ASTRONOMER

Contributing to Apache Airflow

Airflow Summit
8 July 2021

Kaxil Naik
Airflow Committer and PMC Member
OSS Airflow Team @ Astronomer

Who am I?

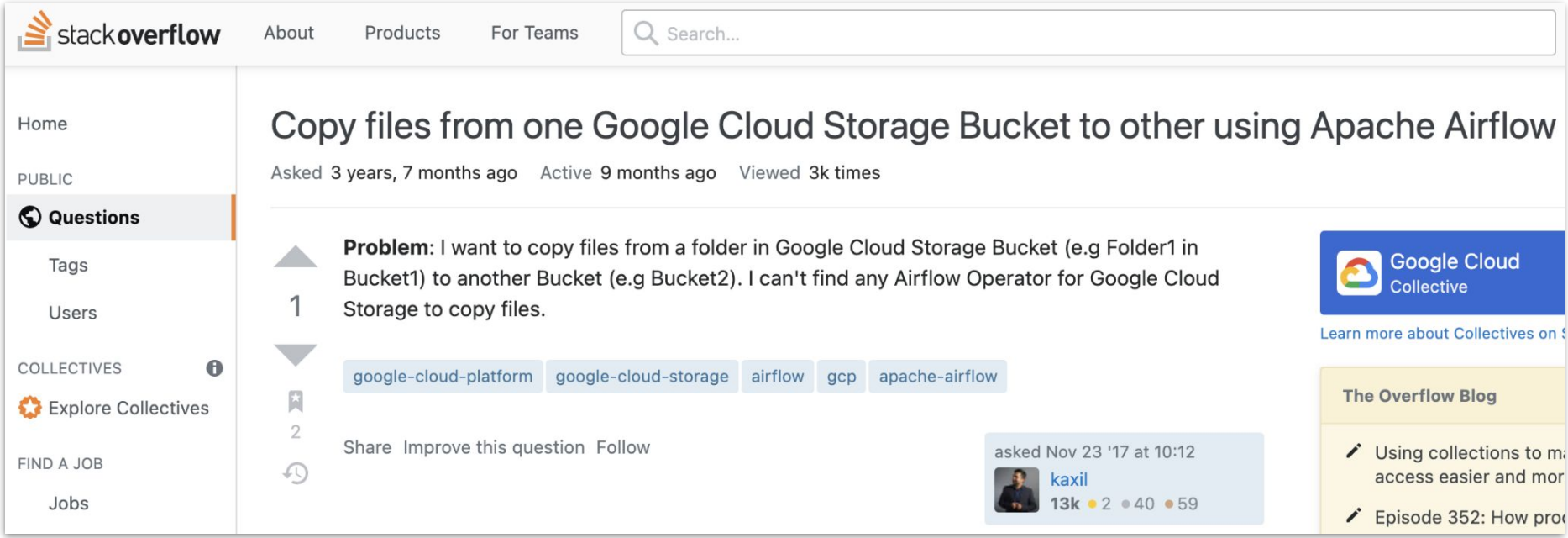
- Airflow Committer & PMC Member
- Manager of Airflow Engineering team @ Astronomer
  - Work full-time on Airflow
- Previously worked at DataReply
- Masters in Data Science & Analytics from Royal Holloway, University of London
- Twitter: https://twitter.com/kaxil
- Github: https://github.com/kaxil/
- LinkedIn: https://www.linkedin.com/in/kaxil/

Agenda

- My Journey with Airflow
- How to start contributing to Airflow
- Contribution Workflow
- Continue committing regularly & steadily to become Airflow Committer. (including talking about current Guidelines of becoming a Committer)
- Different mediums of communication (Dev list, users list, Slack channel, Github Discussions etc)
- Reviewing PRs, Testing and voting on releases
- Dev Environment
- Number of commits does not matter as much as the experience and actual code

My Journey

Motivation to contribute !



https://stackoverflow.com/q/47452879/5691525

Motivation to contribute !

1

I just found a new operator in contrib uploaded 2 hours ago: https://github.com/apache/incubator-airflow/blob/master/airflow/contrib/operators/gcs_to_gcs.py called `GoogleCloudStorageToGoogleCloudStorageOperator` that should copy an object from a bucket to another, with renaming if requested.

