# Business requirements

Input data set contains 3 csv files: airports, carriers, flights



Business description and requirements for source data can be found in Flights\_Data\_Set\_Requirements.xlsx.



1. Data set that should be processed is uploaded to ‘source’ folder in GSI data lake (csv)
2. Target data is stored in ‘Raw’ folder (parquet)

# YOUR TASK

1. **Provide a** **check list** for required data quality checks. Please use the following excel format for this purpose:



1. **Implement DQ using Jupyter Notebook**
   1. Create new notebook in your environment (you should receive a personal link to environment)
   2. Implement required DQ checks

Approach for implementing DQ checks:

Create and use share functions for similar DQ checks.

E.g. create a function to check PK uniqueness with the following arguments

* path to parquet file
* list of columns to be unique.

Call this function as many times as many tables we do require to pass this check.

Do the same for other checks if they are similar.

In case you have a unique DQ check, you can implement it in any way you’d like.

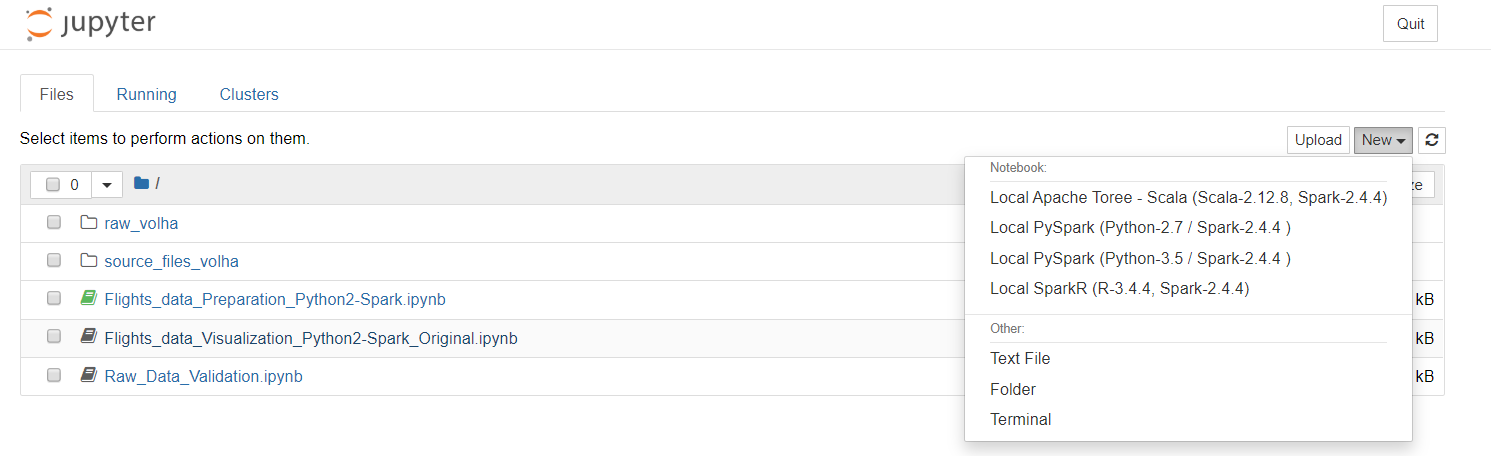
* 1. As a result, please print a table with DQ results. It should look like table below (slight changes acceptable)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| # | Table | DQ check | Column | Status | Bad Data |
| 1 | Carrier | Uniqueness | Code | Failed | List of values that did not pass your check.  E.g. code= ‘ABC’ |
| … | … | … |  | … | … |

# GSI structure

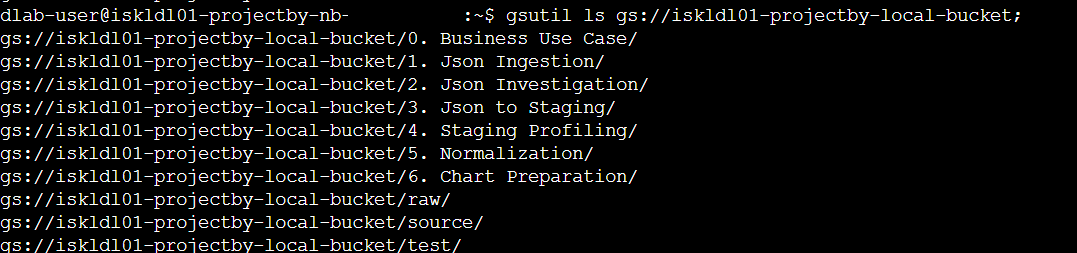
Working with GSI data lake is organized through terminal.

