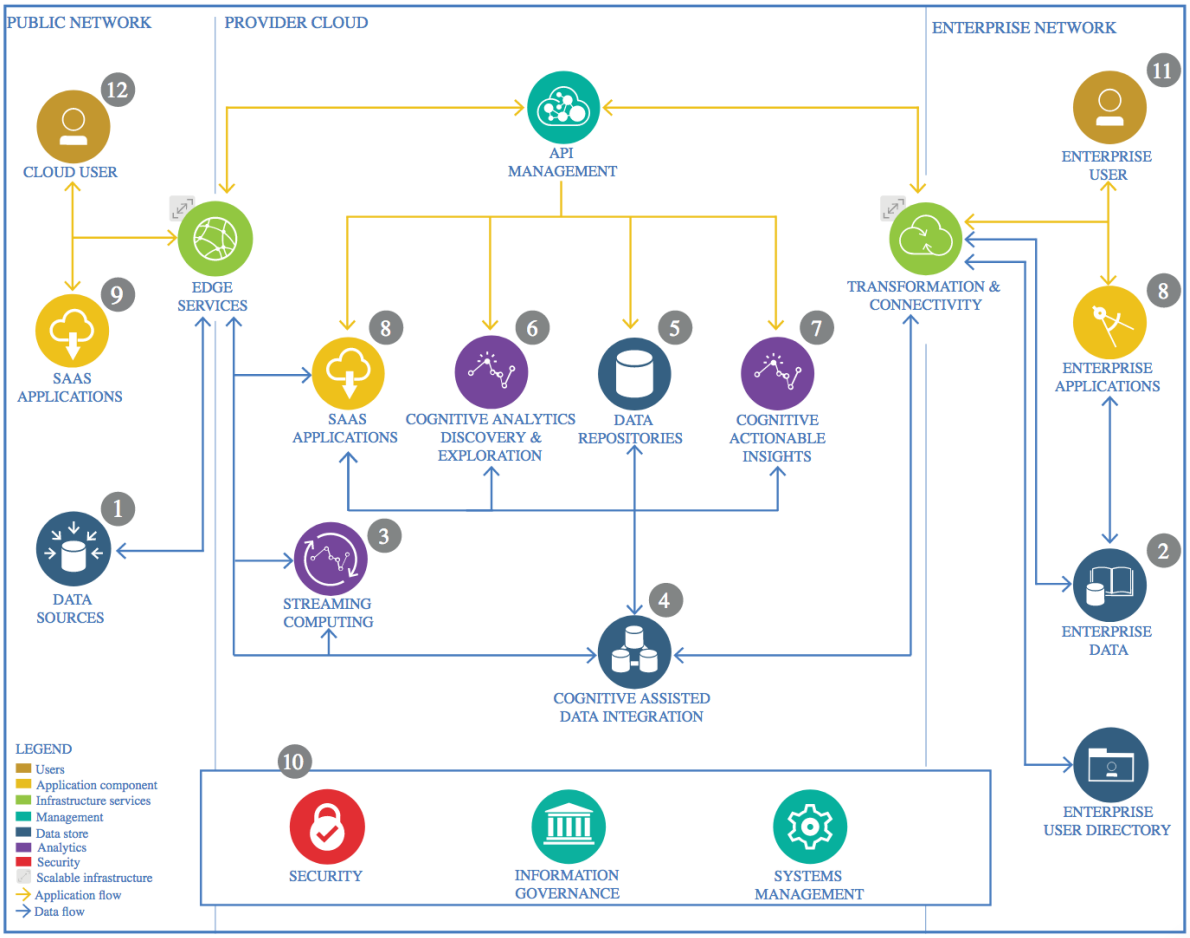
The Lightweight IBM Cloud Garage Method for Data Science

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

Kaggle API for downloading the "Wikipedia Movie Plots" dataset.

1.1.2 Justification

The Kaggle API is a convenient tool for accessing and retrieving datasets from Kaggle, ensuring ease of data acquisition and reproducibility.

1.2 Enterprise Data

1.2.1 Technology Choice

NA

1.2.2 Justification

As a personal project there is no enterprise data used.

1.3 Streaming analytics

1.3.1 Technology Choice

NA

1.3.2 Justification

The project doesn't involve real-time processing or streaming analytics as all data is locally stored. If real-time data processing were a requirement, technologies like Apache Kafka or Apache Flink might be considered.

1.4 Data Integration

1.4.1 Technology Choice

Pandas

1.4.2 Justification

Pandas is a powerful library for data manipulation, providing tools for cleaning, transforming, and merging datasets. It serves as an effective choice for data integration tasks within the Python ecosystem.

1.5 Data Repository

1.5.1 Technology Choice

NA

1.5.2 Justification

The project uses data stored on local csv files. In a production setting, a database system like mySQL or MongoDB might be considered for structured storage.

1.6 Discovery and Exploration

1.6.1 Technology Choice

NLTK, Gensim, Matplotlib and **PyLDAvis**

1.6.2 Justification

NLTK and Gensim are chosen for text processing and topic modeling, providing tools for uncovering patterns in textual data. Matplotlib and PyLDAvis are selected for visualization, aiding in the exploration and interpretation of the data.

1.7 Actionable Insights

1.7.1 Technology Choice

Scikit-Learn for machine learning tasks (Random Forest Regressor)

1.7.2 Justification

Scikit-Learn is a widely used machine learning library, and the Random Forest Regressor is employed to derive insights and predictions from the data, contributing to actionable insights for predicting movie ratings

1.8 Applications / Data Products

1.8.1 Technology Choice

NA

1.8.2 Justification

The project focuses on exploratory data analysis and machine learning modeling. If it was part of a larger system or application, deployment frameworks like Flask or Django might be considered for creating and serving data products.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

NA

1.9.2 Justification

The project doesn't address security, information governance, or systems management explicitly. For enterprise-scale applications, technologies like Apache Ranger for security and governance or Apache Airflow for workflow management might be considered.