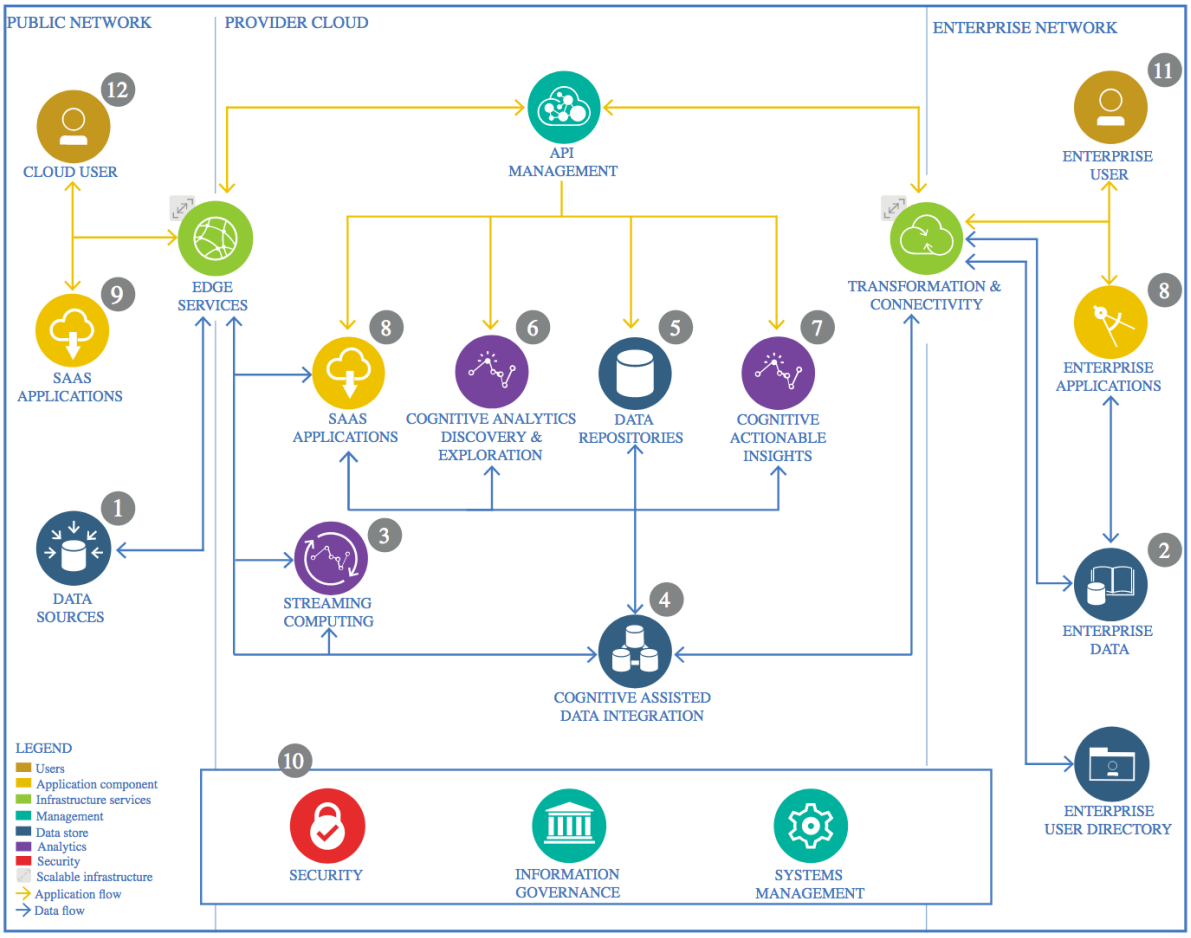
The Lightweight IBM Cloud Garage Method for Data Science

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

To gather the required data for the project, I chose a CSV file containing the Visits of the different websites from the SaaS Vendor INFOnline.

1.1.2 Justification

The reason for using the CSV file is that parts of the historical data has been stored only on an Oracle database and has not been migrated to Foundry in our company, yet. Moreover, the API from the Vendor does only provide data from later stage, but to feed the neural network, I learnt that the more data available the better.

1.2 Enterprise Data

1.2.1 Technology Choice

Our data is mainly stored in Foundry.

1.2.2 Justification

Since we are working with Palantir's Foundry, our data is already cloud-based stored and easily accessible and available there. It's also working with most of the languages available for Machine Learning.