Share Edit Delete Flag

edited Aug 22 '19 at 11:21

answered Nov 23 '17 at 10:14

kaxil

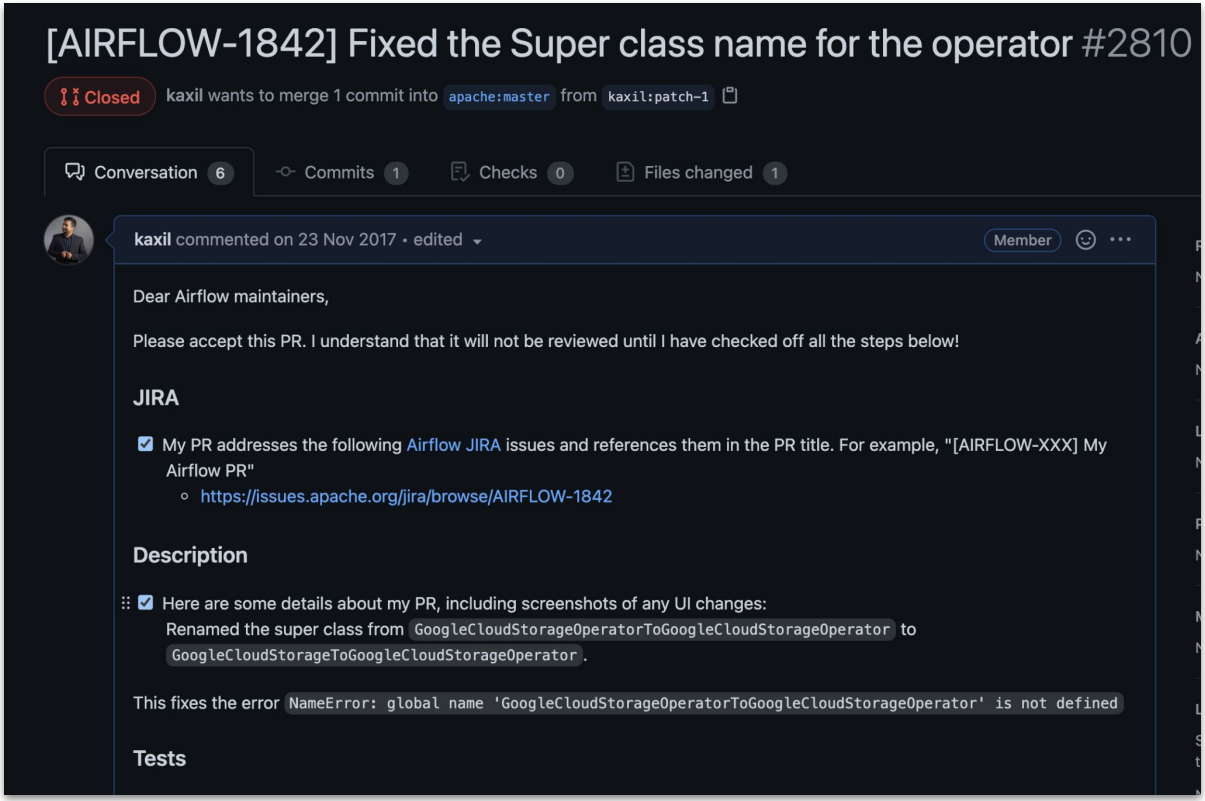
13k ● 2 ● 40 ● 59

Add a comment

https://stackoverflow.com/a/47452939/5691525

But it didn’t work …

Fixed it - My First PR



[AIRFLOW-1842] Fixed the Super class name for the operator #2810

Closed kaxil wants to merge 1 commit into `apache:master` from `kaxil:patch-1`

Conversation 6 Commits 1 Checks 0 Files changed 1

kaxil commented on 23 Nov 2017 • edited — Member

Dear Airflow maintainers,

Please accept this PR. I understand that it will not be reviewed until I have checked off all the steps below!

JIRA

- [x] My PR addresses the following Airflow JIRA issues and references them in the PR title. For example, "[AIRFLOW-XXX] My Airflow PR"
  - https://issues.apache.org/jira/browse/AIRFLOW-1842

