gcp$_{toolkit}$

Release 0.0.1

Oct 10, 2020

Contents

gcp_toolkit is a Python package that makes it easier to perform common operations on **Google Cloud Platform**, such as moving data from a BigQuery table to Storage or loading it into a **pandas** Data Frame.

CHAPTER 1

Contents

1.1 Tutorial

With an activated Python 3 virtual env, clone the repository into your project root folder, install required libraries and copy the package from inside:

```
git clone https://github.com/dougpm/gcp_toolkit.git && \
cp -r gcp_toolkit/gcp_toolkit gcp_toolkit2 && \
cp gcp_toolkit/requirements.txt . && \
pip install -r requirements.txt && \
rm -rf gcp_toolkit && \
mv gcp_toolkit2 gcp_toolkit
```

1.1.1 io module

Using the IO class:

```
import gcp_toolkit as gtk

io = gtk.IO('your-bucket-name', 'your-dataset-name')
```

This automatically creates google.cloud.storage and google.cloud.bigquery Client instances, but you can pass your own to the constructor if you need to specify details.

Note: You must have Create Table permissions on the specified dataset.

Loading data from BigQuery into a pandas Data Frame:

```
df = io.bq_to_df('SELECT fields FROM dataset.table_name')
```

Loading data from pandas Data Frame into BigQuery:

```
io.df_to_bq(df, 'dataset.table_name')
```

Loading data from Storage bucket into pandas Data Frame:

```
df = io.bucket_to_df('path/to/bucket/files/files_prefix*')
```

Moving data from BigQuery to Storage:

```
df = io.bq_to_bucket('SELECT fields FROM dataset.table_name',
                     'path/to/files/file_name')
```

Note: The above may fail occasionally due to the table being too big to be extracted to a single file. In that case, you must add a '*' wildcard to the file name, like so:

```
df = io.bq_to_bucket('SELECT fields FROM dataset.table_name',
                     'path/to/files/file_name*')
```

1.2 gcp_toolkit

1.2.1 gcp toolkit

io module

class gcp_toolkit.io.**IO**(*bucket_name*, *staging_dataset*, *bq_client=None*, *storage_client=None*)

Bases: object

bq_to_bucket(*query*, *path_to_file*)

Runs query in BigQuery and stores results in Storage

bq_to_df(*query*, *use_builtin=False*)

Runs a query in BigQuery and loads the results into a pandas Data Frame

bucket_to_bq(*path_to_file*, *table_id*, *schema=[]*, *csv_delimiter=','*)

Loads a csv from Storage into a BigQuery table

bucket_to_df(*path_to_file*)

Reads a file or a group of files matching a pattern into a pandas Data Frame

df_to_bq(*df*, *table_id*, *schema=[]*)

Loads a pandas Data Frame into a BigQuery table.

utils module

gcp_toolkit.utils.**change_table_schema**(*table_id*, *new_fields*, *bigquery_client=None*)

Updates a table schema with new fields

gcp_toolkit.utils.**convert_pandas_gbq_schema**(*schema*)

Converts a BigQuery schema in the format of a list of dicts, used by pandas gbq into a list of SchemaFields

gcp_toolkit.utils.**create_bucket_folder**(*bucket_name*, *folder_name*, *storage_client=None*)

Creates a Folder in Storage

Python Module Index

g

Index