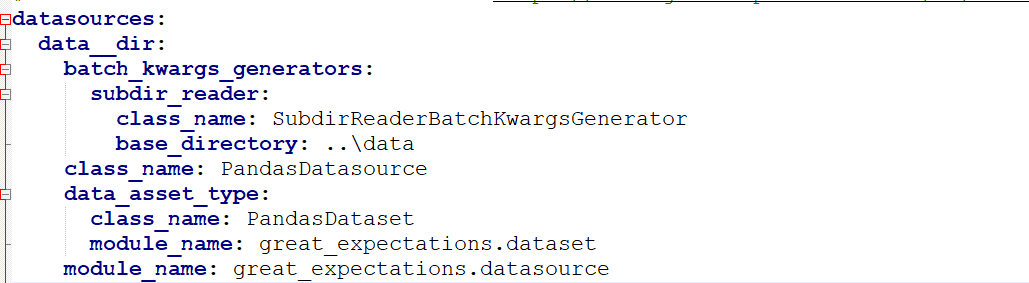
Great Expectation First Take

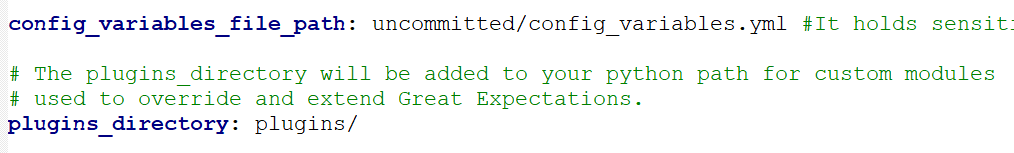
What is cls? Similar to self but cls is used when the method is a class method

What is @DocInherit? It overrides function’s \_\_get\_\_ method with one that will replace the local docstring with the docstring from its parent. It is defined in Dataset.uitl.

Basically, we need to edit the great\_expectations.yml file and push this to the repo and the pipeline or the agent will be able to execute the jobs. Below are the content of the great\_expectation.yml.



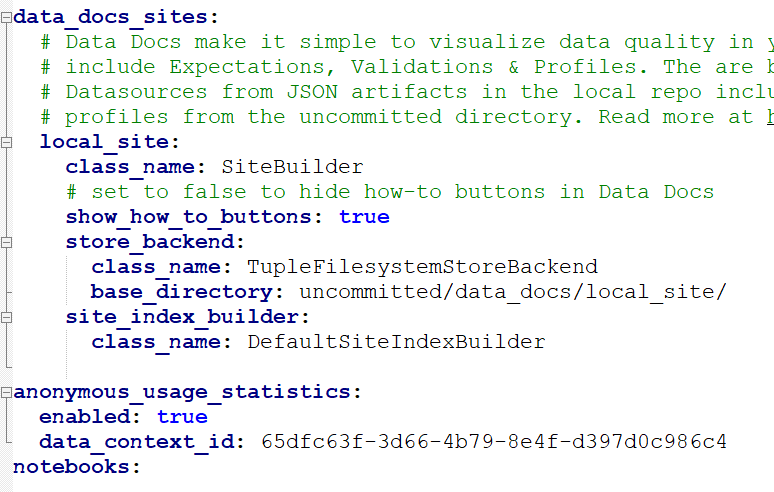
The datasources is to define the data source and extract the batch of the data for processing. In the demo, I used files (csv files) in the files system so the class\_name is PandasDataset. It works for more than csv file. Parquet, xls, xlsx, json and other common extensions are recognized. However, all batch\_kwargs will be extracted out and will be used for creating pandas dataframe for later use (for the cases of local files).



