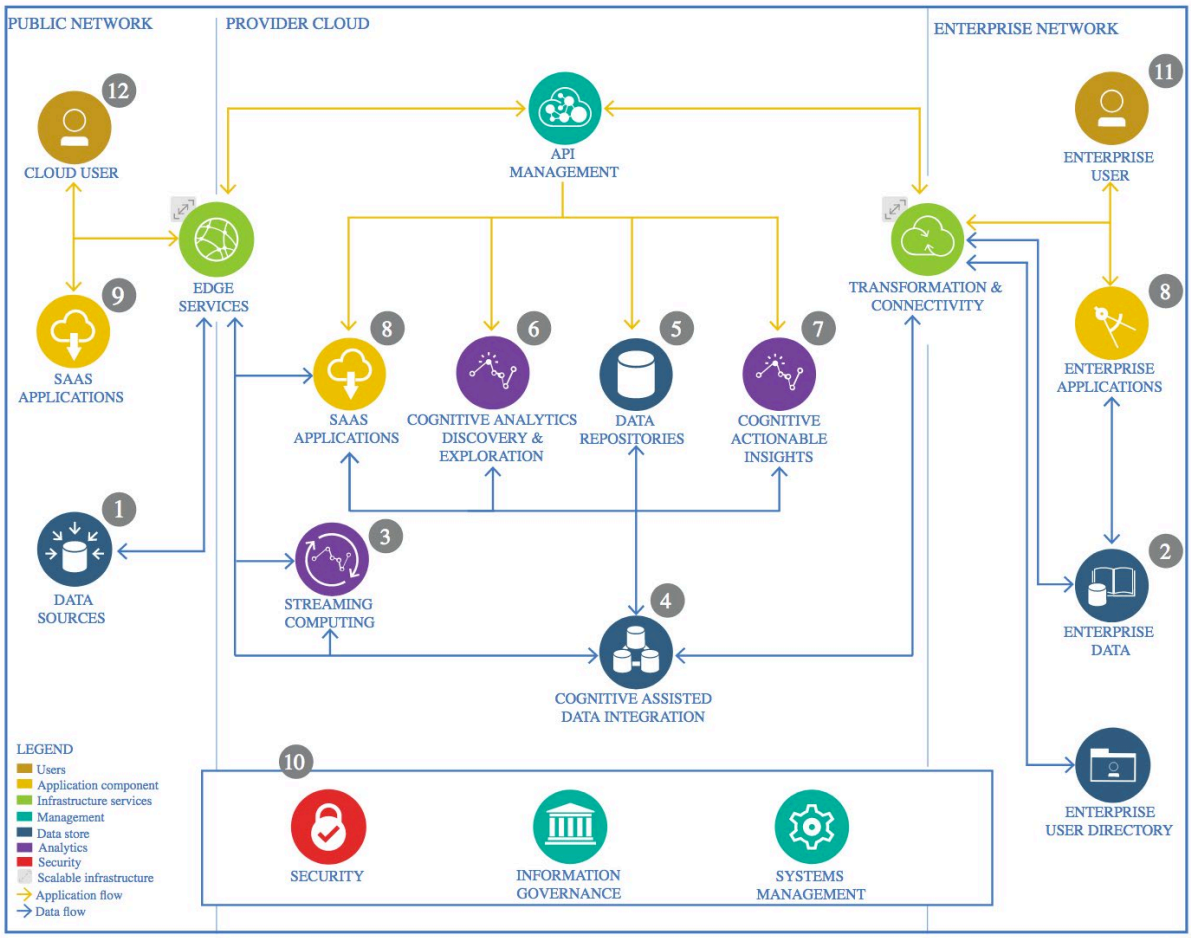
The Lightweight IBM Cloud Garage Method for Data Science

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

The data was obtained from Kaggle's open database collection.

1.1.2 Justification

Kaggle's database has a vast amount of data on various fields starting from medical science related data to economic 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

The choice of dataset came from an open database where contributors could upload data, that could be useful for others. As the data collected was sufficient for analysis purposes, there was no need for any other Enterprise data to be used.

1.3 Streaming analytics

1.3.1 Technology Choice

The data was stored in GitHub, 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

Google Colaboratory was used for developing the Jupyter Notebooks. Data Integration aspect of the data manipulation was taken care of with the use of data frame 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. The usage of Pandas makes it very easy to use SQL-like functions and work on the data in a speedy manner and enables working on various forms of data with no limit on size.

1.5 Data Repository

1.5.1 Technology Choice

GitHub Repository was used to store all the assets.

1.5.2 Justification

GitHub simplifies the process of working with other people and makes it easy to collaborate on projects. It also provides storage solutions for projects on the cloud. Hence, GitHub was chosen for Data Repository.

1.6 Discovery and Exploration

1.6.1 Technology Choice

The primary language used to code out the exploration and analysis of data was Python. Data analysis and exploration was done using Pandas and the python libraries- SciKit-Learn and NumPy. For visualization, Matplotlib and Seaborn were 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 its 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

1.7.1 Technology Choice

Matplotlib and Seaborn libraries were utilized to make data plots. To get meaningful insights from the data plots, Statmodels library was used.

1.7.2 Justification

Pandas makes it very convenient to obtain statistical description of the data frame. Plotting with Matplotlib and Seaborn helped in further insight.
Meaningful insights were made with the help of Statmodels modules like SARIMAX, tsa, ARIMA, etc.

1.8 Applications / Data Products

1.8.1 Technology Choice

There is no data product created in this project.

1.8.2 Justification

This project analyses the sales of the organization in the past few years and predicts future sales. This can be helpful to attain insight regarding the business and obtain solutions to enhance the business and make the sales further profitable.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

None were used.

1.9.2 Justification

Since the project is about analysis and prediction alone, none of the management setups were required.