The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document Template

# Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

## Data Source

### Technology Choice

Understanding data is one of the most important part when designing any machine learning algorithm.The data was downloaded from Kaggle (<https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>). CSV (coma separated values format).

123 KB of data.

### Justification

The reason to download from Kaggle was availability and ease of use. The CSV file provided is a common format for table data, separator by ‘,’.

## Enterprise Data

### Technology Choice

GitHub repository

### Justification

To available for every person every time on the repository

## Streaming analytics

### Technology Choice

NA

### Justification

NA

## Data Integration

### Technology Choice

Not used.

### Justification

Not used.

## Data Repository

### Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

### Justification

Please justify your technology choices here.

## Discovery and Exploration

### Technology Choice

Jupyter Notebooks the following Python 3.6 libraries were used for Data Exploration and Visualization: - Pandas, Matplotlib and Seaborn.

### Justification

Because I feel familiar with it and easy to use specifily with jupyter notebook you can know the parameter and read the documentation of it. The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, data visualization, machine learning, statistical modeling, and much more.

## Actionable Insights

### Technology Choice

The following Python 3.6 libraries: -

Pandas , numpy, sklearn and Tensoflow.

In Classifications, we will use following 2 Techniques to train our model and predict:

1. Random Forest

2. Support Vector Machine

### Justification

We use sklearn library because is most common libraries that introduce the predicted model, We choose Random Forest because power to handle a large data set with higher dimensionality. for SVM because ususlly use for two classes.

We gone use F1 performance indicator because better measure of the incorrectly classified cases

## Applications / Data Products

### Technology Choice

A Jupyter notebook based report was generated.

### Justification

As only the correlating factors needed to be identified Jupyter notebook based report was consider sufficient.

## Security, Information Governance and Systems Management

### Technology Choice

NA.

### Justification

NA.