Usability and Debuggability
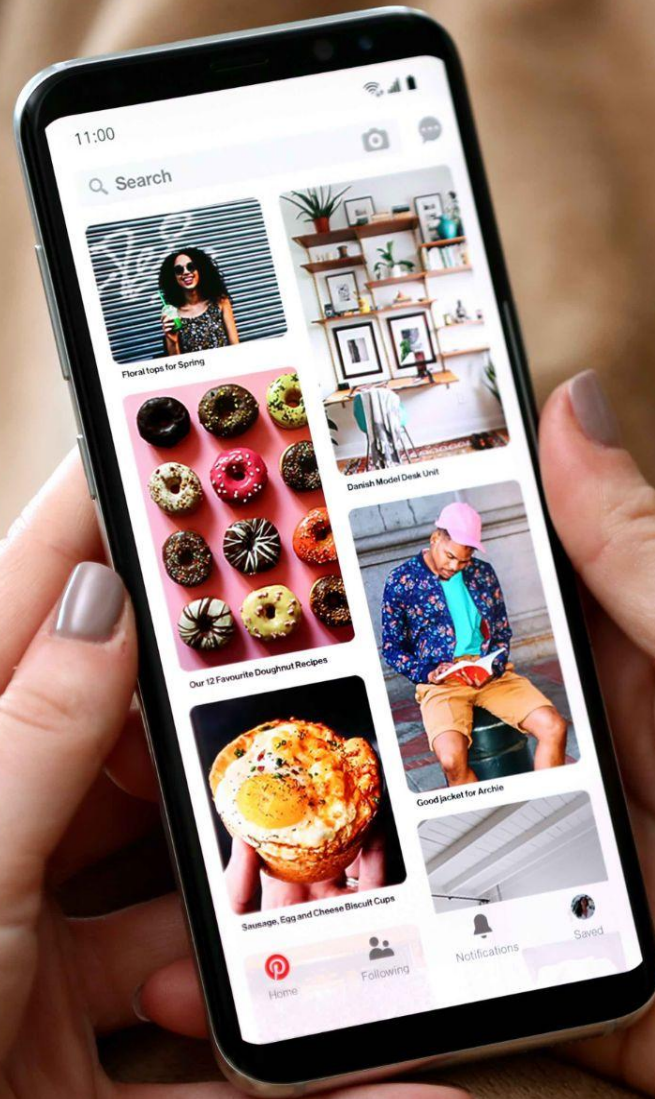## Airflow at Pinterest

Ace Haidrey, Dinghang Yu, Yulei Li
Workflow Platform

2022/05/25

Pinterest

Agenda

- Workflow Discoverability and Searchability

- Workflow Debuggability

- Workflow Profiling

Pinterest - a one-stop shop for inspiration, new products and new possibilities.

Scale

1.3K+
ACTIVE USERS

14K+
DAILY DAG RUNS

131K+
DAILY TASK INSTANCES

160K+
DAILY UI/API REQUESTS

Discovering and Searching

Recently Visited Dags

Favorite Dags

Is there a way I can save a subset of dags to display first?

Dag Template Type - Search Templates

Is there a filter I can setup for my team to access dags we care for?

DAGs

ⓘ You recently visited DAGs: migrated.canary.jss_pinball_wor...

★ Favourites | asdfasdf ❌ ▲ | ⓘ To sea

Currently Appli | **platform**

Add New ➕

Share Template

my_dags

asdfasdf ➤ ✎ 🗑

	ⓘ	DA
✐	On	migrated.canary.jss_pinball_workflow_w_k8s_pii
✐	On	migrated.canary.jss_pinball_workflow_w_k8s_pii_merged_pod

---

Create a new User DAG Search Template. ✕

Add Filter ▾

NOTE: Logical and operator is when you select multiple filters.

✕ Owner	Starts with ▾	Owners
✕ Dag Id	Starts with ▾	Dag Id
✕ Tier	Starts with ▾	Tier
✕ Project	Starts with ▾	Project

Template Name

NOTE: Template name is auto-formatted to all lower cases with max length 40 characters, spaces are replaced by underscore (_).

Cancel | Share | Preview | Create | Create and Apply

Cross Cluster Search

There are 3 clusters. How do I find the dag I need to view without checking each?

