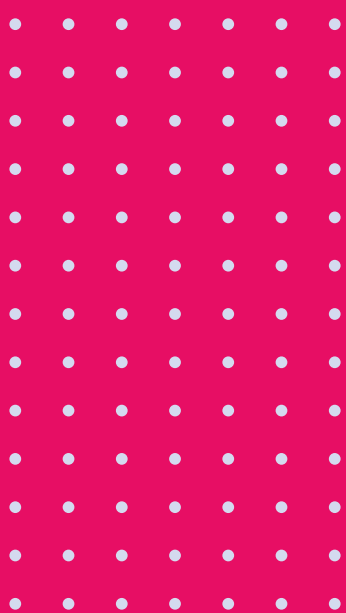




Data pipelines in Clinical Trials: Leveraging Apache Airflow for Automated Task Management (SM04)

Kostiantyn Zvieriev

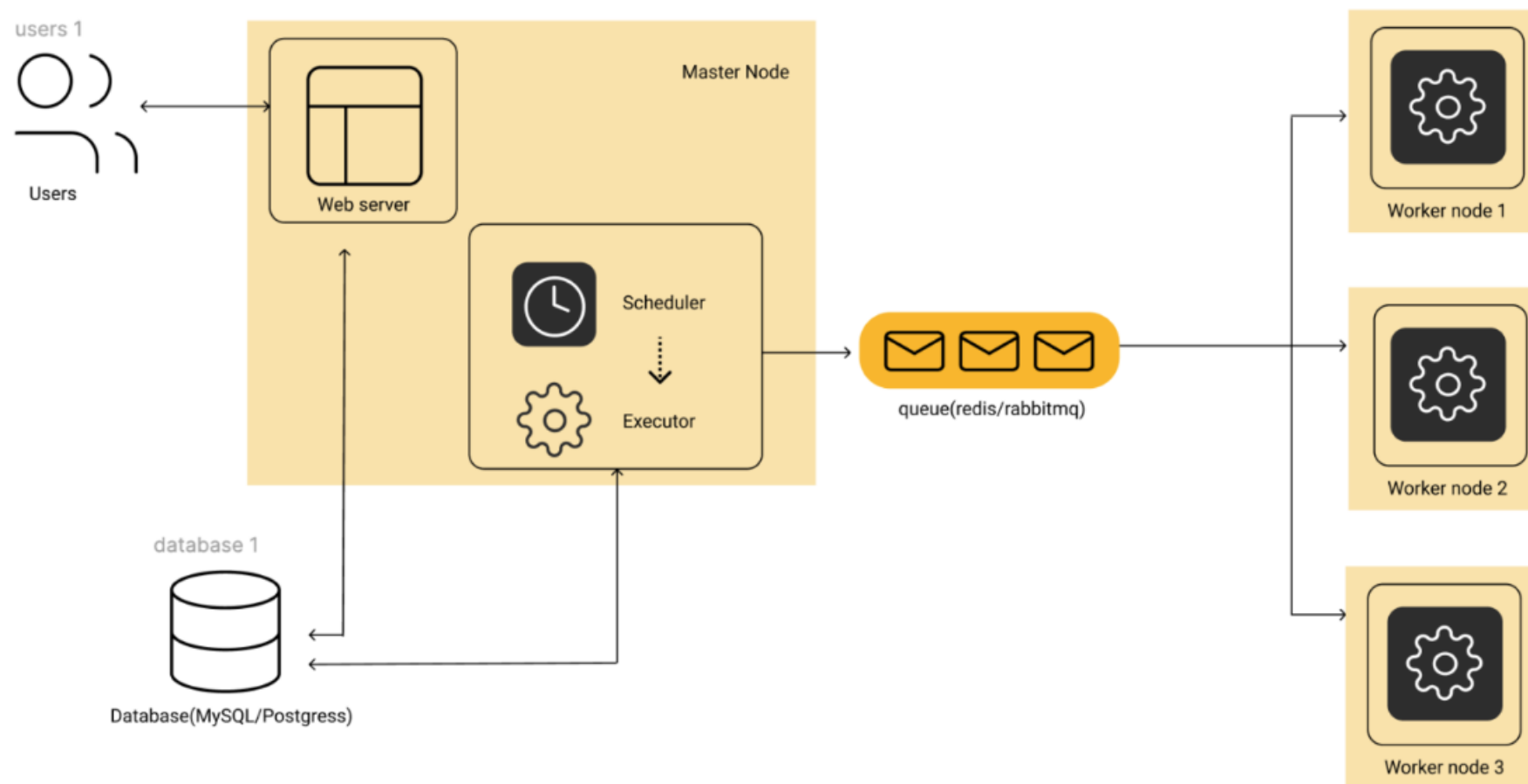
11-NOV-2024

What is Apache Airflow?



