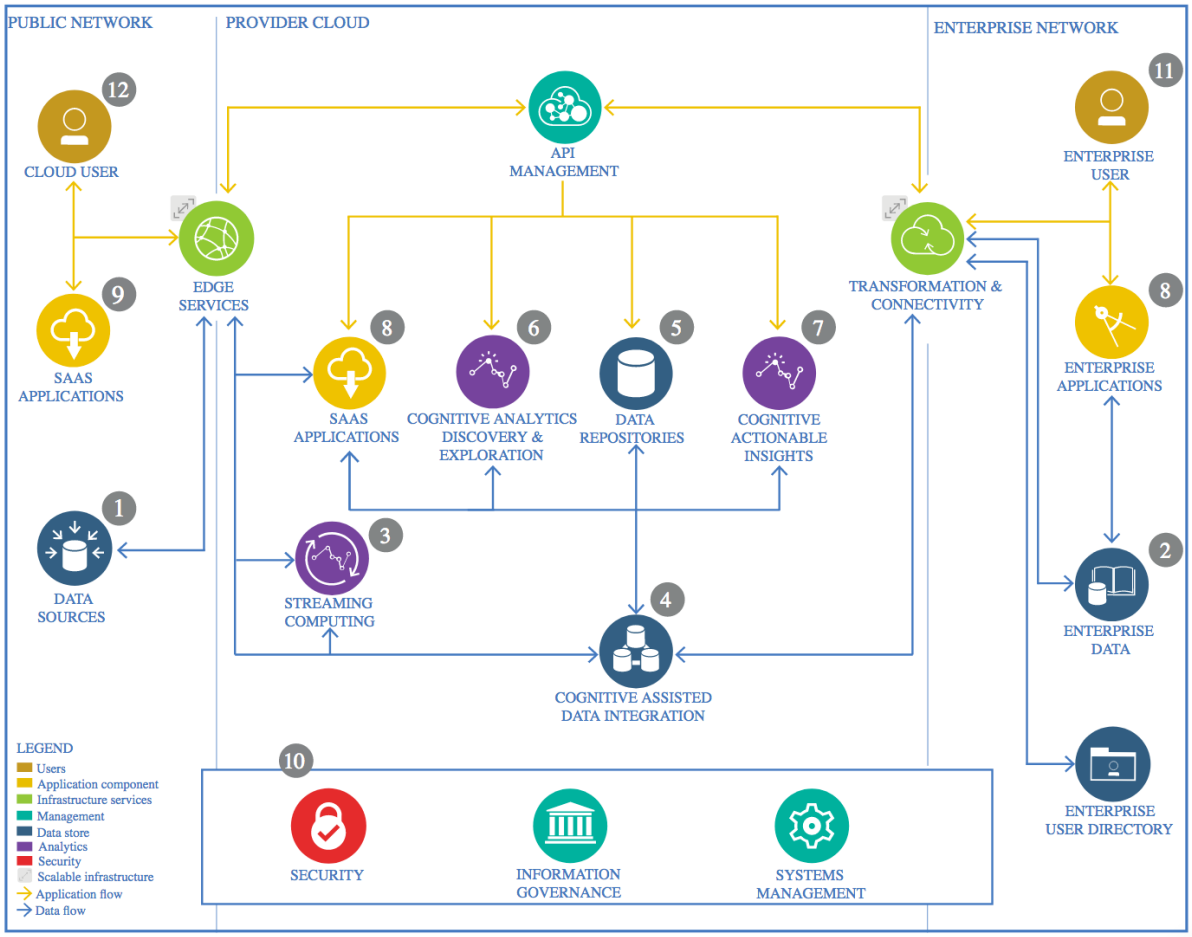
The Lightweight IBM Cloud Garage Method for Data Science

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

Data Source is a CSV extract – this extract is originally sourced from the http://www.bom.gov.au/climate/data/?ref=ftr

Sydney Airport – Station number 066037 – is used to extract historical temperature (Mean – Maximum temperature).

1.1.2 Justification

XGBoost and LSTM were chosen as the non – deep learning and deep learning algorithms.

Initial parameters were chosen randomly with the plan to change these through iterations using R2 as a guide.

Process steps were:

1. Data Extraction
2. Load
3. Exploratory Data Analysis
4. Machine learning
5. Performance measurement
6. Plot for Stakeholder (with predicted vs actual values)

Single node system used for training due to small dataset size.

XGBoost was chosen as the algorithm can handle seasonality and trends in time-series data. It is an ensemble learning method that combines the predictions of multiple weak models (decision trees) to create a strong predictive model.

An LSTM (Long Short-Term Memory) is a recurrent neural network (RNN) variant commonly used in time series prediction. LSTMs are designed to handle data sequences with long-term dependencies, making them suitable for time series prediction [1].

LSTM – initially models were trained on LSTM layer = 50, EPOCHS = 50. This was gradually increased to 500 – showing the best performance (based on R2 scores). Increasing Lags also helped – although lags of 12 improved performance while lower values didn't make much difference.

XGBOOST -

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

Please justify your technology choices here.

1.3 Streaming analytics

1.3.1 Technology Choice
Not applicable

1.3.2 Justification
Not applicable

1.4 Data Integration

1.4.1 Technology Choice
TBC

1.4.2 Justification
TBC

1.5 Data Repository

1.5.1 Technology Choice
Data Repository is Object Storage to ensure open data access.

1.5.2 Justification
This storage chosen as data size is minimal, no need for security (no confidential information)

1.6 Discovery and Exploration

1.6.1 Technology Choice
Juptyer notebook for Data Discovery and Exploration.

1.6.2 Justification
Minimal data – only local computer needed to process the learning algorithms.

1.7 Actionable Insights

1.7.1 Technology Choice
N/A

1.7.2 Justification
N/A

1.8 Applications / Data Products

1.8.1 Technology Choice

Data Products are just a timeseries forecast (plot) of the temperature.

1.8.2 Justification

N/A

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

Publicly available dataset is used – hence no need for security.

1.9.2 Justification

As above.

References

[1]: https://medium.com/@matthew1992/time-series-with-lstm-090cb8d16a59