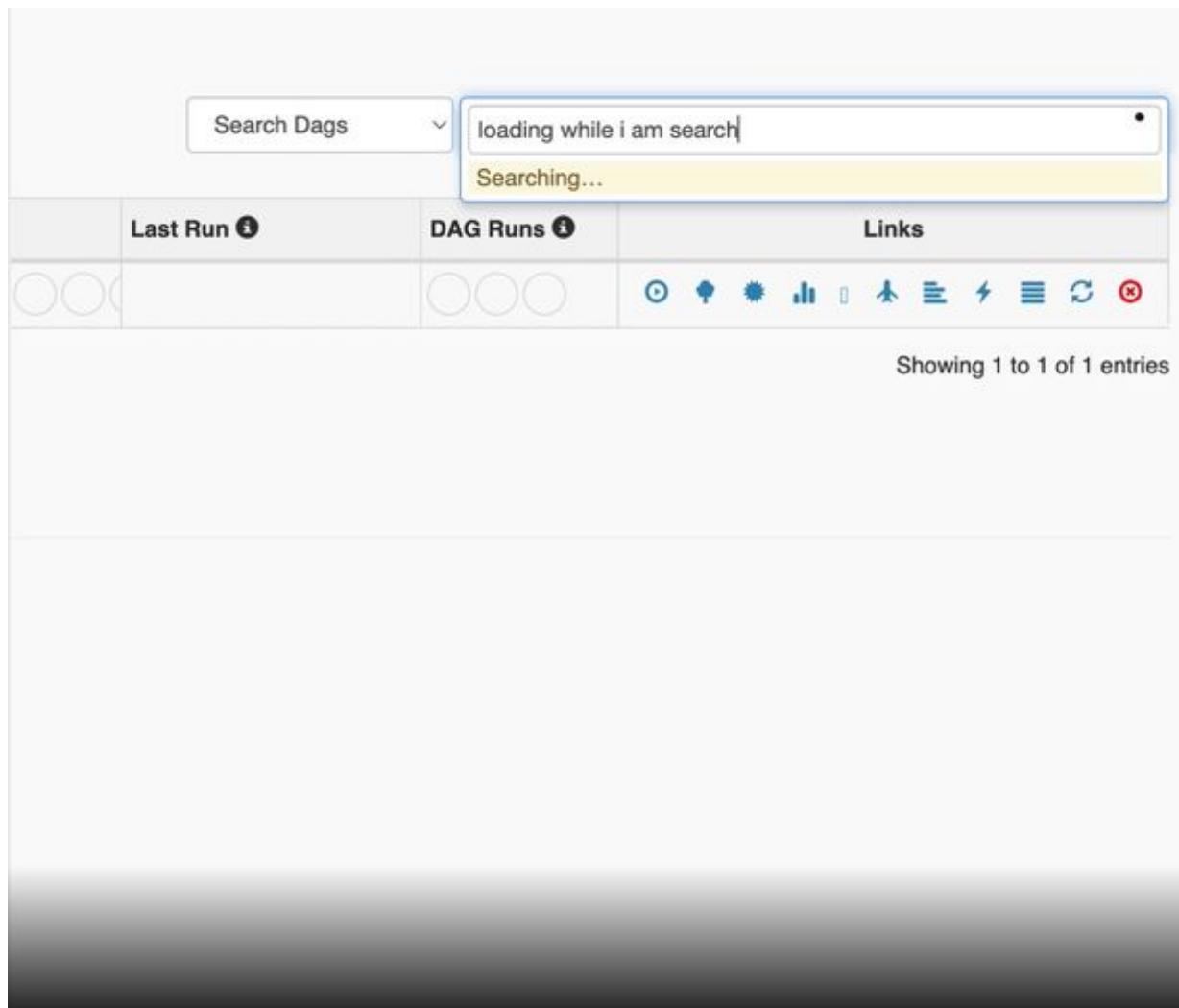# Tour Guides

Spinner has all these cool features, but how to know when to use and what to use?
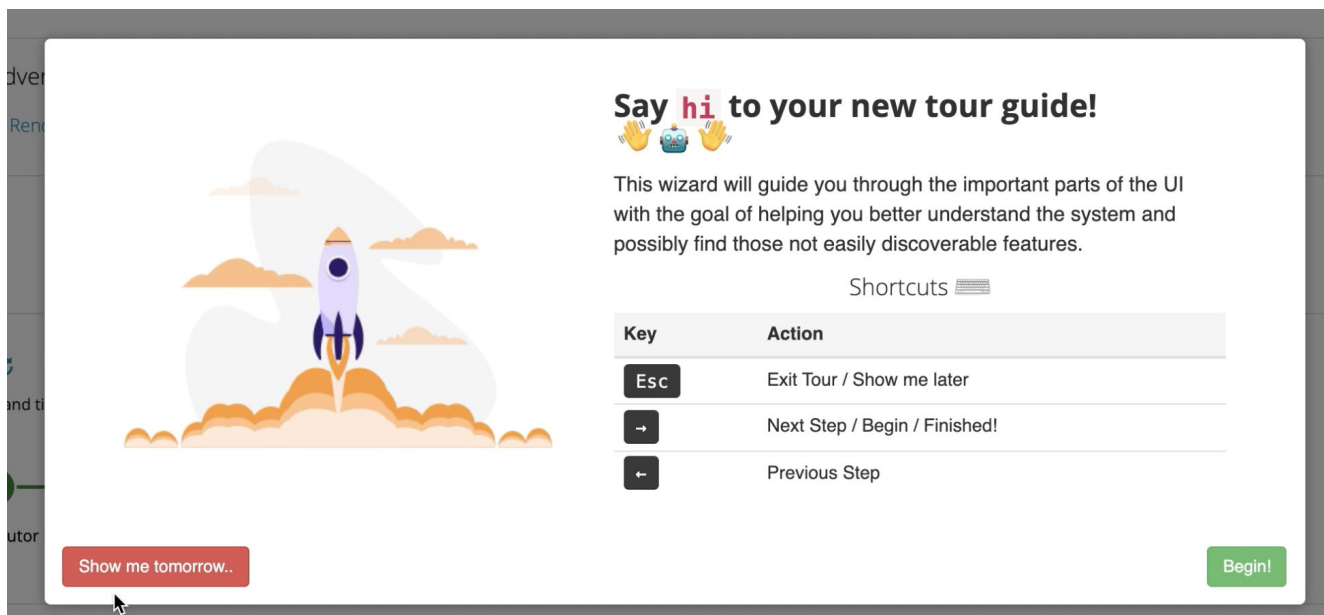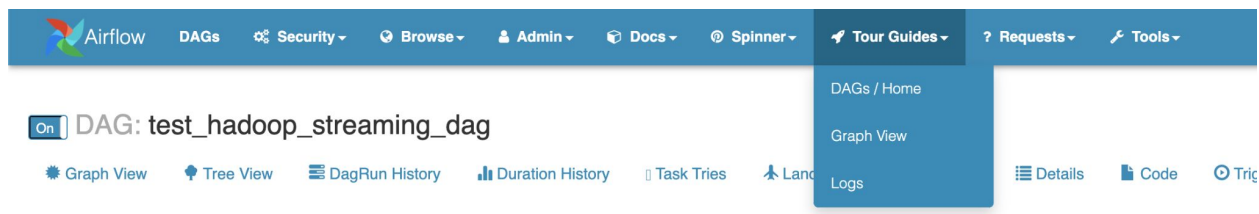
Tour Guides Demo

