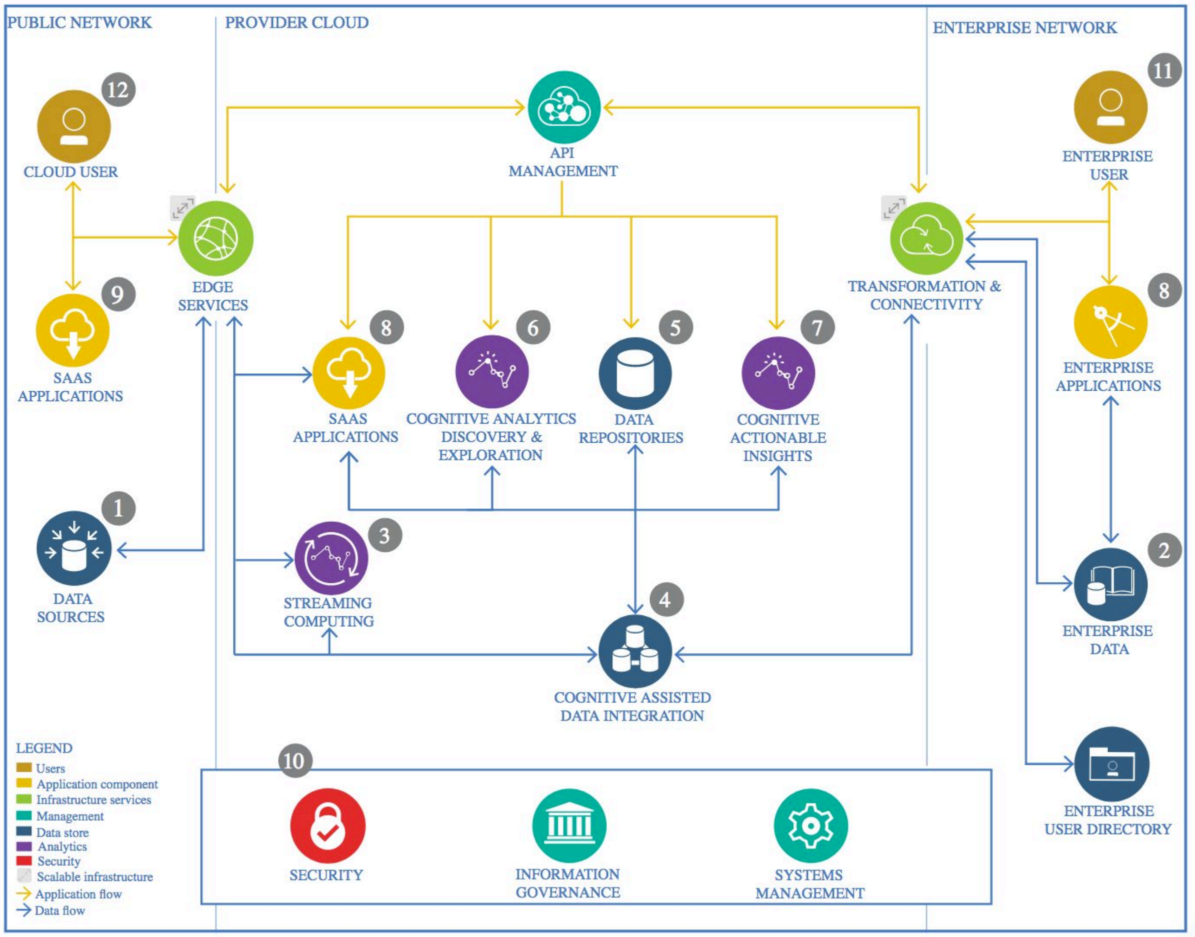
The Lightweight IBM Cloud Garage Method for Data Science

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

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 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

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 data frame libraries such as Pandas, Apache Spark, PyArrow 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 Apache Spark makes it very easy to use SQL 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

I used IBM Cloud Storage to store my assets. Both the csv and parquet dataset, and the associated Notebooks for the project are 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 exploration were done using Apache Spark SQL and the python libraries- SciKit-Learn, NumPy and PySpark. For visualization, Seaborn 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 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

PySpark, Apache Spark SQL and Seaborn libraries were utilized to make data plots and to get meaningful insights from the data plots.

1.7.2 Justification

Apache Spark SQL with PySpark makes it very convenient to obtain statistical description of the data frame. Plotting with Seaborn helped in further insight and hence helped in choosing the best IoT device for later stages of processing.

1.8 Applications / Data Products

1.8.1 Technology Choice

I have used Jupyter Notebooks provided by IBM Watson Studio to create my data product.

1.8.2 Justification

Jupyter Notebooks are far easier to use and 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

Since the created data product was far simple and did not needed any complex management setup.