Folder-based Datalake Usage for University Physics Student: Sharing the Experience

Sparisoma Viridi

Nucler Physics and Biophysics Research Division, Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Indoensia

Abstract

Some students, that are performing research while attending university or college, must deal with a lot of data that combines information from multiple modalities, e.g. text, audio, or images. Even when working only using data from one modality, e.g. text, sometimes they have to process set of original data to obtain a parameter, then several parameters will be obtained from various sets of data. Furthermore, from this parameters, typically using modeling, a new parameter can be extracted. We can label the new parameter as second-level parameter, while parameters obtained from original data as first-level parameters. It sounds easy to obtain the first-level parameters and also the second-level ones, but how about the third-level, fourth-level, etc? Or when the supervisor asks the student to repeat the observation for additional original data? Then, the whole processes must be repeated again. There might also aditional requests for visualization from previous original data compared to additional data. How the students can manage that? Using spreadsheet to save and analyze data is efficient when the original data is a structured data with single modality. For more complex data with multiple modalities and mixture of structured and unstructred data, a concept known as datalake, that is stored in shared folder, could be one of the possible answers. In this session the use of datalake for original data files stored on GitHub and processed using Python with Jupyter Notebook is discussed and some examples are shared.

Keywords: physics research, multimodal data, datalake