The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document

1 Architectural Components Overview

1.1 Data Source

1.1.1 Technology Choice

A collection of thousands of chest X-ray scans in jpeg format of variable size. Final assumption was that they will be staged using a Cloud Object Store but for sake of simplicity I run my sample extracting the scans into arrays (features and labels) using the Keras preprocessing and I can stage the arrays in the Object Store.

Sample dataset used to train and validate the model from Kaggle: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia.

1.1.2 Justification

The raw data will need to be accessible by Jupyter Notebooks to let data scientists to fit the supervised learning model, review and fine tune.

1.2 Enterprise Data

1.2.1 Technology Choice

Real-time scans will be submitted via a simple REST-HTTP endpoint which will return the predictions. Batch wise sync from the enterprise data centers will update the cloud to fit the learning model.

1.2.2 Justification

Because the model will be consumed by a wide range of institutions it is necessary to implement an open standard to deal with the largest possible range of enterprise applications and their data store. Regulatory aspects will be taken care accordingly.

1.3 Streaming analytics

1.3.1 Technology Choice

I don’t see such a compelling demand for a streaming analytics solution because the endpoint will simply return predictions.

1.3.2 Justification

The model needs load balancing which is different than digesting a high volume of big data in real time.

1.4 Data Integration

1.4.1 Technology Choice

In my lab I will use the Jupyter Notebook to run ETL but in a real production scenario I would recommend adopting a solution like IBM DataStage.

1.4.2 Justification

I need to load, quality control and transform a large set of scans.

1.5 Data Repository

1.5.1 Technology Choice

A Cloud Object Store.

1.5.2 Justification

The storage cost has a primary impact for a huge data store of image scans.

1.6 Discovery and Exploration

1.6.1 Technology Choice

JupyterLab, python, scikit-learn, pandas, NumPy, matplotlib, Keras and Tensorflow.

1.6.2 Justification

There's a minimal need of visualizations while I need state of the art metrics.

1.7 Actionable Insights

1.7.1 Technology Choice

Keras and Tensorflow.

1.7.2 Justification

They are a Deep Learning framework of reference for the type of model which I need to build.

1.8 Applications / Data Products

1.8.1 Technology Choice

D3.

1.8.2 Justification

I assume to deliver to a REST endpoint.

1.9 Security, Information Governance and Systems Management

1.9.1 Technology Choice

Open.

1.9.2 Justification

Open.