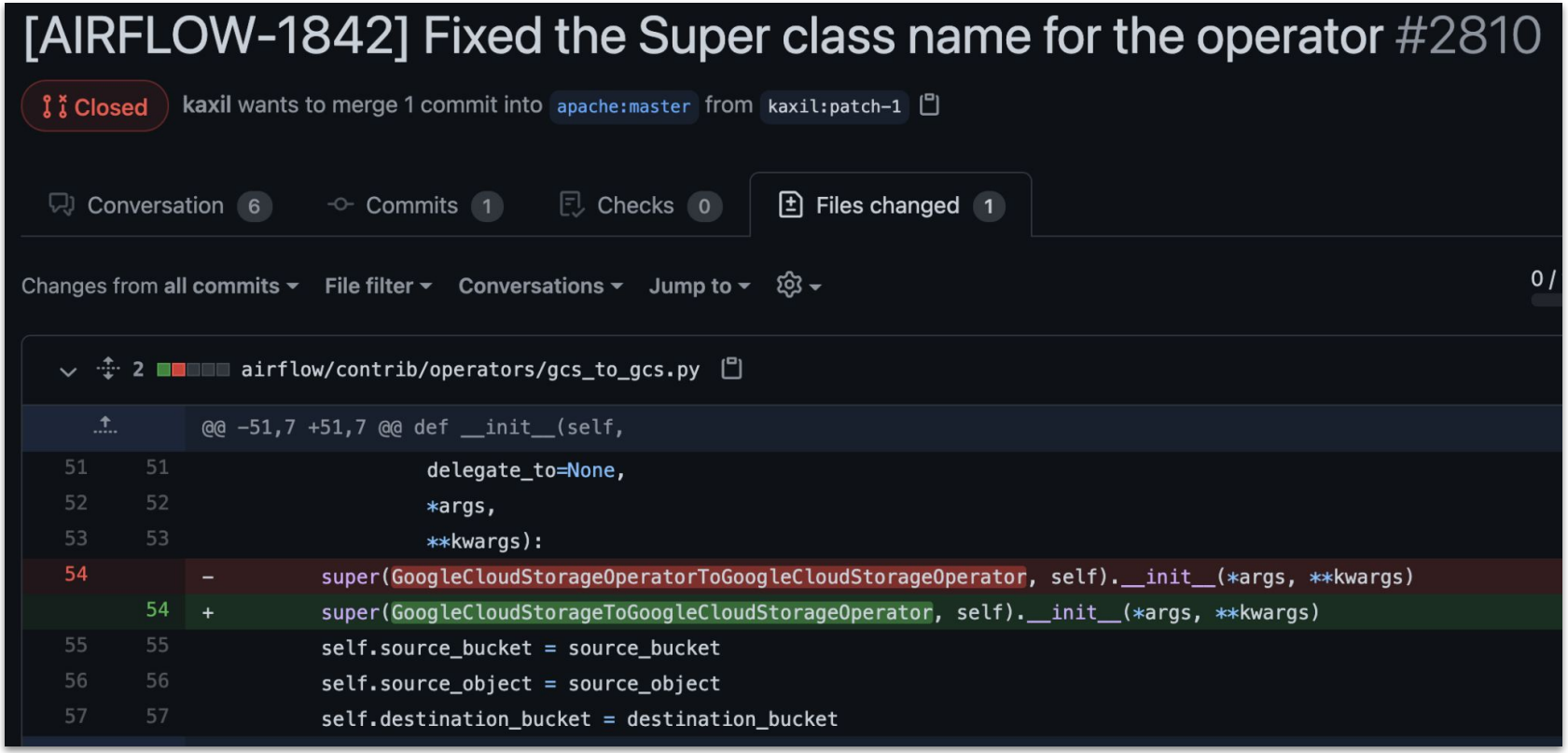
Description

- [x] Here are some details about my PR, including screenshots of any UI changes:
  Renamed the super class from `GoogleCloudStorageOperatorToGoogleCloudStorageOperator` to `GoogleCloudStorageToGoogleCloudStorageOperator`.

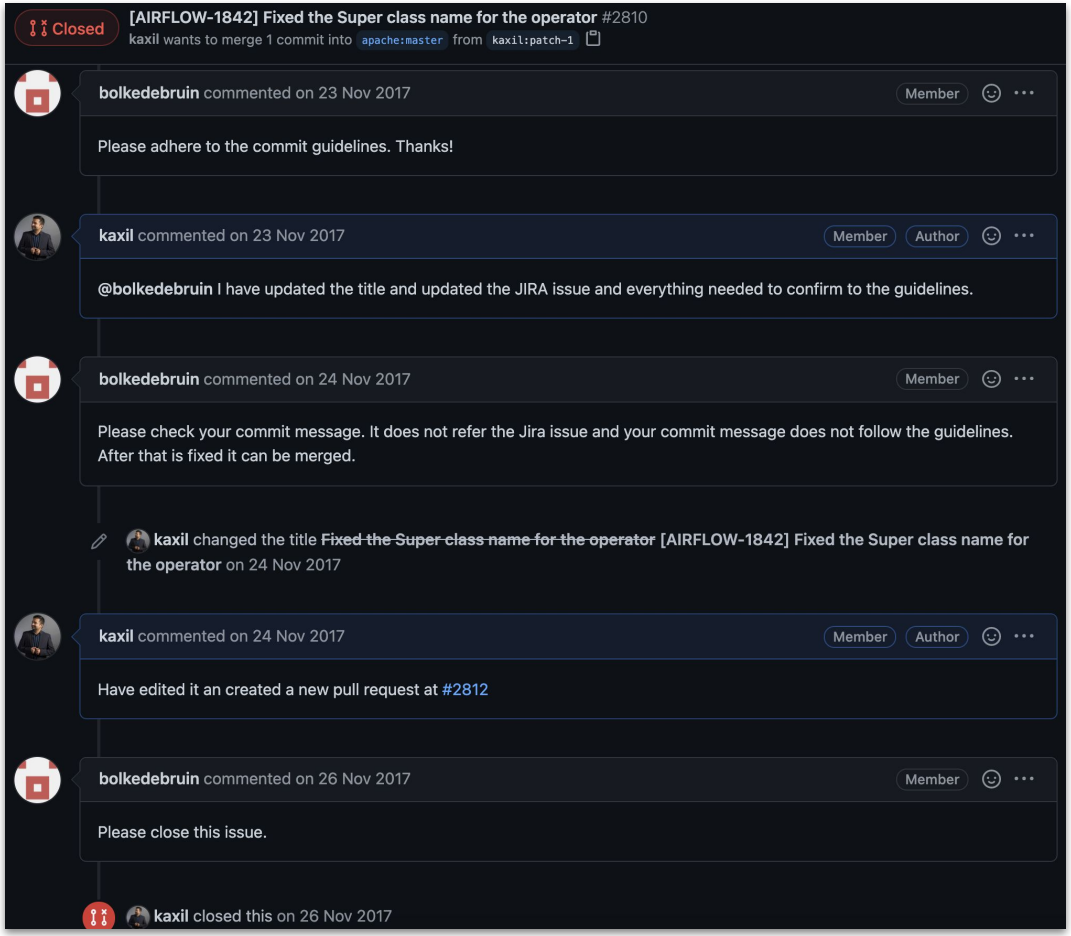
This fixes the error `NameError: global name 'GoogleCloudStorageOperatorToGoogleCloudStorageOperator' is not defined`

