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



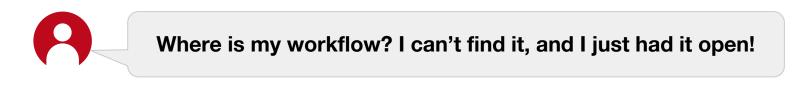
Pinterest - a one-stop shop for inspiration, new products and new possibilities.

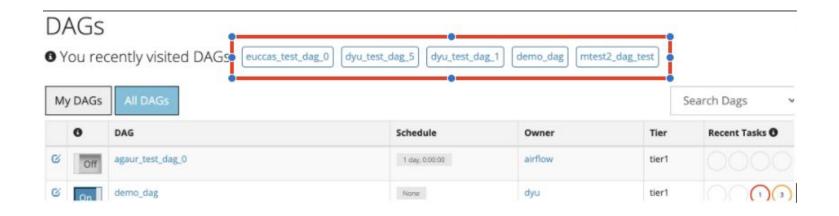


Scale

Discovering and Searching

Recently Visited Dags





Favorite Dags

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

DAGs

• You recently visited DAGs: No item to show yet...

*	Favourite	es	ZA	Airflow	DAGs	¢° Security -	Ø Browse +	🛔 Admin -	Docs-	⑦ Spinne	r 🗸 🖸 Tour Guides 🗸
	0	DAG	DAG I	Import E	rrors (80)						
C	On	ace_ixia_readable_dump_workflow	DA	Gs							
Ø	On	actalike_eval_expansion_v4 Click this star to Favourit	🚯 Yo	u rece	ntly visit	ed DAGs:	ace_ixia_readable	_dump_workflow	ads_optimiza	tion.check_co	nversio) consumer_ap
Ø.	On	actalike_expansion_v4	\star Fa	avourites	Select	a DAG Search T	emplate	•	To search DAGs	or go back to A	Il DAGs, click on the Favourite
						DAGs that yo	u have chosen	as your favo	urite by click	ing on the	star next to their na
					AG						Schedule
						pp_strike_syster	m			*	0 10 * * *
			©. [On	ser_role_ba	ackup				*	@hourly
				< 1 aused DA							

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...

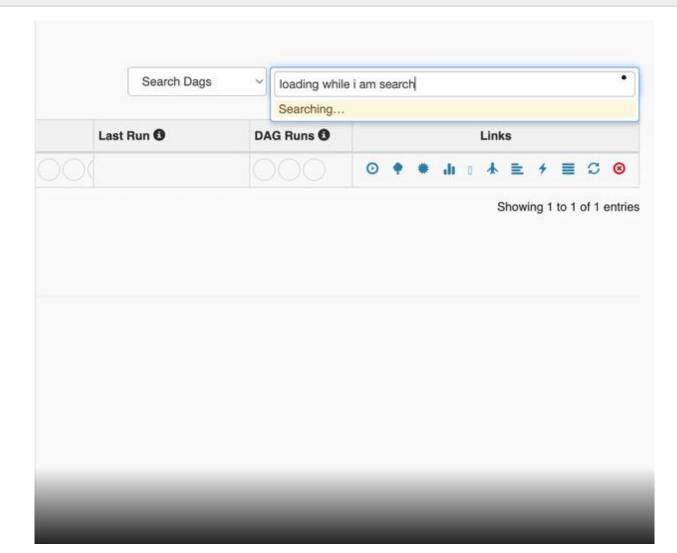
*	Favour	ites	asdfasdf	8 -	• • • • • • • • • • • • • • • • • • •			
Currently Appli		Appli	Add New 🕂		olatfo			
		Abbu	my_dags Share Template					
	0	DA	asdfasdf	Ó 🖉 🛈				
©.	On	mig	ated.canary.jss_pinbal	_workflow_w_k8s_pii				
C	On	mig	ated.canary.jss_pinbal	_workflow_w_k8s_pii_merged_pd	bd			

× Owner	Starts with 🔹] [Owners
× Dag Id	Starts with •		Dag Id
× Tier	Starts with •		Tier
× Project	Starts with •		Project
emplate Nam	e		

Cross Cluster Search



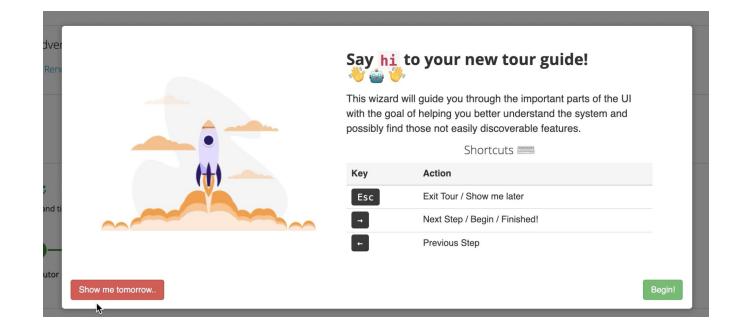
There are 3 clusters. How do I find the dag I need to view without checking each?



