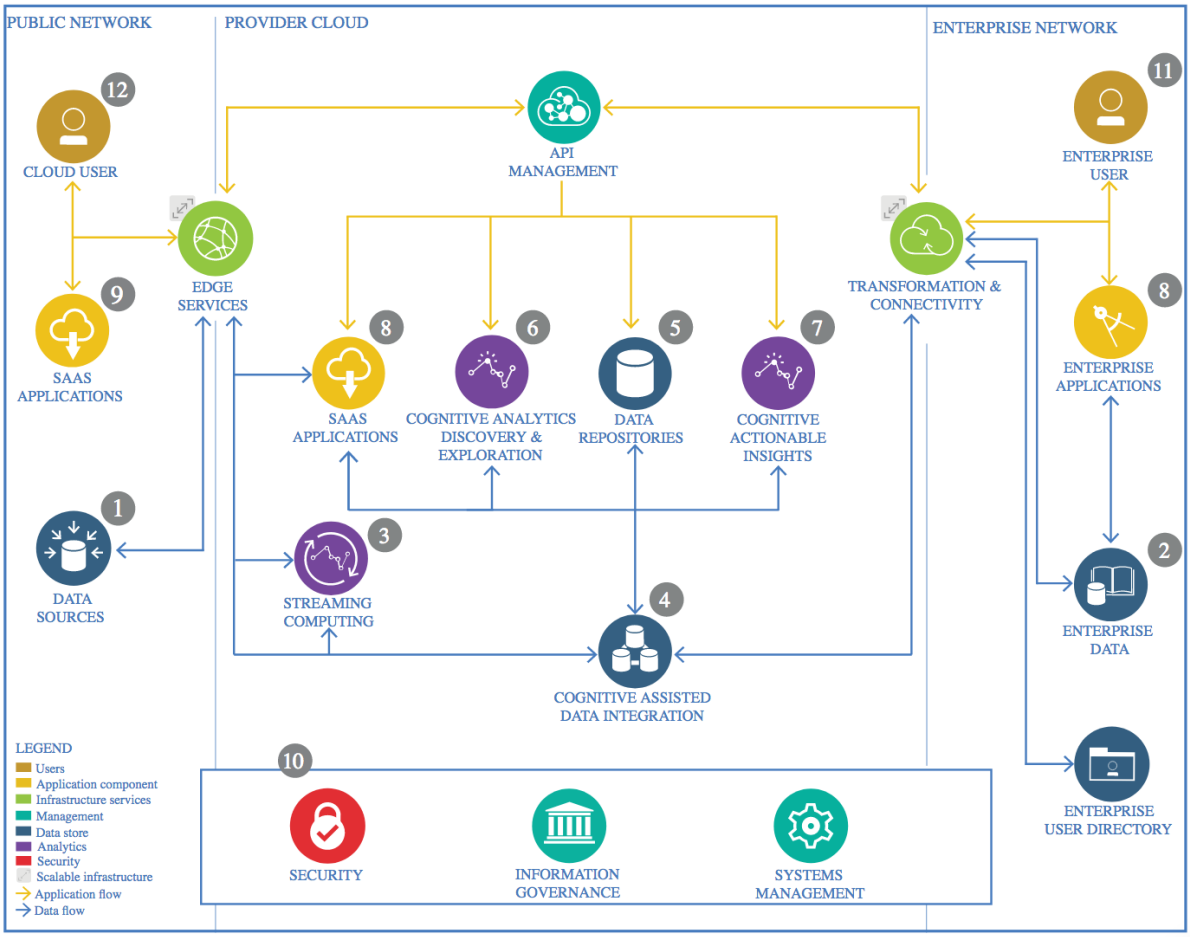
The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document Template

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

1.1.1 Technology Choice

We used data from the **UC Irvine Machine Learning Repository**. This dataset derived from *Chicco and Jurman (2020)*. The dataset includes cases of 299 patients, and it was collected in 2015. The data is CSV format.

