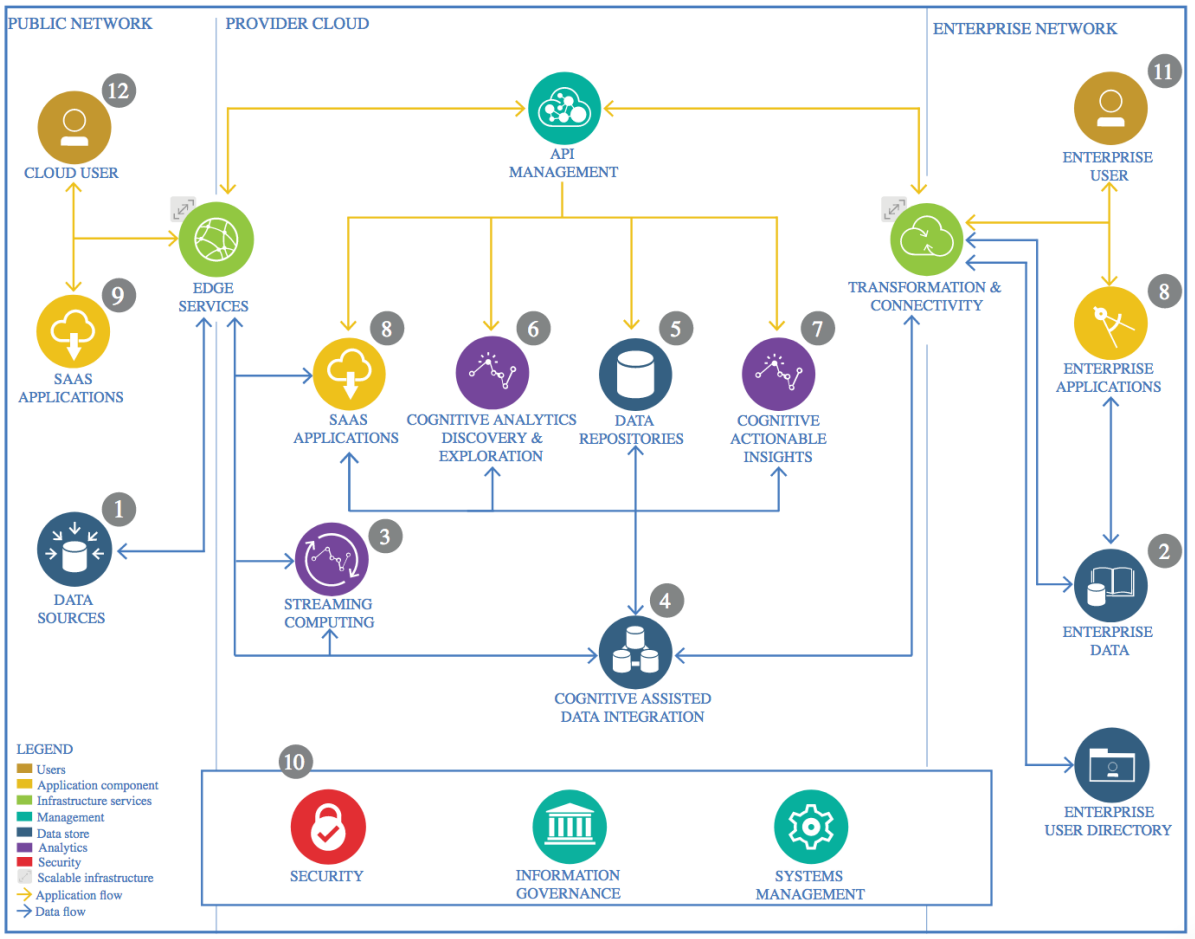
IBM Coursera Advanced Data Science Capstone – Prediction of House prices

