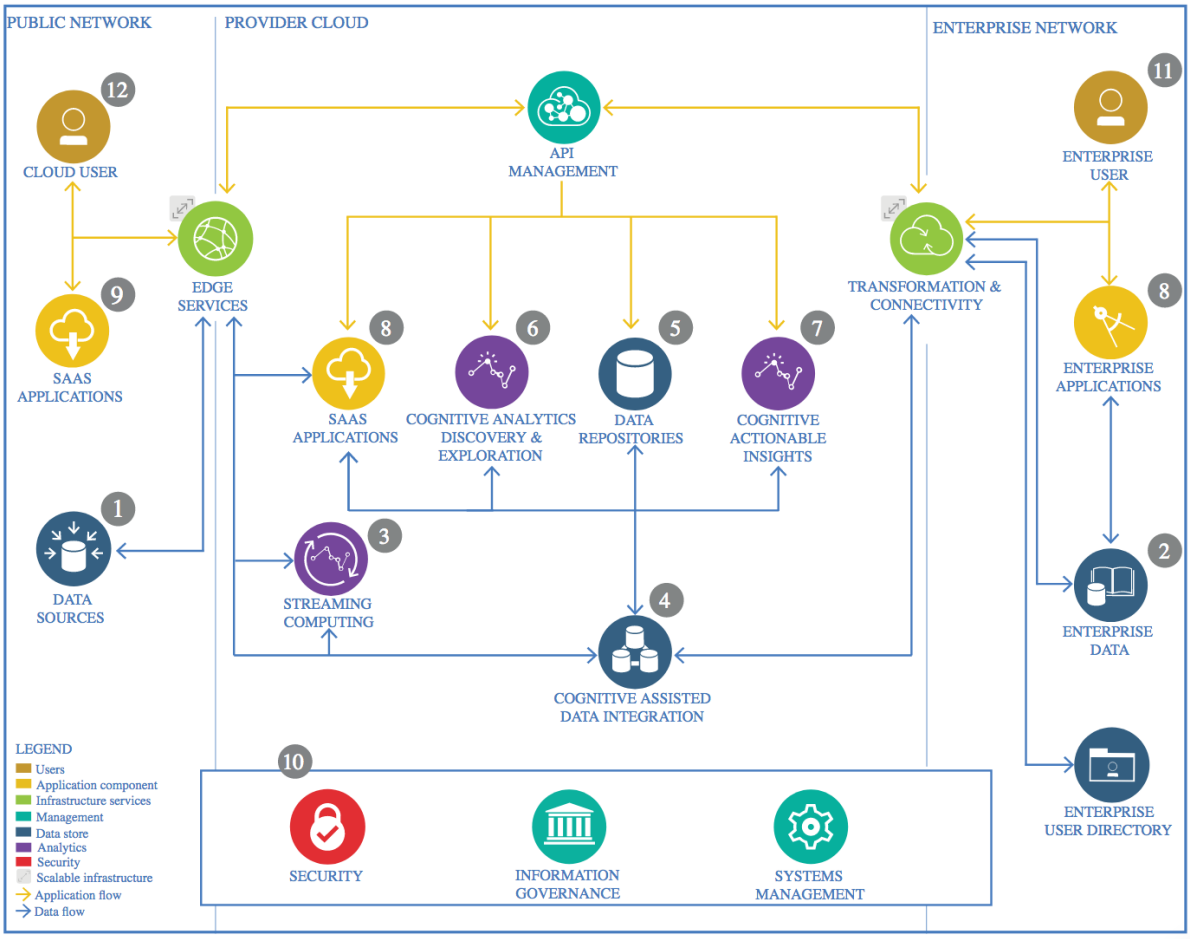
The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

1.1.1 Technology Choice

I obtained this dataset from Kaggle.
Dataset Link:
https://www.kaggle.com/edalrami/19000-spotify-songs

1.1.2 Justification

Kaggle has a vast collection of datasets ranging from wide fields like medical science, financial databases, etc. They have all formats of datatype (such as csv, json, annotated images, etc) providing a rich variety of choices.

1.2 Enterprise Data

1.2.1 Technology Choice

No enterprise data was needed here as only an open dataset was used for analysis.

1.2.2 Justification

I used an open database which was freely available on Kaggle. 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

Not Applicable.

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 csv files which were loaded into Jupyter Notebook. Data Integration aspect of the data manipulation was done 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 seamlessly allow data analysis and manipulation. It also makes it an easy task to load in the files in to notebook to perform analysis on it.

1.5 Data Repository

1.5.1 Technology Choice

IBM Cloud Object Storage.

1.5.2 Justification

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

1.6 Discovery and Exploration

1.6.1 Technology Choice

The main language used for the exploration and analysis of data was Python. The Data analysis and exploration was done using the python libraries, such as scikit-learn, numpy and pandas. For visualization, Matplotlib was used.

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 narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

1.7 Actionable Insights

1.7.1 Technology Choice

Data visualization with matplotlib has provided with certain actionable insights that can be taken such as the factors responsible for popularity of a particular song, the correlation between various parameters.

1.7.2 Justification

The Jupyter Notebook is an open-source web application that allows us to create and share documents that contain live code, equations, visualizations and narrative text. The Uses include: data cleaning and transformation, data visualization, machine learning, and much more.

1.8 Applications / Data Products

1.8.1 Technology Choice

The chosen models are Logistic Regression, KNN and SVM. Jupyter Notebook takes in CSV data file and outputs the results to predict the popularity class of songs.
Evaluation metric used – Accuracy. The accuracy of these models was compared to look at their performance.

1.8.2 Justification

Jupyter Notebooks are easy to use and also easy to share 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 discovered during EDA.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

Not Applicable.

1.9.2 Justification

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