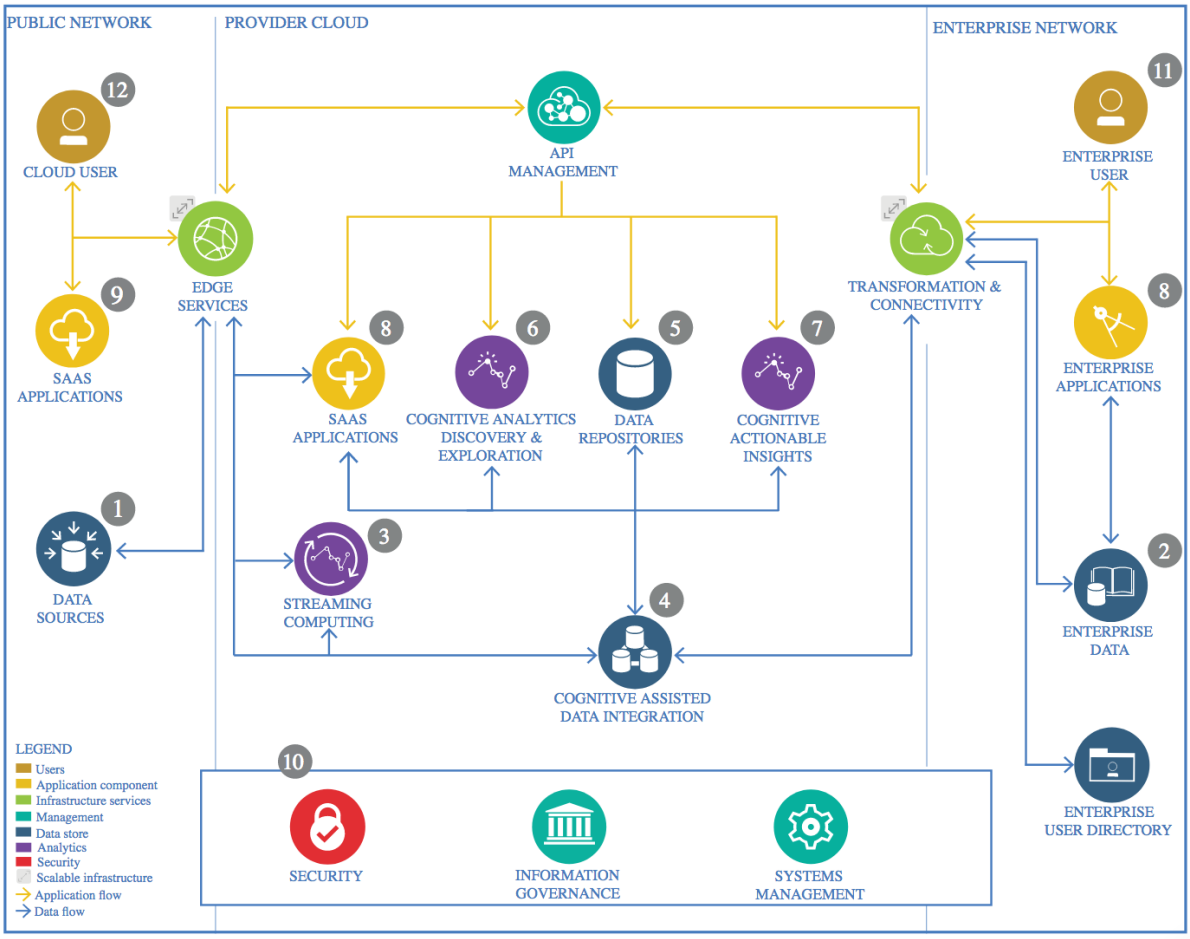
The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document for project 'NYC Taxi trip Duration Prediction'

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

1.1.1 Technology Choice

For the data acquisition part, we use the following data sources:
Either a subset of the Sample at :
https://www.kaggle.com/c/nyc-taxi-trip-duration/data

Or the entire dataset at:
https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_

1.1.2 Justification

Several other sources have been investigated, but they lacked good API or broad span over the range of data

1.2 Enterprise Data

1.2.1 Technology Choice and Data Quality Assessment:

The data being used in the project is public. No in-house or enterprise sources were used.

