The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document Template

# Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

## Data Source

### Technology Choice

External date will be used in the form of CSV, JSON, GeoJSON and text files. The data gathered from AWS cloud server, US Census Bureau and other relevant websites. There are two types of datasets collected:

Static datasets

1. USA county population
2. USA state abbreviations
3. Hospital capacity in USA
4. Geographical boundaries of US states
5. Geographical boundaries of US counties
6. Geographical locations of US counties

Dynamic datasets

1. Number of Covid 19 cases in US counties
2. Number of Covid 19 cases in US states

### Justification

These data sources contain the most up-to-date and relevant information about my project.

## Enterprise Data

### Technology Choice

The dynamic data (number of Covid 19 cases in US counties and states) is hosted on AWS cloud server.

### Justification

The data on AWS cloud server is updated regularly.

## Streaming analytics

### Technology Choice

At the moment, the dynamic data (number of Covid 19 cases in US counties and states) is updated and downloaded manually from the AWS cloud server.

### Justification

To simplify the process of the data.

## Data Integration

### Technology Choice

Data is cleaned through several steps.

First all the datasets are checked for missing and null values. Then these values were either removed or replace by appropriate values based on their feature. For example, the names of some counties were unknown in some of the reported covid 19 cases, so these unknown values were replaced by the county with the highest population in the same state.

Secondly, the consistency of the county names was checked in all the data sets. For example, there are some naming conventions were the county name has the word “parish” attached to it. So all the county names in all the data sets were unified.

Finally, all the data sets were joined together into more specific data sets:

1. df\_hospitals 🡪 contains information about the known hospital capacities in some counties as well as information about the counties themselves such as location and population
2. df\_ cases 🡪 contains information about number of cases in each county with known hospital capacity
3. df\_cases\_all 🡪 contains information about the number of cases in all counties
4. df\_cases\_date\_total 🡪 contains the number of cases and deaths in each day in USA
5. df\_cases\_county\_total 🡪 contains information about the number of cases in each county in each day

Feature transformation methods:

1. imputing 🡪 as explained above, unknown counties with reported cases were replaced by the county name with the highest population in the same state. This is justified by the level spread in the high population counties.
2. Scaling 🡪 scaling the number of cases using StandardScaler and MinMaxScaler in the DBSCAN and LSTM models, respectively.

### Justification

All these methods mentioned above simplified the process of data in the used models. Also, it allowed for joining the dataframes and extract useful information in the Data Exploration step.

## Data Repository

### Technology Choice

A local hard drive is used to store all the data as well as the produced dataframes and models.

### Justification

To ease the access, manipulation and visualization of the data.

## Discovery and Exploration

### Technology Choice

1. Folium map 🡪 they is used to visualize the population, number of hospitals and number of covid 19 cases in US counties. They is also used in the DBSCAN model to visualize the evolution in the number of clusters of case in USA counties each day.
2. Time-series charts 🡪 they are used to visualize the increase in the number of cases each day in some counties.
3. Linear regression plots and bar charts 🡪 They are used to explore the relationship between some features such as number of cases and deaths, cases and population, and number of hospitals and cases in US counties.
4. Correlation matrix 🡪 it is used to study the correlation between some important features such as cases, deaths, number of hospitals, number of beds in hospitals, and average ventilation usage in hospitals in all counties.

### Justification

These methods are relevant to critically explore all the data sets and identify potential/interesting trends to study using machine/deep learning models.

## Actionable Insights

### Technology Choice

The unsupervised machine learning model DBSCAN is used to study the spread of cases in all the counties over specific time frames (7 days). This model is used to study the change in the number of clusters of cases in these time frames and visualize this evolution of cases on a folium map. This model is optimized and evaluated using the silhouette coefficient due to the absences of a ground truth. The optimization was done by varying the eps distance value in the DBSCAN model to maximize the silhouette coefficient in all the clusters.

The deep learning model LSTM is used to predict the number of cases in all the counties. The model is trained and tested over time frame windows (3 days) in the total number of cases each day. The model is optimized by doing a mini-grid search over several optimizer and loss function that maximize the R2 score. The best model is chosen to predict the number of cases in a one window time frame (3 days) in the future.

### Justification

The first model (DBSCAN) provides insights about the spread of the covid 19 virus in the US counties. The second model (LSTM) is very interesting as it can understand the evolution in the number of cases and predict future cases. The model is only as good as the available data which is not sufficient at the moment.

## Applications / Data Products

### Technology Choice

A product will be in a form of a predictor that can explain how similar data can behave over time. It is not a commercial product, but it will have an impact on studying the data itself.

Also, a PDF report is generated using the Jupyter notebooks (ETL, DataExploration\_and \_FeatureEngineering, and ModelTraining\_and\_Evaluation).

### Justification

It is mainly a research endeavor to better understand how this data behave in similar situations in the future. A PDF report is useful to understand the whole process (from extracting the data to building the machine learning models).

## Security, Information Governance and Systems Management

### Technology Choice

There is not an exact security protocol over the data and the models since they are located on a local hard drive. It will be shared in public.

### Justification

To ease the access to the data to people who might find it valuable.