Debuggability

Unified Testing

How can I test changes end to end?

- Single script to bring up a local testing container

User's personal Spinner System
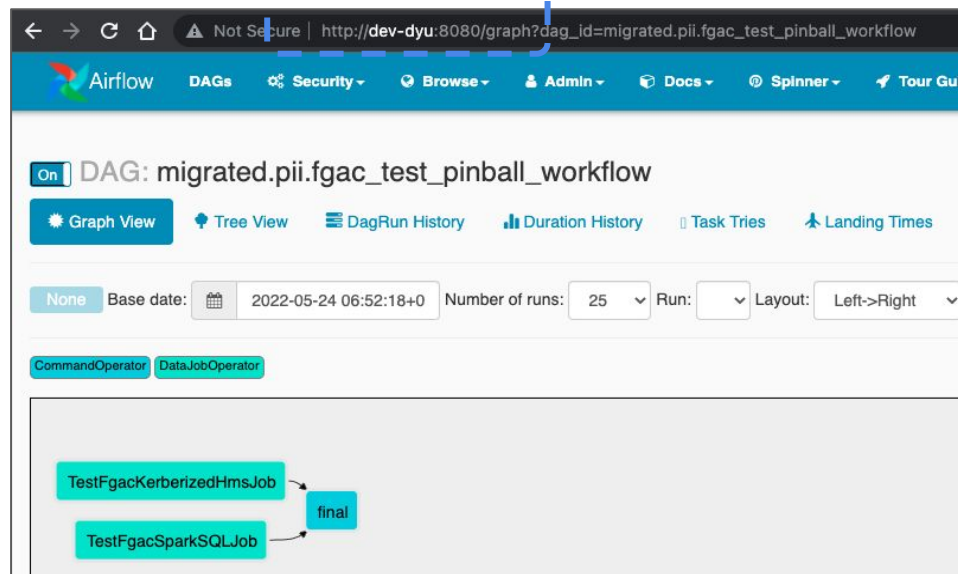
Managed Testing Envs

How can I test experimental changes in prod env without affecting product dags?

