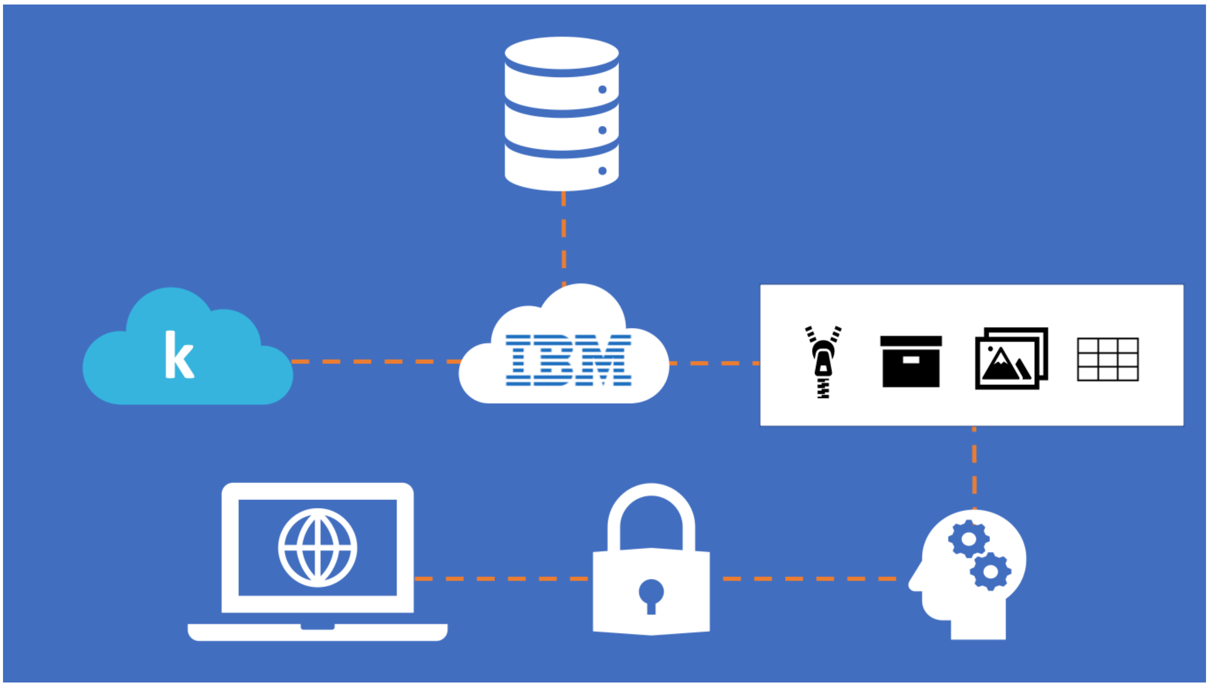
Jonah Winninghoff
February 11, 2021

Architecture Decision Record

Image Classification of X-Ray Chests



Contents

1.0 Architecture Decision Record

1.0.1 What is an ADR?

The Architecture Decision Record (ADR) is to capture every key decision for each part of relevant coding process. It does not describe what, and how, coding executes but rather the abstract of it—the reason of it. For this project, the ADR is to document the *whys.* These documents are to understand better why the image classification is architected this way.

1.1 Data Source

By definition, the data source is a location of which data being used comes from. The nomenclature of this term is databases, which commonly used for many data scientists and analysts—that is, relational database management systems.

1.1.1 Technology Choice

In order to build the image classification that accurately detects what type of image is, the unstructured dataset containing images is necessary. Pranav Raikote who is R&D engineer at Bengaluru, Karnataka, India publishes the images released by the University of Montreal in Kaggle platform. This platform is a community where data scientists and machine learning engineers collect data and work on a variety of projects. It is as well open-source data. The Kaggle API in IBM Watson Studio Jupyter Notebook is in use to communicate this platform and retrieve this dataset. This dataset is 153mb in total and it has two different files, which are training and testing. Both have three different subsets, which are Covid-19, viral pneumonia and normal lungs.

1.1.2 Justification

The open-source data in Kaggle lubricates the transparency of work and update on dataset. For example, this dataset may be occurred if the Covid-19 new variants emerge. However, the discretion to do so is entirely up to its owner.

1.2 Enterprise Data

Enterprise data is either datasets or databases shared by the users of institutes or geographic regions, which mainly focuses on resilience of data storage. This resilience is a key to prevent from financial loss experience for all whose involvement in this.

1.2.1 Technology Choice

The IBM Watson Studio Jupyter Notebook is capable of implementing any cloud and has several different programming languages including R and Python. It is often in use for initial data exploration, ETL, data cleansing, exploratory data analysis, feature engineering, model training, model evaluation, and deployment. As mentioned earlier, this notebook utilizes Kaggle API to retrieve this dataset. This process is entirely cloud-based.

1.2.2 Justification

The advantage of using cloud is a resilience of dataset, which has no geographic restrictions and high accessibility. To process this dataset does not require internal power. Generally, some ethical concerns have arisen because of data privacy and security. Certain data information may be highly sensitive so that one needs to take careful consideration how to maintain this information confidential before doing so. In this case, there is no indication of personal information record associated in this dataset.