Apache Airflow

Airflow architecture



Web UI


Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
13:52 UTC
AA

DAGs

All 58
Active 5
Paused 53
Running 0
Failed 3

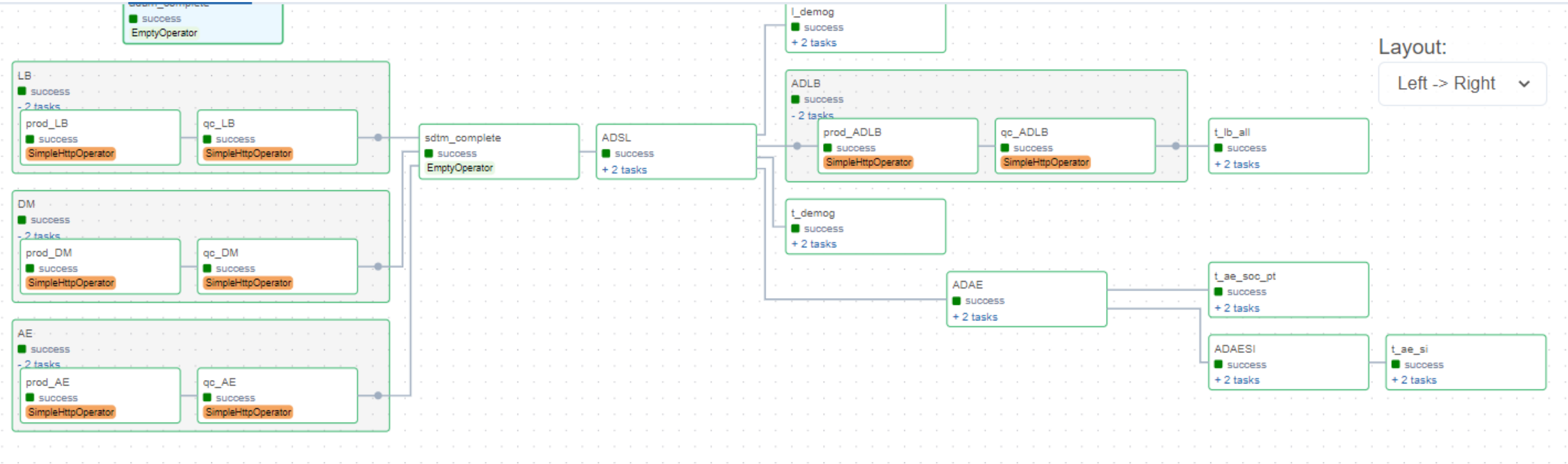
☒ Auto-refresh
↻

i	DAG ↕	Owner ↕	Runs i	Schedule	Last Run ↕ i	Next Run ↕ i	Recent Tasks i	Actions	Links
☒	my1dag	airflow	○ 1 ○ ○	None i	2024-03-18, 13:53:35 i		○ ○ ○ ○ ○ ○ 2 ○ ○ ○ ○ ○ ○ ○ ○	▶ 🗑️	...
☒	my2dag	airflow	○ ○ ○ 5	None i	2024-03-23, 23:27:51 i		○ ○ ○ ○ ○ ○ ○ ○ 1 ○ ○ ○ ○ ○ ○ ○	▶ 🗑️	...
☒	my3dag	airflow	○ 1 ○ 4	None i	2024-03-21, 22:14:57 i		○ ○ ○ ○ ○ ○ 1 ○ ○ ○ ○ ○ ○ ○ ○ ○	▶ 🗑️	...
☒	my4dag	airflow	○ ○ ○ 3	None i	2024-03-24, 17:26:02 i		○ ○ ○ ○ ○ ○ ○ ○ 1 ○ ○ ○ ○ ○ ○ ○	▶ 🗑️	...
☒	my5dag	airflow	○ 10 ○ 2	None i	2024-05-28, 12:07:12 i		○ ○ ○ ○ ○ ○ 9 ○ 1 ○ ○ 16 ○ ○ ○ ○	▶ 🗑️	...

«
<
1
>
»

Showing 1-5 of 5 DAGs

What is workflow?



