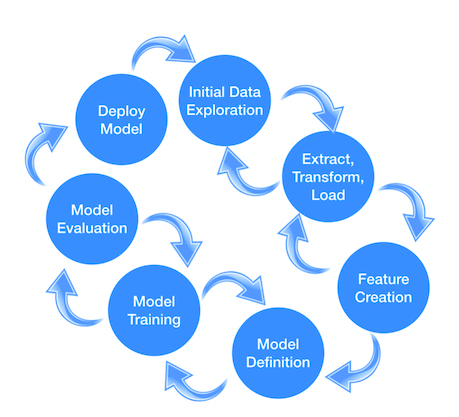
The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document

Project title: Vaccination against COVID vs. new cases of COVID

# Architectural Components Overview



## Initial data exploration

### Technology Choice

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

The use case for my capstone project is analysis of the progress of vaccination against COVID in the world and new cases of COVID. The data set is Data on COVID-19 (coronavirus) by Our World in Data (https://github.com/owid/covid-19-data/tree/master/public/data) which contains data for every country and for the whole world taken from official sources. The goals of my project are: 1) to analyze how the progress of vaccination reduces the number of new cases of COVID; 2) to make a prognosis when the whole population of the world will be vaccinated.

The stage of initial data exploration includes standard procedures, such as: reading CSV file, finding the information about the dataset, creating a new dataset that will be used in this project and saving it in a new CSV file.

The original dataset contains data for all the countries, continents, and the world in whole. The main problem with data in it is that there are quite a lot of NaNs, which need to be corrected.

### Justification

Please justify your technology choices here.

Since the original dataset is not too large (around 40 MB), there is no need to use scalable data technology: Apache Spark, SparkML, etc. So, in this project, I did not use them. I used regular frameworks used data science, such as: numpy, pandas, matplotlib, scipy, sklearn.

In the first stage (initial data exploration), I used pandas for reading the dataset, finding information about it, and for creating the new dataset and saving it. Pandas is the standard framework for these tasks, and since I did not need to use Apache Spark, I used pandas.

The original dataset includes 67 columns with various information. However, not all of them are needed in this project. So, I create a new dataset with columns that will be used in the project: 'location','date','total\_cases','new\_cases','total\_deaths','new\_deaths','icu\_patients', 'hosp\_patients','total\_vaccinations','people\_vaccinated','people\_fully\_vaccinated', 'total\_boosters','new\_vaccinations','population'.

Although this dataset needs some data cleansing since it contains NaNs, I do this later, in the next stage, after extracting data for the whole world and for some individual countries. I do not correct NaNs now because they will be corrected in different ways, depending on where they occur.

## Extract, transform, load (ETL)

### Technology Choice

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

In this stage, I do ETL, and also EDA and data visualization. I use the dataset which I created in the previous stage, and extract data for the whole world, and also some individual countries.

**Data Cleansing**. The main problem with data in this dataset is that there are many NaNs. I use three strategies to deal with them.

1. I remove columns that contain only NaNs.
2. I replace NaNs with zeros.
3. In order not to deal with many zeros in the beginning, I am going to use several datasets: 1) the dataset for analysis and visualization of cases of COVID - since the beginning of pandemic; 2) the dataset for analysis and visualization of the number of people who got vaccinated (by, at least, one injection of vaccine) - since the beginning of vaccination; 3) the dataset for analysis and visualization of the number of people who got fully vaccinated - since the beginning of full vaccination; 4) the dataset for analysis and visualization of the number of total boosters - since the beginning of revaccination.

Also, for the datasets for the whole world, I remove columns "hosp\_patients" and "icu\_patients" since they do not contain any data for the world. In other columns, NaNs will be replaced with zeroes.

Then, I draw temporal diagrams of new cases of COVID in the world (since the beginning of pandemic), total cases, people vaccinated (with, at least, one vaccine dose), people fully vaccinated, and total boosters. I also draw scatter diagrams people vaccinated vs. new cases, people fully vaccinated vs. new cases, and revaccination (total boosters) against Covid in the world. Then, I calculate Pearson correlation coefficient between new cases and people vaccinated, people fully vaccinated, and total boosters.

I did similar procedures for data for two countries: United Arab Emirates, and USA. I used them as example countries. UAE is the country that has the highest number of people vaccinated (99%) and the highest number of people fully vaccinated (92%). This is why I used it as one of the examples.

### Justification

Please justify your technology choices here.

I used numpy for creating new datasets and cleansing data, matplotlib for visualization, and scipy for calculating Pearson correlation coefficient. These frameworks are quite standard for these kind of tasks. They have the functions that I needed. This is why I used them.

## Feature creation

### Technology Choice

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

In this stage, I prepare data for developing models of the progress of vaccination and full vaccination in the world.

1. For machine training, dates need to be transformed from object type into float64 type (date\_delta).
2. Also, I create new datasets for machine learning that contain only people vaccinated / people fully vaccinated, dates, and date\_delta.

### Justification

Please justify your technology choices here.

The main thing here is to transform dates in object type into date\_delta in float64 type in order that data for dates may be used in machine training model. Other data in the dataset are already in float64 type and do not need transformation.

