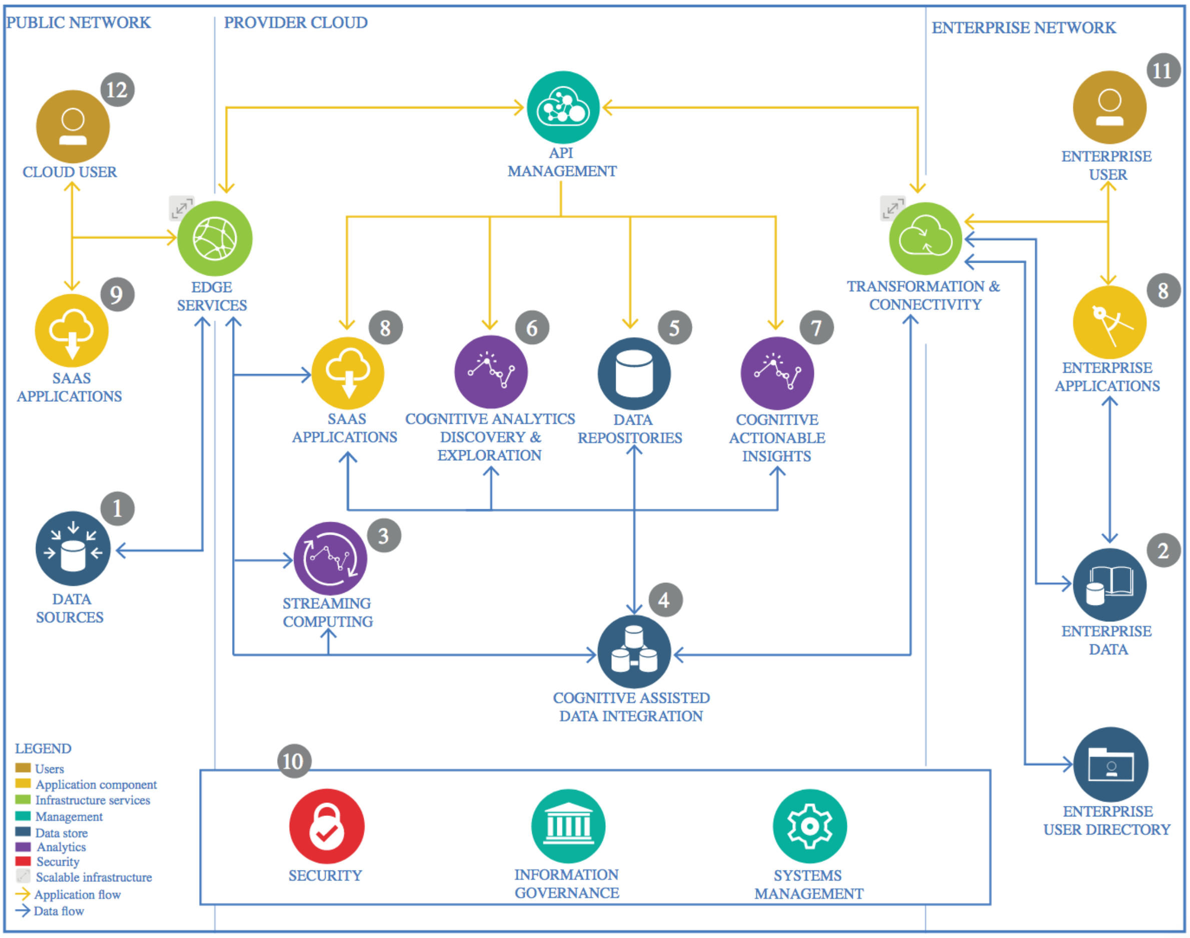
The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document for the IBM Advanced Data Science Capstone Project: **Predicting future electricity stock market prices**

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

1.1.1 Technology Choice

All data we need we find at the web portal SMARD from the Germanys Federal Network Agency (https://www.smard.de). In regard to our business problem, we need following data for our analysis: electricity stock market prices, electricity demand and production. We can download these data for the period from 2016 until now via the link below in csv format. Notice that the demand and production have a resolution of 15 minutes and the prices of 1 hour.

(https://www.smard.de/home/downloadcenter/download_marktdaten/726#!?downloadAttributes=%7B%22selectedCategory%22:5,%22selectedSubCategory%22:17,%22selectedRegion%22:%22DE%22,%22from%22:1514761200000,%22to%22:1546297199999,%22selectedFileType%22:%22CSV%22%7D)

1.1.2 Justification

Cause the data source is from a federal agency we can assume a high quality. The data volume is not so big that it is not manageable as csv.

1.2 Enterprise Data

1.2.1 Technology Choice

All data that is used is public. No in-house or enterprise data were used for this project.

1.3 Data Integration

1.3.1 Technology Choice

Google Drive is used to store the csv files. The data is persistent and not be streamed.

1.3.2 Justification

All computations and modeling are done in Jupyter notebooks within Google Colab. Cause Google Drive is easy reachable from Google Colab it is reasonable to use it.

1.4 Data Repository

1.4.1 Technology Choice

Python pandas DataFrames are used as in-memory store of the data.

1.4.2 Justification

File sizes aren’t that big, that one can’t manage it with pandas fast and easily. Later on, if one use a larger dataset, Apache Spark can be used and the code can be rewritten to operate on RDD instead of pandas DataFrame.

1.5 Discovery and Exploration

1.5.1 Technology Choice

We use Python and the following libraries: Pandas, Numpy. To determine the quality assessment we just use the methods info() and describe(). Furthermore we visualize some lines of the dataset.

1.5.2 Justification

Python, especially in combination with pandas and numpy is great for quick, effortless and effective data explorations and easy to use within Jupyter notebooks. Cause the data is from an federal agency we assume that the quality is perfect. That is why a deeper look than described is fortunately not necessary.

1.6 Actionable Insights

1.6.1 Technology Choice

We use Matplotlib and Seaborn to get some insights from the data.
For feature engineering I use some of my domain knowledge and the correlation matrix to calculate and combine certain features.

1.6.2 Justification

Both libraries are easy to use and very powerful. It is very reasonable to combine features that are similar to each other.

1.7 Applications / Data Products

1.7.1 Technology Choice

We use Google Colab with Python, Sklearn and Keras.
I use the multiple linear/polynomial regression for the machine learning part and a stateful LSTM neural network for the deep learning part.
As model performance indicator we choose the Root Mean Square.

1.7.2 Justification

Google Colab is useable for free and for an unlimited time. Additionally, you can choose a GPU and TPU for calculating neural networks impressive fast. Sklearn is an easy useable machine learning library and compatible with the other libaries. Keras library is flexible and extensible. Cause it is very high level it is appropriate for explaining the model in more detail to the stakeholders.
Since we want to solve an regression problem with many features, the decision is clear to use an linear or polynomial regression model. Cause

we have a timeseries, it make sense to use some LSTM layers in our neural network.
We use the RMSE performance indicator, because we want to punish outliers in the prediction. That is why outliers mean big economic losses.

1.8 Security, Information Governance and Systems Management

1.8.1 Technology Choice

We don’t use any sensitive data.