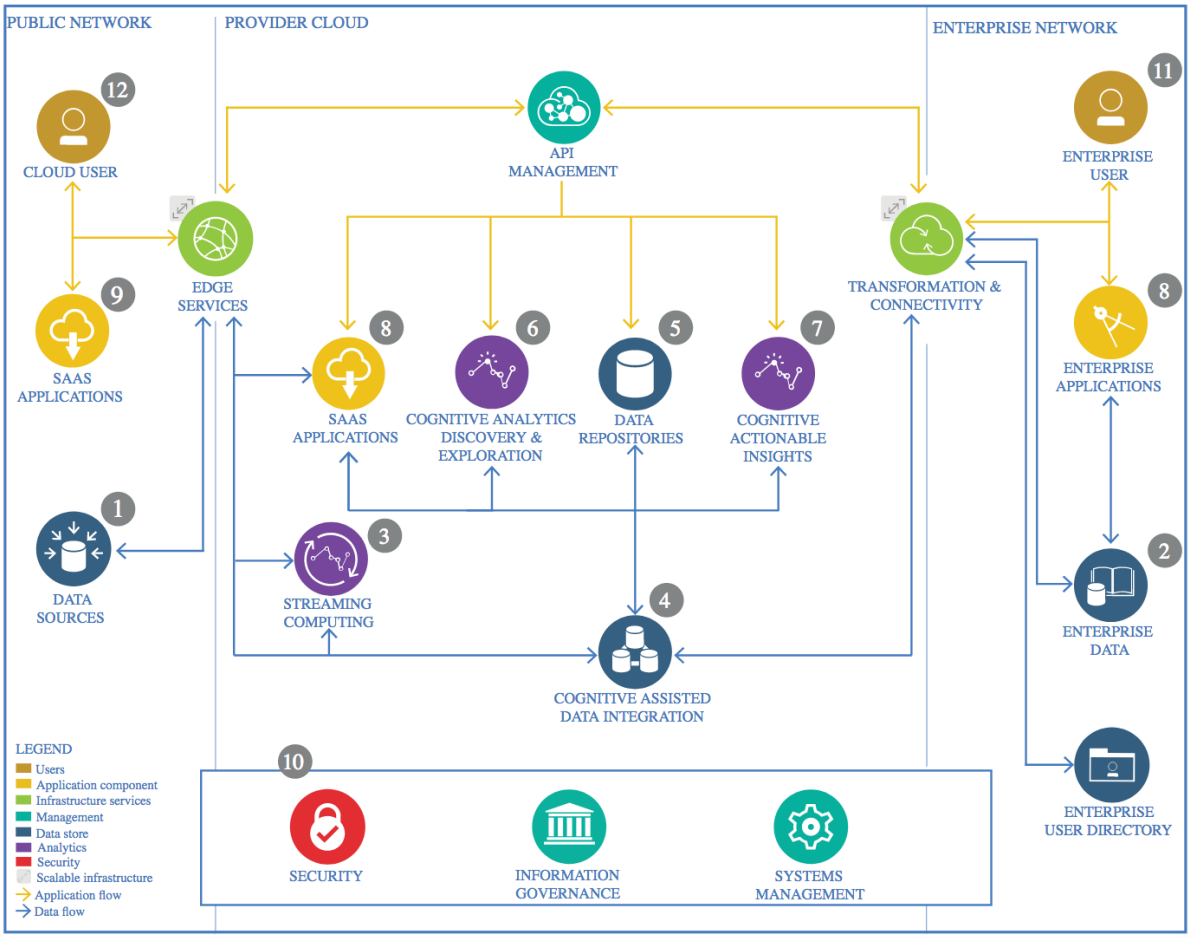
The Lightweight IBM Cloud Garage Method for Data Science

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

To obtain the dataset, I used Kaggle's open database.

1.1.2 Justification

Kaggle's database has a vast amount of data on various fields starting from medical science related data to economical and financial databases. They have all formats of datatype such as csv, json to even annotated images thus providing with rich options to choose from.

1.2 Enterprise Data

1.2.1 Technology Choice

No use of Enterprise Data was needed here as only an open database was used and analyzed.

1.2.2 Justification

My choice of database came from an open database where contributors could upload data they thought could be useful for others. As the data that I collected was sufficient for analysis purposes I did not need any other Enterprise data to be used.

1.3 Streaming analytics

1.3.1 Technology Choice

I stored the data in IBM Cloud Storage, the data was loaded into the associated notebook from this source.

1.3.2 Justification

Streaming analysis was not needed as the data was static and no live data was being collected in real time.

1.4 Data Integration

1.4.1 Technology Choice

I had used IBM Cloud Storage from which it was loaded into Jupyter Notebook. Data Integration aspect of the data manipulation was done there with the use of dataframe libraries such as Pandas and Numpy in Python language.

1.4.2 Justification

Python provides very easy data manipulative libraries like pandas and numpy which seemlessly allow data analysis and manipulation. The association of a IBM Cloud Storage Bucket with a project notebook also makes it an easy task to load in the files in to the notebook to perform analysis on it.

1.5 Data Repository

1.5.1 Technology Choice

I used IBM Cloud Storage to store my assets. Both the csv dataset and the associated Notebook for the project stored in a bucket.

1.5.2 Justification

IBM Storage comes associated with projects that one creates. As it was also free for limited storage, Cloud Storage was a nice option to use as Data Repository.

1.6 Discovery and Exploration

1.6.1 Technology Choice

The main language used to code out the exploration and analysis of data was Python. data analysis and eploration was done using the python libraries, scikit-learn, numpy and pandas. For visualization, Matplotlib was used.

1.6.2 Justification

The ML APIs provided were easy to use in python which is a well established data science programming language. Numpy has C running in it's core which makes all the data manipulation tasks fast, even for large datasets. These libraries also provide a lot of feature engineering functions such Label Encoding, One hot encoding, etc.

1.7 Actionable Insights

Data visualization has provided with certain actionabe insights that can be taken such as the higher initial costs of new customers lead to most of them ending the subscriptions. Therefore cost was one reason that customers would quit.

1.7.1 Technology Choice

I used python and especially the matplotlib library to make data plots and to get meaningful insights from the data plots.

1.7.2 Justification

Python is a well established language used by data scientist and machine learning enigneers. With the abundance of options in its library when it comes to making data visualization charts to get more insights in data, it was an easy choice to use them for my purposes.

1.8 Applications / Data Products

1.8.1 Technology Choice

I have created a Jupyter Notebook provided in IBM Watson Studio to create my data product.

1.8.2 Justification

Jupyter Notebook is far easier to use and also easy to share to the customers who asked for the data analysis. It also allows the author to give explanations next to the code and plotted charts, which makes it easy to help the customers also understand the insights that were made.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

None were used.

1.9.2 Justification

I did not have to use any of these as the created data product was far simple and did not needed any complex management setups.