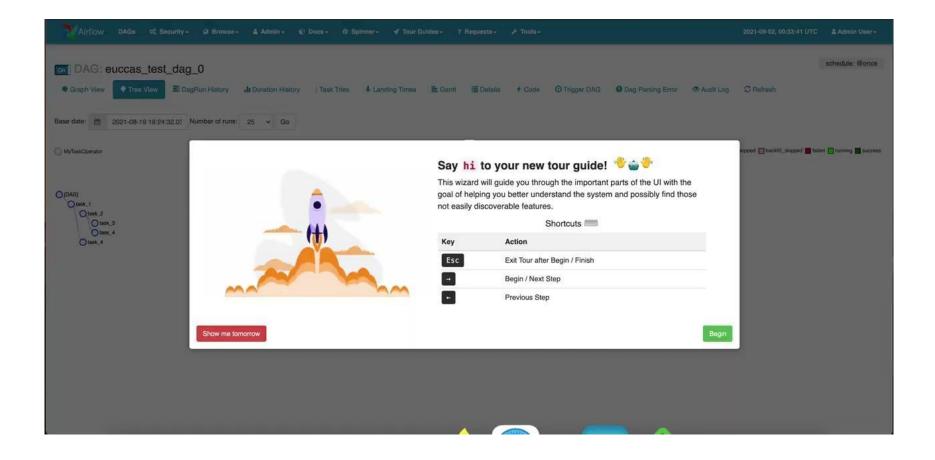
Tour Guides

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

Airflow	DAGs 🕫	Security -		🛔 Admin -	🗊 Docs 🗸	O Spinne O	r,	🛷 Tour Guides -	? R	equests -	≁ Tools+	
								DAGs / Home				
on DAG: te	st_hadoo	op_strea	aming_da	g				Graph View				
Graph View	Tree View	🖀 DagF	Run History	Duration Histo	ory 🛛 Task	Tries 🛧	Land	Logs		🔳 Details	Code	🕑 Triç



