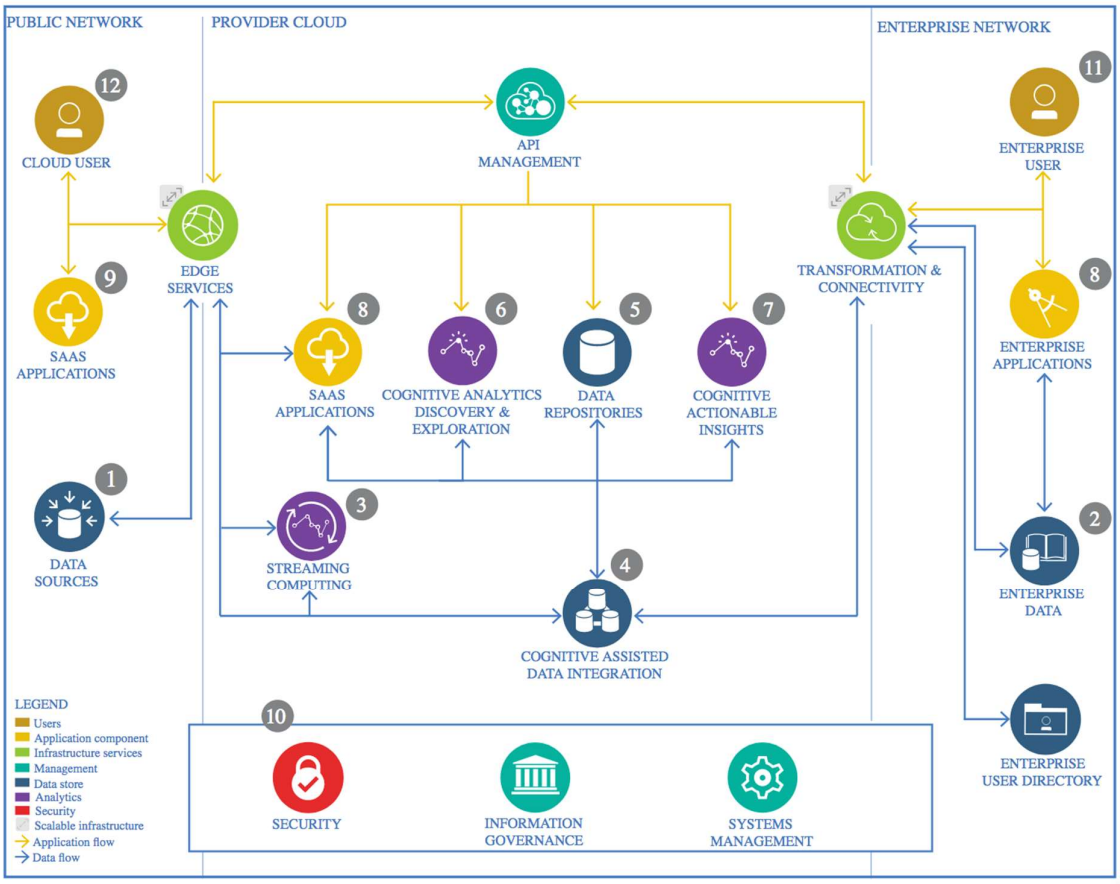
The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document Template for 'Prediction of accessible and affordable Airbnb prices in Singapore'

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

1.1.1 Technology Choice

The dataset is obtained directly from the Airbnb website and updated as of 27 Feb 2020.
http://data.insideairbnb.com/singapore/sg/singapore/2020-02-27/visualisations/listings.csv

The aim of this case study to identify the factors that affect the price affordability of renting a Airbnb property such as location, room type and number of reviews. There will be also be

additional features added to the current dataset such as distance to the nearest train stations as well as distance to the city area or Central Business District (CBD). Accessibility in this project refers to the distance between a particular Airbnb property and the nearest train station.

1.1.2 Justification

I have considered venturing into Airbnb in other countries in remote areas in United States and Europe. However, I chose Singapore due to my familiarity of the location and regions and the data set is constantly updated on their website. The objective of this project is to predict the prices of these Airbnb and identify the most affordable and accessible property to rent.

1.2 Enterprise Data

1.2.1 Technology Choice

The dataset used for this project is available for public access. Hence, there was no in-house or enterprise data being sourced.

1.3 Data Integration

1.3.1 Technology Choice

The data sources are stored in my GitHub repository and imported into Jupyter Notebook. All the files are in CSV format hence there is no cloud storage required as the dataset is quite small

1.3.2 Justification

All the analysis and modeling are done on Jupyter notebook as well as IBM Watsons Studio where the use of Spark is demonstrated for some parts of the analysis.