Apache Spark (PySpark 3.3) was used for all analyses. The data was downloaded from the source link and stored in the IBM Cloud Object storage.

1.1.2 Justification

We used the IBM Cloud Object storage and run PySpark to enhance data access and the efficiency of data computing. IBM Watson tool is a good and powerful API for big data.

1.2 Enterprise Data

1.2.1 Technology Choice

We did not use enterprise data. Our data retrieved from the **UC Irvine Machine Learning Repository.**

1.2.2 Justification

For this Capstone project, we used a small amount of data for the demonstration. Therefore, we did not need to use enterprise data. When we perform a specific project using big data for a real-world application development, we might need the enterprise data.

1.3 Streaming analytics

1.3.1 Technology Choice

No applicable.

1.3.2 Justification

In this project, we only used the data once from our use case for the cloud-based computing using IBM Watson Studio. If we continuously upload small data to IBM platform and do streaming analysis, then we might need to use streaming analytics. In the future, we might use this service.

1.4 Data Integration

1.4.1 Technology Choice

Data integration is an important part of consolidating data from multiple sources. We extracted data from the source in csv file and loaded to the IBM Cloud Object storage and used that csv file. No data integration was applied.

1.4.2 Justification

In this project, we used a single dataset derived from the machine learning repository, in which data contains 299 observations. The whole modeling process used this single dataset and no data integration from other sources was applied.

1.5 Data Repository

1.5.1 Technology Choice

We uploaded and stored our data in IBM object storage.

1.5.2 Justification

The data stored in the IBM object storage can be accessed through different electronic devices and geographic locations with a higher protection and safety.

1.6 Discovery and Exploration

Data Discovery concluded that our data was cleaned and no strong correlation between independent variables, indicating that it is reliable for further analysis.

1.6.1 Technology Choice

We used Apache Spark, Python 3.10 Jupyter Notebook on Watson Studio. Python libraries, such as pandas, numpy, seaborn, and matplotlib, were used for data exploration to check data type, missing values, distribution of features, and correlation between independent variables.

1.6.2 Justification

Python Jupyter Notebook is a powerful tool and environment for data exploration and visualization.

1.7 Actionable Insights

From three Machine Learning and one Deep Learning algorithm, Random Forest model showed a better performance of predicting the likelihood of deceased from heart failure.

1.7.1 Technology Choice

Our dataset include labeled features and it is the best suited to use Supervised Machine Learning algorithms. In Supervised Machine Learning, there are two learning algorithms: Regression and Classification. Regression is good for continuous numerous target variables and classification is good for categorical and discrete target variables. In our case, the target ('DEATH_EVENT') is a binary categorical variable (Survived:0 and Deceased:1). Therefore, we used Supervised Machine Learning classifiers.

1.7.2 Justification

1. Firstly, we used a simple Machine Learning model, which is a Logistic Regression. It is good for binary categorical classification.

2. Then, the Random Forest (RF) algorithm was used. RF is an ensemble learning model that uses bagging. Both classification and regression can be used in RF.

3. Also, Gradient-Boosted Tree (GBT) is a popular Machine Learning model and an ensemble learning model. However, GBT uses boosting and suitable for both classification and regression.

4. For a Deep Learning model, we used Feed Forward Neutral Network, which is also called an Artificial Neutral Network (ANN). This is a basic neutral network and is good for classification and regression.

1.8 Applications / Data Products

1.8.1 Technology Choice

Our data was only used to predict which machine learning and deep learning algorithms were better suited to predict the probability of deceased from heart failure. We did not create or produce any applications from this modeling.

1.8.2 Justification

For a final data product, we need to make a pdf report or code repository. No application or other data products were applied.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

No applicable.

1.9.2 Justification

For this demonstration project, we have no security and information governance. However, if our data is used to deploy for a specific application or purpose, we might need security, information governance and system management.