



GitHub's Airflow Journey Lessons, Mistakes, and Insights

Oleksandr Slyntko

3.0



3.0

About me



Data Engineer at GitHub for past 3,5 years

Have lots of experience building and operating development platforms

Love upgrading software

Airflow at GitHub

1000

active DAGs

70

teams

50k+

tasks per day

15

most contributed repo inside GitHub

GitHub is running Airflow for 9 years

Since 9 years ago

GitHub

- Got acquired by Microsoft
- Introduced GitHub Actions
- Acquired npm, Dependabot and CodeQL
- Introduced GitHub Copilot

Airflow

- Added CHANGELOG
- Added Kubernetes executor
- Graduated from Apache incubator
- Extracted DAG processor from scheduler
- Added RBAC and audit

Since 9 years ago

GitHub

- Got acquired by Microsoft
 - Introduced GitHub Actions
 - Acquired npm, Dependabot and CodeQL
 - Introduced GitHub Copilot
-
- Got 18k pull requests

Airflow

- Added CHANGELOG
 - Added Kubernetes executor
 - Graduated from Apache incubator
 - Extracted DAG processor from scheduler
 - Added RBAC and audit
-
- Got 37k pull requests

11 lessons about growing Airflow usage

How it started

- Needed to get some data for billing in June 2016
- Added more and more DAGs
- Everything is critical and important

What if we monitor all DAGs for failures?

Getting too many alerts

A screenshot of a GitHub issue page titled "Getting too many alerts". The page is displayed in a dark mode browser window. At the top, there's a navigation bar with a GitHub icon, a title bar showing "github.com", and a search bar. Below the header, the issue details are shown: "opened on Aug 9, 2017 · edited by [redacted]". On the right, there are "Edits" and a three-dot menu button. The main content area contains a text block and a chart.

Lately there's been an uptick in alerts generated by [Airflow](#). I looked into the nines data, and we're consistently generating two or more SEV2 alerts per day, and many days it's much more than that. My hunch is that many of these alerts aren't actionable and shouldn't be generated at all or should be SEV3 and some alerts should ping someone besides data-engineering. I've reviewed many of the DAGs listed here and have some proposals below.

Alerts per day

The chart shows the count of alerts per day. The y-axis is labeled "count" and ranges from 5.0 to 17.5. The x-axis represents days, though the specific dates are not labeled. The data shows a general upward trend with significant fluctuations, peaking around 18 alerts on one day.

Day	Count
1	9.0
2	8.0
3	8.0
4	7.5
5	10.0
6	15.0
7	11.0
8	6.0
9	10.5
10	18.0
11	7.0
12	11.0
13	9.0
14	7.0
15	7.0
16	12.0
17	5.0
18	5.0
19	5.0
20	4.0
21	4.0
22	4.0
23	6.0

What if we monitor ~~all~~ some
DAGs for failures?

DAG monitoring is not enough

- You need to monitor underlying infrastructure
- Possibly monitor that DAGs finish at expected time

We need more data for analysis

We need more data for
analysis
Let's add more people

Lesson 1

ETL is too much for a single team

Self-serve for ETL



Make sure code has owners

- Use CODEOWNERS file
- Data owners are not always codeowners
- Someone who will debug and fix the issue when DAG fails
- Someone who can approve code
- Test it in CI and Airflow continuously

Follow Airflow best practices



Provide example DAG

- No one wants to learn the code, it is easier to copy
- Keep it up-to-date and running in your cluster

Provide example DAG

- No one wants to learn the code, it is easier to copy
- Keep it up-to-date and running in your cluster
- Fix deprecations

Provide example DAG

- No one wants to learn the code, it is easier to copy
- Keep it up-to-date and running in your cluster
- Fix deprecations
- Do not create custom DAG class