- A/B testing env

support shadow runs of workflows without duplicating wf definition
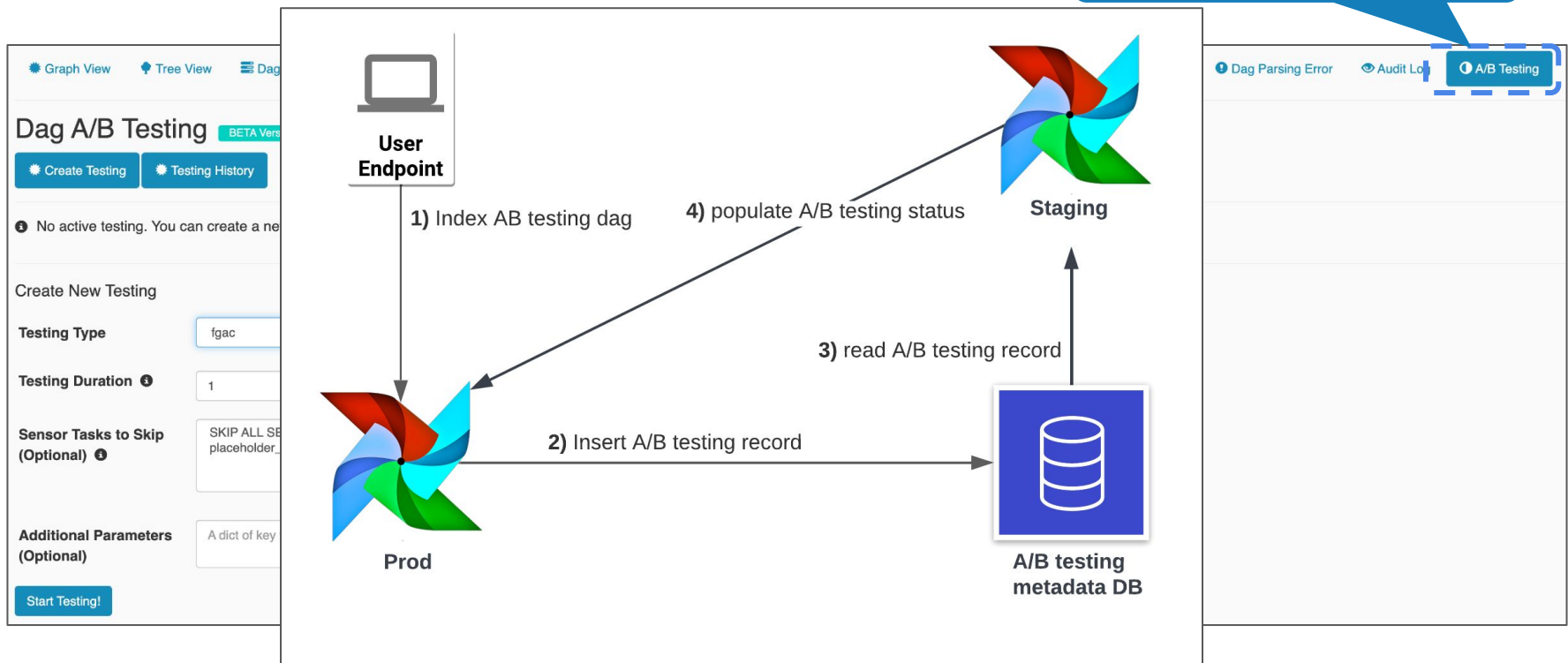
Managed Testing Envs

How can I test in development FGAC dags?

