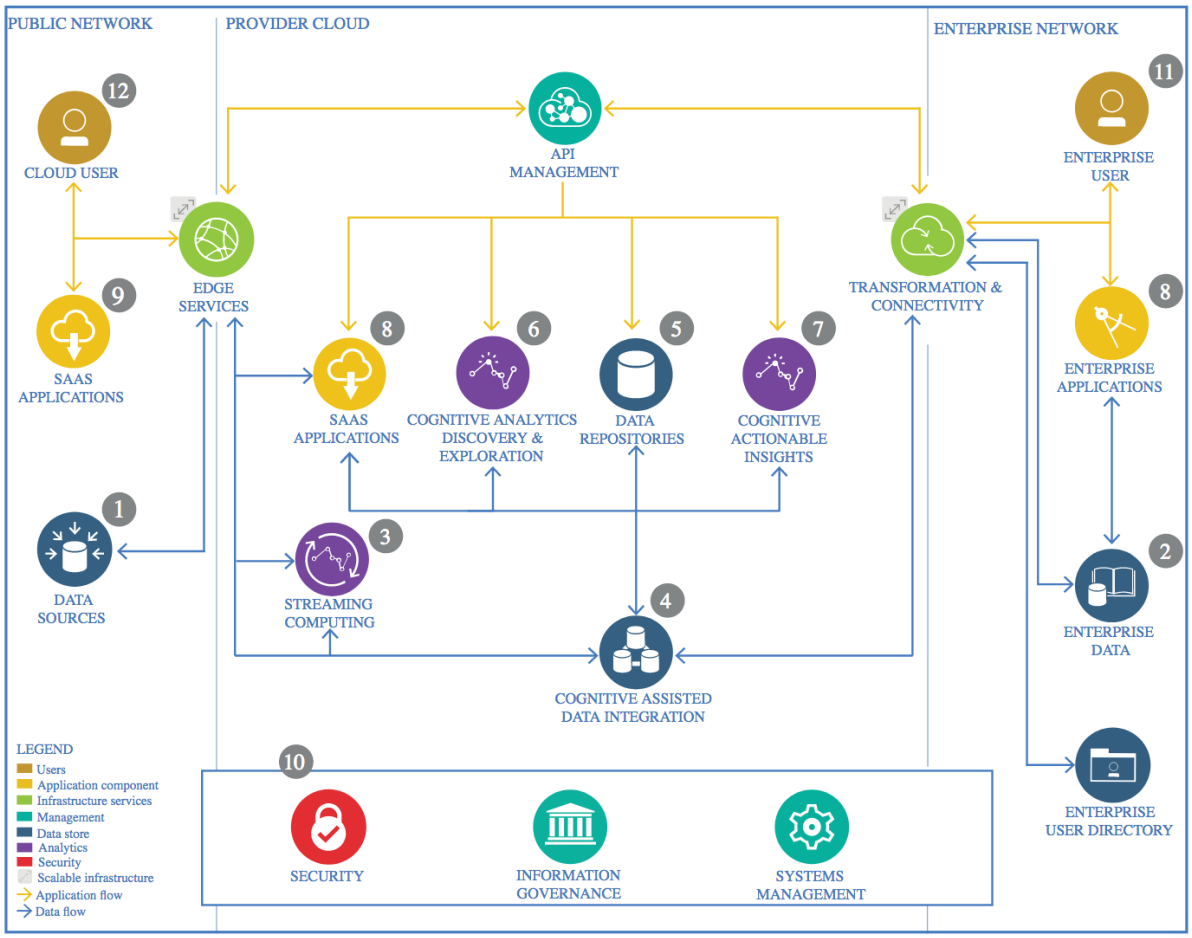
The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document Template

I am going to predict Wine Quality by using Watson Machine Learning.

I selected this example from IBM Call for code challenge: https://developer.ibm.com/patterns/create-and-deploy-a-scoring-model-to-predict-heartrate-failure/. This example will follow me on how to create correct repository and how to deploy application to IBM Watson studio. This example has all the steps described on how I am going to deploy my own application I have created specifically for Capstone project

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

1.1.1 Technology Choice

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

1.1.2 Justification

Red Wine Quality dataset was selected as datasource for this task.
https://www.kaggle.com/maitree/wine-quality-selection

This datasource was taken from Kaggle and I would like to say that it is well organized and all fields are without null and na values so minimum work is needed on ETL step. This dataset is for beginners so it is easy to use and it fits perfect for the start.

