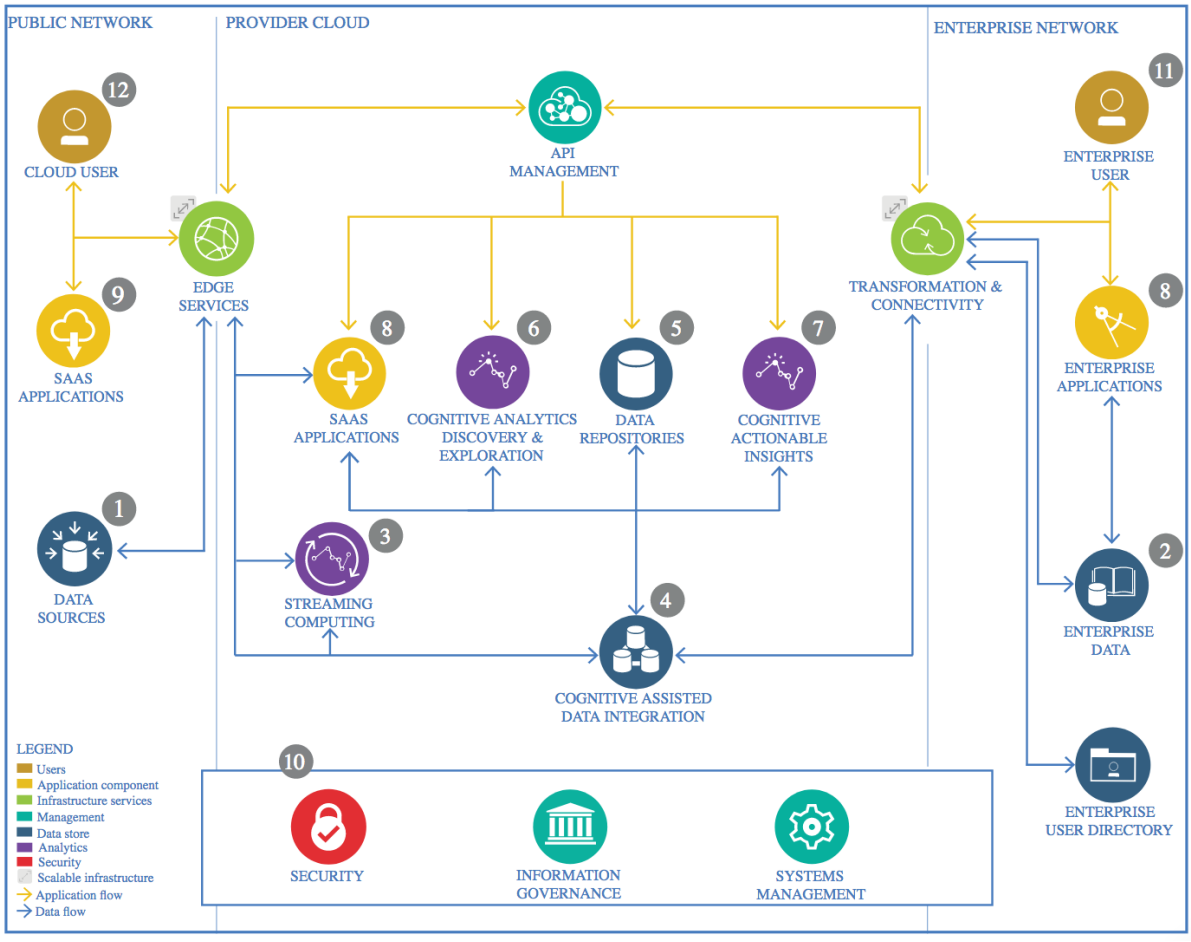
The Lightweight IBM Cloud Garage Method for Data Science

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

The source financial data is taken from Yahoo finance.

1.1.2 Justification

Yahoo finance is a reliable source of financial data.

1.2 Enterprise Data

1.2.1 Technology Choice

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

1.2.2 Justification

DOES NOT APPLY

1.3 Streaming analytics

1.3.1 Technology Choice

The model is trained at each prediction using current data. Before its usage for the network training, the data has been normalized following a z-score normalization (subtraction of the mean and division by the standard deviation).

1.3.2 Justification

Since the model predicts a time series, it is important to update the training as soon as new data is available. In this case, the data is updated every day.

1.4 Data Integration

1.4.1 Technology Choice

As said above, the data is taken from Yahoo finance only. Only feature extraction and engineering has been applied to derive the weekday at each time point.

1.4.2 Justification

The data provided by Yahoo finance was sufficient. The new feature has been engineered to make the model represent possible week seasonality.

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

DOES NOT APPLY

1.6 Discovery and Exploration

1.6.1 Technology Choice

The data discovery has been performed visually using plots, and by visualizing the pandas dataframe.

1.6.2 Justification

This simple methods were sufficient to assess the data quality given the reliability of the data source.

1.7 Actionable Insights

1.7.1 Technology Choice

1) The model has been built in TensorFlow
2) The model has been build using a LSTM neural network
3) The network consist in the LSTM layer made of 32 units, and a final Dense layer with a single unit and no activation function (the modeled quantity is continuous)
4) The standard activation function (tanh) has been used.
5) The MAE has been used as metric to assess model performance
6) The model could be further sensitized to financial news sentiment using an additional NLP model

1.7.2 Justification

1) TensorFlow is a well-known framework and the Keras layer is easy to use
2) The LSTM is a preferred algorithm for prediction of time series
3) 32 units were chosen in order to achieve a low Mean Absolute Error
4) The tanh activation function is the only one supported by cuDNN / tensorflow-gpu
5) The Mean Absolute Error is a reasonable choice to assess the prediction accuracy of time series
6) While technical analysis of stock markets is mostly based on stock close prices and volume, these does not seem sufficient for the network to derive a sufficiently reliable price model

1.8 Applications / Data Products

1.8.1 Technology Choice

This model could be used as aid to perform stock trading.

1.8.2 Justification

The ability to predict a stock price is crucial for stock trading.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

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

1.9.2 Justification

DOES NOT APPLY