- FGAC testing env
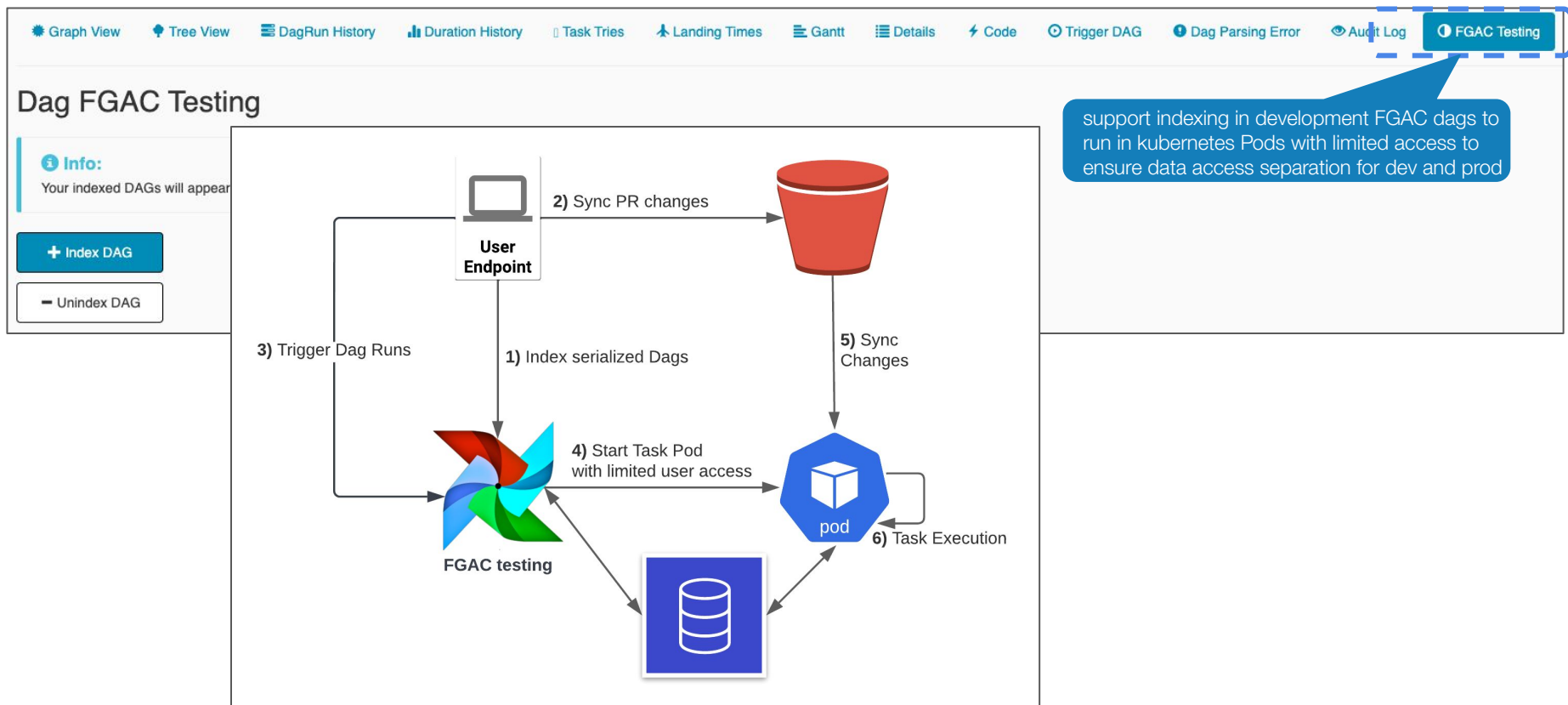
Next Deployment Time

When will my changes be deployed?

2022-05-24, 16:28:16 UTC   **Deployment Schedule**   **👤 Ace Haidrey ▾**

Next estimated deployment time:

spinner-workflows	: in 28 mins
pinboard (migrated)	: in 2 mins
pinboard (native)	: in 1 min

Last deployment finish time:

spinner-workflows	: 15 mins ago
pinboard (migrated)	: 18 hrs 20 mins ago
pinboard (native)	: 22 hrs 24 mins ago

anking_recall_stand

Non Scheduled Reason

My dag did not have an error message, and it still didn't have a run. What can be happening?

New DAG run is not scheduled!

Reason: one of task's `end_date` passed

How to fix

The next scheduled DAG run can't be created because one of its tasks has user-defined `end_date` that already passed. If this is not expected, please update its value.

×

On DAG: migrated.pii.jss_pinball_workflow_w_k8s_pii_merged_pod

schedule: @daily   nonscheduled reason: task_end_date

☀ Graph View   ♠ Tree View   ☰ DagRun History   �ⅠⅠ Duration History   ▯ Task Tries   ✈ Landing Times   ☰ Gantt   ☰ Details   ⚡ Code   ⊙ Trigger DAG   ❶ Dag Parsing Error

Critical Error Messaging Highlighting

I see the logs but it is too verbose. Is there an easier way to identify the error messages?

Debuggability features demoed in Airflow Summit 2021

- Dag Parsing Errors
- Dag Critical Path
- Build Information table

Dag Parsing Errors

This dag was supposed to have a run today but it didn't, anything went wrong?

I see an error message saying my dag is broken, what happened?

Detected 1 parsing error for this dag. See the tab Dag Parsing Error for error details.

Error Type	Dag Filepath	Dag Id	Execution Date	Detected At	Last Updated At	Stacktrace

DagImportError	/pinairflow/dags/spinner_tier_3_0/related_pins/navboost/xperf_navboost_idea_pins.py	None	None	2022-05-24 04:11:46.473727+00:00	2022-05-24 05:23:57.988122+00:00	Traceback (most recent call last):
						File "/usr/local/lib/python3.7/dist-packages/airflow/models/dagbag.py", line 321, in process_file
						  m = imp.load_source(mod_name, filepath)
						File "/usr/lib/python3.7/imp.py", line 171, in load_source
						  module = _load(spec)
						File "<frozen importlib._bootstrap>", line 696, in _load
						File "<frozen importlib._bootstrap>", line 677, in _load_unlocked
						File "<frozen importlib._bootstrap_external>", line 728, in exec_module
						File "<frozen importlib._bootstrap>", line 219, in _call_with_frames_removed
						File "/pinairflow/dags/spinner_tier_3_0/related_pins/navboost/xperf_navboost_idea_pins.py", line 125, in <module>
						  xperf_external_table_aggregator_job, post_agg_ranking_job, _ = create_x_perf_logic(dag, x_perf_args)
						File "/pinairflow/versioned/1.0.20220524040817/plugins/dag_plugins/x_perf/create_x_perf_logic_util.py", line 138, in create_x_perf_logic
						  raise AirflowException("xPerf logic: not enough arguments for the aggregation job!")
						airflow.exceptions.AirflowException: xPerf logic: not enough arguments for the aggregation job!

DAGs

Search Dags

✓ Search Dag Errors

Type dag id or dag filepath to search for dag parsing error records

	ℹ	DAG	Schedule	Owner	Tier	Recent Tasks ℹ	Last Run ℹ	DAG Runs ℹ
✎	On	ace_ixia_readable_dump_workflow	15 12 * * *	ads_review	tier3	2	2021-07-13 12:15 ℹ	219 / 2
✎	On		@weekly		tier3	7	2021-07-04 00:00 ℹ	17 / 8
✎	On		0 0 * * *		tier3	72	2021-07-13 00:00 ℹ	129 / 23
✎	On		0 */6 * * *		tier3	75 1 68	2021-07-14 12:00 ℹ	84 1 3

Build information Table

Were changes deployed at the time my task ran?

