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

Open source CSV file from Kaggle.com.

### Justification

CSV file was the only form which was available in this data set, and it is easy to import with Pandas library.

## Enterprise Data

### Technology Choice

Not used in this project.

### Justification

No need to transfer data to the cloud.

## Streaming analytics

### Technology Choice

Not used in this project.

### Justification

This project does not need real-time analytics capabilities because the data set used does not lose value that quickly.

## Data Integration

### Technology Choice

Data was explored and cleaned by Pandas library. Sklearn library was used to transform the data into more suitable form for machine learning models.

### Justification

Pandas and Sklearn are open-source libraries and they provide versatile tools to manipulate data.

## Data Repository

### Technology Choice

The data is saved in CSV file which is stored in Github repository.

### Justification

The data set does not require updates or constant access so other than storage services are not needed.

## Discovery and Exploration

### Technology Choice

Matplotlib and Seaborn libraries are used for data visualization at data exploration stage and when evaluating machine learning models with different metrics.

### Justification

Matplotlib and Seaborn provide large range of different plot types which are useful for examining how the data is distributed and how classification algorithm performs.

## Actionable Insights

### Technology Choice

Sklearn and Keras have been used to define, train and evaluate machine learning and deep learning models.

### Justification

Sklearn and Keras are both free to use and with Sklearn it is possible to create different kind of machine learning models and with Keras the data set can be examined with a non-linear deep learning model.

## Applications / Data Products

### Technology Choice

A Jupyter notebook about the data processing and modeling will be done and the results are presented in pdf format.

### Justification

A Jupyter notebook provides more detailed information what kind of actions have been taken when knowledge is extracted from the data set. A pdf file provides more understandable and clearer information for the ordinary business user in a compact form.

## Security, Information Governance and Systems Management

### Technology Choice

Not used in this project.

### Justification

The data set used in this project is open source so there is no need to control who has access to the data.