1.3 Streaming analytics

1.3.1 Technology Choice

The data needs only been update on a monthly basis.

1.3.2 Justification

Due to the fact that the prediction is required only on a monthly basis, streaming analytics is not necessary at that point and due to the small size of the data, Apache Spark or similar technologies are not required to store, access or do cluster computing on it.

1.4 Data Integration

1.4.1 Technology Choice

To collect the future data, the process is as followed:
The extraction of the data happens in Snowflake where the data is collected via the vendor's API. This happens on an hourly basis on a very detailed level provided as xml files for the various websites. The second step is to aggregate the data to get daily numbers. This data is then mirrored to Foundry to work with it.

1.4.2 Justification

To work with the data process this way was chosen in the company two years ago where I had no influence on.

1.5 Data Repository

1.5.1 Technology Choice

In Palantir's Foundry, data is usually stored in XML file format and uses Amazon Redshift to store data. All sorts of data types are available to read from or to store in Foundry.

1.5.2 Justification

Foundry provides different storage options such as PostgreSQL, Oracle, HDFS and NoSQL. Since our data set is pretty small and will be in the future in comparison to other Deep Learning examples such as image recognition, even if taking more websites into account, the storage cost and amount will still be very little with around 2000-3000 data points in the time series and around 40 platforms (websites).

1.6 Discovery and Exploration

1.6.1 Technology Choice

Visualization: For the data visualization, I chose matplotlib|pyplot as well as bokeh.
Computer Language: Jupyter/Python
Python Modules used: scikit-learn, pandas, numpy

1.6.2 Justification

Visualisation: The reason for this was that I needed to drill down in the data on a weekday or even daily basis to figure out if there was any sign of a seasonal component to the data.
Computer Language: It is most widely used in the Data Science field and I already had some knowledge on it.
Python Modules used: I chose them because most of the libraries are very helpful for data analysis and model training and evaluation

1.7 Actionable Insights

1.7.1 Technology Choice

Feature Engineering: Aggregating the data from day to weekday, week and month for more insights where part of the feature engineering task.
Data quality assessment: using bokeh and matplotlib for visualizing "uncertainties"

Data Science Model: SARIMAX
Deep Learning Model: LSTM (tensorflow with keras on top of it)
Model performance indicator: R-square, Mean Squared Error

1.7.2 Justification

Feature Engineering: By aggregating the data on a weekday basis, I found out that the data had a seasonal component (on a weekly basis).
Data quality assessment: no adjustments in the data needed to be made. The csv provided by the company did not have any missing values nor outliers related to measuring/tracking issues, when analyzing the data. The dip in the data in Jun 2018 was due to a technical change within our system, where mobile users were not necessarily redirected to the desktop version of the website anymore. All values were in range of one to two standard deviations and the outliers could be explained by the News or Holidays.
Data Science Model: Taking the seasonal component in the data into consideration to decide which model to choose from, I decided for the SARIMA(X) instead of the classical ARIMA model.
Deep Learning Model: I chose the tensorflow framework in combination with keras, since I needed for my time series predicition a neural network that does not forget but has a memory function. Therefore, using an LSTM (an recurrent neural network (RNN)) seemed to be the best fit, since there is usually a strong correlation between past and future values in time series and the predicition isn't too long where RNNs/LSTMs tend to fail.
Model performance indicator: First, I started of with the Mean Squared Error mainly because bigger outliers have a higher impact on the score, which in turn is good for evaluating the model's performance especially for news/publisher websites, who usually struggle with high volatility in the data due to special news events (corona pandemic etc). But when looking at the numbers, I received values that were far too high. Therefore, a performance measurement based on the numbers weren't possible. For that reason, I took the R-squared indicator as a performance measure. Not the best of indicators for a times series, but enough to evaluate the relation between two series (train and test).

1.8 Applications / Data Products

1.8.1 Technology Choice

Providing the predicted values in Foundry, in a format to work with it, and to visualize the data using Looker Reports.

1.8.2 Justification

Making the traffic forecasts easier and more reliable with less manual work involved (copy and pasting numbers and formulas in excel spreadsheets) by providing the data in Foundry in a format to work with it. This could mean taking the predicted values from the forecast or discussing the predicted values and enrich the forecast with the domain knowledge of the people responsible for the numbers.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

No restrictions required to who should see the data within the company.

1.9.2 Justification

No sensible data and the more people know about it the more they discuss it and provide valuable insights for future predictions and forecasts that can be taken into consideration to form better models.