Task Instance: spark_sql_test_task on 2022-05-23T18:30:00+00:00

☀ Task Instance Details   ☀ Rendered Template   ☀ Log   ☀ XCom   ☀ Quick Actions

Log by attempts

1

Try 1 Execution Stages ⟳

Execution pipeline shows the status and time taken for each stage of your task. You can hover over the circles to see stage details and failure root cause analysis.

12.0 Sec	12.0 Sec	1 Min 37.0 Sec	9.0 Sec	1 Min 43.0 Sec	1.0 Sec

In Scheduler   In Executor   Prepare Pod in K8s   Task Runs in Worker   Task Runs in JSS Job   Executes Callbacks   End

Artifacts Information

Workflows have a lot of moving parts, and this table is to inform you what the commit sha for each major component was at the time of your execution. This will help you understand if any change made were in at the time of the run or not.

Artifact	Build Link ⓘ	Deploy Status ⓘ
Pinairflow Build	f88116bdec02bce497d10df9e66a09aa399f5f6e	Lupa Link
Pinboard Build	f027eab8dd24b6a755482a27efab4c284829abad	Lupa Link
Spinner Workflows Build	b8c24d15473fd5d54950441a29ea36b9f2d7b1d9	Lupa Link

Critical Path Finder
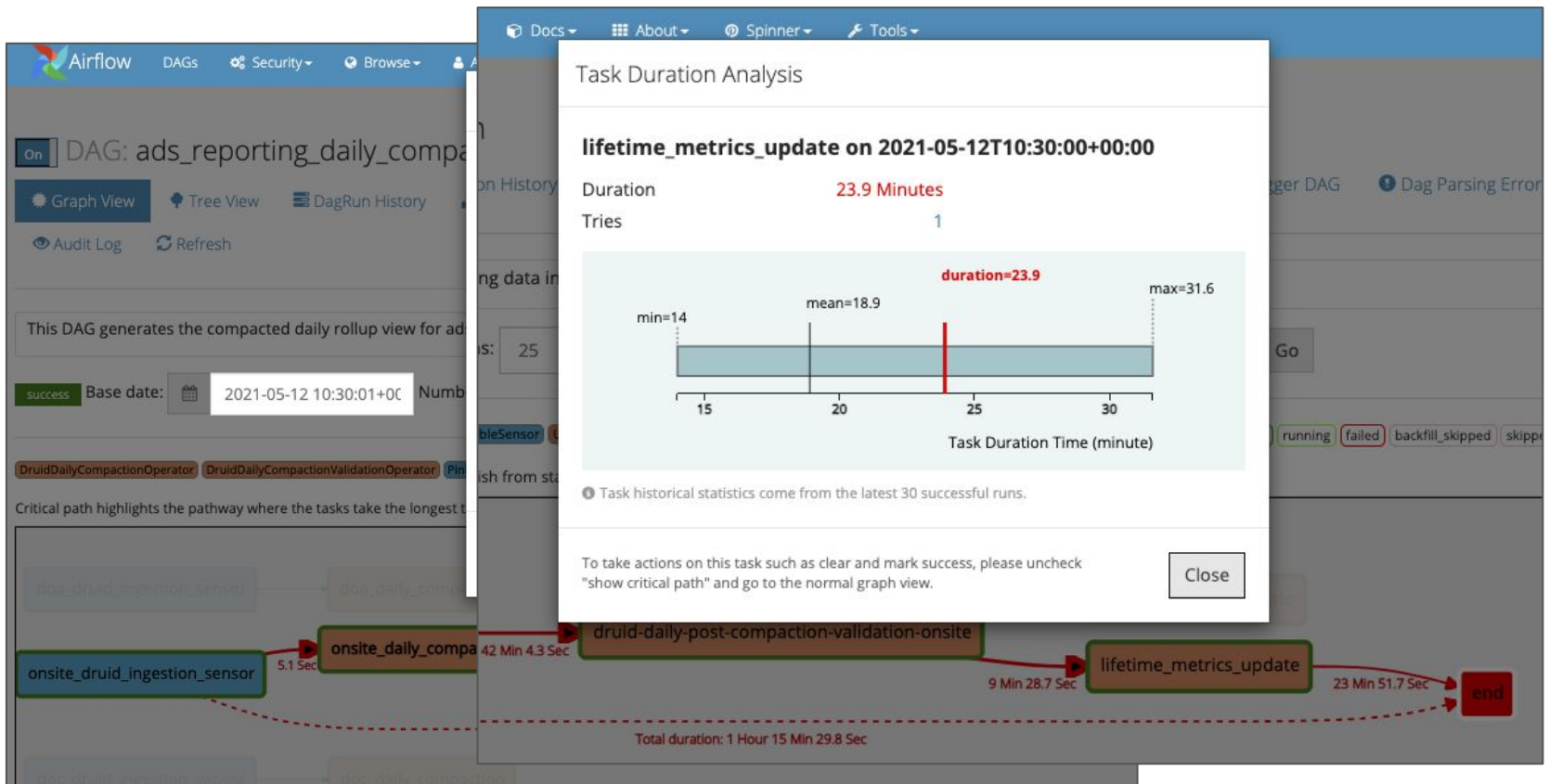
Workflow Profiling

Task Execution Memory Prediction

Each [task]

Prov... ...rride

- No one-size-fit-all solution
- Cause res...

