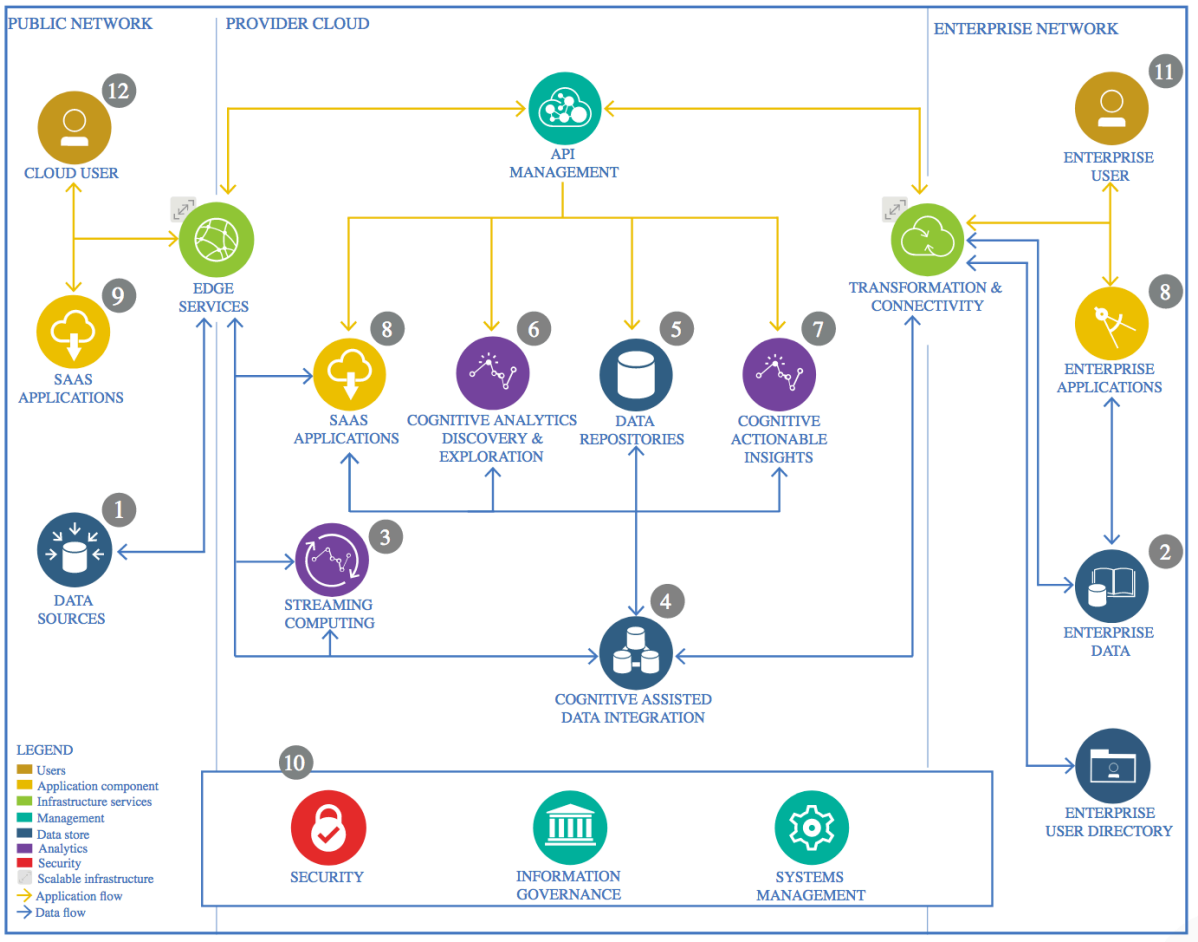
The Lightweight IBM Cloud Garage Method for Data Science

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

1.1.1 Technology Choice

- CSV files of IoV network packets

1.1.2 Justification

The project wanted to examine anomalies in network packet data for IoV (internet of vehicles), and sourcing was found to be difficult. A dataset from the Canadian Institute for Cybersecurity was sourced.

1.2 Enterprise Data

1.2.1 Technology Choice

- N/A

1.2.2 Justification

As this is a locally hosted project, there is no enterprise data available. However, in an enterprise environment, cloud storage could have been used such as object storage in IBM Watson.

1.3 Streaming analytics

1.3.1 Technology Choice

- N/A

1.3.2 Justification

As this project uses static data, there is no need for streaming analytics. However, if this project were to be done with live data, a potential choice would be Snowflake.

1.4 Data Integration

1.4.1 Technology Choice

- Apache Spark
- String Indexer

1.4.2 Justification

Apache spark can be used for big data, in this case over a million rows of network packet data. This allowed for easy extraction, loading, and transformation of the data set for use downstream.

1.5 Data Repository

1.5.1 Technology Choice

- Github

1.5.2 Justification

As this is a locally hosted project, Github is a great choice for storing files, as anyone else looking to use the project's notebooks can find the data present within the same repository.

1.6 Discovery and Exploration

1.6.1 Technology Choice

- Pandas, matplotlib

1.6.2 Justification

Pandas allows for a summary of a dataframe in order to view statistical moments such as mean and standard deviation. This allows for quick insights into the data that we

will be working with in this project, alongside matplotlib for visualization of our data to understand and trends or patterns.

1.7 Actionable Insights

1.7.1 Technology Choice

- N/A

1.7.2 Justification

As this is a locally hosted project, there is no actionable insight apps that were used. If done via cloud, a potential choice would be nodered.

1.8 Applications / Data Products

1.8.1 Technology Choice

- Apache Spark
- Keras
- Jupyter Labs

1.8.2 Justification

A locally hosted project can use these libraries and platform in order to create a viable machine learning product, which can be replicated on any machine.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

- N/A

1.9.2 Justification

As this is a locally hosted project, there is no concern for using security information governance and systems management. If done via cloud, a potential choice would be Google Chronicle .