gwrappy Documentation

Release 0.4.3

Daniel Poon

December 28, 2016

Contents

CHAPTER 1

Indices and tables

- genindex
- modindex
- search

Contents:

1.1 gwrappy

User friendly wrapper for Google APIs.

1.1.1 Features

- **Easily connect to the following Google APIs (more to come eventually)**
  - BigQuery
  - Cloud Storage
  - Drive
  - Gmail
  - Compute Engine

```
# BigQuery
from gwrappy.bigquery import BigqueryUtility
bq_obj = BigqueryUtility()
results = bq_obj.sync_query('my_project', 'SELECT * FROM [foo.bar]')

# Cloud Storage
from gwrappy.storage import GcsUtility
gcs_obj = GcsUtility()
gcs_obj.download_object('bucket_name', 'object_name', 'path/to/write')
gcs_obj.upload_object('bucket_name', 'object_name', 'path/to/read')

# Drive
from gwrappy.drive import DriveUtility
drive_obj = DriveUtility(json_credentials_path, client_id)
drive_obj.download_object('file_id', 'path/to/write')
drive_obj.upload_file('path/to/read')
```

```
# Gmail
from gwrappy.gmail import GmailUtility
gmail_obj = GmailUtility(json_credentials_path, client_id)
gmail_obj.send_email(sender='Daniel Poon', to=['recipient_1@xx.com', 'recipient_2@yy.com'], subject=
```

1.1.2 Installation

```
$ pip install gwrappy
```

1.1.3 Authors

Development Lead

- Daniel Poon <daniel.poon.wenjie@gmail.com>

Contributors

None yet. Why not be the first?

1.1.4 Contributing

Contributions are welcome, and they are greatly appreciated! Every little bit helps, and credit will always be given.

You can contribute in many ways:

Types of Contributions

Report Bugs

Report bugs at https://github.com/danielpoonwj/gwrappy/issues.

If you are reporting a bug, please include:

- Your operating system name and version.
- Any details about your local setup that might be helpful in troubleshooting.
- Detailed steps to reproduce the bug.

Fix Bugs

Look through the GitHub issues for bugs. Anything tagged with "bug" and "help wanted" is open to whoever wants to implement it.

Implement Features

Look through the GitHub issues for features. Anything tagged with "enhancement" and "help wanted" is open to whoever wants to implement it.

Write Documentation

gwrappy could always use more documentation, whether as part of the official gwrappy docs, in docstrings, or even on the web in blog posts, articles, and such.

Submit Feedback

The best way to send feedback is to file an issue at https://github.com/danielpoonwj/gwrappy/issues.

If you are proposing a feature:

- Explain in detail how it would work.
- Keep the scope as narrow as possible, to make it easier to implement.
- Remember that this is a volunteer-driven project, and that contributions are welcome :)

Get Started!

Ready to contribute? Here's how to set up *gwrappy* for local development.

1. Fork the *gwrappy* repo on GitHub.
2. Clone your fork locally:

```
$ git clone git@github.com:your_name_here/gwrappy.git
```

3. Install your local copy into a virtualenv. Assuming you have virtualenvwrapper installed, this is how you set up your fork for local development:

```
$ mkvirtualenv gwrappy
$ cd gwrappy/
$ python setup.py develop
```

4. Create a branch for local development:

```
$ git checkout -b name-of-your-bugfix-or-feature
```

   Now you can make your changes locally.

5. Commit your changes and push your branch to GitHub:

```
$ git add .
$ git commit -m "Your detailed description of your changes."
$ git push origin name-of-your-bugfix-or-feature
```

6. Submit a pull request through the GitHub website.

Pull Request Guidelines

Before you submit a pull request, check that it meets these guidelines:

1. If the pull request adds functionality, the docs should be updated. Put your new functionality into a function with a docstring, and add the feature to the list in README.rst.

1.1.5 History

History

0.4.3 (2016-12-16)

- Bugfix in gwrappy.drive.DriveUtility.upload_file()

0.4.2 (2016-12-16)

- Bugfix in gwrappy.dataproc.DataprocUtility.delete_cluster()

0.4.1 (2016-12-14)

- **Breaking Changes:**
  - gwrappy.dataproc.DataprocUtility takes project_id in its constructor rather than a method parameter.
- gwrappy.dataproc.DataprocUtility operation/job methods return Response objects when wait_finish=True
- pandas dependency removed from requirements.txt as its functionality is limited to specific functions and largely unnecessary otherwise.

0.4.0 (2016-11-21)

- Added gwrappy.dataproc for Google Dataproc
- **Minor Changes**
  - gwrappy.storage.GcsUtility.update_object() added
  - Added ability to set object Acl on upload with gwrappy.storage.GcsUtility.upload_object()

0.3.0 (2016-10-31)

- **Python 3 compatibility**
  - Most API functions were already compatible, most changes were done for the utilities functions.
- **Minor Bugfixes/Changes**
  - BigqueryUtility().poll_resp_list() now doesn't break once an exception is encountered. The respective Error object is returned and job checking is uninterrupted.
  - Fixed int columns being interpreted as float for pandas 0.19.0 when querying to dataframe.

0.2.1 (2016-10-20)

- **Minor Bugfixes:**
  - bigquery.utils.read_sql properly checks kwargs.
  - BigqueryUtility queries with return_type='dataframe' uses inferred dtypes for integer columns to stop pandas from breaking if column contains NaN.

0.2.0 (2016-09-27)

- Added gwrappy.compute for Google Compute Engine.
- **Minor Bugfixes:**
  - drive.DriveUtility.list_files(): Removed fields, added orderBy and filter_exp.
  - bigquery.utils.JobResponse: time_taken in __repr__ for some job types fixed.

0.1.6 (2016-09-08)

- **Added more utilities**
  - utils.month_range: Chunk dates into months.
  - utils.simple_mail: Send basic emails for alerts or testing. *Note*: For greater security and flexibility, do still use the gmail functionality within this package.
  - utils.StringLogger: Simply wrapper for logging with a string handler and convenience functions for retrieving logs as a string.
- Added dateutil as a dependency

0.1.5 (2016-08-30)

- list methods now return a generator for memory efficiency
- **BigQuery:**
  - list_jobs takes 2 new args *projection* and *earliest_date*
- Documentation updates

0.1.4 (2016-08-29)

- gwrappy.errors no longer imports service specific error objects. To access JobError, import it from gwrappy.bigquery.errors
- simple date range generator function added to gwrappy.utils

0.1.3 (2016-08-23)

- **BigQuery:**
  - **JobResponse now only sets time_taken if data is available.**
    - Fixed bug that raised KeyError when wait_finish=False, since endTime was unavailable in the API response.
  - poll_resp_list returns JobReponse objects. Also propagates 'description' attribute if available.

0.1.2 (2016-08-19)

- Bug Fixes
- Documentation updates

0.1.1 (2016-08-16)

- Completed docstrings and amendments to documentation
- Added list_to_html under gwrappy.gmail.utils
- Added tabulate as a dependency

0.1.0 (2016-08-15)

- New and improved version of https://github.com/danielpoonwj/gcloud_custom_utilities
- First release on PyPI.

1.1.6 Credits

This package was created with Cookiecutter and the audreyr/cookiecutter-pypackage project template.

1.2 Installation

1.2.1 Stable release

To install gwrappy, run this command in your terminal:

```
$ pip install gwrappy
```

This is the preferred method to install gwrappy, as it will always install the most recent stable release.

If you don't have pip installed, this Python installation guide can guide you through the process.

1.2.2 From sources

The sources for gwrappy can be downloaded from the Github repo.

