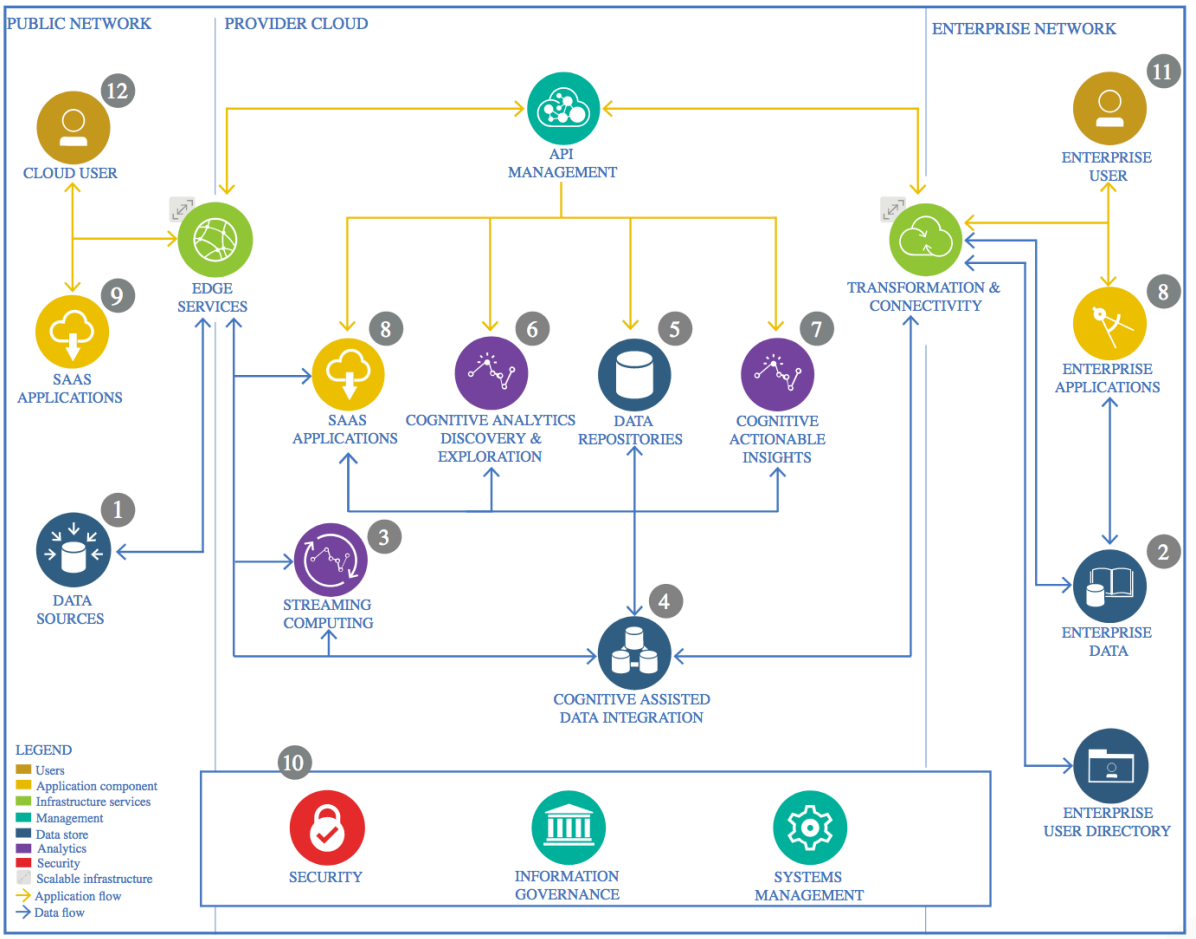
The Lightweight IBM Cloud Garage Method for Data Science

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

I have tried to look what I could use from the architecture above.
And added some topics related to the Capstone requirements.

1.1 Data Source

1.1.1 Technology Choice

The data is available as public dataset on https://www.kaggle.com/uciml/forest-cover-type-dataset as CSV data. See also https://archive.ics.uci.edu/ml/datasets/covertype

1.1.2 Justification

It is a one-time dataset used for machine learning practice which is available as CSV.

1.2 Enterprise Data

1.2.1 Technology Choice
This component is not needed.

1.2.2 Justification
The model to be developed is not meant to be used for an Enterprise.
It is only for training purposes.

1.3 Streaming analytics

1.3.1 Technology Choice
This component is not needed.

1.3.2 Justification
I use a one-time dataset as CSV.

1.4 Data Integration

1.4.1 Technology Choice
Not needed.

1.4.2 Justification
All data needed is available in one CSV dataset, so no need to integrate with other data.

1.5 Data Repository

1.5.1 Technology Choice
Not needed.

1.5.2 Justification
Data used is well described training data. No need to store this in a separate repository.

1.6 Discovery and Exploration

1.6.1 Technology Choice
Jupyter Notebooks (Python) is used with Pandas, seaborn,matplotlib and sklearn as imports.

1.6.2 Justification
We can just use methods etc within Python like distribution plot, violin plot etc based on the imports mentioned above. These methods are sufficient to do exploration of the dataset and get visual insights.

1.7 Actionable Insights

1.7.1 Technology Choice

Not needed.

1.7.2 Justification

Separate insights not needed. We only need to be able to predict the Covertype which is part of the model.

1.8 Applications / Data Products

1.8.1 Technology Choice

Not needed.

1.8.2 Justification

Data used to predict is standalone data.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

Not needed.

1.9.2 Justification

We use publicly available data which is used in many trainings for Machine learning.

1.10 Questions based on Coursera Assignment

1.10.1 Why have I chosen a specific method for data quality assessment?

The source dataset is a dataset which is actually of good quality as it is used for training purposes where not too much data exploration is need to do cleansing etc.
So for me it was more looking into the meaning and to some simple checks like if the binary data was one hot encoded or not.

1.10.2 Why have I chosen a specific method for feature engineering?

Feature engineering was based on domain knowledge and general knowledge, so not really a specific method was needed.

1.10.3 Why have I chosen a specific algorithm?

The randomforestclassifier I already used earlier as part of a beginner competition on Kaggle.
For the neural network I chose to do it as simple as possible. So no real ‘choice’ except that it had to be able to be used as multiple classifier.

1.10.4 Why have I chosen a specific framework?

I have chosen Keras as I am most use to it as part of several training.

1.10.5 Why have I chosen a specific model performance indicator?

I used a default from the sequence model for accuracy and loss which I could use to plot making the performance Visual.