Do linting and code format

- ruff
- SQLFluff
- Unit tests for required params in DAGs



Make backfills simple for users

- DAG to do backfills of other DAGs with all the steps and configs

Check operators and connections

- Know your connection specifics and issues and test them

20%

**Increase in number
of PRs to Airflow
DAGs**

Year comparison, excluding library code

100%

**Increase in number
of issues with
Airflow DAGs**

10%

increase for issues for our team

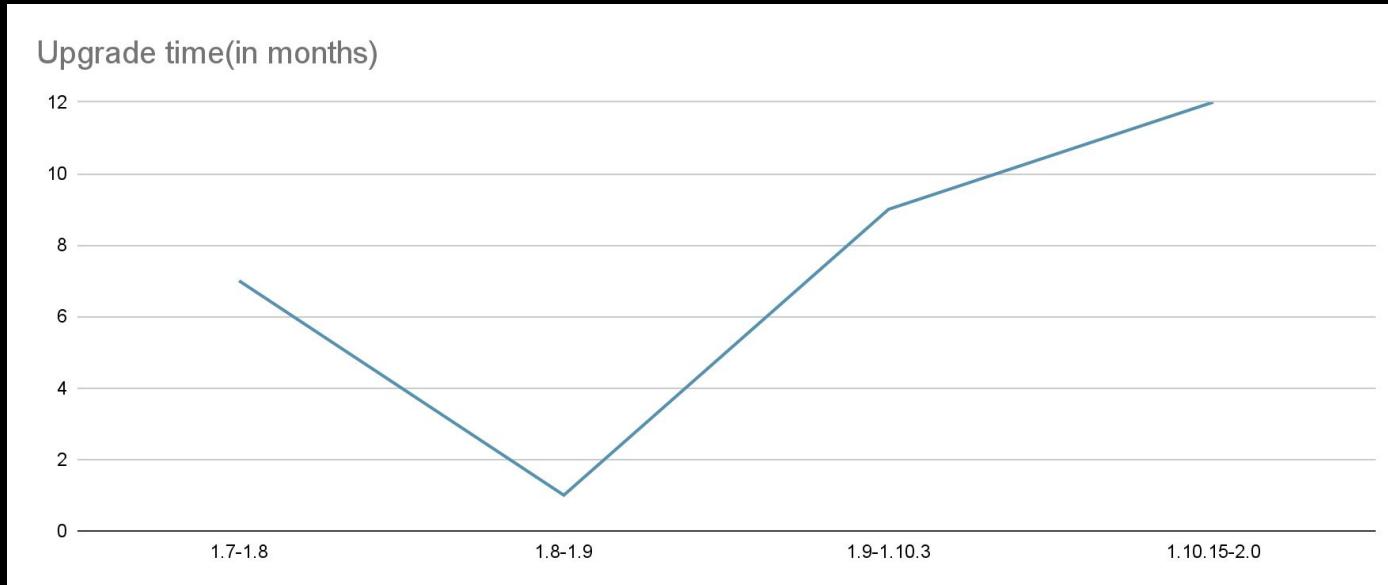
30%

more DAGs

Running a platform



Upgrades



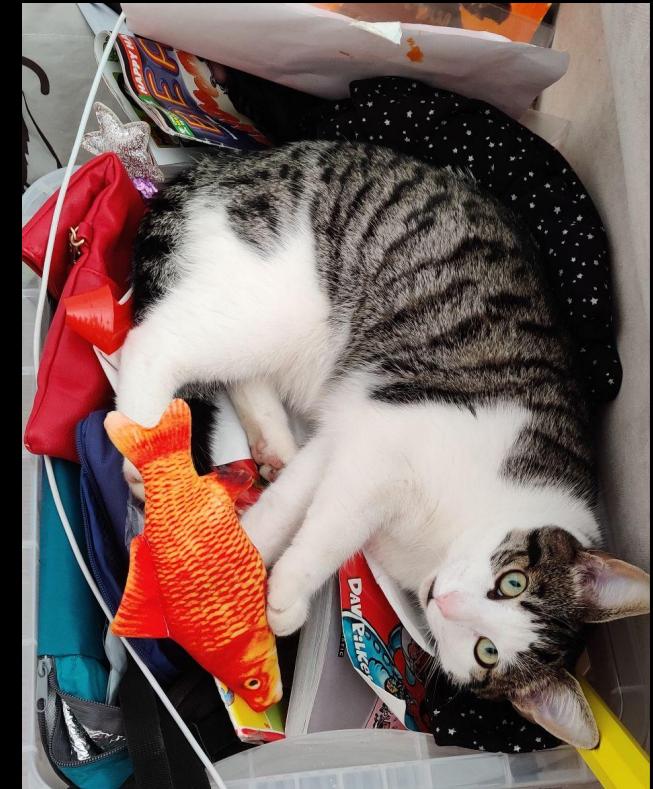
Year long upgrade

The screenshot shows a GitHub issue page with the following details:

- Title:** [Incident] Airflow Upgrade Issues #2435
- Status:** Closed (indicated by a purple button)
- Author:** Data Vertical (represented by a GitHub icon)
- Opened:** Jun 28, 2022 · edited by [redacted]
- Labels:** Edits, ...
- Section:** **Data Vertical**
- Category:** Data Platform
- Section:** **Problem**
- Description:** Listed the issues we are still working on in Mitigation that are causing DAGs to fail and miss SLAs.
- Link:** More details here: https://docs.google.com/document/d/15zggVqfM0hwZUbEDwATmFl991swmxel2Hazt7gLB_V8/edit
- Section:** **Mitigation**
- Section:** Top Priority:
- Items:**
 - Scheduler Throughput is Low: We enabled the HA schedulers and have 4 schedulers running.
 - [Repair Item] Sensors are being stuck in scheduled state but only slowly going into the Sensor pool #2422
 - [Repair Item] Airflow k8 Scheduler Failure Restarts with KubernetesJobWatcher #2420

Use open source knowledge

- Do not copy just code
- Contribute back PRs and issues