Tests

My First PR - Fixes Typo



My First PR - I didn't follow Guidelines !!



https://media.giphy.com/media/KSKvdT1YGCpUIonvSq/giphy.gif

My First “Merged” PR/commit

✓ [AIRFLOW-1842] Fixed Super class name for the gcs to gcs copy operator

Browse files

```
Fixed incorrect Super class name in gcs to gcs
copy operator from `GoogleCloudStorageOperatorToGo
ogleCloudStorageOperator` to
`GoogleCloudStorageToGoogleCloudStorageOperator`.

Closes #2812 from kaxil/patch-2
```

main

upgrade-check/1.4.0 ··· 1.10.0rc1

kaxil authored and bolkedebruin committed on 26 Nov 2017 — 1 parent 87c6c83 commit 4247ff0228afd06b8210321dcf660e6e7aca7b62

Showing 1 changed file with 1 addition and 1 deletion. Unified Split

2 airflow/contrib/operators/gcs_to_gcs.py

```
@@ -51,7 +51,7 @@ def __init__(self,
                 delegate_to=None,
                 *args,
                 **kwargs):
-        super(GoogleCloudStorageOperatorToGoogleCloudStorageOperator, self).__init__(*args, **kwargs)
+        super(GoogleCloudStorageToGoogleCloudStorageOperator, self).__init__(*args, **kwargs)
         self.source_bucket = source_bucket
         self.source_object = source_object
         self.destination_bucket = destination_bucket
```

http://gph.is/15RTH5O

Slowly & Steadily started adding more contributions

Commits on Feb 2, 2018

[AIRFLOW-2053] Fix quote character bug in BQ hook ...
kaxil authored and Fokko committed on 2 Feb 2018 ✓

Commits on Jan 31, 2018

[AIRFLOW-2037] Add methods to get Hash values of a GCS object ...
kaxil authored and Fokko committed on 31 Jan 2018 ✓

Commits on Jan 26, 2018

[AIRFLOW-1943] Add External BigQuery Table feature ...
kaxil authored and Fokko committed on 26 Jan 2018 ✗

[AIRFLOW-2033] Add Google Cloud Storage List Operator ...
kaxil authored and Fokko committed on 26 Jan 2018 ✗

Commits on Jan 25, 2018

[AIRFLOW-2031] Add missing gcp_conn_id in the example in DataFlow doc... ...
kaxil authored and Fokko committed on 25 Jan 2018 ✗

Commits on Jan 12, 2018

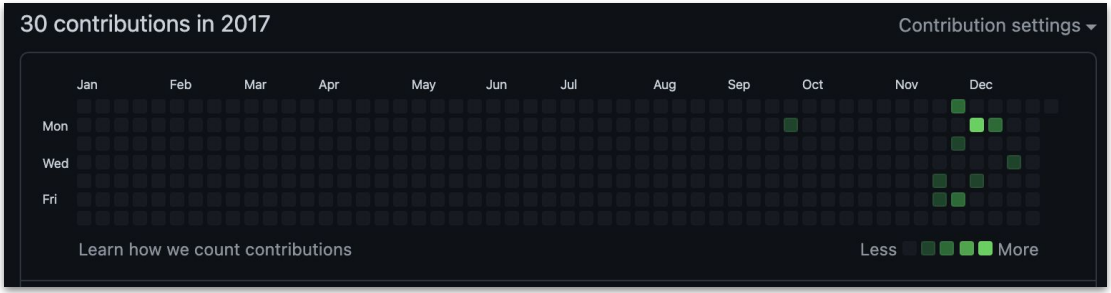
[AIRFLOW-1997] Fix GCP operator doc strings ...
kaxil authored and criccomini committed on 12 Jan 2018 ✓

Commits on Jan 3, 2018

[AIRFLOW-1946][AIRFLOW-1855] Create a BigQuery Get Data Operator ...
kaxil authored and criccomini committed on 3 Jan 2018 ✓

