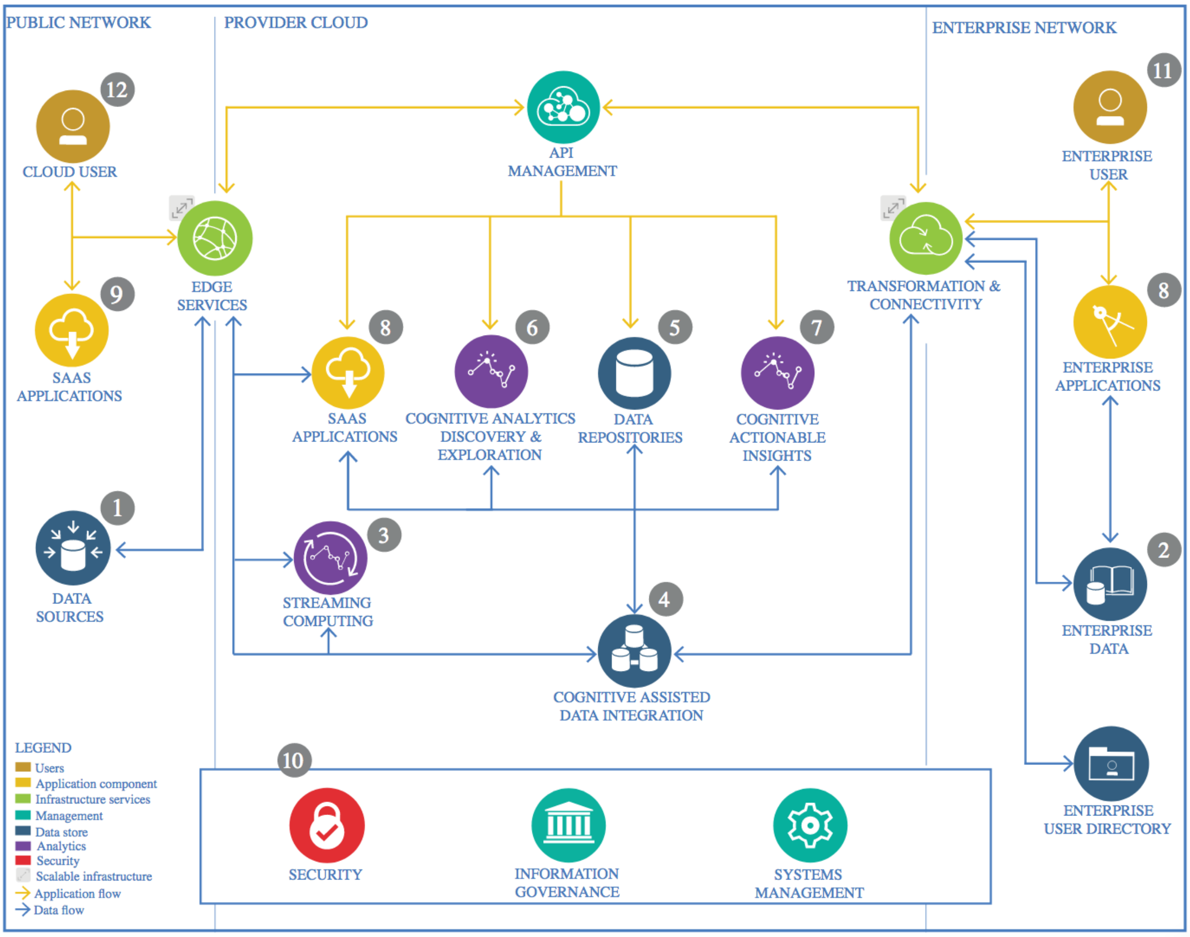
The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document

Online Customer Intention Prediction

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 downloaded from Kaggle (https://www.kaggle.com/roshansharma/online-shoppers-intention)

1.1.2 Justification

Primary reason to download from Kaggle was availability and ease of use.

1.2 Enterprise Data

1.2.1 Technology Choice

GitHub repository

1.2.2 Justification

Up-to-date data would be available on the repository

1.3 Streaming analytics

1.3.1 Technology Choice

NA

1.3.2 Justification

NA

1.4 Data Integration

1.4.1 Technology Choice

Not used

1.4.2 Justification

Not used

1.5 Data Repository

1.5.1 Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

1.5.2 Justification

Please justify your technology choices here.

1.6 Discovery and Exploration

1.6.1 Technology Choice

The following Python 3.6 libraries were used for Data Exploration and Visualization: -
Pandas,
Matplotlib,
Seaborn

1.6.2 Justification

The size of the dataset was the key factor in deciding data exploration tools.
The current data small enough to be processed on a single computer ruling out the need for distributed processing (Spark, pyspark)

1.7 Actionable Insights

1.7.1 Technology Choice

The following Python 3.6 libraries were used for Data Exploration and Visualization: -
Pandas,
Light-GBM,
Keras,
Tensoflow.

1.7.2 Justification

To understand the Correlating features a white-box model was required. Tree based algorithms were identified as a good match. Thus Light-GBM was used.
Neural network based algorithm was used as a reference for the Tree based model. Easiest and Fastest implementation is possible in keras. Tensorflow is the backend.

1.8 Applications / Data Products

1.8.1 Technology Choice

A Jupyter notebook based report was generated

1.8.2 Justification

As only the correlating factors needed to be identified Jupyter notebook based report was consider sufficient.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

None

1.9.2 Justification

NA