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

1. Source: Gorgolewski KJ, Durnez J and Poldrack RA. Preprocessed Consortium for Neuropsychiatric Phenomics dataset [version 2; peer review: 2 approved]. F1000Research 2017, 6:1262 (https://doi.org/10.12688/f1000research.11964.2)
2. Link to zip files https://www.openfmri.org/dataset/ds000030/

### Justification

1. Open-source fMRI repositories
2. Data can be downloaded by different ways: web browser, Node.js, S3, Datalad

## Enterprise Data

### Technology Choice

Not necessary.

### Justification

Data are provided by academic groups

## Streaming analytics

### Technology Choice

Not necessary

### Justification

1. Limited budget
2. Data source is not continuously updated

## Data Integration

### Technology Choice

1. Data were download into an upgraded HP Z640 workstation:

* Processor: Intel Xeon E5-2630 v4 20 CPUs @ 2.2 GHz
* RAM: Samsung 32 GB x 4 (128GB) 2133P DDR4
* MOBO: HP 212A
* GPU: MSI NVIDIA RTX3070 8GB
* HDD: HGST WD 12TB 7200 RPM SATA 6Gb/s 3.5-Inch Enterprise Hard Drive (HUH721212ALE604).
* SSD: Samsung 850 Evo 500 GB

1. Zip files were unzipped with WinRAR 5.91 (64-bit)
2. Gz files were unzipped with gunzip function in Nipype 1.6.0
3. Python 3.8.5 was used as programming language
4. Pip 20.3 and Anaconda3-2020.11 were used to manage packages
5. Python code was written in Jupyter Notebook v2020.12.414227025 extension for VS Code 1.53
6. In order to maintain correct subject and diagnosis information, data were loaded onto Pandas 1.2.0 dataframe
7. Image data were unpacked with Nibabel 3.2.0

### Justification

1. WinRAR on Windows is freely available
2. Nipype is a robust platform to create workflow for fMRI preprocessing.
3. Python 3.8.5 has less package conflict problems compared to 3.7 and 3.9
4. Jupyter Notebook has clear format, intuitive debugger and easy to manipulate.
5. VS Code has an extensive plug-in library, making coding more flexible and easier.
6. Pandas dataframe can be constructed from different file types including tsv file, which is used by the dataset authors to provide additional information. Furthermore, Pandas dataframes can be used to flexibly convert in between datatypes, facilitating data exploration and feeding data to learning models
7. Nibabel can be used to unpack whole or slices of fMRI data, converting image data to ndarrays.

## Data Repository

### Technology Choice

Data were downloaded and stored locally on a HP Z640 Workstation with a HGST WD 12TB 7200 RPM SATA 6Gb/s 3.5-Inch Enterprise Hard Drive (HUH721212ALE604).

### Justification

The only available option due to limited budget.

## Discovery and Exploration

### Technology Choice

1. Pandas 1.2.0 and Numpy 1.20.0 were used as foundations and building blocks for data exploration
2. Matplotlib 3.3.4 was used for data visualization
3. Scikit-learn 0.24.1 was used for machine learning implementation
4. Joblib 1.0.1 was used for parallelism
5. Tensorflow 2 and Keras 2.4.3 were used for deep learning modeling.

### Justification

1. Image data were stored in Numpy ndarrays therefore all primary manipulation needs to be done on Numpy
2. Np arrays can be easily transferred to Pandas dataframes to be combined with non-fMRI data
3. Matplotlib is a robust and easy to use plotting library and similar to R’s ggplot
4. Scikit-learn is the most extensive machine learning platform which includes all the necessary tools from processing to tuning models
5. Keras API of Tensorflow 2 is fast and easy to use and build deep learning models, unlike Pytorch which requires a lot more coding involved.

## Actionable Insights

### Technology Choice

Validation accuracy and loss were used as indicators for model performance

### Justification

The resulting models are not yet ready for realistic prediction due to low prediction accuracy and no learning process

## Applications / Data Products

### Technology Choice

1. A Jupyter notebook
2. Html presentation file converted with Reveal.js
3. ML and DL models
4. Presentation video, recorded with OBS 26.1.1, edited and rendered with VEGAS Pro 18

### Justification

1. Python Script shows that the program was executed successfully without any major problem.
2. Resulting models therefore can be implemented on a different datasets or experiments to further improve learning process and prediction accuracy.
3. OBS is open source and easy to use
4. VEGAS Pro 18 is a robust video editing platform, can edit mkv file, which is output from OBS

## Security, Information Governance and Systems Management

### Technology Choice

* Jupyter Notebooks and other intermediate files were uploaded and stored in Dropbox and Github.
* Presentation video was uploaded to Youtube

### Justification

1. Premium Dropbox provides low-cost, 2TB of cloud storage, making it easy to access working directories from anywhere
2. Github provides free infrastructure to deposit and distribute final products
3. Youtube gives a lot of exposure to my content