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

The data has been obtained from <https://www.kaggle.com/>. All input data are x-ray images of lungs. There are two types of classes for them: normal lungs or pneumonia ill lungs.

### Justification

It is a free and easy to use source of data and thanks to the user feedback it presents, we can have a degree of confidence in the quality of the data.

## Enterprise Data

### Technology Choice

Component not needed.

### Justification

This project uses a small amount of static data, around 1 GB so it is intended to run locally. Because of this, not cloud solutions are needed now.

## Streaming analytics

### Technology Choice

Component not needed.

### Justification

The database to use is a static dataset of images not bound to real-time events.

## Data Integration

### Technology Choice

Pandas library is used to import all the data and separate it in train, test, and validation subsets.

Keras image libraries are used to transform the input image data.

### Justification

Because all the input data consists of images, we can apply image transformation techniques like rotations, shifts and zoom-ins to increase the amount of data without hurting the model performance. Also, because images are integer number matrices, there are going to be normalized between 0 and 1.

## Data Repository

### Technology Choice

Local file system.

### Justification

As stated before, due to the small size of the data it can be stored locally without requiring cloud-based solutions. Furthermore, this data was obtained from <https://www.kaggle.com/> so there is no danger of losing the data.

## Discovery and Exploration

### Technology Choice

We have an x-ray images database of normal and pneumonia ill lungs, each with its class label assigned and organized in the train, test and validation subsets. The objective is to elaborate a deep-learning model using the Keras library that learns to classify an x-ray image of lungs as normal or pneumonia ill.

### Justification

Keras Libraries are used due to being easy and free to implement while having great performance.

## Actionable Insights

### 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.

## Applications / Data Products

### Technology Choice

The data product is a jupyter notebook.

### Justification

The reason for developing a jupyter notebook is due to easy access it provides to be read, modified and ran.

## Security, Information Governance and Systems Management

### Technology Choice

The source code will be open source to anyone that wants to use it or improve it. The dataset is open source provided by <https://www.kaggle.com/>.

### Justification

Because the project uses an open source dataset and it has been developed for academic use, it will be open source. This can help other data-scientist in their career development and goes along the open-source guidelines.