You can either clone the public repository:

```
$ git clone git://github.com/danielpoonwj/gwrappy
```

Or download the tarball:

```
$ curl  -OL https://github.com/danielpoonwj/gwrappy/tarball/master
```

Once you have a copy of the source, you can install it with:

```
$ python setup.py install
```

1.3 Usage

This package is designed to take the pain out of using Google's powerful APIs. Focus on using them instead of getting them to work.

Each API is accessible through a specific Utility found within each subpackage. These Utilities are objects that initiate and authenticate the service objects required to access the each API's functionality.

1.3.1 Authentication

It is **highly** recommended to download the wonderful gcloud SDK [https://cloud.google.com/sdk/] as a complementary tool. For one, it allows you to simplify access to Google Cloud Platform services using Application Default Credentials. Otherwise, credentials would have to be authenticated and a credentials file would have to be stored etc etc.

If the gcloud SDK has been installed and configured to the desired user and project, authentication is seamless.

```
from gwrappy.bigquery import BigqueryUtility
bq_obj = BigqueryUtility()
```

Application Default Credentials are only applicable for services under the Google Cloud Platform. For other services, such as Gmail or Drive, unfortunately the process is *slightly* less elegant. Also, if multiple user credentials are required, this solution may actually be more convenient. Authentication flow with client_secret.json is only required once per service, thereafter the credentials are stored as a value in credentials.json, with the key being the client_id.

```
from gwrappy.bigquery import BigqueryUtility
secret_path = path/to/client_secret.json
cred_path = path/to/credentials.json
client_id = me@gmail.com

bq_obj = BigqueryUtility(
    client_secret_path=secret_path,
    json_credentials_path=cred_path,
    client_id=client_id
)
```

1.3.2 Class Methods

Once the Utility object has been initialized, accessing methods within the object are generally wrapped APi calls. For more information on accepted kwargs, please visit the respective method's documentation.

1.3.3 Working with Response Objects

Some methods return Response objects eg. gwrappy.bigquery.utils.JobResponse. These objects are generally parsing the JSON responses from the API, calculating statistics like time taken and size (if applicable) and converting it to a human-readable format.

Should the original API response be required for custom logging or other reasons, access the **resp** variable within the Response object.

CHAPTER 2

Documentation

2.1 BigQuery

2.1.1 BigqueryUtility

class gwrappy.bigquery.**BigqueryUtility**(***kwargs*)
Initializes object for interacting with Bigquery API.

By default, Application Default Credentials are used.
If gcloud SDK isn't installed, credential files have to be specified using the kwargs *json_credentials_path* and *client_id*.

Parameters

- **max_retries** (*integer*) – Argument specified with each API call to natively handle retryable errors.
- **client_secret_path** – File path for client secret JSON file. Only required if credentials are invalid or unavailable.
- **json_credentials_path** – File path for automatically generated credentials.
- **client_id** – Credentials are stored as a key-value pair per client_id to facilitate multiple clients using the same credentials file. For simplicity, using one's email address is sufficient.

list_projects(*max_results=None*, *filter_exp=None*)
Abstraction of projects().list() method with inbuilt iteration functionality. [https://cloud.google.com/bigquery/docs/reference/v2/projects/list]

Parameters

- **max_results** (*integer*) – If None, all results are iterated over and returned.
- **filter_exp** (*function*) – Function that filters entries if filter_exp evaluates to True.

Returns List of dictionary objects representing project resources.

list_jobs(*project_id*, *state_filter=None*, *show_all=False*, *projection='full'*, *max_results=None*, *earliest_date=None*, *filter_exp=None*)
Abstraction of jobs().list() method with inbuilt iteration functionality. [https://cloud.google.com/bigquery/docs/reference/v2/jobs/list]

Note - All jobs are stored in BigQuery. Do set *max_results* or *earliest_date* to limit data returned.

Parameters

- **project_id** (*string*) – Unique project identifier.
- **state_filter** (*string*) – Pre-filter API request for job state. Acceptable values are "done", "pending" and "running". [Equivalent API param: stateFilter]
- **show_all** (*boolean*) – Whether to display jobs owned by all users in the project. [Equivalent API param: allUsers]
- **max_results** (*integer*) – If None, all results are iterated over and returned.
- **projection** (*string*) – Acceptable values are *'full'*, *'minimal'*. *'full'* includes job configuration.
- **earliest_date** (*datetime object or string representation of datetime in %Y-%m-%d format.*) – Only returns data after this date.
- **filter_exp** (*function*) – Function that filters entries if filter_exp evaluates to True.

Returns List of dictionary objects representing job resources.

list_datasets(*project_id*, *show_all=False*, *max_results=None*, *filter_exp=None*)

Abstraction of datasets().list() method with inbuilt iteration functionality. [https://cloud.google.com/bigquery/docs/reference/v2/datasets/list]

Parameters

- **project_id** (*string*) – Unique project identifier.
- **show_all** (*boolean*) – Include hidden datasets generated when running queries on the UI.
- **max_results** (*integer*) – If None, all results are iterated over and returned.
- **filter_exp** (*function*) – Function that filters entries if filter_exp evaluates to True.

Returns List of dictionary objects representing dataset resources.

list_tables(*project_id*, *dataset_id*, *max_results=None*, *filter_exp=None*)

Abstraction of tables().list() method with inbuilt iteration functionality. [https://cloud.google.com/bigquery/docs/reference/v2/tables/list]

Parameters

- **project_id** (*string*) – Unique project identifier.
- **dataset_id** (*string*) – Unique dataset identifier.
- **max_results** (*integer*) – If None, all results are iterated over and returned.
- **filter_exp** (*function*) – Function that filters entries if filter_exp evaluates to True.

Returns List of dictionary objects representing table resources.

get_job(*project_id*, *job_id*)

Abstraction of jobs().get() method. [https://cloud.google.com/bigquery/docs/reference/v2/jobs/get]

Parameters

- **project_id** (*string*) – Unique project identifier.
- **job_id** (*string*) – Unique job identifier.

Returns Dictionary object representing job resource.

get_table_info(*project_id*, *dataset_id*, *table_id*)

Abstraction of tables().get() method. [https://cloud.google.com/bigquery/docs/reference/v2/tables/get]

Parameters

- **project_id** (*string*) – Unique project identifier.
- **dataset_id** (*string*) – Unique dataset identifier.
- **table_id** (*string*) – Unique table identifier.

Returns Dictionary object representing table resource.

delete_table(*project_id*, *dataset_id*, *table_id*)
Abstraction of tables().delete() method. [https://cloud.google.com/bigquery/docs/reference/v2/tables/delete]

Parameters

- **project_id** (*string*) – Unique project identifier.
- **dataset_id** (*string*) – Unique dataset identifier.
- **table_id** (*string*) – Unique table identifier.

Raises AssertionError if unsuccessful. Response should be empty string if successful.

poll_job_status(*job_resp*, *sleep_time=1*)
Check status of job until status is "DONE".

Parameters

- **job_resp** (*dictionary*) – Representation of job resource.
- **sleep_time** (*integer*) – Time to pause (seconds) between polls.

Returns Dictionary object representing job resource's final state.

Raises JobError object if an error is discovered after job finishes running.

sync_query(*project_id*, *query*, *return_type='list'*, *sleep_time=1*, *dry_run=False*, ***kwargs*)
Abstraction of jobs().query() method, iterating and parsing query results. [https://cloud.google.com/bigquery/docs/reference/v2/jobs/query]

Parameters

- **project_id** (*string*) – Unique project identifier.
- **query** (*string*) – SQL query
- **return_type** (*string*) – Format for result to be returned. Accepted types are "list", "dataframe", and "json".
- **sleep_time** (*integer*) – Time to pause (seconds) between polls.
- **dry_run** (*boolean*) – Basic statistics about the query, without actually running it. Mainly for testing or estimating amount of data processed.
- **useLegacySql** – Toggle between Legacy and Standard SQL.