Commits on Dec 11, 2017

[AIRFLOW-1892] Modify BQ hook to extract data filtered by column ...
kaxil authored and criccomini committed on 11 Dec 2017 ✓



What did I learn while working on OSS

- Writing unit-tests
- Improved Coding skills
- Got to know many companies & devs across the globe
- Improved communication skills
    - Commit messages & PR descriptions
    - Email threads on dev list
    - Presentations

Became Airflow Committer & (P)PMC Member

https://twitter.com/ApacheAirflow/status/993950478785490945

Steered Release for Airflow 1.10.2

From: Kaxil Naik <k...@gmail.com>
To: de...@airflow.apache.org
Subject: [VOTE] Airflow 1.10.2rc1
Date: 2019/01/17 00:04:25
List: dev@airflow.apache.org

Hey all,

I have cut Airflow 1.10.2 RC1. This email is calling a vote on the release,
which will last for 72 hours. Consider this my (binding) +1.

Airflow 1.10.2 RC1 is available at:
https://dist.apache.org/repos/dist/dev/airflow/1.10.2rc1/

apache-airflow-1.10.2rc1-source.tar.gz is a source release that comes
with INSTALL instructions.
apache-airflow-1.10.2rc1-bin.tar.gz is the binary Python "sdist" release.

Public keys are available at:
https://dist.apache.org/repos/dist/release/airflow/KEYS

Only votes from PMC members are binding, but members of the community are
encouraged to test the release and vote with "(non-binding)".

Please note that the version number excludes the `rcX` string, so it's now
simply 1.10.2. This will allow us to rename the artifact without modifying
the artifact checksums when we actually release.

Changes since 1.10.2b2:
[AIRFLOW-3691] Update notice to 2019 (#4503)
[AIRFLOW-3689] Update pop-up message when deleting DAG in RBAC UI (#4505)
[AIRFLOW-3690] Fix bug to set state of a task for manually-triggered DAGs
(#4504)
[AIRFLOW-XXX] Update Contributing Guide - Git Hooks (#4120)
[AIRFLOW-3426] Correct Python Version Documentation Reference (#4259)
[AIRFLOW-3693] Replace psycopg2-binary by psycopg2 (#4508)
[AIRFLOW-2663] Add instructions to install SSH dependencies
[AIRFLOW-XXX] Clean up installation extra packages table (#3750)
[AIRFLOW-XXX] Remove redundant space in Kerberos (#3866)
[AIRFLOW-3086] Add extras group for google auth to setup.py. (#3917)
[AIRFLOW-XXX] Add Kubernetes Dependency in Extra Packages Doc (#4281)

From: Kaxil Naik <k...@gmail.com>
To: de...@airflow.apache.org
Subject: Airflow 1.10.2 is released
Date: 2019/01/23 01:18:38
List: dev@airflow.apache.org

Dear Airflow community,

I'm happy to announce that* Airflow 1.10.2* was just released.

The source release, as well as the binary "sdist" release, are available
here:

https://dist.apache.org/repos/dist/release/airflow/1.10.2/

We also made this version available on PyPi for convenience (`pip install
apache-airflow`):

https://pypi.python.org/pypi/apache-airflow

The documentation is available on:

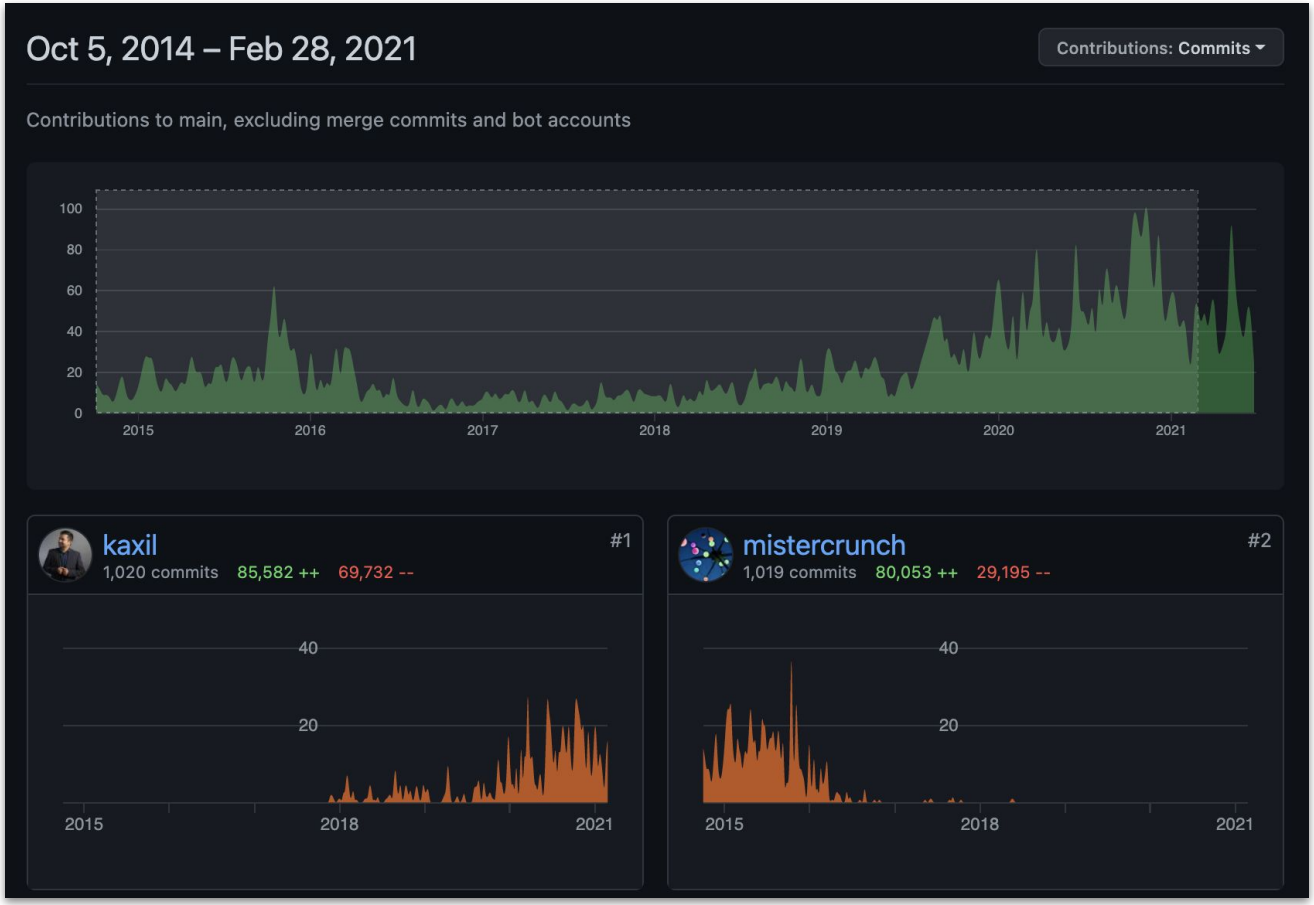
- https://airflow.apache.org/
- https://airflow.apache.org/1.10.2/
- https://airflow.readthedocs.io/en/1.10.2/
- https://airflow.readthedocs.io/en/stable/

Find the CHANGELOG here for more details:

https://airflow.apache.org/changelog.html#airflow-1-10-2-2019-01-19

Cheers,
Kaxil

Became Leading Airflow Committer in Feb 2021



You are next !!

How to start contributing?

How to start contributing?

- Contributing Guidelines: CONTRIBUTING.rst
- Contributing Quick Start Guide: CONTRIBUTORS_QUICK_START.rst
- Good First Issues: https://github.com/apache/airflow/contribute

https://gph.is/g/ZWdK71X

Contribution Workflow

1. Find the issue you want to work on
2. Setup a local dev environment
3. Understand the codebase
4. Write Code & add tests if it’s a code change
5. Run tests locally
6. Create PR and wait for reviews
7. Address any suggestions by reviewers
8. Nudge politely if your PR is pending reviews for a while

Finding issues to work on

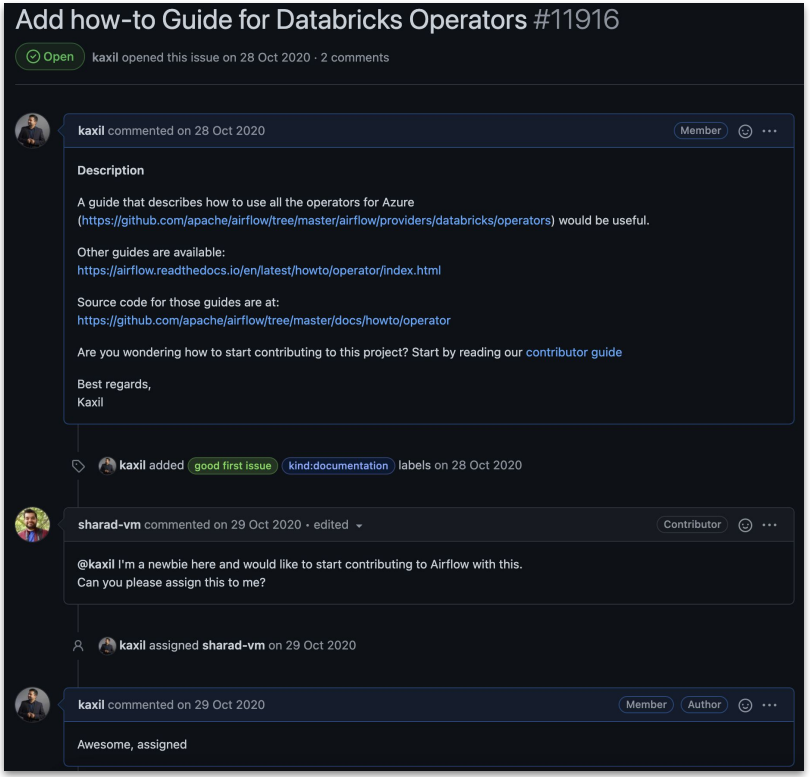
Finding issues to work on

- Start small: the aim should be to understand the process
- Bugs / features impacting you or your work
- Documentation Issues (including Contribution Guides)
  - Missing or outdated info, typos, formatting issues, broken links etc
- Good First Issues: https://github.com/apache/airflow/contribute
- Other open GitHub Issues: https://github.com/apache/airflow/issues

Finding issues to work on - Open Unassigned Issues

If the issue is open and un-assigned, comment you want to work on it.