Tour Guides Demo

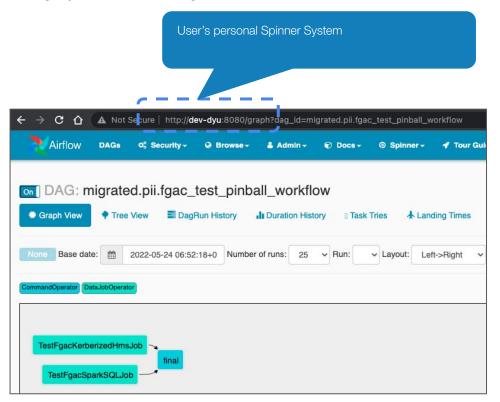


Debuggability

Unified Testing

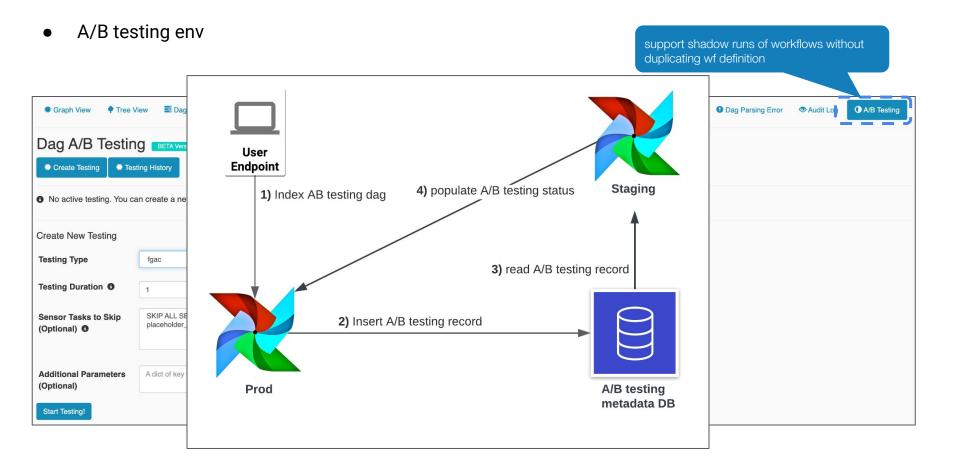


• Single script to bring up a local testing container



Managed Testing Envs

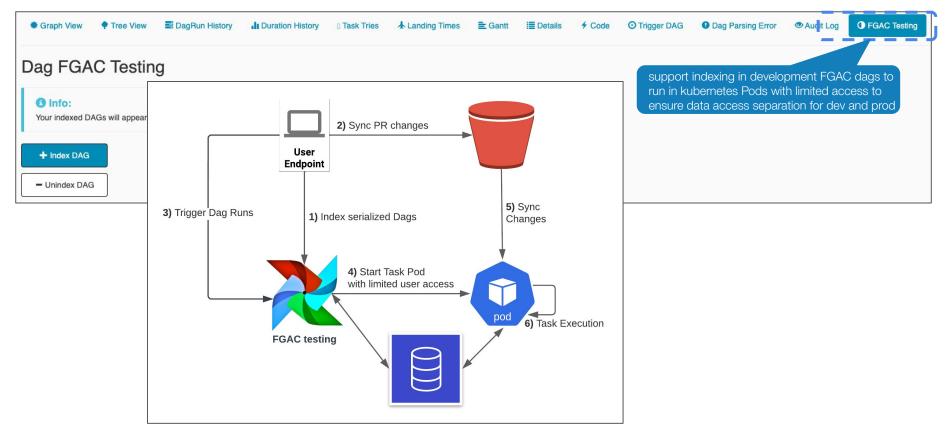
How can I test experimental changes in prod env without affecting product dags?



Managed Testing Envs

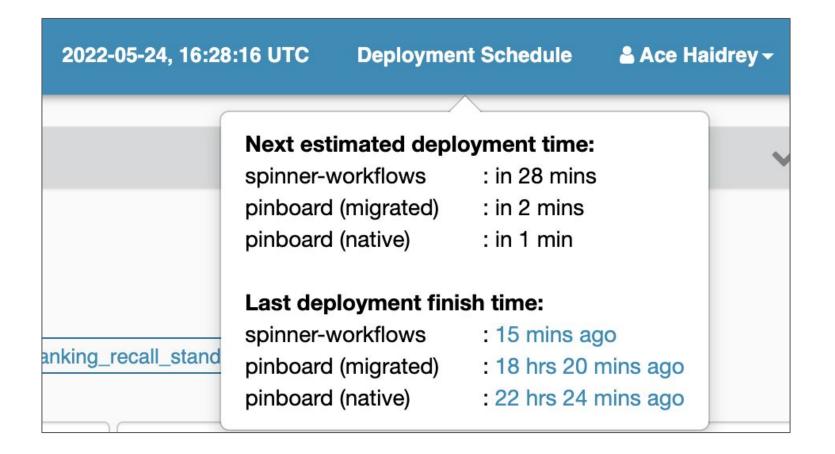


• FGAC testing env



Next Deployment Time





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 r Reason: one of ta How to fix	notscheduled! ask's end_date p	assed										×
The next sche	duled DAG run ca	n't be created because o	one of its tasks has user	-defined end_da	te that already passed	d. If this is not	expected, pleas	se update its	value.			
on DAG: m	nigrated.pii.	jss_pinball_w	orkflow_w_k8	s_pii_merg	ged_pod			sched	dule: @daily	nonse	cheduled reason: task_en	d_date
# Graph View	† Tree View	DagRun History	Duration History	Task Tries	Landing Times	E Gantt	E Details	4 Code	O Trigger D	AG	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?