Returns If not dry_run: result in specified type, JobResponse object. If dry_run: Dictionary object representing expected query statistics.

Raises JobError object if an error is discovered after job finishes running.

async_query(*project_id*, *query*, *dest_project_id*, *dest_dataset_id*, *dest_table_id*, *udf=None*, *return_type='list'*, *sleep_time=1*, ***kwargs*)
Abstraction of jobs().insert() method for **query** job, iterating and parsing query results. [https://cloud.google.com/bigquery/docs/reference/v2/jobs/insert]

Asynchronous queries always write to an intermediate (destination) table.

This query method is preferable over sync_query if:
1. Large results are returned.
2. UDF functions are required.
3. Results returned also need to be stored in a table.

Parameters

- **project_id** (*string*) – Unique project identifier.
- **query** (*string*) – SQL query
- **dest_project_id** (*string*) – Unique project identifier of destination table.
- **dest_dataset_id** (*string*) – Unique dataset identifier of destination table.
- **dest_table_id** (*string*) – Unique table identifier of destination table.
- **udf** (*string or list*) – One or more UDF functions if required by the query.
- **return_type** (*string*) – Format for result to be returned. Accepted types are "list", "dataframe", and "json".
- **sleep_time** (*integer*) – Time to pause (seconds) between polls.
- **useLegacySql** – Toggle between Legacy and Standard SQL.
- **writeDisposition** – (Optional) Config kwarg that determines table writing behaviour.

Returns result in specified type, JobResponse object.

Raises JobError object if an error is discovered after job finishes running.

write_table(*project_id*, *query*, *dest_project_id*, *dest_dataset_id*, *dest_table_id*, *udf=None*, *wait_finish=True*, *sleep_time=1*, ***kwargs*)

Abstraction of jobs().insert() method for **query** job, without returning results. [https://cloud.google.com/bigquery/docs/reference/v2/jobs/insert]

Parameters

- **project_id** (*string*) – Unique project identifier.
- **query** (*string*) – SQL query
- **dest_project_id** (*string*) – Unique project identifier of destination table.
- **dest_dataset_id** (*string*) – Unique dataset identifier of destination table.
- **dest_table_id** (*string*) – Unique table identifier of destination table.
- **udf** (*string or list*) – One or more UDF functions if required by the query.
- **wait_finish** (*boolean*) – Flag whether to poll job till completion. If set to false, multiple jobs can be submitted, repsonses stored, iterated over and polled till completion afterwards.
- **sleep_time** (*integer*) – Time to pause (seconds) between polls.
- **useLegacySql** – Toggle between Legacy and Standard SQL.
- **writeDisposition** – (Optional) Config kwarg that determines table writing behaviour.

Returns If wait_finish: result in specified type, JobResponse object. If not wait_finish: JobResponse object.

Raises If wait_finish: JobError object if an error is discovered after job finishes running.

write_view(*query*, *dest_project_id*, *dest_dataset_id*, *dest_table_id*, *udf=None*, *overwrite_existing=True*, ***kwargs*)

Views are analogous to a virtual table, functioning as a table but only returning results from the underlying query when called.

Parameters

- **query** (*string*) – SQL query
- **dest_project_id** (*string*) – Unique project identifier of destination table.
- **dest_dataset_id** (*string*) – Unique dataset identifier of destination table.
- **dest_table_id** (*string*) – Unique table identifier of destination table.
- **udf** (*string or list*) – One or more UDF functions if required by the query.
- **overwrite_existing** – Safety flag, would raise HttpError if table exists and overwrite_existing=False
- **useLegacySql** – Toggle between Legacy and Standard SQL.

Returns TableResponse object for the newly inserted table

load_from_gcs(*dest_project_id*, *dest_dataset_id*, *dest_table_id*, *schema*, *source_uris*, *wait_finish=True*, *sleep_time=1*, ***kwargs*)

For loading data from Google Cloud Storage.

Abstraction of jobs().insert() method for **load** job.
[https://cloud.google.com/bigquery/docs/reference/v2/jobs/insert]

Parameters

- **dest_project_id** (*string*) – Unique project identifier of destination table.
- **dest_dataset_id** (*string*) – Unique dataset identifier of destination table.
- **dest_table_id** (*string*) – Unique table identifier of destination table.
- **schema** (*list of dictionaries*) – Schema of input data (schema.fields[]) [https://cloud.google.com/bigquery/docs/reference/v2/tables]
- **source_uris** (*string or list*) – One or more uris referencing GCS objects
- **wait_finish** (*boolean*) – Flag whether to poll job till completion. If set to false, multiple jobs can be submitted, repsonses stored, iterated over and polled till completion afterwards.
- **sleep_time** (*integer*) – Time to pause (seconds) between polls.
- **writeDisposition** – Determines table writing behaviour.
- **sourceFormat** – Indicates format of input data.
- **skipLeadingRows** – Leading rows to skip. Defaults to 1 to account for headers if sourceFormat is CSV or default, 0 otherwise.
- **fieldDelimiter** – Indicates field delimiter.
- **allowQuotedNewlines** – Indicates presence of quoted newlines in fields.
- **allowJaggedRows** – Accept rows that are missing trailing optional columns. (Only CSV)

- **ignoreUnknownValues** – Allow extra values that are not represented in the table schema.
- **maxBadRecords** – Maximum number of bad records that BigQuery can ignore when running the job.

Returns JobResponse object

export_to_gcs(*source_project_id*, *source_dataset_id*, *source_table_id*, *dest_uris*, *wait_finish=True*, *sleep_time=1*, ***kwargs*)

For exporting data into Google Cloud Storage.
Abstraction of jobs().insert() method for **extract** job.
[https://cloud.google.com/bigquery/docs/reference/v2/jobs/insert]

Parameters

- **source_project_id** (*string*) – Unique project identifier of source table.
- **source_dataset_id** (*string*) – Unique dataset identifier of source table.
- **source_table_id** (*string*) – Unique table identifier of source table.
- **dest_uris** (*string or list*) – One or more uris referencing GCS objects
- **wait_finish** (*boolean*) – Flag whether to poll job till completion. If set to false, multiple jobs can be submitted, repsonses stored, iterated over and polled till completion afterwards.
- **sleep_time** (*integer*) – Time to pause (seconds) between polls.
- **destinationFormat** – (Optional) Config kwarg that indicates format of output data.
- **compression** – (Optional) Config kwarg for type of compression applied.
- **fieldDelimiter** – (Optional) Config kwarg that indicates field delimiter.
- **printHeader** – (Optional) Config kwarg indicating if table headers should be written.

Returns JobResponse object

copy_table(*source_data*, *dest_project_id*, *dest_dataset_id*, *dest_table_id*, *wait_finish=True*, *sleep_time=1*, ***kwargs*)

For copying existing table(s) to a new or existing table.
Abstraction of jobs().insert() method for **copy** job.
[https://cloud.google.com/bigquery/docs/reference/v2/jobs/insert]

Parameters

- **source_data** – Representations of single or multiple existing tables to copy from.
- **source_date** – dictionary or list of dictionaries
- **dest_project_id** (*string*) – Unique project identifier of destination table.
- **dest_dataset_id** (*string*) – Unique dataset identifier of destination table.
- **dest_table_id** (*string*) – Unique table identifier of destination table.
- **wait_finish** (*boolean*) – Flag whether to poll job till completion. If set to false, multiple jobs can be submitted, repsonses stored, iterated over and polled till completion afterwards.