A committer will assign that issue to you. Then it is all yours.



Finding issues to work on - Improving Documentation

- If you faced an issue with docs, better fix it for future readers
- Documentation PRs are the best first contributions
- Missing or outdated info, typos, formatting issues, broken links etc
- No need of writing unit tests
- Examples:
  - https://github.com/apache/airflow/pull/16275
  - https://github.com/apache/airflow/pull/13462
  - https://github.com/apache/airflow/pull/15265

Setup a local dev environment

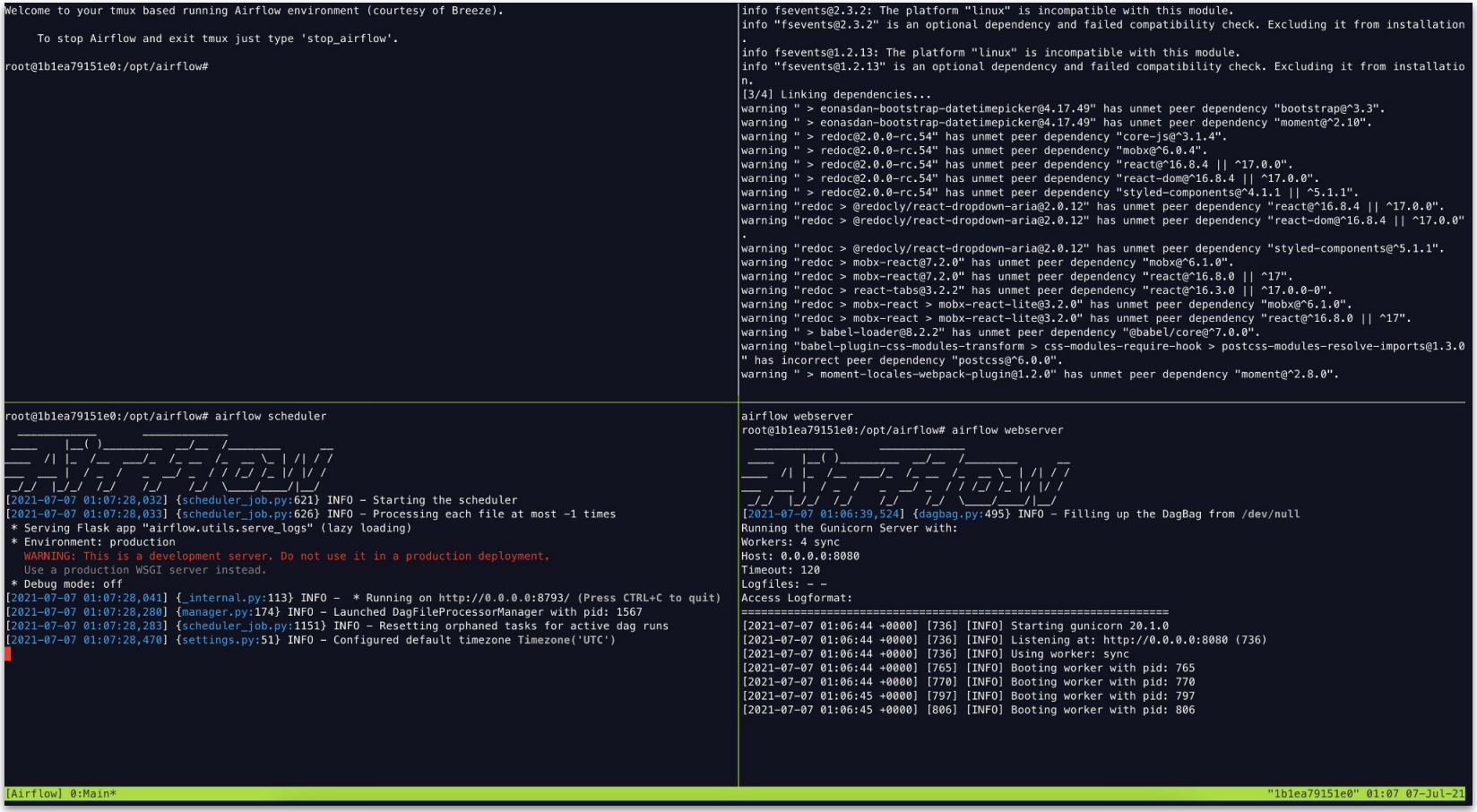
SetUp Local Development Environment

- Fork Apache Airflow repo & clone it locally
- Install pre-commit hooks (link) to detect minor issues before creating a PR
  - Some of them even automatically fix issues e.g ‘black’ formats python code
  - Install pre-commit framework: `pip install pre-commit`
  - Install pre-commit hooks: `pre-commit install`
- Use `breeze` - a wrapper around docker-compose for Airflow development.
  - Mac Users: Increase resources available to **Docker for Mac**
  - Check Prerequisites: https://github.com/apache/airflow/blob/main/BREEZE.rst#prerequisites
  - Setup autocomplete: `./breeze setup-autocomplete`