1.2 Enterprise Data

1.2.1 Technology Choice

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

1.2.2 Justification

For working with this task I selected several approaches so for some basic classification algorithms I divided data on 2 parts: Training and testing so model was evaluated.
For deep learning I used K-fold cross validation approach and I did not divide my dataset to 2 parts

Enterprise data come from the REST API of the deployed model on IBM Watson studio.

1.3 Streaming analytics

1.3.1 Technology Choice

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

1.3.2 Justification

No Streaming analytics in this application is presented

1.4 Data Integration

1.4.1 Technology Choice

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

1.4.2 Justification

For the Capstone project I decided to use next technologies:
First of all it will be IBM Watson studio. There I will create Notebook with support of Python3 and Spark 2.1 together with Python2 and Spark 2.1 for SytemML.

I decided to use Python since I know this language. Also I will use next several Python libraries:

1. NumPy
2. Matplotlib
3. Seaborn
4. Scicit learn
5. Pandas
6. Keras
7. SystemMl
8. Apache Spark
9. SystemML

All those technologies makes available to use different approaches for estimating of quality of wine.

1.5 Data Repository

1.5.1 Technology Choice

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

1.5.2 Justification

As a data repository I selected CSV comma separated file. I choose 2 datasets. First dataset is for red wine, I try to predict quality of wine by using threshold of good/bad wines. In other example I use also CSV comma separated file with joined dataset of white and red wines together and there by using Deep Learning approach I try to classify exactly quality of wine.

1.6 Discovery and Exploration

1.6.1 Technology Choice

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

1.6.2 Justification

I created several diagrams by matplotlib and seaborn in the ETL step. Also for the further data exploration I have used Apache Spark SQL and pandas framework to see basic dataset analytics and shema. Besides this I have implemented StandardScaler from SciKit learn framework in order to optimize dataset

1.7 Actionable Insights

1.7.1 Technology Choice

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

1.7.2 Justification

This current dataset is for beginners so not a lot of work is needed here in order to explore it and create features. All data variables has float type and basic classifier is needed in order to classify quality of the wine. I set up threshold for the good wine I select all quality of wine more than 7 or equal to 7 and all wines that has quality less then 7 are considered bad wines. So it is basic binary classifier like Logistic Regression.

I used F1 score as a quality metric in order to evaluate model performance. I tried PCA in order to do feature engineering.

Besides this I implemented Depp Learning approach using Keras framework. In this approach I classified quality of wine not by using threshold but by determine exactly quality of wine. Also I have used joined dataset of red and white wines.
I used R2 regression metrics for evaluation of my model as well as MSE and MAE metrics.

First method I will describe here:

SystemML (LogisticRegression):
Here I used SystemML framework together with Pandas dataframe. As a preprocessing I dropped 3 columns: quality, density and pH since those columns I think does not give many values to the dataset. I decided to do this based on the ETL step. And quality is a label so I dropped it also. Then I added new column label with threshold 7, it means that all wines with quality more than 7 or equal to 7 have 1 and all other wines have 0 in this column.

I used LogisticRegression classification algorithm in this framework. Winequality-red dataset is used in this notebook

Results:
LogisticRegression score: 0.861742

PySpark (LinearRegression):
I tried to use Linear Regression in order to predict Red Wine quality using Linear Regression library frin PySpark framework and model is not very good. I think because Linear Regression is not suitable for this kind of task. I used Apache Spark and pyspark regression algorithm. Normalizer did not help.

Results:
RMSE: 0.736556
r2: 0.176016

PySpark(Gradient-Boosted Trees (GBTs)):
I tried to use this algorithm and works pretty well. Also Apache Spark was used together with this classification algorithm. I also used Pipeline with Normalizer as part of data engineering process.

Results:
F1: 0.8901

Keras:
Here I used feed-forward network in order to try to predict quality of wine itself. I used Pandas and K-fold cross validation algorithm in order to get correct prediction values. Also I used Standard Scaler in order to standardize the data. Unfortunately because data is not enough I think the results are not good:

Results:
MSE: 0.4849
MAE: 0.5514
R2 score: 0.3619

1.8 Applications / Data Products

1.8.1 Technology Choice

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

1.8.2 Justification

In order to deploy this application I selected Watson Cloud Storage as a repository platform. IBM Watson dataplatform and IBM cloud together with Machine Learning service of IBM. It is possible to request REST API in order to use this model.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

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

1.9.2 Justification

No security