How much memory should I allocate to my task?

My task failed with OOM error, how do I fix it?

How much memory should I allocate to my task?

4. execute task

executor

1.
memory allocation

task pod

My task failed with OOM error, how do I fix it?

memory allocation predictor

2. read memory utilization
historical data

5. record task memory utilization

5. record task execution oom

task memory utilization
metastore

Workflow Cost Display

How much computing resources do each of my workflow run use?

- Collect resource usage data from yarn cluster
  - Offline pipeline to convert the resource usage data into $

- A measurable metric to take actions on
  - Bring awareness to users
  - Target wasted resources
  - Drive workflow optimization

Average cost for each dagrun

Generate the report with customized filter (date range, task list)

Choose to view cost by vcore, memory or the two combined

Workflow Cost Demo

Incident Manager Tooling

Incident Management Admin Tooling

Incident Management Smart Retries

Is there a way for retrying all failed jobs if the platform had an issue?

Allow Auto Retry *	0
	Allow failed tasks to retry or not, 0 means disallow, 1 means allow

Triggering Dag Clarification

> **I try to trigger my dag manually but the conf doesn't work right?**

ℹ️ Click here to learn **how to use the Configuration JSON option**.

If you want to pass extra configurations to the DAG run, you can put the configuration in the form. Please note the type of your configuration must be a dictionary, and other formats are disallowed. You can take look at the below examples or read on the open source doc to better understand the usage.

A **valid** configuration example:

```
{ "input_key": "input_value" }
```

Both the **key** and the **value** are enclosed in **double quotes (")** and everything is wrapped in **curly braces ({})**.

An **invalid** configuration example:

```
"input_key": "input_value"
```

How can I **access** this configuration **in my DAG?**:

Configuration JSON:

```
{ "input_key": "input_value" }
```

Using `input_key` in the code:

```python
from airflow import DAG
from airflow.operators.bash import BashOperator
from airflow.utils.dates import days_ago

dag = DAG("example_parameterized_dag", schedule_interval=None, start_date=days_ago(2))

parameterized_task = BashOperator(
    task_id='parameterized_task',
    bash_command="echo value: {{ dag_run.conf['input_key'] }}", # ==> using the given "input_key"
    dag=dag,
)
```

> **Note**
>
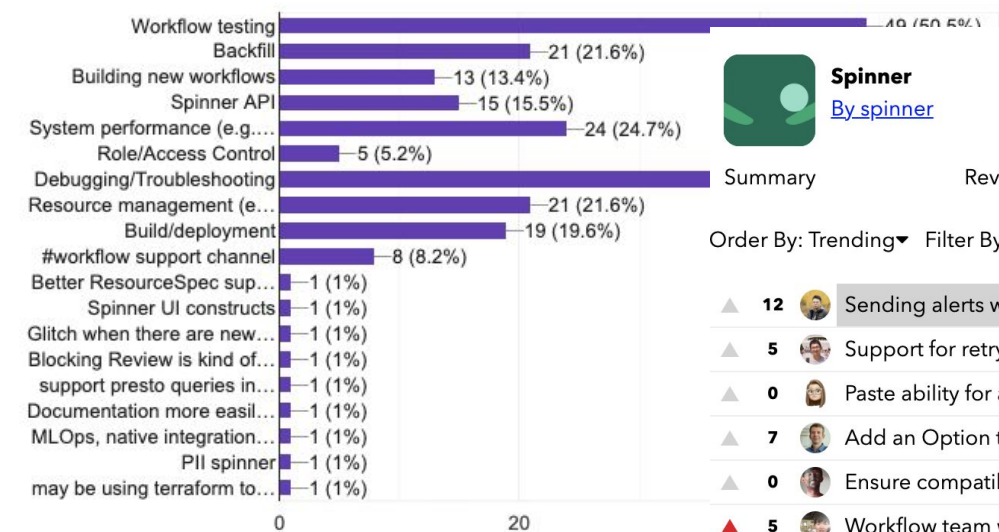> The parameters from `{{ dag_run.conf }}` can only be used in a template field of an operator.