- **sleep_time** (*integer*) – Time to pause (seconds) between polls.
- **writeDisposition** – (Optional) Config kwarg that determines table writing behaviour.

Returns JobResponse object

load_from_string(*dest_project_id*, *dest_dataset_id*, *dest_table_id*, *schema*, *load_string*, *wait_finish=True*, *sleep_time=1*, ***kwargs*)

For loading data from string representation of a file/object.
Can be used in conjunction with gwrappy.bigquery.utils.file_to_string()

Parameters

- **dest_project_id** (*string*) – Unique project identifier of destination table.
- **dest_dataset_id** (*string*) – Unique dataset identifier of destination table.
- **dest_table_id** (*string*) – Unique table identifier of destination table.
- **schema** (*list of dictionaries*) – Schema of input data (schema.fields[]) [https://cloud.google.com/bigquery/docs/reference/v2/tables]
- **load_string** (*string*) – String representation of an object.
- **wait_finish** (*boolean*) – Flag whether to poll job till completion. If set to false, multiple jobs can be submitted, repsonses stored, iterated over and polled till completion afterwards.
- **sleep_time** (*integer*) – Time to pause (seconds) between polls.
- **writeDisposition** – (Optional) Config kwarg that determines table writing behaviour.
- **sourceFormat** – (Optional) Config kwarg that indicates format of input data.
- **skipLeadingRows** – (Optional) Config kwarg for leading rows to skip. Defaults to 1 to account for headers if sourceFormat is CSV or default, 0 otherwise.
- **fieldDelimiter** – (Optional) Config kwarg that indicates field delimiter.
- **allowQuotedNewlines** – (Optional) Config kwarg indicating presence of quoted newlines in fields.
- **allowJaggedRows** – Accept rows that are missing trailing optional columns. (Only CSV)
- **ignoreUnknownValues** – Allow extra values that are not represented in the table schema.
- **maxBadRecords** – Maximum number of bad records that BigQuery can ignore when running the job.

Returns JobResponse object

write_federated_table(*dest_project_id*, *dest_dataset_id*, *dest_table_id*, *schema*, *source_uris*, *overwrite_existing=True*, ***kwargs*)

Imagine a View for Google Cloud Storage object(s).
Abstraction of jobs().insert() method for **load** job.
[https://cloud.google.com/bigquery/docs/reference/v2/jobs/insert]

Parameters

- **dest_project_id** (*string*) – Unique project identifier of destination table.
- **dest_dataset_id** (*string*) – Unique dataset identifier of destination table.
- **dest_table_id** (*string*) – Unique table identifier of destination table.
- **schema** (*list of dictionaries*) – Schema of input data (schema.fields[]) [https://cloud.google.com/bigquery/docs/reference/v2/tables]
- **source_uris** (*string or list*) – One or more uris referencing GCS objects
- **overwrite_existing** – Safety flag, would raise HttpError if table exists and overwrite_existing=False
- **sourceFormat** – (Optional) Config kwarg that indicates format of input data.
- **skipLeadingRows** – (Optional) Config kwarg for leading rows to skip. Defaults to 1 to account for headers if sourceFormat is CSV or default, 0 otherwise.
- **fieldDelimiter** – (Optional) Config kwarg that indicates field delimiter.
- **compression** – (Optional) Config kwarg for type of compression applied.
- **allowQuotedNewlines** – (Optional) Config kwarg indicating presence of quoted newlines in fields.

Returns TableResponse object

update_table_info(*project_id*, *dataset_id*, *table_id*, *table_description=None*, *schema=None*)

Abstraction of tables().patch() method. [https://cloud.google.com/bigquery/docs/reference/v2/tables/patch]

Parameters

- **project_id** (*string*) – Unique project identifier.
- **dataset_id** (*string*) – Unique dataset identifier.
- **table_id** (*string*) – Unique table identifier.
- **table_description** (*string*) – Optional description for table. If None, would not overwrite existing description.
- **schema_fields** –
- **schema** (*list of dictionaries*) – Schema fields to change (schema.fields[]) [https://cloud.google.com/bigquery/docs/reference/v2/tables]

Returns TableResponse

poll_resp_list(*response_list*, *sleep_time=1*)

Convenience function for iterating and polling list of responses collected with jobs wait_finish=False.

If any job fails, its respective Error object is returned to ensure errors would not break polling subsequent responses.

Parameters

- **response_list** (*list of dicts or JobResponse objects*) – List of response objects
- **sleep_time** (*integer*) – Time to pause (seconds) between polls.

Returns List of JobResponse or Error (JobError/HttpError) objects representing job resource's final state.

2.1.2 Misc Classes/Functions

class gwrappy.bigquery.utils.**JobResponse**(*resp*, *description=None*)
Wrapper for Bigquery job responses, mainly for calculating/parsing job statistics into human readable formats for logging.

Parameters

- **resp** (*dictionary*) – Dictionary representation of a job resource.
- **description** – Optional string descriptor for specific function of job.

class gwrappy.bigquery.utils.**TableResponse**(*resp*, *description=None*)
Wrapper for Bigquery table resources, mainly for calculating/parsing job statistics into human readable formats for logging.

Parameters

- **resp** (*dictionary*) – Dictionary representation of a table resource.
- **description** – Optional string descriptor for table.

gwrappy.bigquery.utils.**read_sql**(*read_path*, ***kwargs*)
Reads text file, performing string substitution using str.format() method if necessary.

Parameters

- **read_path** – File path containing SQL query.
- **kwargs** – Key-Value pairs referencing {key} within query for substitution.

Returns Query string.

gwrappy.bigquery.utils.**bq_schema_from_df**(*input_df*)
Derive Bigquery Schema from Pandas Dataframe object.

Parameters **input_df** – Pandas Dataframe object

Returns List of dictionaries which can be fed directly as Bigquery schemas.

gwrappy.bigquery.utils.**file_to_string**(*f*, *source_format='csv'*)
Specifically for BigqueryUtility().load_from_string()

Parameters

- **f** (*file path, list of lists/dicts, dataframe, or string representation of json list*) – Object to convert to string.
- **source_format** (*string*) – Indicates format of input data. Accepted values are "csv" and "json".

Returns String representation of object/file contents

2.2 Cloud Storage

2.2.1 GcsUtility

class gwrappy.storage.**GcsUtility**(***kwargs*)
Initializes object for interacting with Google Cloud Storage API.

By default, Application Default Credentials are used.

If gcloud SDK isn't installed, credential files have to be specified using the kwargs *json_credentials_path* and *client_id*.

Parameters

- **max_retries** (*integer*) – Argument specified with each API call to natively handle retryable errors.
- **chunksize** (*integer*) – Upload/Download chunk size
- **client_secret_path** – File path for client secret JSON file. Only required if credentials are invalid or unavailable.
- **json_credentials_path** – File path for automatically generated credentials.
- **client_id** – Credentials are stored as a key-value pair per client_id to facilitate multiple clients using the same credentials file. For simplicity, using one's email address is sufficient.

list_buckets(*project_id*, *max_results=None*, *filter_exp=None*)
Abstraction of buckets().list() method with inbuilt iteration functionality. [https://cloud.google.com/storage/docs/json_api/v1/buckets/list]

Parameters

- **project_id** (*string*) – Unique project identifier.
- **max_results** (*integer*) – If None, all results are iterated over and returned.
- **filter_exp** (*function*) – Function that filters entries if filter_exp evaluates to True.

