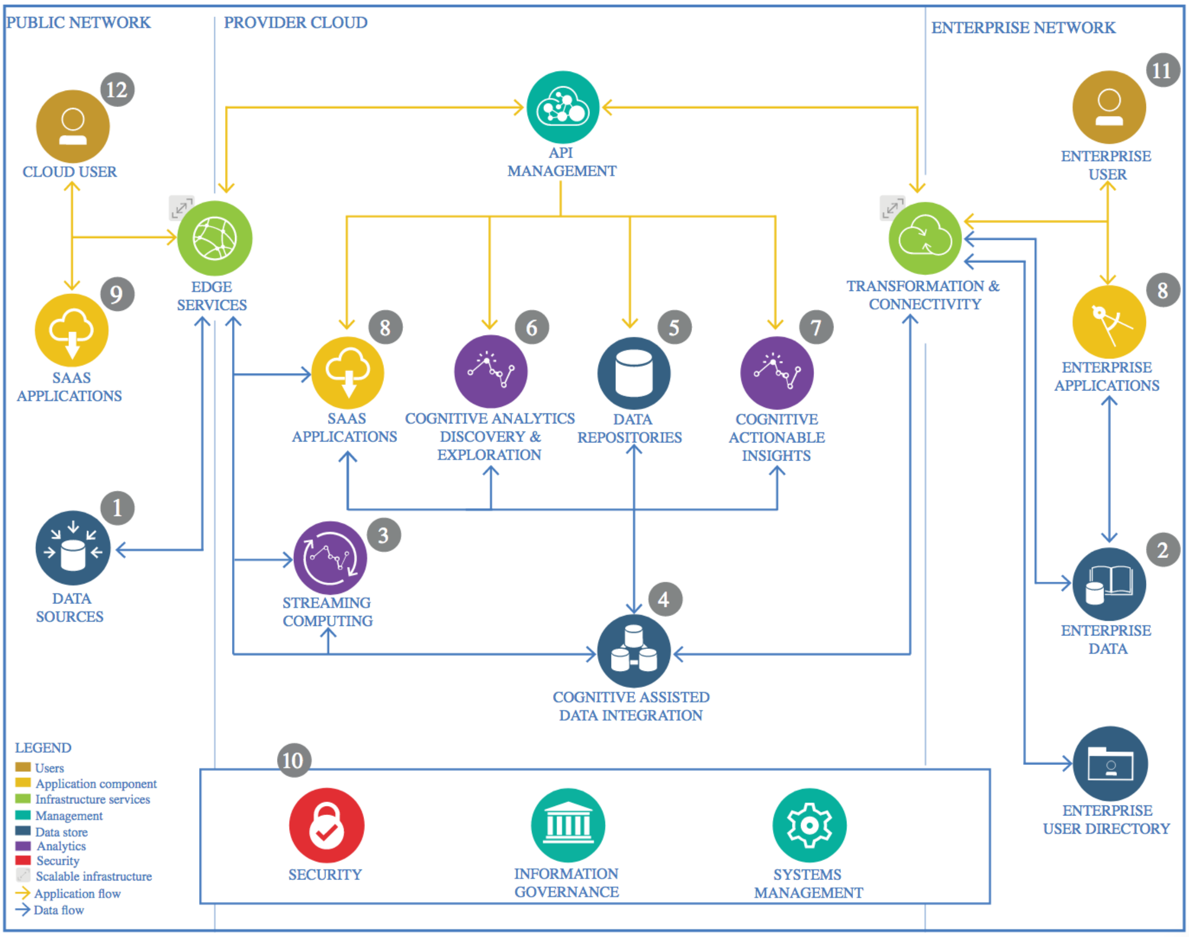
The Lightweight IBM Cloud Garage Method for Data Science

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

CSV file format is used as data source of model.

Recruit Restaurant Visitor Forecasting | Kaggle

1.1.2 Justification

In business case, insights about expected visitors count is not needed in real time. But only in scheduled resource planning meeting. As csv is universal file format, before planning meetings data can be extracted from any relational database systems in csv format.

1.2 Enterprise Data

1.2.1 Technology Choice

NA

1.2.2 Justification

Because it imparts additional costs to move data to cloud, data will be extracted from enterprise systems when needed.

1.3 Streaming analytics

1.3.1 Technology Choice

NA

1.3.2 Justification

Since the analytics need not to be real-time, no streaming analytics setup is created.

1.4 Data Integration

1.4.1 Technology Choice

python, pandas

1.4.2 Justification

Since all the process is handled in python data integration is also done in python. Pandas library which offers efficient support to handle structured data is also employed for this purpose.

1.5 Data Repository

1.5.1 Technology Choice

Object Storage: CSV

1.5.2 Justification

Because of cheap and ease, transformed data is stored in csv file.

1.6 Discovery and Exploration

1.6.1 Technology Choice

- Jupyter notebook
  - Matplotlib
  - Seaborn
  - Numpy
  - Plotly
  - Prettytable

1.6.2 Justification

Jupyter notebook is interactive, easy to use and offers best way to document code. Thus jupyter notebook is standard choice. Python libraries are offers good community support and creates professional graphs and table representation.
Apache Spark is not used because data can be handled by single machine.

1.7 Actionable Insights

1.7.1 Technology Choice

Feature Engineering: pandas, numpy, scipy, sklearn
Model Selection: sklearn, keras, tensorflow, hyperopt
Model Evaluation: numpy, sklearn

1.7.2 Justification

Feature Engineering:

- Pandas offer convenient and efficient way to create feature from different datatypes such as timestamp, strings
- Numpy, Scipy allows us to perform vast number of mathematical operations such as calculating mean, sum, log etc. without loss of speed.
- Sklearn provides different kinds of labelers, scalers to transform numeric and categorical data with better efficiency.

Model Selection:

- Sklearn offers all kinds of traditional machine learning algorithms and is easy to tweak. Gradient Boost Regressor was chosen from different available options.
- Keras, tensorflow supports building and training deep leaning seamlessly. LSTM encoder-decoder model was built to predict time series of visitors.
- Hyperopt was used to tweak the parameters of machine learning models to improve their performance

Model Evaluation:

- Root mean squared logarithmic error was calculated using numpy and sklearn library. RSMLE is robust to outliers and measures relative error. Thus, it was chosen as evaluation metric.

1.8 Applications / Data Products

1.8.1 Technology Choice

Prediction in csv format will be analysis team to build visuals and embed in report.

1.8.2 Justification

Easy to read format and can be loaded back into enterprise system or visualization softwares such as tableu, qlickview.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

NA

1.9.2 Justification

No additional process will be required as data will be loaded back into same system.