Architectural Decisions Document for COVID-19 Predictions

# Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

**Data quality assessment**

Assessing and improving data quality should be the first step of any machine learning project.

This includes checking for consistency, accuracy, compatibility, completeness, timeliness, and duplicate or corrupted records.

In this Project, I removed duplicates for Provinces and Countries

**Feature engineering**

Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

Performance measures I have chosen (MAE, MSE)

predictive Models I used SVM (Support Vector Machine) and Polynomial regression

The JHU COVID-19 data frim Github has selected

**Algorithm**

I have chosen Linear Regression algorithm for predicting COVID-19 confirmed cases for World

Predictive Models I used are SVM (Support Vector Machine) and Polynomial regression

**Framework**

I have chosen COVID-19 dataset and used Pandas data frame to store the csv data from JHU github

Scikit-Learn is a powerful Python library for machine learning coding and explicitly used in the creation of Models. It has built on other libraries such as SciPy, Numpy, and matplotlib. It is a highly efficient tool for classification, regression, and clustering like statistical modeling.

**Model performance indicator**

I have choosen MAE (Mean Absolute Error) as it shoes the average of the difference between the Original Values and the Predicted Values. It gives us the measure of how far the predictions were from the actual output.

## Data Source

### Technology Choice

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

**John Hopkins University time series COVID19 Data Set from Github CSV Files**

<https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv>

https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse\_covid\_19\_data/csse\_covid\_19\_time\_series/time\_series\_covid19\_deaths\_global.csv

https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse\_covid\_19\_data/csse\_covid\_19\_time\_series/time\_series\_covid19\_recovered\_global.csv

https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse\_covid\_19\_data/csse\_covid\_19\_daily\_reports/04-18-2020.csv

### Justification

CSV files

## Enterprise Data

### Technology Choice

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

* **IBM Watson Studio Jupyter Notebooks, scikit-learn, pandas**

### Justification

Please justify your technology choices here.

## Streaming analytics

### Technology Choice

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

### Justification

Please justify your technology choices here.

## Data Integration

### Technology Choice

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

* **IBM Watson Studio Jupyter Notebooks, scikit-learn, pandas**

### Justification

Please justify your technology choices here.

## Data Repository

### Technology Choice

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

**IBM Watson Studio**

### Justification

Please justify your technology choices here.

## Discovery and Exploration

### Technology Choice

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

**Jupyter notebook**

### Justification

Please justify your technology choices here.

## Actionable Insights :

## **Predict COVID-19 Confirmed cases and display graphs by World, US**

### Technology Choice

**Jupyter Notebook**

### Justification

Please justify your technology choices here.

## Applications / Data Products : **COVID19Prediction.ipynb**

### Technology Choice

**Jupiter Notebook, PowerPoint**

### Justification

Please justify your technology choices here.

## Security, Information Governance and Systems Management

### Technology Choice

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

### Justification

Please justify your technology choices here.