Returns List of dictionary objects representing bucket resources.

list_objects(*bucket_name*, *max_results=None*, *prefix=None*, *projection=None*, *filter_exp=None*)
Abstraction of objects().list() method with inbuilt iteration functionality. [https://cloud.google.com/storage/docs/json_api/v1/objects/list]

Parameters

- **bucket_name** (*string*) – Bucket identifier.
- **max_results** (*integer*) – If None, all results are iterated over and returned.
- **prefix** (*string*) – Pre-filter (on API call) results to objects whose names begin with this prefix.
- **projection** – Set of properties to return.
- **filter_exp** (*function*) – Function that filters entries if filter_exp evaluates to True.

Returns List of dictionary objects representing object resources.

get_object(*bucket_name*, *object_name*, *projection=None*)
Abstraction of objects().get() method with inbuilt iteration functionality. [https://cloud.google.com/storage/docs/json_api/v1/objects/get]

Parameters

- **bucket_name** (*string*) – Bucket identifier.
- **object_name** (*list or string*) – Can take string representation of object resource or list denoting path to object on GCS.
- **projection** – Set of properties to return.

Returns Dictionary object representing object resource.

update_object(*bucket_name*, *object_name*, *predefined_acl=None*, *projection=None*, ***object_resource*)
Abstraction of objects().update() method. [https://cloud.google.com/storage/docs/json_api/v1/objects/update]

Parameters

- **bucket_name** (*string*) – Bucket identifier.
- **object_name** (*list or string*) – Can take string representation of object resource or list denoting path to object on GCS.
- **predefined_acl** – Apply a predefined set of access controls to this object.
- **projection** – Set of properties to return.
- **object_resource** – Supply optional properties [https://cloud.google.com/storage/docs/json_api/v1/objects/insert#request-body]

Returns Dictionary object representing object resource.

delete_object(*bucket_name*, *object_name*)
Abstraction of objects().delete() method with inbuilt iteration functionality. [https://cloud.google.com/storage/docs/json_api/v1/objects/delete]

Parameters

- **bucket_name** (*string*) – Bucket identifier.
- **object_name** (*list or string*) – Can take string representation of object resource or list denoting path to object on GCS.

Raises AssertionError if unsuccessful. Response should be empty string if successful.

download_object(*bucket_name*, *object_name*, *write_path*)
Downloads object in chunks.

Parameters

- **bucket_name** (*string*) – Bucket identifier.
- **object_name** (*list or string*) – Can take string representation of object resource or list denoting path to object on GCS.
- **write_path** (*string*) – Local path to write object to.

Returns GcsResponse object.

Raises HttpError if non-retryable errors are encountered.

upload_object(*bucket_name*, *object_name*, *read_path*, *predefined_acl=None*, *projection=None*, ***object_resource*)
Uploads object in chunks.

Optional parameters and valid object resources are listed here [https://cloud.google.com/storage/docs/json_api/v1/objects/insert]

Parameters

- **bucket_name** (*string*) – Bucket identifier.
- **object_name** (*list or string*) – Can take string representation of object resource or list denoting path to object on GCS.
- **read_path** (*string*) – Local path of object to upload.
- **predefined_acl** – Apply a predefined set of access controls to this object.
- **projection** – Set of properties to return.

- **object_resource** – Supply optional properties [https://cloud.google.com/storage/docs/json_api/v1/objects/insert#request-body]

Returns GcsResponse object.

Raises HttpError if non-retryable errors are encountered.

2.2.2 Misc Classes/Functions

class gwrappy.storage.utils.**GcsResponse**(*description*)

Wrapper for GCS upload and download responses, mainly for calculating/parsing job statistics into human readable formats for logging.

Parameters **description** – String descriptor for specific function of job.

load_resp(*resp*, *is_download*)

Loads json response from API.

Parameters

- **resp** (*dictionary*) – Response from API
- **is_download** (*boolean*) – Calculates time taken based on 'updated' field in response if upload, and based on stop time if download

2.3 Google Drive

2.3.1 DriveUtility

class gwrappy.drive.**DriveUtility**(*json_credentials_path*, *client_id*, ***kwargs*)

Initializes object for interacting with Bigquery API.

Parameters

- **client_secret_path** – File path for client secret JSON file. Only required if credentials are invalid or unavailable.
- **json_credentials_path** – File path for automatically generated credentials.
- **client_id** – Credentials are stored as a key-value pair per client_id to facilitate multiple clients using the same credentials file. For simplicity, using one's email address is sufficient.
- **max_retries** (*integer*) – Argument specified with each API call to natively handle retryable errors.
- **chunksize** (*integer*) – Upload/Download chunk size

get_account_info(*fields=None*)

Abstraction of about().get() method. [https://developers.google.com/drive/v3/reference/about/get]

Parameters **fields** (*list or ", " delimited string*) – Available properties can be found here: https://developers.google.com/drive/v3/reference/about

Returns Dictionary object representation of About resource.

list_files(*max_results=None*, ***kwargs*)

Abstraction of files().list() method with inbuilt iteration functionality. [https://developers.google.com/drive/v3/reference/files/list]

Parameters

- **max_results** (*integer*) – If None, all results are iterated over and returned.
- **orderBy** – List of keys to sort by. Refer to documentation.
- **spaces** – A comma-separated list of spaces to query within the corpus. Supported values are 'drive', 'appDataFolder' and 'photos'.
- **q** – A query for filtering the file results. Reference here: https://developers.google.com/drive/v3/web/search-parameters

Returns List of dictionary objects representing file resources.

get_file(*file_id*, *fields=None*)
Get file metadata.

Parameters

- **file_id** (*string*) – Unique file id. Check on UI or by list_files().
- **fields** (*list or ", " delimited string*) – Available properties can be found here: https://developers.google.com/drive/v3/reference/about

Returns Dictionary object representing file resource.

download_file(*file_id*, *write_path*, *page_num=None*, *output_type=None*)
Downloads object.

Parameters

- **file_id** (*string*) – Unique file id. Check on UI or by list_files().
- **write_path** (*string*) – Local path to write object to.
- **page_num** (*integer*) – Only applicable to Google Sheets. Check **gid** param in URL.
- **output_type** (*string. 'list' or 'dataframe'*) – Only applicable to Google Sheets. Can be directly downloaded as list or Pandas dataframe.

Returns If Google Sheet and output_type specified: result in selected type, DriveResponse object. Else DriveResponse object.

Raises HttpError if non-retryable errors are encountered.

upload_file(*read_path*, *overwrite_existing=True*, ***kwargs*)
Creates file if it doesn't exist, updates if it does.

Parameters

- **read_path** (*string*) – Local path of object to upload.
- **overwrite_existing** (*boolean*) – Safety flag, would raise ValueError if object exists and overwrite_existing=False
- **kwargs** – Key-Value pairs of Request Body params. Reference here: https://developers.google.com/drive/v3/reference/files

Returns DriveResponse object.

2.3.2 Misc Classes/Functions

class gwrappy.drive.utils.**DriveResponse**(*description*)
Wrapper for Drive upload and download responses, mainly for calculating/parsing job statistics into human readable formats for logging.

Parameters **description** – String descriptor for specific function of job.

load_resp(*resp*, *is_download=False*)
Loads json response from API.

Parameters

- **resp** (*dictionary*) – Response from API
- **is_download** (*boolean*) – Calculates time taken based on 'modifiedTime' field in response if upload, and based on stop time if download

2.4 Gmail

2.4.1 GmailUtility

class gwrappy.gmail.**GmailUtility**(*json_credentials_path*, *client_id*, ***kwargs*)
Initializes object for interacting with Bigquery API.