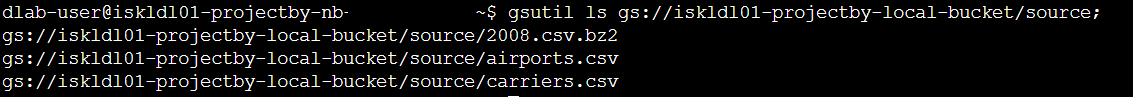
1. Click on Jupyter link -> you will be redirected to Jupyter UI
2. Click on new -> terminal to open it

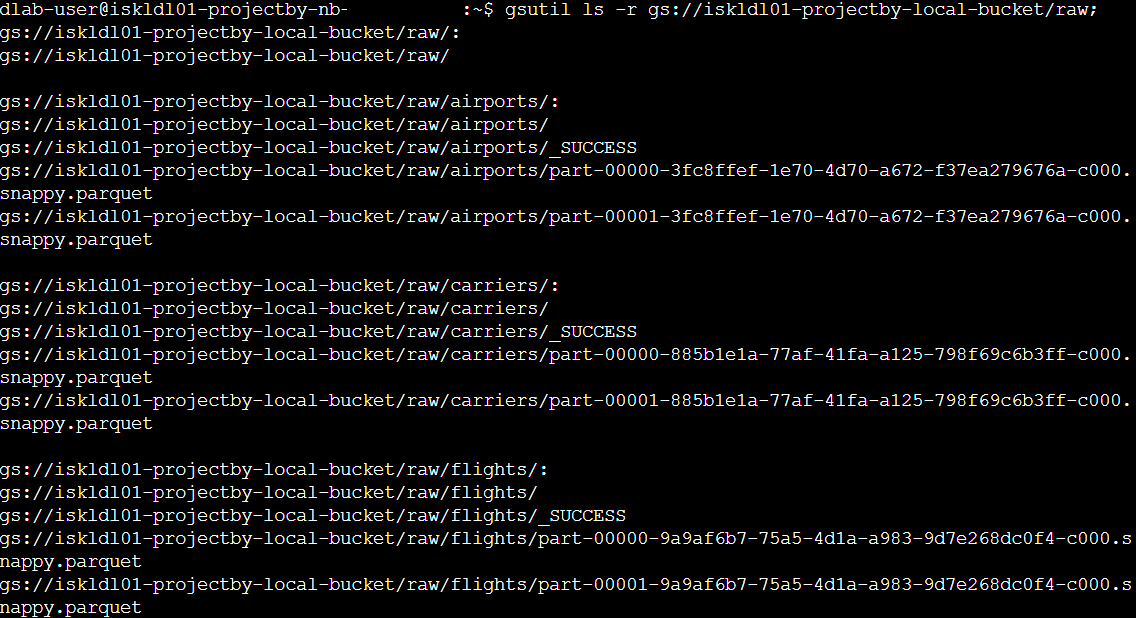


1. Use gsutil commands to work with data lake <https://cloud.google.com/storage/docs/gsutil/commands/ls>
2. Name of bucket where all the data for the task is stored: i**skldl02-projectby-local-bucket**
3. You should be able to see 2 folders under the bucket:

* ‘source’ containing 3 source files required for processing data
* ‘raw’ –contains target data



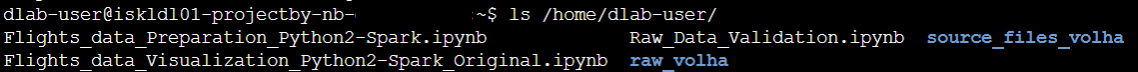




1. Please note that GSI files can ONLY be accessed through terminal and gsutil commands.

**DO NOT ADD/MODIFY/DELETE ANY FILE IN BUCKET** because they are shared between all members

1. On the Jupyter UI you can see data stored locally for dlab user. Default location is /home/dlab-user. This data can also be accessed through the terminal as well as through the UI.



You can work with files (upload/delete) in your personal environment with no restrictions.

1. To create new notebook please go to New -> Local PySpark.