I decided not to use data normalization. On the one hand, data for people vaccinated and people fully vaccinated are quite big (several milliards) and in principle machine training models should perform better when data is normalized. However, in the draft notebook, I tried both ways – models trained on normalized data and models trained on non-normalized data. Performance of models trained on normalized data was not any better than models trained on non-normalized data. On the other hand, after model training and evaluation, I used the models for making prognoses when all people in the world will be vaccinated and all people in the world will be fully vaccinated. Models trained on normalized data require some additional calculations on this step because the number of the world population (milliards people) needs to also be normalized in the same way as training data. So, I decided to avoid that.

## Model definition

### Technology Choice

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

The main task for machine learning in this project is to develop machine learning models for the progress of vaccination (the number of people vaccinated) and full vaccination (the number of people fully vaccinated) in the world, which then will be used for making prognoses when vaccination and full vaccination in the world will be completed if the progress of vaccination and full vaccination continue with the same speed as now. I did not develop machine learning model for the progress of re-vaccination in the world because the dataset contains only data of total boosters, not data of people who got boosters. Some people had more than one booster vaccine, and so the number of total boosters does not give information about the number of people who got booster vaccines. Also, this data cannot be used to predict when all people in the world will get booster vaccines.

The number of people vaccinated and the number of people fully vaccinated depend on two variables: dates and new vaccinations. However, it would be hard to use the number of new vaccinations in machine learning models due to a couple factors:

1. New vaccinations include people who got vaccinated with the first dose of vaccine, people who got vaccinated with the second dose of vaccine, and people who got vaccinated with booster vaccine. Since the number of new vaccinations does not tell how many of people got what kind of vaccination, it is hard to use this number in machine learning models for predicting the number of people vaccinated and people fully vaccinated.

2. The number of people vaccinated and people fully vaccinated depends on the number of new vaccinations not directly but in a cumulative way. That is, the number of people vaccinated is the sum of people who got the first dose of vaccine for all the days of vaccination, and the number of people fully vaccinated is the sum of people who got the second dose of vaccine (or the first dose if the vaccine requires only one dose) for all the days of vaccination.

Due to these reason, I did not use the number of new vaccinations in the machine models of progress of vaccination and of full vaccination. Instead, I used only time/date, that is, date\_delta.

Since the machine models need to predict the number of people and since the data that I am using are labeled, I need to use regression. To begin with, I chose linear regression models and also polynomial regression models with different degrees polynomials (quadratic, cubic, etc.). Other kinds of non-linear regression (exponential, logarithmic, sigmoid, etc.) would not work for these models because the diagrams of growing the number of people vaccinated and people fully vaccinated with time, show that other functions (except linear and polynomial) will not be fitting.

Since my models are very simple – the number of people vaccinated and fully vaccinated depends only on time, and since the datasets are quite small (containing data only for about 400 days), there is no need to use deep learning models. So, I used only machine learning models in this project.

### Justification

Please justify your technology choices here.

For machine learning models, I used sklearn (functions LinearRegression and PolynomialFeatures). I also tried Lasso and Ridge, but they did not give any improvement of performance of the models.

## Model training

### Technology Choice

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

Since the models are simple, their training was also simple. I split the datasets into train and test datasets and then fit the models. I also used visualization.

Since I did not need to use deep learning models, the models were trained on CPU. I did not need to use GPU or TPU for training. I also did not need to do hyperparameter tuning. Since the datasets are quite small, I did not need to use computer clasters.

### Justification

Please justify your technology choices here.

I used sklearn for model training, as well as for model definition and model evaluation. I did not need frameworks for deep learning since I did not use deep learning models. I also did not need to use Apache SystemML because the datasets are quite small. Since the size of the datasets depends on the number of days of vaccination, it will not become so large as that it would require the use of scalable data science technologies, such as Apache Spark and Apache SystemML.

## Model evaluation

### Technology Choice

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

For model evaluation, I used r2 score because it seems to be the most convenient metric for regression models. It gives scores between 0.0 and 1.0 with the best possible score being 1.0. So, it is convenient to evaluate a model performance and to compare it with performances of other models, using r2 score.

As it could be expected, the linear models for both vaccination (with the 1st dose of vaccine) and full vaccination give the lowest r2 score, and polynomial models give higher scores, which become higher with raising of the degree of polynomial.

However, although polynomial regressions with high degree of polynomial (beginning at about degree = 6) give good performance in the range of the datasets used for training and testing, they are not appropriate for making prognoses since they either grow or drop very speedily at higher values for time. This means that such models will either give too optimistic or too pessimistic predictions (either all the population of the world will be vaccinated / fully vaccinated in just a few days or there will be no progress in vaccination at all). Both kinds of scenarios are obviously wrong and therefore such models are not useful.

### Justification

Please justify your technology choices here.

I used r2\_score function from sklearn for evaluating models. I also used matplotlib for visualization. I draw diagrams both for seeing how well the models approximate the datasets and also how they go on higher values for dates.

## Model deployment

### Technology Choice

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

Model deployment includes:

1) three already run, static notebooks created in Jupyter Notebook – initial notebook, notebook for ETL, EDA, and visualization, and notebook for machine learning models;

2) presentations for stakeholders and for peers which contain diagrams and results from the notebooks;

3) video with the presentations.

### Justification

Please justify your technology choices here.

Choice of the ways of model deployment was determined by the requirements for this project. Those three model deployment ways were required by the assignment.