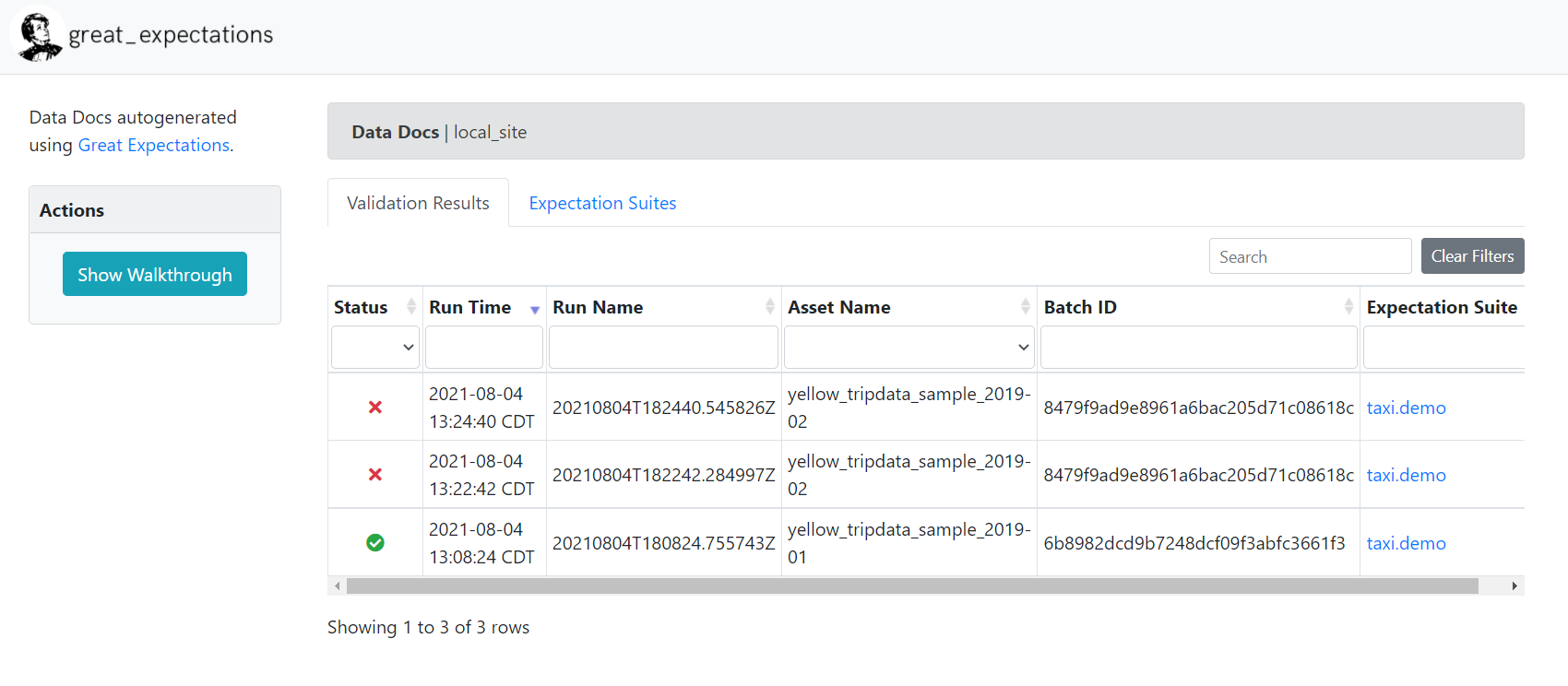
Config\_variables\_file\_path is the directory for saving config\_variables.yml which holds sensitive secrets such as database credential. Plugins is to add more features or for more customizations.

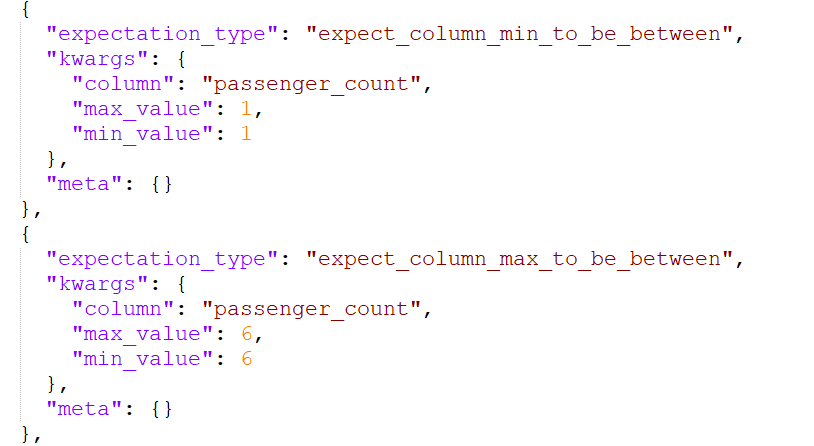


These stores are directories to store different outcomes or parameters. Expectation\_store stores the json files for different suites. The json files explain what expectation types were used. Expectation types are like metrics before. For example, max and min values. And we can set these values and apply to only specific columns. Validation stores validation results. Like how many expectations are tested, how many of them failed and what are the unexpected values in the expectations. Validation results are in json format. Checkpoints are the previous data. Mostly should be the good ones so that we can use them to validate new data.



Data docs are the html file of the expectation suite results, they are the visualizations of the outcome. There is home page for them and can monitor different projects or suites. Home page is like below.



Metrics and tests are expectations in this library. The collection of them is an expectation suite. See below. 

From my own experiences, it does not work like user drop off a file then the output would come out. It went through some python codes in notebook. It would be better if modifying the pipeline to operate those python modules once the target datafile is pushed to the repo. I will upload the notebook later.