

a. Confusion Matrix
b. Accuracy
c. Precision
d. Recall
e. ROC-AUC curve

### 1.7.2 Justification

Data quality assessment:

a. Descriptive and Exploratory Data Analysis is easy to understand and interpretation

Feature Engineering:

a. Model performs better if data get preprocessed and feature engineered

Algorithms:

a. Applied supervised classification algorithms (Logistic Regression, Random Forest & Deep Learning) since the data has label/target information
b. To better understand the correlated features, linear and tree-based algorithms selected as best algorithms

Framework used in the project:

a. Scikit-Learn is open source and most machine learning algorithms already prebuilt. So we can use it as plug and play

b. Open source, easiest and fastest implementation is possible with Keras (backend tensorflow) for deep learning

Model performance indicators used in the project:

a. The solution what we used to detect credit card defaulter is based on classification with supervisor learning. So, the model performance metrics are confusion matrix with accuracy, precision, recall and AUC curve values.

## 1.8 Applications / Data Products

### 1.8.1 Technology Choice

Generate a report based on Jupyter notebook

### 1.8.2 Justification

Jupyter notebook can contain code, text, and visuals. Others can easily reproduce report using notebook.

## 1.9 Security, Information Governance and Systems Management

### 1.9.1 Technology Choice

Not applicable for this capstone project

### 1.9.2 Justification

Not applicable for this capstone project