SetUp Local Development Environment - Breeze

- Airflow CI uses **breeze** too so it allows reproduction locally
- Allows running Airflow with different environments (different Python versions, different Metadata db, etc):
  - `./breeze --python 3.6 --backend postgres --postgres-version 12`
- You can also run a local instance of Airflow using:
  - `./breeze start-airflow --python 3.6 --backend postgres`
- You can then access the Webserver on http://localhost:28080

SetUp Local Development Environment - Breeze



Understand the Codebase

Understand the Codebase

- apache/airflow is mono-repo containing:
  - Apache Airflow Python package
  - More than 60 Providers (Google, Amazon, Postgres, etc)
  - Container image
  - Helm Chart
- Each of these items are released and versioned separately

Understand the Codebase

- Do not try to understand the entire codebase at once
- Get familiar with the directory structure first
- Dive into the source code related to your issue
- Similar to: If you are moving to a new house, you would try to first get familiar with your immediate neighbours and then others. (unless you have memory like Sheldon Cooper !!!)

http://gph.is/2F2nUVb

Understand the Codebase - Directory Structure

Area	Paths (relative to the repository root)
Core Airflow Docs	docs/apache-airflow
Stable API	airflow/api_connexion
CLI	airflow/cli
Webserver / UI	airflow/www
Scheduler (incl. Dag Parsing)	airflow/jobs/scheduler_job.py
Executors	airflow/executors
DAG Serialization	airflow/serialization
Helm Chart (& it’s tests)	chart
Container Image	Dockerfile
Tests	tests

Understand the Codebase - Directory Structure

Area	Paths (relative to the repository root)
Providers	airflow/providers
Core Operators	airflow/operators
Core Hooks	airflow/hooks
Core Sensors	airflow/sensors
DB Migrations	airflow/migrations
ORM Models (Python Class -> DB Tables)	airflow/models
Secrets Backend	airflow/secrets
Configuration	airflow/configuration.py
Permission Model	airflow/www/security.py
All Docs (incl. docs for Chart & Container image)	docs

Understand the Codebase - Areas

- Get expertise in a certain area before diving into a different one.

Easy	**Medium**	**Complex (core)**
Docs	Webserver	Scheduler & Dag Parsing
CLI	Helm Chart	Executors
Operators / Hooks / Sensors (Providers)	Dockerfile	Configuration
Stable API	Secrets Backend	Permission Model
	DB Migrations	

Write Code & add docs & tests

Write code, add tests and docs

- You can re-use lot of logic of the code in the “same area”
- If adding a feature, find previously merged PRs in same area to get an idea
- Check out Coding style and best practices in `CONTRIBUTING.rst`
- Add Tests and Docs
  - The `tests` directory has same structure as `airflow`.
  - E.g If code file is `airflow/providers/google/cloud/operators/bigquery.py`, tests for it should be added at `tests/providers/google/cloud/operators/test_bigquery.py`
  - Docs for it would be at `docs/apache-airflow-providers-google/operators/cloud/bigquery.rst`

Write code, add tests and docs

- You can re-use lot of logic of the code in the “same area”
  - e.g. if you are adding a new Operator for GCP, you can check how current GCS operators are implemented or find a PR that added a operator in Google Provider & see what all they added.
- Check out Coding style and best practices in `CONTRIBUTING.rst`
- Add Tests and Docs
  - The `tests` directory has same structure as `airflow`.
  - E.g If code file is `airflow/providers/google/cloud/operators/bigquery.py`, tests for it should be added at `tests/providers/google/cloud/operators/test_bigquery.py`

Run tests locally

Create PR and wait for reviews

Address Reviews

Nudge for reviews

Communication Channels

Becoming a Committer

Links / References

Links

- Airflow
  - Repo: https://github.com/apache/airflow
  - Website: https://airflow.apache.org/
  - Blog: https://airflow.apache.org/blog/
  - Documentation: https://airflow.apache.org/docs/stable/
  - Slack: https://s.apache.org/airflow-slack
  - Twitter: https://twitter.com/apacheairflow

- Contact Me:
  - Twitter: https://twitter.com/kaxil
  - Github: https://github.com/kaxil/
  - LinkedIn: https://www.linkedin.com/in/kaxil/

Thank You!