Furthermore, the University of Montreal gives acknowledgement for releasing this dataset and it can help them to extend their medical research.

1.3 Streaming analytics

Streaming analytics is, in definition, assesses and engages in coding tasks based on real-time data, which data continuously changes over time. The data can be from either the Internet of Things (IoT), transactions, web interactions, mobile devices, cloud applications, and machine sensors.

1.3.1 Technology Choice

Perhaps, this dataset might be considered as a streaming since it is on cloud application as indicated by above. Even so, there are three reasons why streaming analytics remains inapplicable. Firstly, this dataset is immovable, and it does not change for at least nine months. Secondly, when this dataset extracts, the coding work undertakes batch processing rather than streaming. Finally, if the image classification implements, this model is in controlled environment. In other words, the x-ray chest images originate from the laboratory.

1.4 Data Integration

The meaning of term *Data Integration* is the process of mingling data from a variety of sources, which commonly used for querying relational databases. For example, one might encounter several different datasets across the database and these need to mingle using primary and foreign keys.

1.4.1 Technology Choice

As mentioned earlier, this dataset contains 153mb of x-ray chest images and it is unstructured. This assessment is not actionable, so data integration is required. This dataset unzips and extract, which turns into filesystem. The `pathlib` is in use to convert filesystem into object. Also, it utilizes to separate training and testing datasets. Finally, the `flow_from_directory` imported by TensorFlow using Keras API automates to categorize each subset of both training and testing datasets based on how directory structures.

1.4.2 Justification

The `pathlib` is standard procedure for data transformation, which does not require careful consideration. However, choosing the `flow_from_directory` technique needs to take some considerations by identifying how directory structures. This approach is, without doubt, justified because this directory is as follows:

```
…/data/train
        …/Covid
            …/001.png
            …/002.jpg
        …/Pneumonia
            …/001.jpeg
            …/002.jpg
        …/Normal
            …/001.jpg
            …/002.png
```

1.5 Data Repository

The data repository is, by definition, persistent storage for data. This persistent storage can either be on cloud or a local desktop.

1.5.1 Technology Choice

As mentioned earlier, this dataset is retrieved from Kaggle platform. Also, this dataset zips and persists in the IBM Cloud Pak storage. This dataset is treated as an asset in the Advanced Data Science project at IBM Cloud Pak for Data platform.

1.5.2 Justification

Having data backup is the best data science practice in case if the owner decides to remove this dataset. The data loss has significant impact on a very function of image classification—one that needs data to learn. The cloud-based storage is preferrable since it is global and accessible. As discussed earlier, the data privacy should not be a concern, so this strategy is more sensible.

1.6 Discovery and Exploration

The discovery and exploration are another term to describe Exploratory Data Analysis, which refers to components allowing for visualization and summary statistics. In other words, this procedure undergoes data mining process by searching for correlations in the data.

1.6.1 Technology Choice

There are five different exploration data analyses to comprehend what this dataset looks like and how it behaves. The first step is to randomly select nine different images with label. This is to identify if x-ray images show any difference between three subsets: normal, pneumonia, and Covid. The next one is to plot number of images in each subset using bar chart. Thirdly, the average image is to find each average number per pixel in total for every subset. This result shows clear distinction between average Covid, pneumonia, and normal lung images. Difference image is to ensure if this distinction differs enough from each other. Finally, the standard deviation image is to determine how far this number is from average per pixel in image for each subset.

1.6.2 Justification

Every analysis has a purpose. For example, the first analysis is to familiarize with what this dataset looks. The second analysis is this bar chart. It is to ensure that total number of each subset in this dataset is not too even. This result proves that it is relatively even, so it is unlikely to have an impact on building image classification. But this result also implies that the sampling size is not large enough. The `ImageDataGenerator` can mitigate this problem. Average, difference, and standard deviation images are to test if the distinction is clear from each other and image classification model is trainable.

1.7 Actionable Insights

The actionable insights are—by definition—to train and evaluate either machine learning or deep learning models. This includes hyperparameter tuning, statistics, and several more.

1.7.1 Technology Choice

Given that the model is in training to be image classification, the Convolutional Neural Network (CNN) is an obvious choice. The first model is to construct several different layers using TensorFlow tethered to Keras API. The first layer is to make this dataset more noise by randomly flipping, rotating, and zooming them. The second layer is to normalize every image by rescaling them from 0-255 to 0-1. The next several layers are to repeat same procedure by calculating tensors between max pooling and rectified linear activation. The next layers are to use Dropout rate and to convert tensors into 1-dimensional array. The final two layers are rectified linear and softmax activations, respectively. The second model is to improve accuracy and loss results by adding random brightness since several images existing in this dataset are sometimes too dim or too bright. The `Adam` is a gradient descent that is in use for this CNN as well.

1.7.2 Justification

This CNN technique operates this classification more efficient and agile. Without this, the computation may be time consuming due to plethora of parameters. As mentioned earlier, the sampling size of this dataset is too small. The `ImageDataGenerator` augments images that help mitigate potential biases. Not only does it improve this algorithm performance, it becomes more resistant to overfitting. This generator stabilizes the convergence of categorical cross-entropy loss. The out-of-sample accuracy is 94% and its loss is 0.14. This gradient descent helps converge faster and more stable.