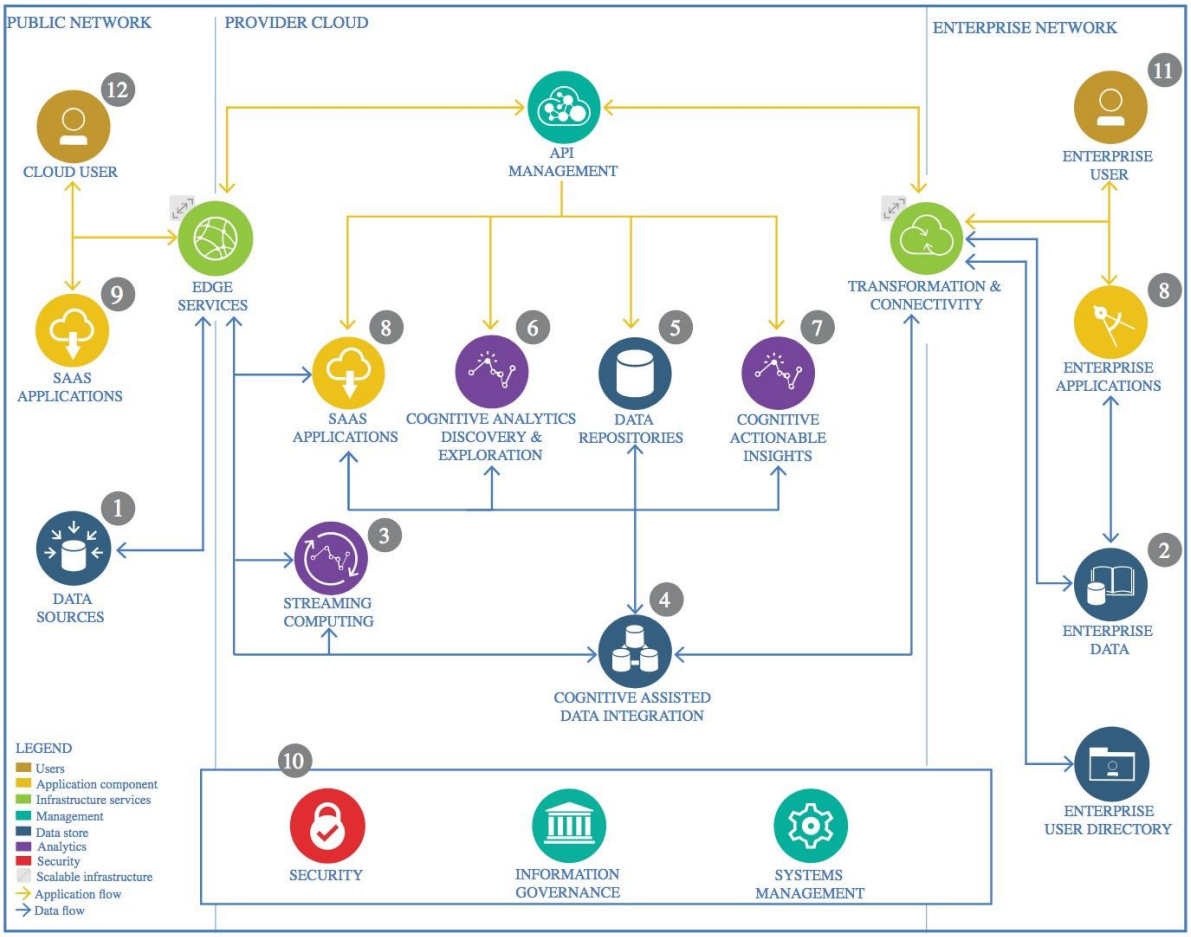
The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document Template
Breast Cancer Classification

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 downloaded from https://www.kaggle.com/yasserh/breast-cancer-dataset

1.1.2 Justification
The dataset is public. Easy to download and use.

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
The dataset contains all necessary information.

1.5 Data Repository

1.5.1 Technology Choice
The dataset is uploaded to IBM Cloud.

1.5.2 Justification
For model development is used Jupyter Notebook on IBM Watson Studio which have access to IBM Cloud to read data.

1.6 Discovery and Exploration

1.6.1 Technology Choice
Jupyter Notebook on IBM Watsin Studio Python3.7
libraries:

- Pandas
- Numpy
- Matplotlib
- Seaborn

1.6.2 Justification

IBM Watson Studio is a good resource for developing. Easy to share and collaborate. Python and its libraries are open source with widespread developer community. Really good supported and actively updated tools.
Apache Spark is not used because the dataset is too small. Single computer is enough to process the data.

1.7 Actionable Insights

1.7.1 Technology Choice

Feature Engineering: pandas, numpy, scklearn, imblearn
Model Selection: scklearn, keras, tensorflow
Model evaluation: scklearn

1.7.2 Justification

Open source. Really good supported and actively updated tools. Pandas, numpy, scklearn, keras and imblearn have all necessary methods:

Feature Engineering

- Pandas to easily find null values and duplicates in the dataset
- sklearn MinMaxScaller() to normalize
- pandas corr() and cor_matrix to define high correlated features
- imblearn SMOTE to create a balanced dataset

Model Selection

The task is binary classification (Breast Cancer Classification). A target is 1 as 'Malignant' and 0 as 'Benign'. 4 models were chosen:
Logistic Regression (scklearn) - the best and commonly used model for binary classification. The output of Logistic Regression is a binary value (0 or 1).
Decision Tree (scklearn) - Each node is test of an attribute/feature, each branch presents full test. At the end (final leaf) we have class (0 or 1). So, this algorithm can be easily interpreted as set of rules.
Gradient Boosting (scklearn) - ensembled algorithm which uses decision tree models. Each model learns from the output of previous model - corrects the prediction errors made by prior models. This method usually has a good accuracy but can be easily overfitted. *Neural Network* (keras) - the output layer with sigmoid activation function returns estimated probability which can be easily converted to 0/1 output using threshold.

Model Evaluation

For evaluation of Binary Classification was used Accuracy, Precision and F1 score.
Accuracy does not always correctly demonstrate if the model is good or not for classification task (especially for unbalanced dataset) but it's always great to have good Accuracy. Recall and Precision could work better but we should decide on what question we want to ask. What is better for us to have some False 'Malignant ' or False 'Benign'. Obviously to have False 'Malignant' is better because it's better to check than to miss. So, Precision works for this goal. From other side we do not want to miss real Malignant cases, so Recall metric is also important for us. We should decrease False Benign.

F1 is harmonic mean of Precision and Recall and can help us to choose what classifier is better in common.

1.8 Applications / Data Products

1.8.1 Technology Choice

Model deployment: PDF Report, Jupyter Notebook
Available on https://github.com/omatveyuk/ibm_advanced_ds

1.8.2 Justification

Easy to read and follow a model process from data exploration to the model evaluation. Jupyter notebook can be easily updated and used for a client dataset if the dataset contains the same features.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

N/A

1.9.2 Justification

N/A