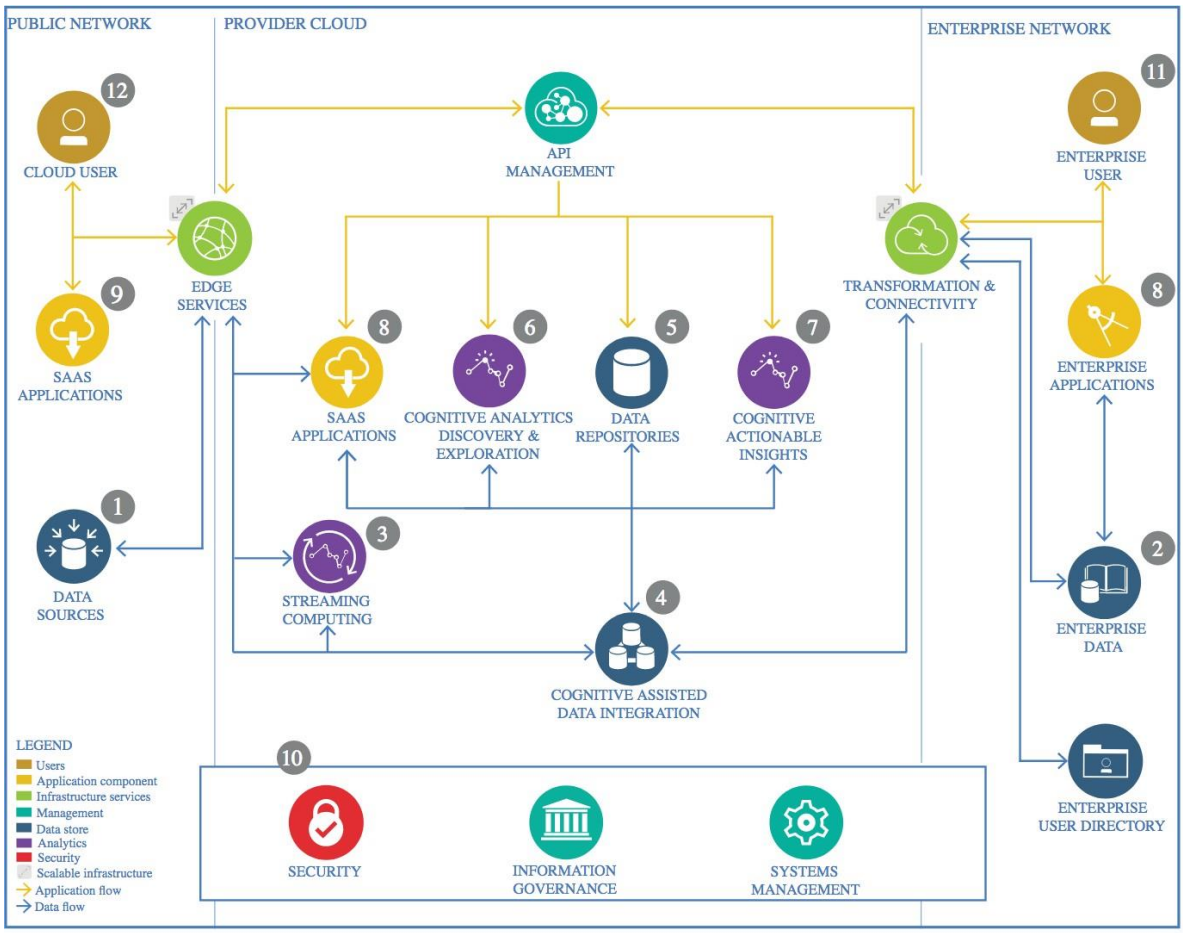
The Lightweight IBM Cloud Garage Method for Data Science

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

1.1.1 Technology Choice

The Data source of data is in the form of CSV file.

1.1.2 Justification

The source has provided the data in CSV form and can be easy handled in python.

1.2 Enterprise Data

1.2.1 Technology Choice

There is no need for enterprise data.

1.2.2 Justification

The data does not need to be share or update except from the source.

1.3 Streaming analytics

1.3.1 Technology Choice

There is no need for streaming analytics.

1.3.2 Justification

The data provided update ones a month and not need for data to be in real time

1.4 Data Integration

1.4.1 Technology Choice

For data ingestion, preparation and integration I only used the Jupyter Python Notebook on the cloud from IBM Watson.

1.4.2 Justification

The technology for using in data integration must support CSV file since our data is in the CSV file. Since the size of data is not large as of now, still we are using Apache Spark.

1.5 Data Repository

1.5.1 Technology Choice

The IBM Cloud has been used for storing CSV file of this project.

1.5.2 Justification

I used python spark data frame for temporary persistent data storage. The csv file I uploaded to the IBM Cloud data platform so that it is accessible from the Watson Studio. The trained models and history files are saved on the file system from my IBM cloud profile.

1.6 Data Quality Assessment

1.6.1 Choice

As the data is in text format. Before we talk about feature engineering, as always, we need to do some data pre-processing or wrangling to remove unnecessary characters, symbols and tokens.

1.6.2 Justification

The words from text will be used as variable to model to predict sentiment outcome. Also, we will remove rows which has “nan” values.

1.7 Feature Engineering

1.7.1 Choice

The text data is first pre-processed and then will be transform using TF-IDF method, the string column which has only “0” or “1” value will be transform to integer 0 and 1.

1.8 Algorithm

1.8.1 Choice

In this prediction the linear support vector machine method will be used for predicting sentiment. Which will be create using pyspark method

1.8.2 Justification

Because the problem is to predict the value of outcome base on text and to classify into sentiment type. The Linear Support Vector Machine method is been use in this model.

1.9 Discovery and Exploration

1.9.1 Technology Choice

Pyspark dataframe will be used for used as input in machine learning model. Pyspark and nltk library will be use for creation of model. Mathplotlib will be use for plotting graph of prediction value of model and ground truth in order to visualize the result and accuracy of the model.

1.9.2 Justification

These libraries are open source and support with IBM Cloud and can be call with python.

1.10 Actionable Insights

1.10.1 Technology Choice

The libraries like pyspark, nltk will be used for text pre-processing and creating model.

1.10.2 Justification

Pyspark has many the library of machine learning methods.

1.11 Performance indicator

1.11.1 Choice

The F1 score method will be use in order to determine the performance of model.

1.11.2 Justification

The F1 score method has been the method for determine the performance of Linear support vector machine model.

1.12 Applications / Data Products

1.12.1 Technology Choice

There is no data product in this project.

1.12.2 Justification

The model has created and prediction base on project. There is no need for data product.

1.13 Security, Information Governance and Systems Management

1.13.1 Technology Choice

There is no technology needed in this process.

1.13.2 Justification

The data that has been use in this project is public data.