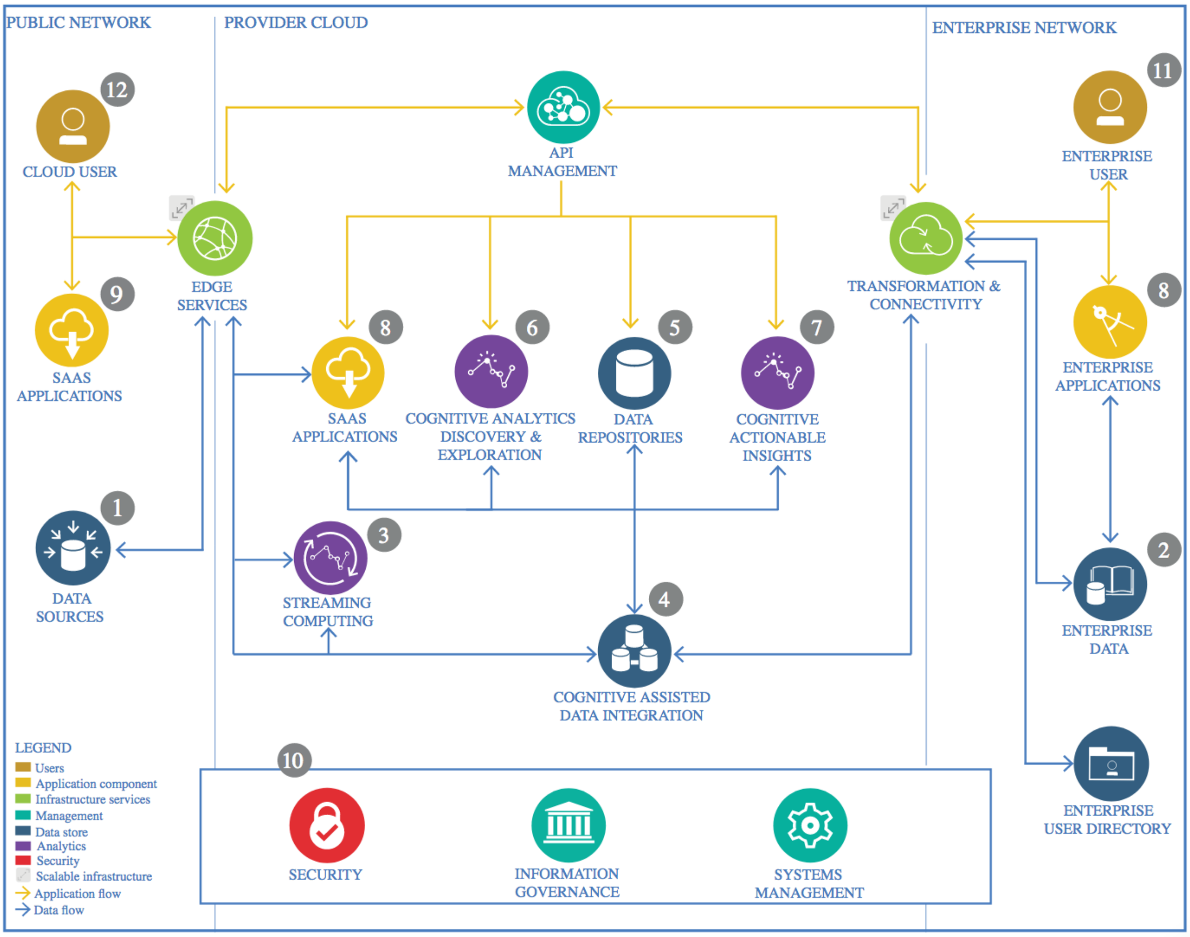
The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document Template

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

Project Title: User Rating Prediction from Review Text

1.1 Data Source

1.1.1 Technology Choice

The data set from the Amazon's Books review was used for the rating prediction and sentiment analysis.

1.1.2 Justification

The selected data source provides a high quality data set that can be used for the project with minimal preprocessing.

1.2 Enterprise Data

1.2.1 Technology Choice

Not applicable.

1.2.2 Justification

As I am using the freely available data and no enterprise data was required for the project.

1.3 Streaming analytics

1.3.1 Technology Choice

Not Applicable.

1.3.2 Justification

No streaming data was required for the project.

1.4 Data Integration

1.4.1 Technology Choice

Not applicable.

1.4.2 Justification

A single data source was used from http://jmcauley.ucsd.edu/data/amazon/

1.5 Data Repository

1.5.1 Technology Choice

The University of Calfornia San Diego (UCSD) maintains a repository of Amazons' reviews data set. The data set was downloaded directly from the source into the notebook's working directory in IBM Cloud.

1.5.2 Justification

The data could be easily maintained and used from the working directory.

1.6 Discovery and Exploration

1.6.1 Technology Choice

A simple analysis was performed to explore the data using python. The exploration mainly required the understanding of number of instances, distribution of class instances, number of features, etc.

The provided data was of high quality, i.e. did not have any missing values. However, for some reviews the length for text was too long. Therefore, I used only the reviews with the length between 10 and 500.

I also tried balance the instances in each class but that does not improve anything rather the performance of the model was deteriorated.

1.6.2 Justification

Python data frames provide handy tools for the required data discovery and exploration.

1.7 Actionable Insights

1.7.1 Technology Choice

To get the actionable insights, I used the following deep learning models.

- GRU based model
- LSTM based model
- 1D CNN based model

An embedding layer was used in all these models to get the high dimensional representation of word vectors that keep the context of words.

The models were evaluated using accuracy. The loss function used for the rating prediction was categorical cross entropy while for sentiment analysis I used binary cross entropy.

1.7.2 Justification

Both RNNs (GRU, LSTM) and Convolutional Neural Networks (1D CNN) are widely used in literature for sequence models and give a performance.

1.8 Applications / Data Products

1.8.1 Technology Choice

The final product is presented as Jupyter Notebook.

1.8.2 Justification

The Jupyter Notebook is a good choice where all the steps of the project are documented and presented.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

This project is publicly available.

1.9.2 Justification

The project is done for educational purpose so does not require any explicit mention of security and information governance.