Data Quality Assessment and its Justifications:

1.The ratio of data to errors
It allows you to track how the number of known errors – such as missing, incomplete or redundant entries – within a data set corresponds to the size of the data set. If you find fewer errors while the size of your data stays the same or grows, you know that your data quality is improving.

2. Number of empty values
Empty values – which usually indicate that information was missing or recorded in the wrong field — within a data set are an easy way to track this type of data quality problem.

3.Data storage costs
If you are storing data without using it, it could be because the data has quality problems. If, conversely, your storage costs decline while your data operations stay the same or grow, you're likely improving the data quality front.

4.4. Data time-to-value
Calculating how long it takes your team to derive results from a given data set is another way to measure data quality. While a number of factors (such as how automated your data transformation tools are) affect data time-to-value, data quality problems are one common hiccup that slows efforts to derive valuable information from data.

1.3 Data Integration

1.3.1 Technology Choice

IBM Object Storage is used to store the scraped data from the raw sources in CSV format. The data is persistent and not being streamed.

1.3.2 Justification

All computations and modeling are done using Jupyter notebooks within IBM Cloud, therefore IBM cloud is a good fit to store unstructured data, because it can be accessed

from there. Files aren't that big, and shall they grow in size, they can be naturally split 'horizontally' .
The information in files has to be prepared and normalized first (for instance, names of the districts should be the same), so there is no point in, say, inserting this raw data into an RDBMS.

1.4 Data Repository

1.4.1 Technology Choice

Python pandas Data Frames are being used as in-memory store of the data. No Database is being used.

1.4.2 Justification

Later on, if the dataset becomes larger, apache spark can be used and the code rewritten to operate on RDD instead of pandas Data Frames.

1.5 Discovery and Exploration

1.5.1 Technology Choice

We use Python, pandas, matplotlib, seaborn and Folium to visualize all the data we have gathered and understand them better. We use it to create new features: proximity to parks score, affordability score, and criminality level score.

1.5.2 Justification

Python is great for data analysis and quick ideas prototyping. There are many powerful third-party libraries for statistical analysis which are free of charge. There are also many Python programmers available on the market.

1.6 Actionable Insights

1.6.1 Technology Choice

Looking deeper at the data ,we see that the dataset is made publicly available over the range of 10 years. So either we use the full 300 gb dataset containing millions of records using IBM cloud or only select a subset of data to implement a small model in the Local system. Either way, we can implement any one of them efficiently by the Technology available.

1.7 Applications / Data Products

1.7.1 Technology Choice

We use Random Forest Regressor as a machine learning model to solve this problem.
We use Keras to build models for the prediction of trip duration i

1.7.2 Justification

Random Forest Regressor , an ensemble model is versatile, dynamic and also the
Keras library is flexible, extensible and allows for model serialization. It is also high-level enough to explain the details of our model to the stakeholders.

1.8 Model Architecture

1.8.1 Technology Choice

The final model architecture was chosen to be a Feed Forward neural network with 17 input gates in the first layer and 10 epochs for training. The selected architecture has shown very good accuracy on the test data in persistent manner.

1.9 Model Training

1.9.1 Technology Choice

The data set was split into 70% for training and 30% used for testing purposes.

1.10 Model Deployment

1.10.1 Technology Choice

Random Forest Regressor is more than sufficient to implement the model over a small scale with great prediction capacity.
Keras models can be serialized and deployed later on different cloud solutions.
Google Cloud Functions was eventually chosen as deployment environment.

1.10.2 Justification

The cost factor was the main driving factor for this decision.

1.11 Security, Information Governance and Systems Management

1.11.1 Technology Choice

Once we integrate user profiles with their preferences, such sensitive information has to be protected.

1.11.2 Justification

Due to GDPR, we will strip away all relevant customer information and anonymize their preferences and choices.