Architectural Decisions Document

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

Table of Contents

1 Architectural Components Overview 1
1.1 Data Source 3
1.1.1 Technology Choice *4*
1.1.2 Justification *4*
1.2 Enterprise Data 4
1.3 Streaming analytics 4
1.4 Data Integration 4
1.4.1 Technology Choice *4*
1.4.2 Justification *4*
1.5 Data Repository 4
1.5.1 Technology Choice *4*
1.5.2 Justification *4*
1.6 Discovery and Exploration 5
1.6.1 Technology Choice *5*
1.6.2 Justification *5*
1.7 Actionable Insights 5
1.7.1 Technology Choice *5*
1.7.2 Justification *5*
1.8 Applications / Data Products 5
1.8.1 Technology Choice *5*
1.8.2 Justification *5*
1.9 Security, Information Governance and Systems Management 6

1.1 Data Source

The dataset used in this notebook is an open dataset that can be found on the next link:
https://www.kaggle.com/achyutanandaparida/dataset%20from%20%20house%20sales%20in%20king%20county,%20usa/notebooks

This dataset contains house sale prices for King County. It includes homes sold between May 2014 and May 2015.

id : A notation for a house

date: Date house was sold

price: Price is prediction target

bedrooms: Number of bedrooms

bathrooms: Number of bathrooms

sqft_living: Square footage of the home

sqft_lot: Square footage of the lot

floors :Total floors (levels) in house

waterfront :House which has a view to a waterfront

view: Has been viewed

condition :How good the condition is overall

grade: overall grade given to the housing unit, based on King County grading system

sqft_above : Square footage of house apart from basement

sqft_basement: Square footage of the basement

yr_built : Built Year

yr_renovated : Year when house was renovated

zipcode: Zip code

lat: Latitude coordinate

long: Longitude coordinate

sqft_living15 : Living room area in 2015(implies-- some renovations) This might or might not have affected the lotsize area

sqft_lot15 : LotSize area in 2015(implies-- some renovations)

1.1.1 Technology Choice

I use an open external data source in CSV format and pandas to import.

1.1.2 Justification

The CSV format separates each of the values with commas, making it very easy to import using pandas.

1.2 Enterprise Data

Not Applicable.

1.3 Streaming analytics

Not Applicable.

1.4 Data Integration

1.4.1 Technology Choice

Jupyter notebook in IBM Cloud Pak for Data.

1.4.2 Justification

IBM Cloud Pak for Data is an environment that brings together everything a data scientist needs for developing projects, including open source tools.

1.5 Data Repository

1.5.1 Technology Choice

IBM Cloud Object Storage

1.5.2 Justification

IBM Cloud Object Storage provides a free plan and it is easy to integrate into projects.

1.6 Discovery and Exploration

1.6.1 Technology Choice
Jupyter Notebooks, numpy, pandas, matplotlib.

1.6.2 Justification
Jupyter Notebook is an open source web application which allows us to create and share documents containing either live code or visualizations, narrative texts etc.
Pandas offers data structure and tools for effective data manipulation and analysis. It provides facts, access to structured data. The primary instrument of Pandas is the two dimensional table consisting of column and row labels, which are called a data frame. It is designed to provided easy indexing functionality.
The NumPy library uses arrays for its inputs and outputs and fast array processing can be done with this using minor coding.
The Matplotlib package is the most well known library for data visualization. It is great for making graphs and plots. The graphs are also highly customizable.

1.7 Actionable Insights

1.7.1 Technology Choice
Seaborn.

1.7.2 Justification
Seaborn is a high level visualization library. It is based on Matplotlib. It's very easy to generate various plots including regression plots and boxplot used for my project, to have insights regarding the correlation between different features and the target feature.

1.8 Applications / Data Products

1.8.1 Technology Choice
The chosen model for this project was Multiple Linear Regression with Ridge regularization. I am using sklearn for model evaluation using train-test split. Matplotlib for visualization.

1.8.2 Justification
Multiple linear regression is used to explain the relationship between one continuous target y variable and two or more predictor x variables. In my case, I am using several features to predict one single value, which is the price of houses.
Ridge regression prevents overfitting.

Separating data into training and testing sets is an important part of model evaluation.
We use the test data to get an idea how our model will perform in the real world.
Matplotlib helped with visualization of the distribution of predicted values versus training and test data.

1.9 Security, Information Governance and Systems Management

Not Applicable.