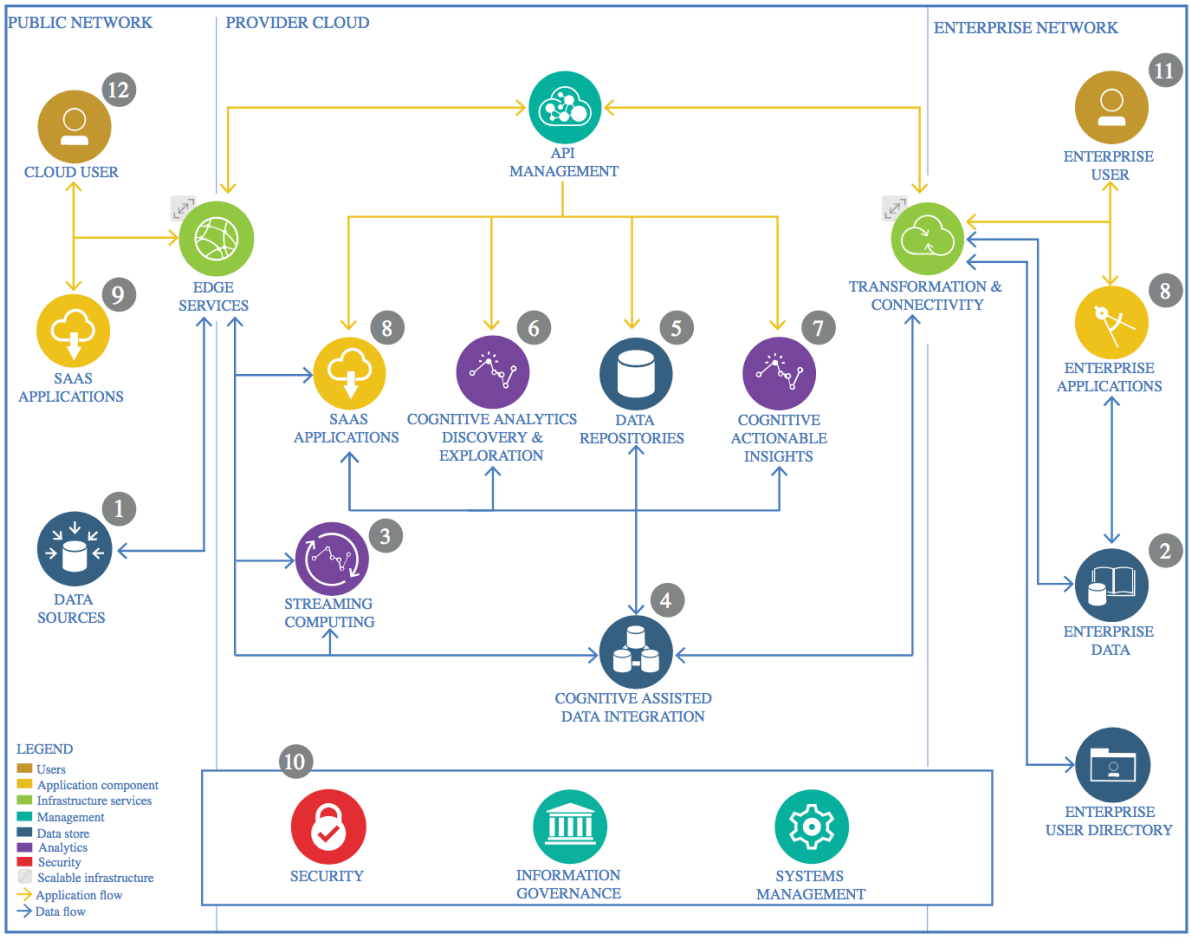
The Lightweight IBM Cloud Garage Method for Data Science

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 Datasets.

1.1.2 Justification

Kaggle has a wide range of datasets for various use cases.

1.2 Enterprise Data

1.2.1 Technology Choice

NA

1.2.2 Justification

Does not apply to my use case.

1.3 Streaming analytics

1.3.1 Technology Choice

NA

1.3.2 Justification

The dataset is not real time.

1.4 Data Integration

1.4.1 Technology Choice

NA

1.4.2 Justification

The dataset perhaps has already been integrated and saved in .csv format

1.5 Data Repository

1.5.1 Technology Choice

https://www.kaggle.com/bhrt97/hr-analytics-classification?select=train_LZdllcl.csv

1.5.2 Justification

The given dataset is appropriate, and its size is large which can give a good model performance.

1.6 Discovery and Exploration

1.6.1 Technology Choice

Most of the employees belong to the Marketing & Sales department and the promoted ones are also mostly from Marketing & Sales

The company has most people in their early 30's and 5 – 8 years of experience

The class label is highly imbalanced (approximately 9:1) and hence an up-sampling approach was used

1.6.2 Justification

NA

1.7 Actionable Insights

1.7.1 Technology Choice

Created three new features from given features. Namely Total Score, Awards_KPI, Experience.

Performed MinMax Scaling to normalize the data.

1.7.2 Justification

These approaches increased model performance by significant amount.

1.8 Applications / Data Products

1.8.1 Technology Choice

The following model can be used to pick ideal candidates for promotion and reducing the workload on traditional approaches.

1.8.2 Justification

NA

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

NA

1.9.2 Justification

NA