File "/pinairflow/dags/test_tier_1_0/dyu_test_dag_diff_ error s.py", line 81, in python_knox_issue
<pre>primary_key = AutoKnox(knox_key).get_primary()</pre>
File "/usr/local/lib/python3.7/dist-packages/knox/initpy", line 401, ininit
<pre>self.cached_key = self.register_and_get(timeout=timeout)</pre>
File "/usr/local/lib/python3.7/dist-packages/knox/initpy", line 359, in register_and_get
raise Knox Exception (message)
knox.Knox Exception: Knox failed to get key jobservice:does_not_exit: Client says "t=2022-02-04T01:30:35+0000 lvl=eror msg=" Error getting key from daemon (hit timeout after 5 seconds);
logs for details (most recent error: open /var/lib/knox/v0/keys/jobservice:does_not_exit: no such file or directory)"". See /var/log/knox_client.log for details.
[2022-02-04 01:30:35,474+00:00] [1644 MainThread] {logging_mixin.py:90} INF0 - [2022-02-04 01:30:35,474+00:00] {stage_base.py:183} INF0 - Exiting stage TaskExecStage for task {'environme
r-test', 'dag_id': 'dyu_test_dag_diff_errors', 'task_id': 'test_knox_issue', 'execution_date': '2022-02-04T01:28:37.155957+00:00', 'try_number': 1}
[2022-02-04 01:30:36,165+00:00] [1644 Thread-1] {logging_mixin.py:90} INF0 - [2022-02-04 01:30:36,164+00:00] {task_exec_stage.py:98} INF0 - Done Root Cause Analysis for stage TaskExecSta
Error Category KNOX_ISSUE for task {'environment': 'spinner-test', 'dag_id': 'dyu_test_dag_diff_error s', 'task_id': 'test_knox_issue', 'execution_date': '2022-02-04T01:28:37.155957+00:
umber': 1}
[2022-02-04 01:30:36,648+00:00] [1644 MainThread] {taskinstance.py:1513} INFO - Marking task as FAILED.
[2022-02-04 01:30:36,780+00:00] [1644 MainThread] {logging_mixin.py:90} INF0 - [2022-02-04 01:30:36,780+00:00] {stage_base.py:162} INF0 - Entering stage CallbackStage for task {'environm
er-test', 'dag_id': 'dyu_test_dag_diff_ <mark>error</mark> s', 'task_id': 'test_knox_issue', 'execution_date': '2022-02-04T01:28:37.155957+00:00', 'try_number': 1}
[2022-02-04 01:30:37,539+00:00] [1644 MainThread] {configuration.py:246} WARNING - section/key [smtp/smtp_user] not found in config
[2022-02-04 01:30:37,539+00:00] [1644 MainThread] {logging_mixin.py:90} INF0 - [2022-02-04 01:30:37,539+00:00] {configuration.py:246} WARNING - section/key [smtp/smtp_user] not found in
[2022-02-04 01:30:37,773+00:00] [1644 MainThread] {email.py:143} INFO - Sent an alert email to ['dyu@pinterest.com']
[2022-02-04 01:30:37,773+00:00] [1644 MainThread] {logging_mixin.py:90} INF0 - [2022-02-04 01:30:37,773+00:00] {email.py:143} INF0 - Sent an alert email to ['dyu@pinterest.com']
[2022-02-04 01:30:38,084+00:00] [1644 MainThread] {callback_utils.py:555} INFO - Log job execution record to workflow_job_executions kafka topic upon job completion
[2022-02-04 01:30:38,085+00:00] [1644 MainThread] {logging_mixin.py:90} INF0 - [2022-02-04 01:30:38,084+00:00] {callback_utils.py:555} INF0 - Log job execution record to workflow_job_exe
a topic upon job completion
[2022-02-04 01:30:38,169+00:00] [1644 MainThread] {logging_mixin.py:90} INF0 - [2022-02-04 01:30:38,169+00:00] {stage_base.py:183} INF0 - Exiting stage CallbackStage for task {'environme
r-test', 'dag_id': 'dyu_test_dag_diff_ error s', 'task_id': 'test_knox_issue', 'execution_date': '2022-02-04T01:28:37.155957+00:00', 'try_number': 1}
[2022-02-04 01:30:38,539+00:00] [1644 MainThread] {taskinstance.py:1327} INFO - [JobExecutionRecord] Updating final status of JobExecutionRecord in the db
[2022-02-04 01:30:38,555+00:00] [1644 MainThread] {taskinstance.py:1350} INFO - [JobExecutionRecord] Updated final status of JobExecutionRecord in the db: JobExecutionRecord(workflow_nam
ag_diff_errors, workflow_cluster=spinner-test, job_name=test_knox_issue, instance_id=manual_2022-02-04701:28:37.155957+00:00 try_number=1, status=failed, application_id_string=, jss_id
tart_time=2022-02-04 01:30:30, end_time=2022-02-04 01:30:38.545561, project=workflow, tier=tier1, data={"pinairflow_build": "8ed9cb07bbea8dc99f274f968cc5cda9c615bf0f", "pinboard_build":
be3ac0b9dff9af8b5ecda7e01be8c", "spinner_workflows_build": "75bffae10897d829070430db87375781e304d523"})
[2022-02-04 01:30:47,124+00:00] [1517 MainThread] {logging_mixin.py:90} INF0 - [2022-02-04 01:30:47,123+00:00] {resource_utilization.py:70} INF0 - TASK_ID PROCESS: test_knox_issue PEAK_R
55 <i>A</i> Mi

