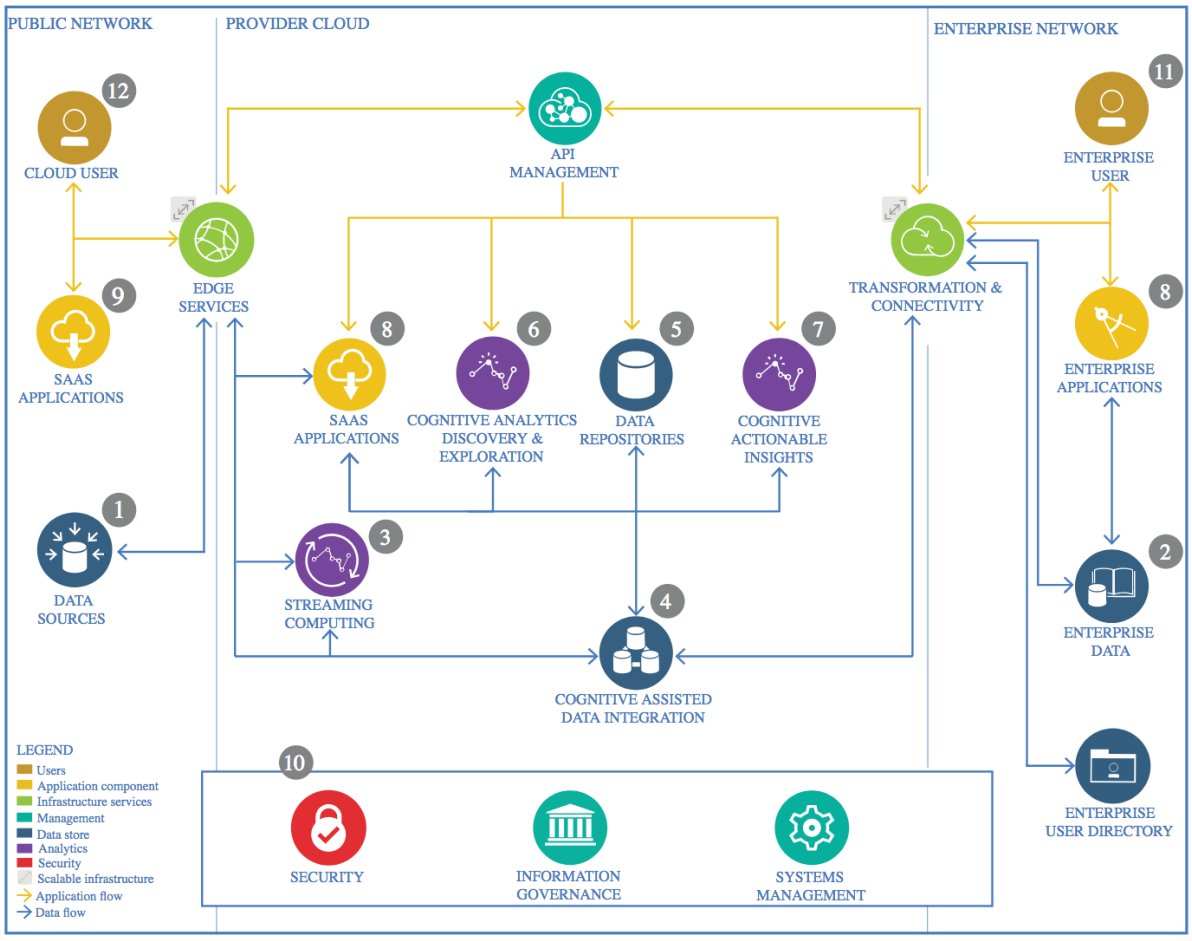
The Lightweight IBM Cloud Garage Method for Data Science
Architectural Decisions Document

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source
https://www.kaggle.com/ronitf/heart-disease-uci

1.1.1 Technology Choice
The data source for this project is a CSV file obtained from Kaggle. This data file contains heart disease data contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date.
The CSV data file is loaded to the IBM Cloud Object storage and then Python is being used for extracting and processing the data from the Object storage. We will also use Spark SQL to connect to the Object Store and handle it like a database.

1.1.2 Justification

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

1.2 Enterprise Data

1.2.1 Technology Choice

Not applicable

1.2.2 Justification

Not applicable

1.3 Streaming analytics

1.3.1 Technology Choice

Not applicable

1.3.2 Justification

Not applicable

.

1.4 Data Integration

1.4.1 Technology Choice

IBM Watson Studio, IBM Cloud Object Storage

1.4.2 Justification

IBM Cloud Object Storage provides a free plan and easy to integrate into IBM Watson Studio projects.

1.5 Data Repository

1.5.1 Technology Choice

IBM Cloud Object storage

1.5.2 Justification

IBM Cloud Object Storage is used as the Data storage solution.

1.6 Discovery and Exploration

1.6.1 Technology Choice

Apache Spark , Jupyter Notebook with Python 2.5 and Spark 2.3 , IBM Watson Studio and IBM Object Storage as our data repository

1.6.2 Justification

The Jupyter notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and texts.

1.7 Actionable Insights

1.7.1 Technology Choice

None as of now.

1.7.2 Justification

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

1.8 Applications / Data Products

1.8.1 Technology Choice

None as of now.

1.8.2 Justification

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

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

Not applicable

1.9.2 Justification

At present our 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.