1.4 Data Repository

1.4.1 Technology Choice

Spark and Pandas data frames are used in importing and accessing the datasets. Hence, there are no database being utilized for this project. The data is stored inside the Github repository and imported into Jupyter notebook.

1.4.2 Justification

The dataset is collected and analyzed based on latest monthly update on the Airbnb website.

1.5 Discovery and Exploration

1.5.1 Technology Choice

We explore the data using the essential Python libraries such as Pandas, NumPy, Matplotlib and Seaborn to obtain the descriptive statistics and identify any missing values. As the dataset consist of geospatial information i.e. the latitude and longitude of the Airbnb, we also used GeoPandas and Folium to derive a choropleth map of the locations of these Airbnb services. Furthermore, we can perform text analytics and develop a word cloud on the names of each Airbnb property to identify the common words used by the hosts in publicizing their services.

1.5.2 Justification

These are the libraries commonly used in Python for data manipulation and visualization during the exploratory data analysis phase. In addition, the processing time and execution is fast and efficient.

1.6 Actionable Insights

1.6.1 Technology Choice

The dataset illustrates the prices of the Airbnb properties in Singapore. We can identify that properties in the Central region are more expensive as compared to the other regions. This is simply due to its ease of accessibility to the city area and the CBD (Central Business District).

1.6.2 Justification

The use of data visualization allows important insights to be gathered thus influencing the outcome of the prediction model. By analyzing the price range of different neighborhoods and room types, we can able to identify the properties which are in demand for rental and short-stays. This is suitable for expatriates to settle while being work-based in Singapore.

1.7 Preprocessing and Feature Engineering

1.7.1 Technology Choice

For this dataset, I have identified plenty of outliers from the variables. Hence, one of the preprocessing done was to remove the outliers for the geolocation variables i.e. latitude and longitude. Secondary, I have to filter out the values under the “minimum nights” columns which have over 365 days.

I have also added a new variable to the existing dataset to calculate the distance (in km) of each Airbnb property to the nearest train station. In Singapore context, it is referred to as the MRT (Mass Rapid Transit) and LRT (Light Rapid Transit)

Finally, I applied the using of Label Encoding using sci-kit learn preprocessing model to encode the categorical variables such as

1.7.2 Justification

1.8 Model Architecture

1.8.1 Technology Choice

Using the Python sci-kit learn libraries, I created four regression models to predict the prices.

1.8.2 Justification

As the predictor variable consists of continuous values, the development of the model place a greater emphasis on flexibility rather than interpretability. Using a conventional linear regression model is not suitable for this dataset as the variables are weakly correlated with one another.

1.9 Model Training

1.9.1 Application of regressors and learning algorithm

When training a model, the data is split into a training and a test set. The training set is used to train the model whereas the test set is used to assess the ability of the system to generalize (Ciaburro, 2019).

The model is trained using cross-validation of 5-folds and 5 splits. As most of the variables are weakly correlated with each other, multi-linear regression was not used. Instead the following models were trained:

- Kernel SVR (Support Vector Regression)
- Random Forest
- Ridge Regression
- Extreme Gradient Boosting (XG Boost)
- Light Gradient Boosting (Light GBM)

Kernel SVR is a form of supervised learning algorithm derived from support vector machines.

Ridge regression is a method of L_2 norm regularization by which size of the coefficient of each independent variable are given a shrinkage penalty. It estimates the coefficient by using the Residual Sum of Square and then adding the penalties.

Cross-validation of each model is also performed in predicting the error rate of each learning algorithm. Using 5-fold is used obtain an optimal error estimate for each regressor. One key advantage of using a cross-validation is to help in minimizing the risk of selecting points near to the hyperplane.

Application of deep learning technique

An artificial neural network was also trained on the model using 2 hidden layers and applying the "ReLU" activation for on each layer. The network architecture is defined as below:

- Input Layer: 10 (No of Columns)
- Hidden Layer 1: 128 Neurons, activation ReLU
- Hidden Layer 2: 64 Neurons, activation ReLU
- Output layer: 1 Neuron, activation ReLU

1.9.2 Model Improvements

We can improve the model performance by using decreasing the learning rate of the boosting algorithms as well as increase the maximum depth of training data learned. We can also re-train the model using lesser independent variables based on the feature importance.

1.10 Model Evaluation

As this is a mixed regression model, the metrics that were used to evaluate the performance of the model is based on the following scores:

- Root Mean Squared Error

Root Mean Squared Error (RMSE) is calculated based on the validation data and similar to the standard error of estimate in linear regression.

- Mean Absolute Error

The mean absolute error (MAE) obtains the magnitude of the average absolute error.

- R^2 score / Adjusted R^2 score

The R^2 score also known as the coefficient of determination measures the amount of variance of the predictive model.