Debuggability features demoed in Airflow Summit 2021

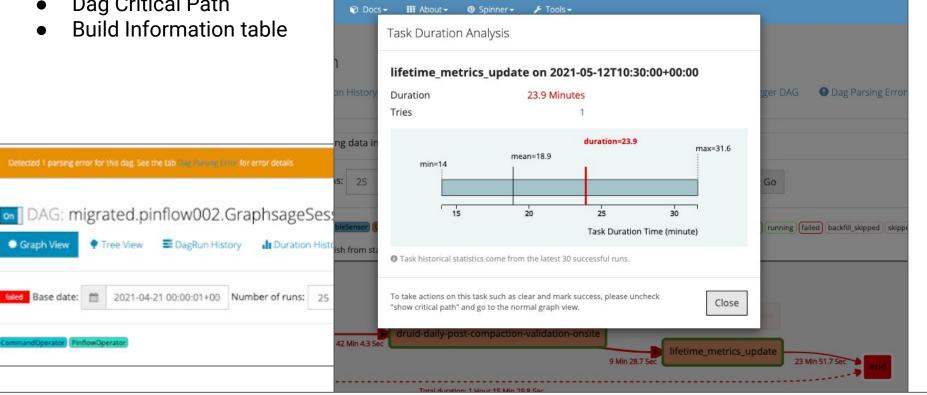
- **Dag Parsing Errors** •
- **Dag Critical Path**

Tree View

Build Information table

E DagRun History

2021-04-21 00:00:01+00



Artifacts Information

Graph View

failed Base date:

nandOperator PinflowOperat

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

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 table an immunitient 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/mp.yr", line 171, in load_source module = _load(spec) File "cfrozen importlibbootstrap>", line 696, in _load File "cfrozen importlibbootstrap>", line 677, in _load_unlocked File "cfrozen importlibbootstrap>", line 219, in _call_with_frames_removed File "cfrozen importlibbootstrap>", line 219, in _call_with_frames_removed File "cfrozen importlibbootstrap>", line 3_0/related_pins/nxtperf_navboost/gea_pins.py", line 125, in <module> xperf_extran_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 AirflowExceptions.kr/lowExceptions.kr/ent logic: not enough arguments for the aggregation job!"</module>

				Search Dags Search Dag Errors		Type dag id or dag filepath to search for dag parsing error records				
	0	DAG	Schedule	Owner	Tier	Recent Tasks	Last Run 🕄	DAG Runs 🕄		
©.	On	ace ivia roadable dume workflow	15 12 * * *	ade review	tier3	2000000	2021-07-13 12:15	219 2		
©.	On	and a subscript of	@weekly	1	tier3	00000000	2021-07-04 00:00	17 8		
©.	On	and the summary of the	00***	peuses.	tier3	n 0000000	2021-07-13 00:00	(12) (2)		
C	On		D */6 * * *	£	tier3		68 2021-07-14 12:00	84 1 3		

Build information Table



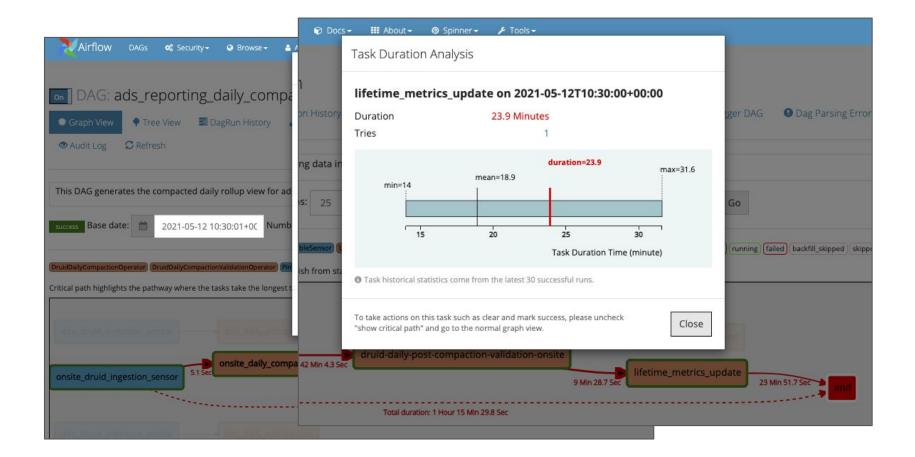
Artifacts Information

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 0
Pinairflow Build	f88116bdec02bce497d10df9e66a09aa399f5f6e	Lupa Link
Pinboard Build	f027eab8dd24b6a755482a27efab4c284829abad	Lupa Link
Spinner Workflows Build	b8c24d15473fd5d54950441a29ea36b9f2d7b1d9	Lupa Link

