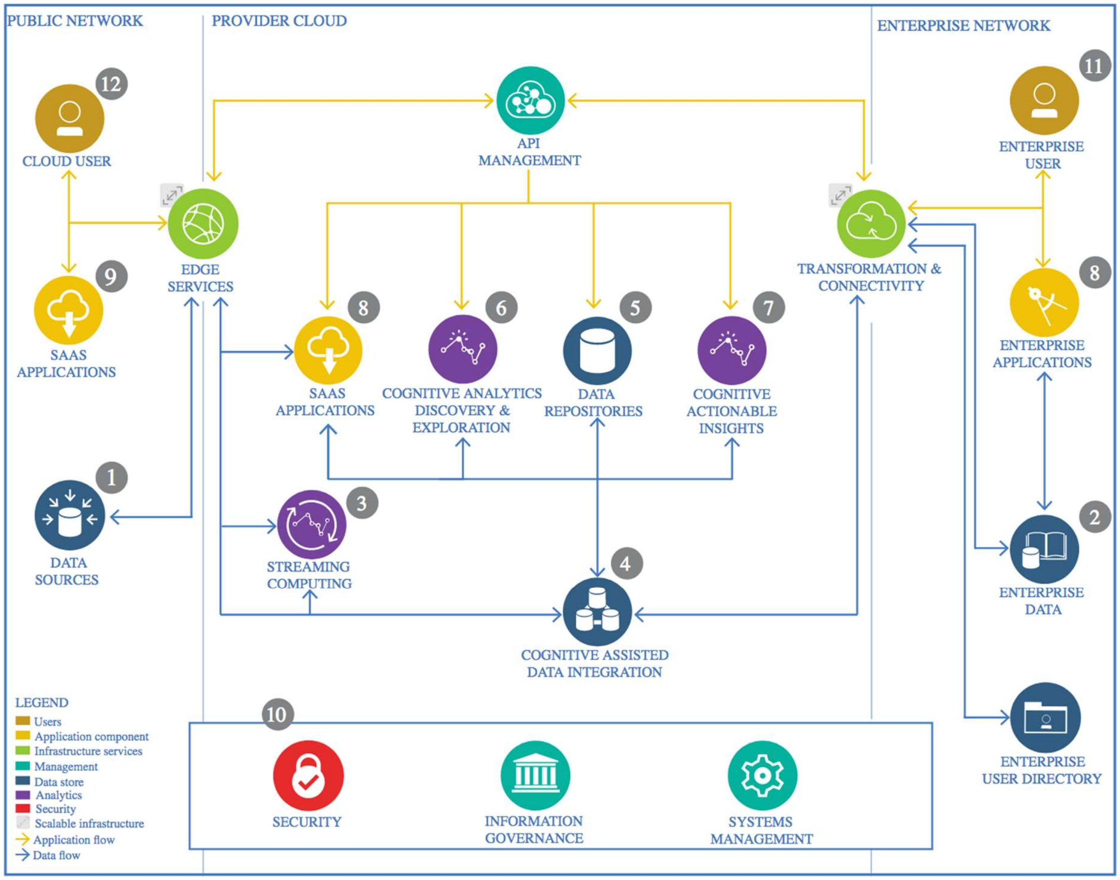
The Lightweight IBM Cloud Garage Method for Data Science

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

1.1.1 Technology Choice

An external data source (open Case) in csv format.

1.1.2 Justification

The csv format is easy to convert into a data frame to identity all the features. Then apply all the operations according to requirements.

1.2 Enterprise Data

Cloud based solutions tend to extend the enterprise data model. Therefore, it might be necessary to continuously transfer subsets of enterprise data to the cloud or access those in real-time through a VPN API gateway.

1.2.1 Technology Choice

IBM Watson, Anaconda & Kaggle

1.2.2 Justification

Availability of jupyter notebooks, easily work on pandas and apache spark, and storage data facility.

1.3 Streaming analytics

1.3.1 Technology Choice

Apache spark.

1.3.2 Justification

No memory problem. Run on parallel clusters.

1.4 Data Integration

Extract Transform Load is applied on each csv file to make one data frame. Then perform Data Cleansing, feature engineering.

1.4.1 Technology Choice

Apache Spark, Pandas.

1.4.2 Justification

Pre-processing of the data can easily be done on the above-mentioned technology.

1.5 Data Repository

1.5.1 Technology Choice

GitHub

1.5.2 Justification

It is a reliable storage place where you can save your data permanently.

1.6 Discovery and Exploration

Energy Consumption in Netherland. Both Electricity and Gas. The main features are delivery percent of energy, smart meters, low tariff power consumption, active connection and annual consumption, the target variable. The data is comprised of ten years.

1.6.1 Technology Choice

IBM and Kaggle Cloud. Sub technology involves Jupyter Notebook, Python, scikit learn, pandas, matplotlib.

1.6.2 Justification

To analyze the data for model developing, some features are visualized through Histogram (Smart Meters), bar chart (Smart Meters Spreading over the time), and scatter (annual consumption prediction) and line plot (solar energy trend). First, I used IBM cloud but due to limit problem, I shifted to Kaggle cloud where I worked without any concerns.

1.7 Actionable Insights

Predicting Annual Consumption of Energy in the Netherland.

1.7.1 Technology Choice

Kaggle Cloud. Sub tools are Python, SciPy, nonlinear regression analysis.

1.7.2 Justification

First analyzed the annual consumption data over the ten years' time. Visualize the data through scatter plot. After visualizing, it is obvious that the trend is nonlinear (Logarithmic) and the Machine Learning Regression technique is useful. Finally, the machine learning algorithm is evaluated through coefficient of determination (R2-score), mean absolute error (MAE) and mean squared error (MSE).

1.8 Applications / Data Products

The designed model can be used for many future purposes.

1.8.1 Technology Choice

D3

1.8.2 Justification

The energy providing companies can use this model to make an application which assist the companies to plan according to market energy utilization. This model be a part of an application not the whole.

1.9 Security, Information Governance and Systems Management

Data Privacy

1.9.1 Technology Choice

IBM & Kaggle Cloud.

1.9.2 Justification

Data privacy is very important. In Business, the way you designed the model or your plan is the only difference between you and your competitors.