Parameters

- **client_secret_path** – File path for client secret JSON file. Only required if credentials are invalid or unavailable.
- **json_credentials_path** – File path for automatically generated credentials.
- **client_id** – Credentials are stored as a key-value pair per client_id to facilitate multiple clients using the same credentials file. For simplicity, using one's email address is sufficient.
- **max_retries** (*integer*) – Argument specified with each API call to natively handle retryable errors.

get_profile()
Abstraction of users().getProfile() method. [https://developers.google.com/gmail/api/v1/reference/users/getProfile]

Returns Dictionary object representing authenticated profile.

get_message(*id*, *format='full'*)
Abstraction of users().messages().get() method. [https://developers.google.com/gmail/api/v1/reference/users/messages/get]

Parameters

- **id** (*string*) – Unique message id.
- **format** (*string*) – Acceptable values are 'full', 'metadata', 'minimal', 'raw'

Returns Dictionary object representing message resource.

get_draft(*id*, *format='full'*)
Abstraction of users().drafts().get() method. [https://developers.google.com/gmail/api/v1/reference/users/drafts/get]

Parameters

- **id** (*string*) – Unique message id.
- **format** (*string*) – Acceptable values are 'full', 'metadata', 'minimal', 'raw'

Returns Dictionary object representing draft resource.

list_messages(*max_results=None*, *full_messages=True*, ***kwargs*)
Abstraction of users().messages().list() method with inbuilt iteration functionality. [https://developers.google.com/gmail/api/v1/reference/users/messages/list]

Parameters

- **max_results** (*integer*) – If None, all results are iterated over and returned.
- **full_messages** (*boolean*) – Convenience toggle to call self.get_message() for each message returned.
- **q** – A query for filtering the file results. Can be generated from gwrappy.gmail.utils.generate_q

Returns List of dictionary objects representing message resources.

list_drafts(*max_results=None*, *full_messages=True*, ***kwargs*)
Abstraction of users().drafts().list() method with inbuilt iteration functionality. [https://developers.google.com/gmail/api/v1/reference/users/drafts/list]

Parameters

- **max_results** (*integer*) – If None, all results are iterated over and returned.
- **full_messages** (*boolean*) – Convenience toggle to call self.get_draft() for each message returned.
- **q** – A query for filtering the file results. Can be generated from gwrappy.gmail.utils.generate_q

Returns List of dictionary objects representing draft resources.

create_draft(*sender*, *to*, *subject*, *message_text*, *attachment_file_paths=None*)
New draft based on input parameters.

Parameters

- **sender** (*string*) – Name of sender
- **to** (*string or list*) – One or more recipients.
- **subject** – Subject text
- **message_text** (*string, dict, or list of dicts*) – Message string, or one or more dict representations of message parts. If dict, keys required are **type** and **text**.
- **attachment_file_paths** (*string or list*) – One or more file paths of attachments.

Returns API response.

send_draft(*draft_id*)
Send unsent draft.

Parameters **draft_id** – Unique draft id.

Returns API Response.

send_email(*sender*, *to*, *subject*, *message_text*, *attachment_file_paths=None*)
Send new message based on input parameters.

Parameters

- **sender** (*string*) – Name of sender
- **to** (*string or list*) – One or more recipients.
- **subject** – Subject text
- **message_text** (*string, dict, or list of dicts*) – Message string, or one or more dict representations of message parts. If dict, keys required are **type** and **text**.

- **attachment_file_paths** (*string or list*) – One or more file paths of attachments.

Returns API response.

get_attachments(*message_id*)
Get message attachments.

Parameters **message_id** – Unique message id. Can be retrieved and iterated over from list_messages() method.

Returns Dictionary with parsed dates and attachment_data (ready to write to file!). Duplicate handling and overwriting logic **should** be handled externally when iterating over list of messages.

2.4.2 Misc Classes/Functions

gwrappy.gmail.utils.**generate_q**(***kwargs*)
Generate query for searching messages. [https://support.google.com/mail/answer/7190]

Parameters **kwargs** – Key-Value pairs. Descriptive flags like *has* or *is* can take lists. Otherwise, list values would be interpreted as "OR".

Returns String representation of search q

gwrappy.gmail.utils.**list_to_html**(*data*, *has_header=True*, *table_format=None*)
Convenience function to convert tables to html for attaching as message text.

Parameters

- **data** (*list of lists*) – Table data
- **has_header** (*boolean*) – Flag whether data contains a header in the first row.
- **table_format** (*dictionary*) – Dictionary representation of formatting for table elements. Eg. {'table': "border: 2px solid black;"}

Returns String representation of HTML.

2.5 Compute Engine

2.5.1 ComputeEngineUtility

class gwrappy.compute.**ComputeEngineUtility**(*project_id*, ***kwargs*)
Initializes object for interacting with Compute Engine API.

By default, Application Default Credentials are used.

If gcloud SDK isn't installed, credential files have to be specified using the kwargs *json_credentials_path* and *client_id*.

Parameters

- **project_id** – Project ID linked to Compute Engine.
- **max_retries** (*integer*) – Argument specified with each API call to natively handle retryable errors.

- **client_secret_path** – File path for client secret JSON file. Only required if credentials are invalid or unavailable.
- **json_credentials_path** – File path for automatically generated credentials.
- **client_id** – Credentials are stored as a key-value pair per client_id to facilitate multiple clients using the same credentials file. For simplicity, using one's email address is sufficient.

get_project()
Abstraction of projects().get() method. [https://cloud.google.com/compute/docs/reference/latest/projects/get]

Returns Project Resource

list_regions(*max_results=None*, *filter_str=None*)
Abstraction of regions().list() method with inbuilt iteration functionality. [https://cloud.google.com/compute/docs/reference/latest/regions/list]

Parameters

- **max_results** (*integer*) – If None, all results are iterated over and returned.
- **filter_str** – Check documentation link for more details.

Returns Generator for dictionary objects representing resources.

list_zones(*max_results=None*, *filter_str=None*)
Abstraction of zones().list() method with inbuilt iteration functionality. [https://cloud.google.com/compute/docs/reference/latest/zones/list]

Parameters

- **max_results** (*integer*) – If None, all results are iterated over and returned.
- **filter_str** – Check documentation link for more details.

Returns Generator for dictionary objects representing resources.

list_instances(*zone_id=None*, *max_results=None*, *filter_str=None*)
Abstraction of instances().list() method with inbuilt iteration functionality. [https://cloud.google.com/compute/docs/reference/latest/instances/list]

Parameters

- **zone_id** – Zone name. If None, all Zones are iterated over and returned.
- **max_results** (*integer*) – If None, all results are iterated over and returned.
- **filter_str** – Check documentation link for more details.

Returns Generator for dictionary objects representing resources.

list_addresses(*region_id=None*, *max_results=None*, *filter_str=None*)
Abstraction of addresses().list() method with inbuilt iteration functionality. [https://cloud.google.com/compute/docs/reference/latest/addresses/list]

Parameters

- **region_id** – Region name. If None, all Regions are iterated over and returned.
- **max_results** (*integer*) – If None, all results are iterated over and returned.
- **filter_str** – Check documentation link for more details.

Returns Generator for dictionary objects representing resources.

list_operations(*operation_type*, *location_id=None*, *max_results=None*, *filter_str=None*)
Choose between region or zone operations with operation_type.

Abstraction of zoneOperations()/regionOperations().list() method with inbuilt iteration functionality.

https://cloud.google.com/compute/docs/reference/latest/zoneOperations/list

https://cloud.google.com/compute/docs/reference/latest/regionOperations/list

Parameters

- **operation_type** – 'zone' or 'region' type operations.
- **location_id** – Zone/Region name. If None, all Zones/Regions are iterated over and returned.
- **max_results** (*integer*) – If None, all results are iterated over and returned.
- **filter_str** – Check documentation link for more details.