Wrap Up

Identifying What Users Want

Focus Areas

Select the top 3 areas you'd like enhanced

97 responses

Workflow testing — 49 (50.5%)
Backfill — 21 (21.6%)
Building new workflows — 13 (13.4%)
Spinner API — 15 (15.5%)
System performance (e.g.... — 24 (24.7%)
Role/Access Control — 5 (5.2%)
Debugging/Troubleshooting
Resource management (e... — 21 (21.6%)
Build/deployment — 19 (19.6%)
#workflow support channel — 8 (8.2%)
Better ResourceSpec sup... — 1 (1%)
Spinner UI constructs — 1 (1%)
Glitch when there are new... — 1 (1%)
Blocking Review is kind of... — 1 (1%)
support presto queries in... — 1 (1%)
Documentation more easil... — 1 (1%)
MLOps, native integration... — 1 (1%)
PII spinner — 1 (1%)
may be using terraform to... — 1 (1%)

0   20

User Surveys

Request Forms

Spinner
By spinner

[Edit] [Open]

Summary   Reviews   Analytics   **Requests**   Uptime

Order By: Trending▾ Filter By: Open▾

+ Create a New Request

▲	12	Sending alerts when a dagrun or workflow can't finish within its SLO	›
▲	5	Support for retrying a job with new parameters
▲	0	Paste ability for a comma separated task list into the backfill form
▲	7	Add an Option to Turn off Alerts for backfill runs
▲	0	Ensure compatible artifacts are used in jobs that depend on multiple deployment pipeli…
▲	5	Workflow team will create JIRA tickets for the requests have more upvotes! (>=5)
▲	3	Differentiate `Success` with `Marked as Success`
▲	3	Support send_dag_level_fail_email param similar to send_dag_level_failure_message (…
▲	3	Frequent pinboard deploys to both migrated & native environments
▲	2	Differentiate the PROD Spinner UI and devapp Spinner UI
▲	2	support multi-partition sensor
▲	2	Add a `[DAG Final]` node in Graph View
▲	1	add Callback hook for native spinner when job is marked failed/success
▲	1	Run access control scripts automatically for spinner users

Wrap Up

- Our goal is to make the Spinner workflow system more usable
  - Improve the workflow discoverability for our multi-environment setup
  - Streamline the debugging experience
  - Profile workflow and surface data on UI
  - …

- Meanwhile, free the system admins from heavy & repetitive user support routine
  - Drop in our user support load

number of user questions
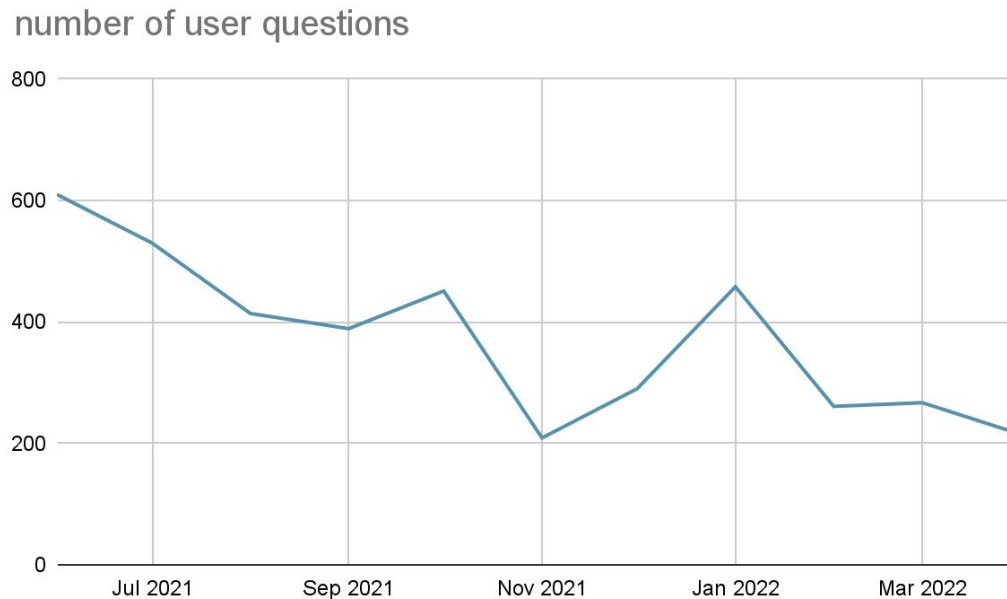
Future Work

- Improvements to the cost report to include other infra costs
- Search by the task_id, not just the dag_id
- Features to further reduce noise and identify errors in the logs
  - Improvement in log filtering tooling
  - Link source code in error messages
- Upgrade Airflow Version to 2.0+
- Blog post regarding UI feature development
- And More!

Additional Resources

- Airflow Summit 2021:
  - [Pinterest's Migration Journey](#)
  - [Usability Improvements for Debugging](#)
- Airflow Summit 2020:
  - [Airflow as the next gen of workflow system at Pinterest](#)
- Tech Blog
  - [Spinner: Pinterest's Workflow Platform](#)
  - [Spinner: The Mass Migration to Pinterest's New Workflow Platform](#)

Thank you!

Ace Haidrey
Software Engineer

Ashim Shrestha
Site Reliability Engineer

Apoorv Gaur
Software Engineer

Dinghang Yu
Software Engineer

Evan Li
Engineering Manager

Hannah Chen
Technical Program Manager

Hilda Johny
Software Engineer

Kuo-Kai Hsieh
Software Engineer

Numan Aral
Software Engineer

Yulei Li
Software Engineer

The Faces Behind The Work