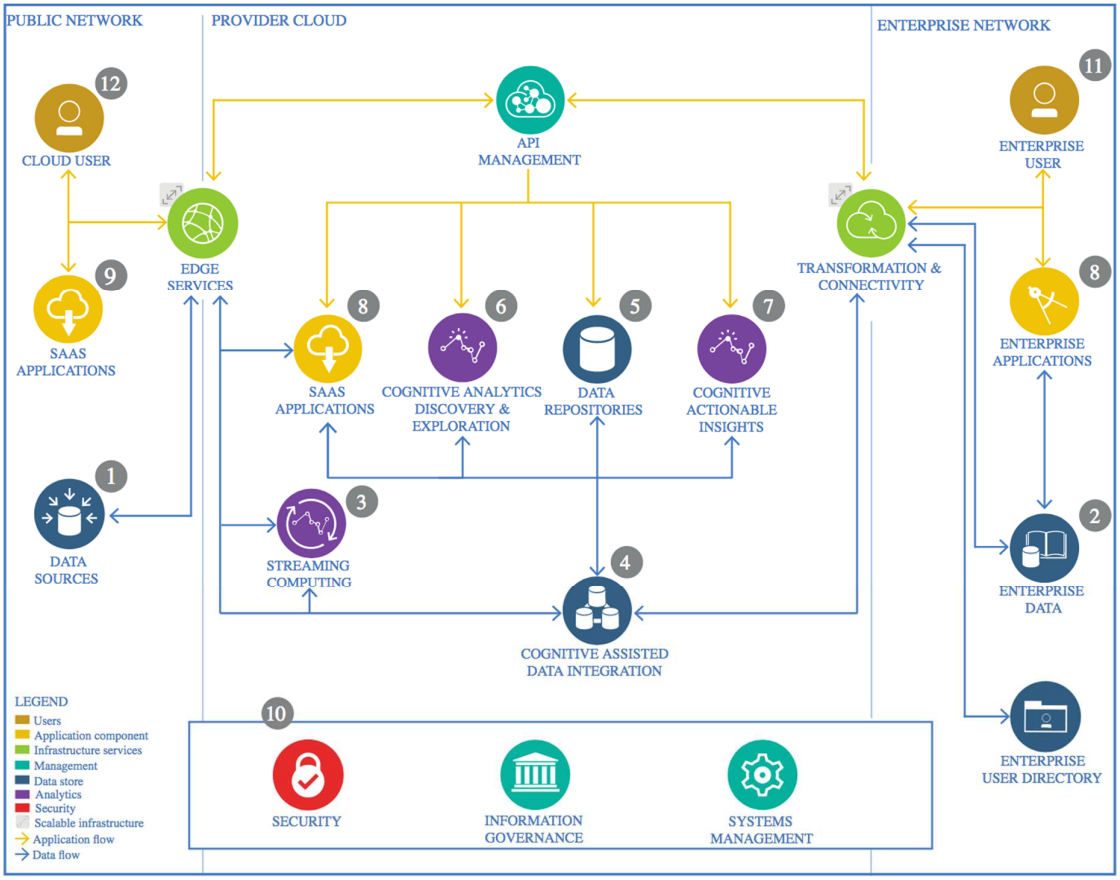
The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document

1 Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

1.1 Data Source

1.1.1 Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

ETL Section:

Data Source: COVID19 data is taken from John Hopkins University, The Center for Systems Science and Engineering (CSSE).

Data is properly transformed and proper features are checked for validity.

Data Analysis Findings:

1. From worldwide perspective, the confirmed cases are the highest range and many outliers found in boxplot.

2. This analysis and machine learning model will predict COVID-19 deaths in Malaysia.

3. From time plots, the biggest jump is from October 2020 onwards.

4. By extracting individual cases by day, deaths peaked in April 2020 and October 2020 onwards.

5. Separate charts by Month and Day are plotted to see the distribution.

Feature Engineering:

1. The dataset is transformed using diff method to extract per day for Confirmed, Recovered, Deaths and Active Cases.

2. Due to diff method, I imputed a NaN value as 0.0 as starting point.

3. Created separate month and day features from Date feature.

1.1.2 Justification

Please justify your technology choices here.

To ensure data is consistent and frequently updated to reflect the current status of the pandemic.

1.2 Enterprise Data

1.2.1 Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

Not applicable.

1.2.2 Justification

Please justify your technology choices here.

Not applicable.

1.3 Streaming analytics

1.3.1 Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

Not applicable.

1.3.2 Justification

Please justify your technology choices here.

Not applicable.

1.4 Data Integration

1.4.1 Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

Not applicable.

1.4.2 Justification

Please justify your technology choices here.

Not applicable.

1.5 Data Repository

1.5.1 Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

Not applicable.

1.5.2 Justification

Please justify your technology choices here.

Not applicable.

1.6 Discovery and Exploration

1.6.1 Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

Python code in Jupyter Notebook.

1.6.2 Justification

Please justify your technology choices here.

Easy to code offline and do experimentation.

1.7 Actionable Insights

1.7.1 Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

Numpy, Pandas, Scikit-Learn, XGBoost, Keras libraries are used to do machine learning.

Model Definition: Model performance will use RMSE to measure errors. MSE and R2 are for references.

Four algorithms will be used: Linear Regression, Extra-Trees Regression, XGBoost and Deep Neural Networks.

Model Training: All models are training using local computer within Jupyter Notebook.

Model Evaluation: Root Mean Squared Error (RMSE), Mean Squared Error (MSE) and R-Squared (R2) are used to measure performances.

Model Deployment: The model can be deployed in a website to do the prediction.

1.7.2 Justification

Please justify your technology choices here.

These packages are open source, free and fully supported by creators and community.

Linear Regression: The first and simplest model to implement.

Extra-Trees Regression: Using randomized decision trees to control over-fitting

XGBoost Regression: One of the top algorithms used for higher prediction accuracy with hyperparameters tuning.

Deep Neural Network: The most complex of all and longest training time for accuracy. Keras framework is used for easy Sequential modeling.

The root-mean-square error (RMSE) is a used measure of the differences between values (sample or population values) predicted by a model or an estimator and the values observed.

1.8 Applications / Data Products

1.8.1 Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

Predict estimated death cases per day.

1.8.2 Justification

Please justify your technology choices here.

To find out Covid19 death rate over a long time period and in cyclical waves.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

Not applicable.

1.9.2 Justification

Please justify your technology choices here.

Not applicable.