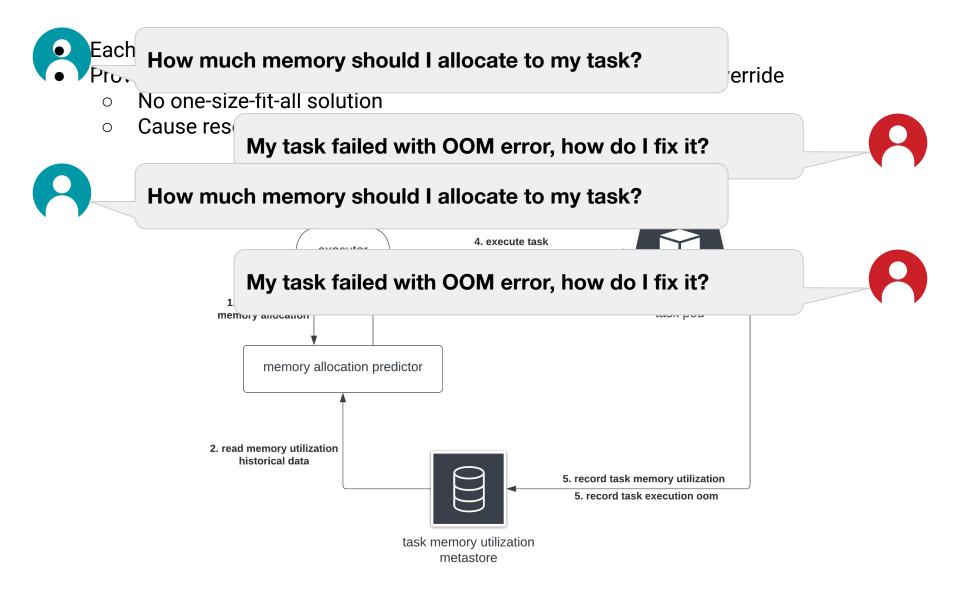
Critical Path Finder



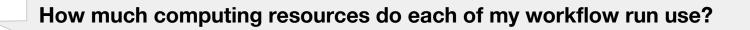


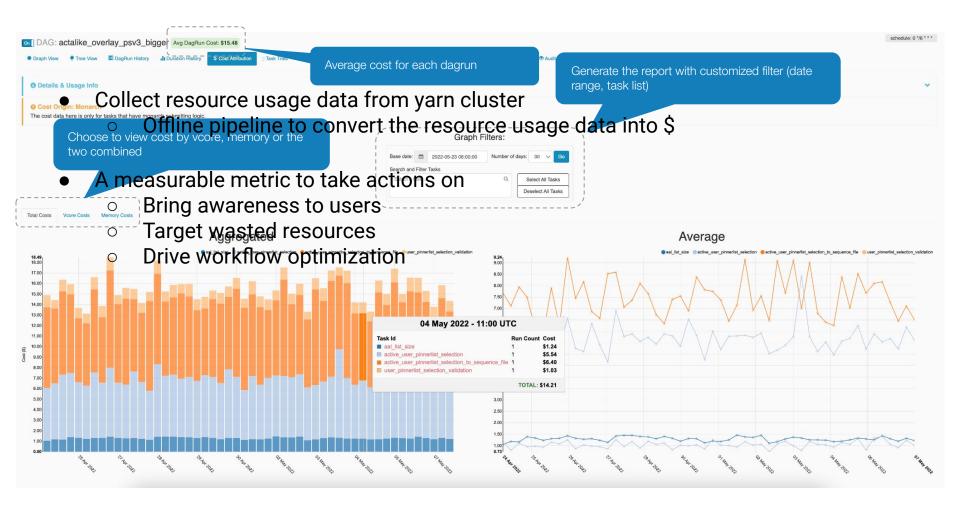
Workflow Profiling

Task Execution Memory Prediction



Workflow Cost Display





Workflow Cost Demo



Incident Manager Tooling

I have a few workflow jobs suddenly start to fail today, is there something wrong in Spinner?

