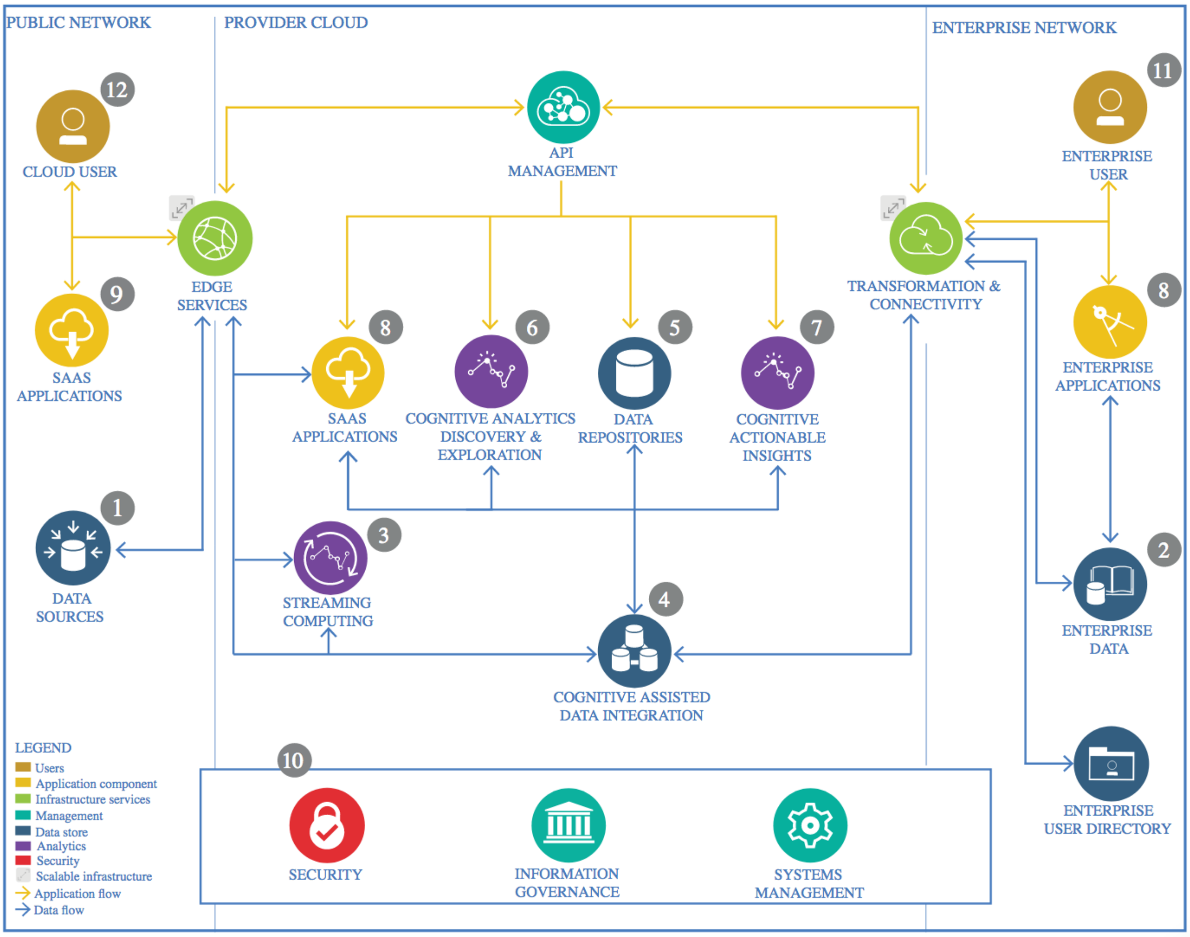
Architectural Decisions Document (ADD)

The Lightweight IBM Cloud Garage Method for Data Science

Project:

Sentiment Analysis

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

1.1.1 Technology Choice

The heart of the this project is the dataset “IMDB Movies Reviews” distributed by Kaggle.

1.1.2 Justification

The data set is from a public website.

1.2 Enterprise Data

1.2.1 Technology Choice

We will use Jupyter Notebooks on IBM Watson to analyze the data and run models.

1.2.2 Justification

Jupyter Notebooks are sufficient for this project that does not require deployment. The Jupyter Notebooks can be shared with other Data Scientists in the department.

1.3 Streaming analytics

1.3.1 Technology Choice

The data set is a finished product that does not require real-time streaming.

1.3.2 Justification

The data set is text and was generated by audience itself not in an automated production workflow allowing for streaming.

1.4 Data Integration

1.4.1 Technology Choice

No data integration applied.

1.4.2 Justification

The draw data set used for this project comes from a single csv file and was already itntegrated.

1.5 Data Repository

1.5.1 Technology Choice

The data is stored as csv file and can be loaded from the GitHub repository. It was uploaded to IBM Cloud Object Storage and connected to the Watson Studio project. From there it can be loaded as streamingbody into the notebook.

1.5.2 Justification

The data is light enough to be stored and loaded either way described above.

1.6 Discovery and Exploration

1.6.1 Technology Choice

An ETL notebook was created analyzing the four columns with Pandas data frames and MatPlotLib. A more detailed exploration was started in the main notebook with Pandas data frames and seaborn plots.

1.6.2 Justification

The data are given as continuous values in a csv file perfectly suited for Pandas data frames.

1.7 Actionable Insights

1.7.1 Technology Choice

The data required lots of preprocessing because it contains url, html tags, special symbols.

1.7.2 Justification

The SciKit-Learn library for Python is the gold standard for preprocessing and machine learning.

1.8 Applications / Data Products

1.8.1 Technology Choice

SciKit-Learn (in-memory Machine Learning)
Keras (in-memory Deep Learning)
IBM Cloud Machine Learning for deployment of a pre-trained model.

1.8.2 Justification

SciKit-Learn was run in the Watson Studio 1-CPU environment. Scaling the data increased especially the Support Vector Regression task significantly.
Keras was run in GPU mode to expedite the process. A Keras2DML wrapper and deployment to SystemML and Apache Spark servers would be a good choice for larger data sets.
The Cloud deployment makes it easy to predict values from anywhere with the API and access credentials.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

The deployed IBM Cloud model will be removed within 2 days.

1.9.2 Justification

Cleaning up the Cloud environment.