Simplify setup and development

- Codespaces/devcontainers - same setup inside container

Ephemeral dev environments

- One staging is not enough
- It is hard to test upgrades on single environment



alex-slynsko 1:11 PM

.airflow canary <https://github.com/github/airflow-sources/pull/42018>

x #42018 Use different image for postgres

```
## Overview [ 🐕 ](https://github.githubassets.com/images/icons/emoji/octocat.png "
    🐕")
```

This PR replaces postgres image in hope it will fix canary.

Validations

Post-PR Checklist

- If any files **outside** of the `dags` directory were modified, run `.airflow deploy` in `#data-engineering-ops`.

github/airflow-sources | Sep 18th | Added by GitHub



Hubot APP 1:11 PM

Canary deployment started for <https://github.com/github/airflow-sources/pull/42018> by alex-slynsko.



FlowBee APP 1:11 PM

@alex-slynsko is rolling out an airflow canary for `replace-postgres` using image tag `downstream-0dd30cd`. This may take several minutes.

@alex-slynsko's deploy of `replace-postgres` to `canary` is complete. The canary is available at <https://airflow-canary.githubapp.com/alex-slynsko>

Simplify CI

- It should be easy to change Airflow versions
- The test setup should be simple

Upgrade packages

- Use Dependabot

Time to major upgrade

4

months

Reduced issues by

20%

PR number increased by

30%

Things to do



Test operators and
connections in DAG



Make sure you have dev
environment to test



Fix deprecations

Things to do

Contribute to Airflow



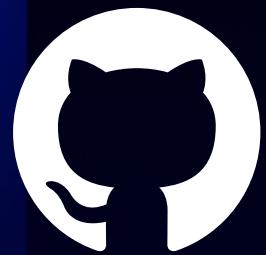
Test operators and connections in DAG



Make sure you have dev environment to test



Fix deprecations



Thank you



@alex-slyntko