Returns Generator for dictionary objects representing resources.

get_operation(*operation_type*, *location_id*, *operation_name*)
Choose between region or zone operations with operation_type.

Abstraction of zoneOperations()/regionOperations().get() method.

https://cloud.google.com/compute/docs/reference/latest/zoneOperations/get

https://cloud.google.com/compute/docs/reference/latest/regionOperations/get

Parameters

- **operation_type** – 'zone' or 'region' type operations.
- **location_id** – Zone/Region name.
- **operation_name** – Operation name.

Returns ZoneOperations/RegionOperations Resource.

poll_operation_status(*operation_type*, *location_id*, *operation_name*, *end_state*, *sleep_time=0.5*)
Poll operation to until desired end_state is achieved. eg. 'DONE' when adding addresses.

Parameters

- **operation_type** – 'zone' or 'region' type operations.
- **location_id** – Zone/Region name.
- **operation_name** – Operation name.
- **end_state** – Final status that signifies operation is finished.
- **sleep_time** – Intervals between polls.

Returns ZoneOperations/RegionOperations Resource.

get_address(*region_id*, *address_name*)
Abstraction of addresses().get() method. [https://cloud.google.com/compute/docs/reference/latest/addresses/get]

Parameters

- **region_id** – Region name.
- **address_name** – Address name.

Returns Addresses Resource.

add_address(*region_id*, *address_name*)
Abstraction of address.insert() method with operation polling functionality. [https://cloud.google.com/compute/docs/reference/latest/addresses/insert]

Parameters

- **region_id** – Region name.
- **address_name** – Address name.

Returns RegionOperations Resource.

delete_address(*region_id*, *address_name*)
Abstraction of address.delete() method with operation polling functionality. [https://cloud.google.com/compute/docs/reference/latest/addresses/delete]

Parameters

- **region_id** – Region name.
- **address_name** – Address name.

Returns RegionOperations Resource.

get_instance(*zone_id*, *instance_name*)
Abstraction of instances().get() method. [https://cloud.google.com/compute/docs/reference/latest/instances/get]

Parameters

- **zone_id** – Zone name.
- **instance_name** – Instance name.

Returns Instances Resource.

start_instance(*zone_id*, *instance_name*)
Abstraction of instances().start() method with operation polling functionality. [https://cloud.google.com/compute/docs/reference/latest/instances/start]

Parameters

- **zone_id** – Zone name.
- **instance_name** – Instance name.

Returns ZoneOperations Resource.

stop_instance(*zone_id*, *instance_name*)
Abstraction of instances().stop() method with operation polling functionality. [https://cloud.google.com/compute/docs/reference/latest/instances/stop]

Parameters

- **zone_id** – Zone name.
- **instance_name** – Instance name.

Returns ZoneOperations Resource.

delete_instance(*zone_id*, *instance_name*)
Abstraction of instances().delete() method with operation polling functionality. [https://cloud.google.com/compute/docs/reference/latest/instances/delete]

Parameters

- **zone_id** – Zone name.
- **instance_name** – Instance name.

Returns ZoneOperations Resource.

2.6 Dataproc

2.6.1 DataprocUtility

class gwrappy.dataproc.**DataprocUtility**(*project_id*, ***kwargs*)
Initializes object for interacting with Dataproc API.

By default, Application Default Credentials are used.
If gcloud SDK isn't installed, credential files have to be specified using the kwargs *json_credentials_path* and *client_id*.

Parameters

- **project_id** – Project ID linked to Dataproc.
- **client_secret_path** – File path for client secret JSON file. Only required if credentials are invalid or unavailable.
- **json_credentials_path** – File path for automatically generated credentials.
- **client_id** – Credentials are stored as a key-value pair per client_id to facilitate multiple clients using the same credentials file. For simplicity, using one's email address is sufficient.

list_clusters(*max_results=None*, *filter=None*)
Abstraction of projects().regions().clusters().list() method with inbuilt iteration functionality. [https://cloud.google.com/dataproc/docs/reference/rest/v1/projects.regions.clusters/list]

Parameters

- **max_results** (*integer*) – If None, all results are iterated over and returned.
- **filter** (*String*) – Query param [https://cloud.google.com/dataproc/docs/reference/rest/v1/projects.regions.cl parameters]

Returns List of dictionary objects representing cluster resources.

get_cluster(*cluster_name*)
Abstraction of projects().regions().clusters().get() method. [https://cloud.google.com/dataproc/docs/reference/rest/v1/project

Parameters **cluster_name** – Cluster name.

Returns Dictionary object representing cluster resource.

diagnose_cluster(*cluster_name*)
Abstraction of projects().regions().clusters().diagnose() method. [https://cloud.google.com/dataproc/docs/reference/rest/v1/p

Parameters **cluster_name** – Cluster name.

Returns Dictionary object representing operation resource.

list_operations(*max_results=None*, *filter=None*)
Abstraction of projects().regions().operations().list() method with inbuilt iteration functionality. [https://cloud.google.com/dataproc/docs/reference/rest/v1/projects.regions.operations/list]

Parameters

- **max_results** (*integer*) – If None, all results are iterated over and returned.

- **filter** (*String*) – Query param [https://cloud.google.com/dataproc/docs/reference/rest/v1/projects.regions.o parameters]

Returns List of dictionary objects representing operation resources.

get_operation(*operation_name*)
Abstraction of projects().regions().operations().get() method. [https://cloud.google.com/dataproc/docs/reference/rest/v1/proj

Parameters **operation_name** – Name of operation resource.

Returns Dictionary object representing operation resource.

poll_operation_status(*operation_resp*, *sleep_time=3*)
Abstraction of projects().regions().operations().get() method. [https://cloud.google.com/dataproc/docs/reference/rest/v1/proj

Parameters

- **operation_resp** – Representation of operation resource.
- **sleep_time** – If wait_finish is set to True, sets polling wait time.

Returns Dictionary object representing operation resource.

create_cluster(*zone*, *cluster_name*, *wait_finish=True*, *sleep_time=3*, ***kwargs*)
Abstraction of projects().regions().clusters().create() method. [https://cloud.google.com/dataproc/docs/reference/rest/v1/proj

Parameters

- **zone** – Dataproc zone.
- **cluster_name** – Cluster name.
- **wait_finish** – If set to True, operation will be polled till completion.
- **sleep_time** – If wait_finish is set to True, sets polling wait time.
- **config_bucket** – Google Cloud Storage staging bucket used for sharing generated SSH keys and config.
- **network** – Google Compute Engine network to be used for machine communications.
- **master_boot_disk** – Size in GB of the master boot disk.
- **master_num** – The number of master VM instances in the instance group.
- **master_machine_type** – Google Compute Engine machine type used for master cluster instances.
- **worker_boot_disk** – Size in GB of the worker boot disk.
- **worker_num** – The number of VM worker instances in the instance group.
- **worker_machine_type** – Google Compute Engine machine type used for worker cluster instances.
- **init_actions** – Google Cloud Storage URI of executable file(s).

Returns Dictionary object or OperationResponse representing cluster resource.

delete_cluster(*cluster_name*, *wait_finish=True*, *sleep_time=3*)
Abstraction of projects().regions().clusters().delete() method. [https://cloud.google.com/dataproc/docs/reference/rest/v1/proj

Parameters

- **cluster_name** – Cluster name.
- **wait_finish** – If set to True, operation will be polled till completion.
- **sleep_time** – If wait_finish is set to True, sets polling wait time.