3	Airflow	/ DAGs 📽 Security	/• 🛛 Browse•	🔒 Admin v	🕄 Docs -	🎟 About -	🛛 Spinner -	🗲 Too	5*	20	21-05-06, 20:24:54 UT	°C 🛛 🛔 Euccas Chen 🛨
DA	G Import	Errors (54)										~
E	o Atten	tion Workflow Use	ers !									×
		IM-8464 Incident: Ta Increased task failures ca							00:00 Omitigated at 2021-05-06T13:20:45+00:0 Details in slack #im-8464			
(reported	III Spinner Change Fre	eze is In Effect !!!	🌢 started at 2	:021-05-05T12:0							
		Production Change Freez workflow changes in spin							reeze to end on Monday, May 10, 2021 at 10an you!	n PT. During the chang	e freeze Spinner does	not allow
	AGs											
D	AGS								Search:			
	0	DAG			Schedule	e Owner		Tier Re	cent Tasks 🖲	Last Run	DAG Runs 3	
G	On	ace_ixia_readable_dump_wo	rkflow									Links
	17. Contraction 19. Contractio 19. Contraction 19. Contraction 19. Contraction 19. Contractio				1512***	ads-revie	w t	tier3	00000000000	2021-05-05 12:15 🕄		Links
C		actalike_eval_expansion_v4			1512*** @weekly			tier3 (2 tier3 (7)0000000000)0000000000	2021-05-05 12:15 2021-04-25 00:00 3	000	
G	On	actalike_eval_expansion_v4 actalike_expansion_v4				ralekar	t					⋻⋪⋕⋴⋴⋩⋸⋭≣ <mark>∁⊗</mark>
	On			nains	@weekly	ralekar	t	tier3		2021-04-25 00:00 3		⋑⋪⋪⋼⋏⋺⋭⋽⋽⋑
Ø	On On On	actalike_expansion_v4		nains	@weekly 0 12 * * *	ralekar pnunez zhechen		tier3 (13		2021-04-25 00:00 3 2021-05-05 12:00 3		●℃≣₫≦⊁₩₩₽₩₽ ●℃≣₫≦₩₩₽₽ ●℃≣₫≦⊁₩₩₽₽©

Incident Management Admin Tooling

Ad	d New In	ncident											
Status *					1.000	ported	d, or closed						
Reporter *					ro	pt							
Name *					Na	me							
De:	All Incidents Search+												
Jira	Record										ord Count: 5		
			Id I	Status I	Reporter I	Name İ	Description	1	Jira 1	Severity I	Allow Auto Retry I	Created At I	Updated At
Sev	- a		377	closed	euccaschen	Incident: Task failed with OOM error (exit code -9) in Spinner	Increased task failures caused by OOM in Spinner. Workflow tasks failed with error: INFO settings for Spinner native tasks, you can RETRY your previously failed tasks and the new	- Task exited with return code -9. Details in slack #im-8464. [Update 3:07PT] we have just bumped the default memory memory settings should be sufficient!	IM- 8464	1	No	2021-05-06 20:18:13+00:00	2021-05-07 20:22:40+00:00
Imp	- a	. 7	376	closed	euccaschen	III Spinner Change Freeze is In Effect III	Production Change Freeze for Knit Con has taken effect on Wednesday, May 5, 2021 at 12 noon PT. We expect the change freeze to end on Monday, May 10, 2021 at 10am PT. During the change freeze Spinner does NOT allow landing workflow changes in spinner-workflows repo without an approved Change Management ticket, and Pinbaard deploy stops. For more details please consult #workflow. Thank you!					2021-05-05 21:22:12+00:00	2021-05-10 17:05:41+00:00
im	- a		375	closed	euccaschen	Incident Notification is enabled in Spinner	Ongoing incidents that impact Spinner workflows will be shown here! You can close this banner using the x button on the right side.				No	2021-04-30 05:22:55+00:00	2021-05-05 19:06:05+00:00
Sta	- a		311	closed	euccaschen	Spinner web down due to endpoint-register issue in canary	spinner.pinadmin.com is not accessible, with "upstream request timeout" error.				No	2021-04-26 06:54:40+00:00	2021-05-04 01:59:38+00:00
	- a	Q IP 309 Issue euccaschen [Drill] Service discovery for MySQL problem blocked spinner spinner new tasks cannot be scheduled.				problem blocked spinner	spinner new tasks cannot be scheduled.		IM- 8138	1	No	2021-04-22 05:32:38+00:00	2021-05-05 18:28:26+00:00
Allo									_				
L					Allow	failed tasks to retry or not, 0 means o	lisallow, 1 means allow						
Sa	ve 🖹 (÷											

