Architectural Decisions Document
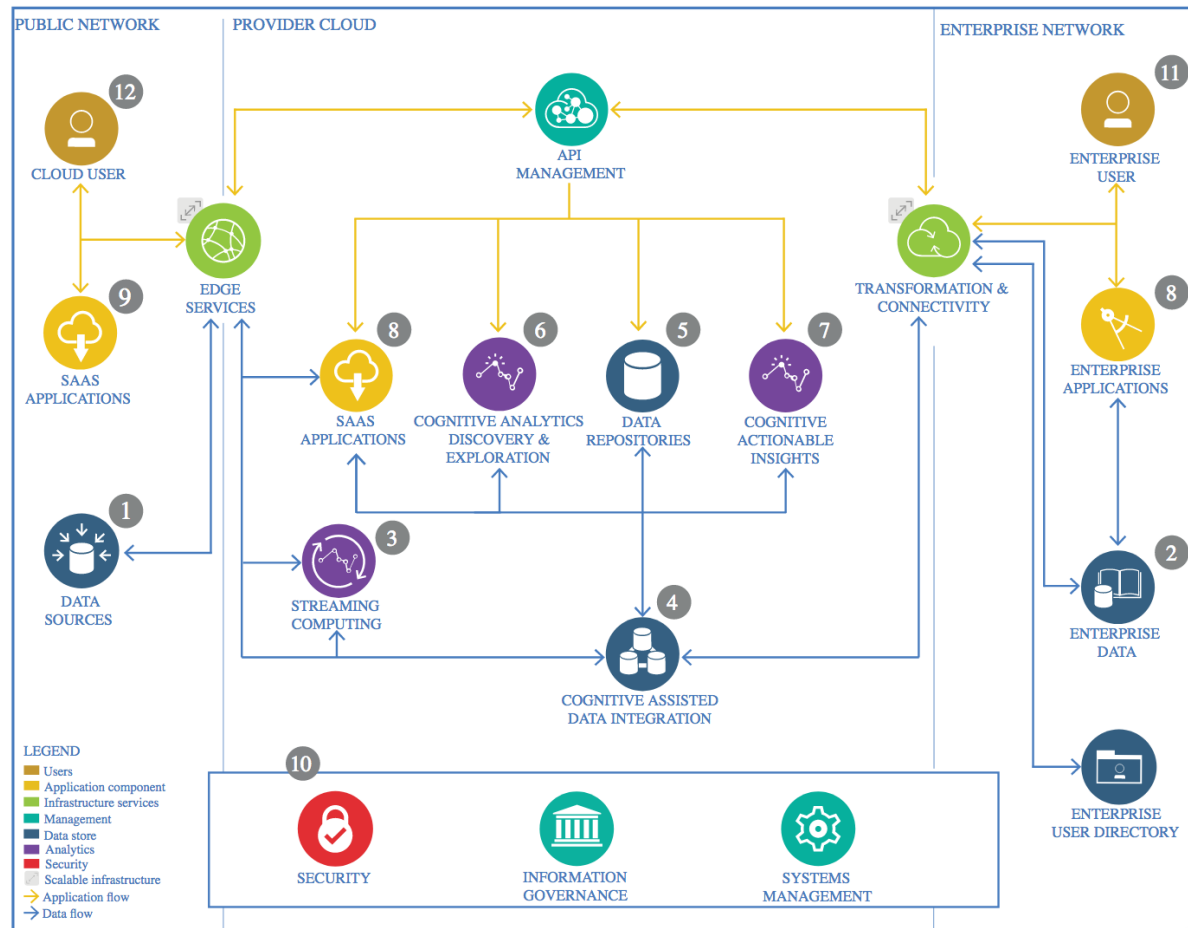
1 Architectural Components Overview

IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

The data source for this project is a CSV file obtained from Kaggle. The data source is the file sample.csv. It contains the data for this example in comma-separated values (CSV) format. The number of columns is 5, and the number of rows is 150.

The variables are:

sepal_length: Sepal length, in centimeters, used as input.
sepal_width: Sepal width, in centimeters, used as input.
petal_length: Petal length, in centimeters, used as input.
petal_width: Petal width, in centimeters, used as input.
class: Setosa, Versicolor, or Virginica, used as the target.

1.1.1   Technology Choice

The CSV data file is loaded to the IBM Watson Studio Notebook and then Python is being used for extracting and processing the data from the Object storage.

1.1.2   Justification

1) For this Capstone project, we currently have not yet explored the need for a SQL database due to the high cost of such databases. However this is something that can be looked into in the future if the need arises.

1.1.3   Technology Choice

Not applicable as of now.

1.1.4   Justification

At present as we do not see any Enterprise data needs for this project. As of now this is being envisioned as a stand-alone model / application / product. However this is something that can be used for finding the correct plant species.

1.2   Streaming analytics

1.2.1   Technology Choice

Not applicable as of now.

1.2.2   Justification

My present use-case does not have the need for streaming analytics.

1.3   Data Integration

1.3.1   Technology Choice

At present Data Integration is pretty straight forward and is mostly manual. The data is made available to us as csv or Excel files containing the data. This is downloaded and loaded into our Jupyter Notebook using native pandas or Tensor flow libraries.

1.3.2   Justification

My model is a very simple one with no requirement for frequent and automatic data updates via data integration frameworks.

1.4   Data Repository

1.4.1   Technology Choice

Presently data is loaded to the notebook and is not stored in IBM data storage.

1.4.2   Justification

IBM Cloud Object storage can be used as the Data storage solution for the following reasons:

1) Most of the data is being expected in the form of Data files. Hence storing the data files becomes easier in an object storage.

1.5   Discovery and Exploration

1.5.1   Technology Choice
For the Discovery and exploration of the data and our model, we are currently using the following.
- IBM Watson Studio as cloud based platform for building and deploying the solution
- IBM Object Storage as the data repository
- Python and Tensor flow libraries

1.5.2   Justification
Most of the tools and technology used in the project are Open source and IBM Watson studio is one of the best a cloud based platform which allows us to build and deploy AI solutions built using these Open Source tools.

1.6   Actionable Insights

1.6.1   Technology Choice
None as of now.

1.6.2   Justification
There are no actionable insights at present. However this can be relevant in the future as we enhance our model / data product.

1.7   Applications / Data Products

1.7.1   Technology Choice
None as of now.

1.7.2   Justification
There are no applications / Data Products in scope of our project at present. This is something that will be addressed as part of the enhancement scope in the future.

1.8   Security, Information Governance and Systems Management

1.8.1   Technology Choice
Not applicable as of now. There is however some Systems Management that is required which is being performed manually on a need basis.

1.8.2   Justification
At present my model is stand-alone model which does not interact or integrate with any external application or systems. Hence there is no real need or scope for implementing Security, Information Governance and Systems Management. This is also something which will need to be considered in the future as we enhance our model and start streaming data and also start interacting with external systems / data products.