Returns Dictionary object or OperationResponse representing cluster resource.

list_jobs(*cluster_name=None*, *job_state='ACTIVE'*, *max_results=None*, *filter=None*)
Abstraction of projects().regions().jobs().list() method with inbuilt iteration functionality. [https://cloud.google.com/dataproc/docs/reference/rest/v1/projects.regions.jobs/list]

Parameters

- **cluster_name** – Cluster name, if unset, will return jobs from all clusters.
- **job_state** – Category of jobs to return. [https://cloud.google.com/dataproc/docs/reference/rest/v1/projects.reg
- **max_results** (*integer*) – If None, all results are iterated over and returned.
- **filter** (*String*) – Query param [https://cloud.google.com/dataproc/docs/reference/rest/v1/projects.regions.jo parameters]

Returns List of dictionary objects representing cluster resources.

get_job(*job_id*)
Abstraction of projects().regions().jobs().get() method. [https://cloud.google.com/dataproc/docs/reference/rest/v1/projects.re

Parameters **job_id** – Job Id.

Returns Dictionary object representing job resource.

poll_job_status(*job_resp*, *sleep_time=3*)

Parameters

- **job_resp** – Representation of job resource.
- **sleep_time** – If wait_finish is set to True, sets polling wait time.

Returns Dictionary object representing job resource.

submit_spark_job(*cluster_name*, *main_class*, *wait_finish=True*, *sleep_time=5*, ***kwargs*)
Abstraction of projects().regions().jobs().submit() method. [https://cloud.google.com/dataproc/docs/reference/rest/v1/project

Body parameters can be found here: [https://cloud.google.com/dataproc/docs/reference/rest/v1/projects.regions.jobs#resourc job]

Parameters

- **cluster_name** – The name of the cluster where the job will be submitted.
- **main_class** – The name of the driver's main class.
- **wait_finish** – If set to True, operation will be polled till completion.
- **sleep_time** – If wait_finish is set to True, sets polling wait time.
- **args** – The arguments to pass to the driver.
- **jar_uris** – HCFS URIs of jar files to add to the CLASSPATHs of the Spark driver and tasks.
- **file_uris** – HCFS URIs of files to be copied to the working directory of Spark drivers and distributed tasks.
- **archive_uris** – HCFS URIs of archives to be extracted in the working directory of Spark drivers and tasks.
- **properties** – A mapping of property names to values, used to configure Spark.

Returns Dictionary object or JobResponse representing job resource.

submit_pyspark_job(*cluster_name*, *main_py_uri*, *wait_finish=True*, *sleep_time=5*, ***kwargs*)
Abstraction of projects().regions().jobs().submit() method. [https://cloud.google.com/dataproc/docs/reference/rest/v1/project

Body parameters can be found here: [https://cloud.google.com/dataproc/docs/reference/rest/v1/projects.regions.jobs#resourc
job]

Parameters

- **cluster_name** – The name of the cluster where the job will be submitted.
- **main_py_uri** – The HCFS URI of the Python file to use as the driver.
- **wait_finish** – If set to True, operation will be polled till completion.
- **sleep_time** – If wait_finish is set to True, sets polling wait time.
- **args** – The arguments to pass to the driver.
- **python_uris** – HCFS file URIs of Python files to pass to the PySpark framework.
- **jar_uris** – HCFS URIs of jar files to add to the CLASSPATHs of the Python driver and tasks.
- **file_uris** – HCFS URIs of files to be copied to the working directory of Python drivers and distributed tasks.
- **archive_uris** – HCFS URIs of archives to be extracted in the working directory of Spark drivers and tasks.
- **properties** – A mapping of property names to values, used to configure PySpark.

Returns Dictionary object or JobResponse representing job resource.

2.6.2 Misc Classes/Functions

class gwrappy.dataproc.utils.**OperationResponse**(*resp*)
Wrapper for Dataproc Operation responses, mainly for calculating/parsing statistics into human readable formats for logging.

Parameters **resp** (*dictionary*) – Dictionary representation of an operation resource.

class gwrappy.dataproc.utils.**JobResponse**(*resp*)
Wrapper for Dataproc Job responses, mainly for calculating/parsing statistics into human readable formats for logging.

Parameters **resp** (*dictionary*) – Dictionary representation of an operation resource.

2.7 General Utilities

2.7.1 Date Manipulation

gwrappy.utils.**timestamp_to_datetime**(*input_timestamp*, *tz=None*)
Converts epoch timestamp into datetime object.

Parameters

- **input_timestamp** (*long*) – Epoch timestamp. Microsecond or millisecond inputs accepted.
- **tz** – String representation of timezone accepted by pytz. eg. 'Asia/Hong_Kong'. If param is unfilled, system timezone is used.

Returns timezone aware datetime object

gwrappy.utils.**datetime_to_timestamp**(*input_datetime*, *date_format='%Y-%m-%d %H:%M:%S'*, *tz=None*)

Converts datetime to epoch timestamp.

Note - If input_datetime is timestamp aware, it would first be localized according to the tz parameter if filled, or the system timezone if unfilled.

Parameters

- **input_datetime** (*datetime object or string representation of datetime.*) – Date to convert.
- **date_format** – If input is string, denotes string datetime format to convert from.
- **tz** – String representation of timezone accepted by pytz. eg. 'Asia/Hong_Kong'. If param is unfilled, system timezone is used.

Returns timezone aware datetime object

gwrappy.utils.**date_range**(*start*, *end*, *ascending=True*, *date_format='%Y-%m-%d'*)

Simple datetime generator for dates between start and end (inclusive).

Parameters

- **start** (*datetime object or string representation of datetime.*) – Date to start at.
- **end** (*datetime object or string representation of datetime.*) – Date to stop at.
- **ascending** (*boolean*) – Toggle sorting of output.
- **date_format** – If input is string, denotes string datetime format to convert from.

Returns generator object for naive datetime objects

gwrappy.utils.**month_range**(*start*, *end*, *full_months=False*, *month_format='%Y-%m'*, *ascending=True*, *date_format='%Y-%m-%d'*)

Simple utility to chunk date range into months.

Parameters

- **start** – Date to start at.
- **end** – Date to end at.
- **full_months** – If true, only data up till the last complete month would be included.
- **month_format** – Format for month key.
- **date_format** – If start and end is string, denotes string datetime format to convert from.
- **ascending** – Sort date list ascending

Returns dictionary keyed by month (in the format specified by month_format) with values being the list of dates within that month.

gwrappy.utils.**simple_mail**(*send_to*, *subject*, *text*, *send_from=None*, *username=None*, *password=None*, *server='smtp.gmail.com'*, *port=587*)

Simple utility mail function - only text messages without attachments.

Note - In Gmail you'd have to allow 'less secure apps to access your account'. Not recommended for secure information/accounts.

Parameters

- **send_to** (*list or string*) – Email recipients
- **subject** – Email Subject
- **text** – Email Body
- **send_from** – Name of sender
- **username** – Login username
- **password** – Login password
- **server** – Mail server
- **port** – Connection port

class gwrappy.utils.**StringLogger**(*name=None*, *level=20*, *formatter=None*, *ignore_modules=None*)

Simple logging wrapper with a string buffer handler to easily write and retrieve logs as strings.

Parameters

- **name** – Name of logger
- **level** – Logging level
- **formatter** – logging.Formatter() object
- **ignore_modules** – list of module names to ignore from logging process

get_logger()

Return instantiated Logger.

Returns logging.Logger object

get_log_string()

Return logs as string.

Returns logged data as string

g

J

L

M

O

P

R

S

T

U

W