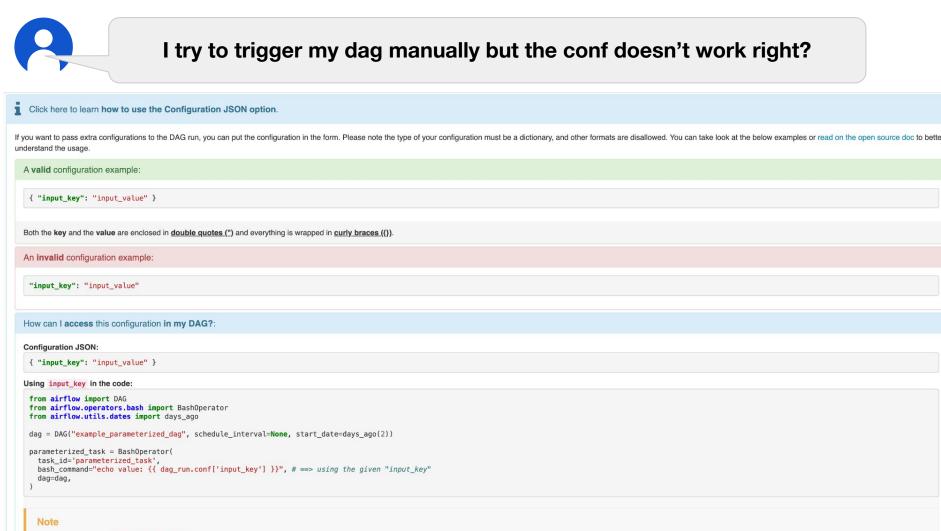
Incident Management Smart Retries



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



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

Wrap Up

Identifying What Users Want

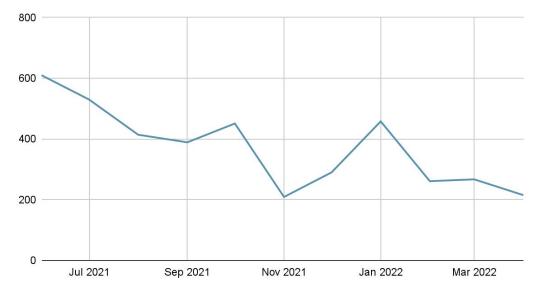
Focus Areas

Request Forms Select the top 3 areas you'd like enhanced 97 responses 40 (50 5%) Workflow testing Backfill -21 (21.6%) Building new workflows Spinner Edit Open -13(13.4%)Spinner API -15 (15.5%) By spinner -24 (24.7%) System performance (e.g.... Role/Access Control -5 (5.2%) Summarv Reviews Analytics Requests Uptime Debugging/Troubleshooting Resource management (e... -21 (21.6%) Build/deployment -19 (19.6%) Order By: Trending▼ Filter By: Open▼ + Create a New Request #workflow support channel -8 (8.2%) Better ResourceSpec sup... -1 (1%) Sending alerts when a dagrun or workflow can't finish within its SLO 12 Spinner UI constructs -1 (1%) Glitch when there are new... -1 (1%) Support for retrying a job with new parameters . Blocking Review is kind of ... -1 (1%) support presto queries in... -1 (1%) Paste ability for a comma separated task list into the backfill form 0 0 Documentation more easil... -1 (1%) MLOps, native integration... -1 (1%) 7 Add an Option to Turn off Alerts for backfill runs . PII spinner -1 (1%) Ensure compatible artifacts are used in jobs that depend on multiple deployment pipeli... . 0 may be using terraform to... -1 (1%) 20 Workflow team will create JIRA tickets for the requests have more upvotes! (>=5) 5 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 3 **User Surveys** Differentiate the PROD Spinner UI and devapp Spinner UI . 2 support multi-partition sensor 2 Add a `[DAG Final]` node in Graph View 2 . add Callback hook for native spinner when job is marked failed/success . 1 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
 - o ...
 - 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
 - <u>Usability Improvements for Debugging</u>
- Airflow Summit 2020:
 - <u>Airflow as the next gen of workflow system at Pinterest</u>
- 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