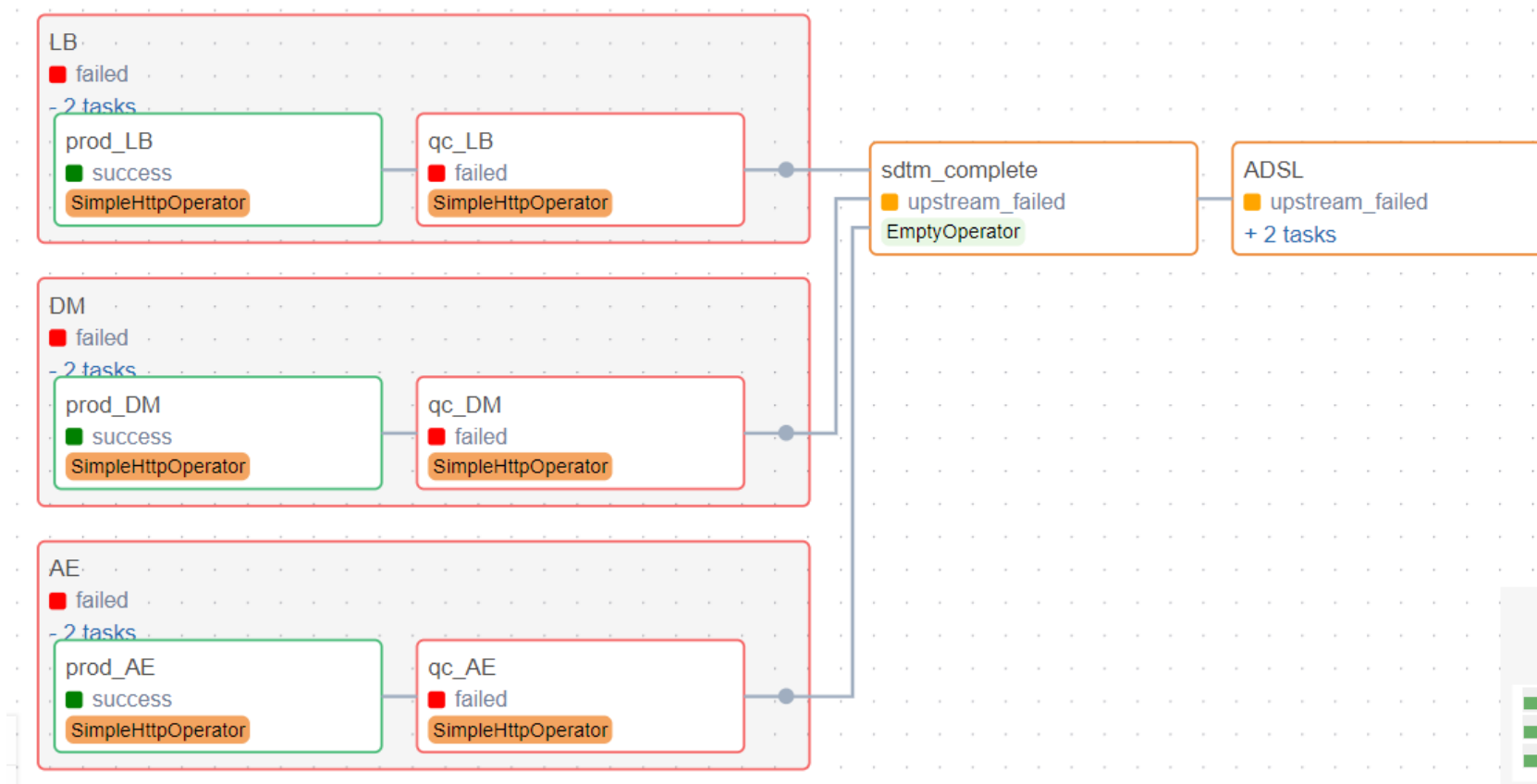
Creating tasks

```
with DAG(
    dag_id="my1dag",
    schedule_interval=None
) as dag1: #declaring instance of DAG
    #creating task t1
    t1 = BashOperator(
        task_id="print_date",
        bash_command="date",
    )
    #creating task t2
    t2 = BashOperator(
        task_id='second_task',
        bash_command='echo 2'
    )

    #setting up dependency between them
    t1 >>t2
```

Data pipelines in Clinical Trials: Leveraging Apache Airflow for Automated Task Management (SM04)

What if the task fails?



We can re-run failed tasks

25

Ma

Clear and Retry

Task: LB.qc_LB

Run: manual__2024-05-08T16:46:37.557280+00:00

Include:

Past

Future

Upstream

Downstream

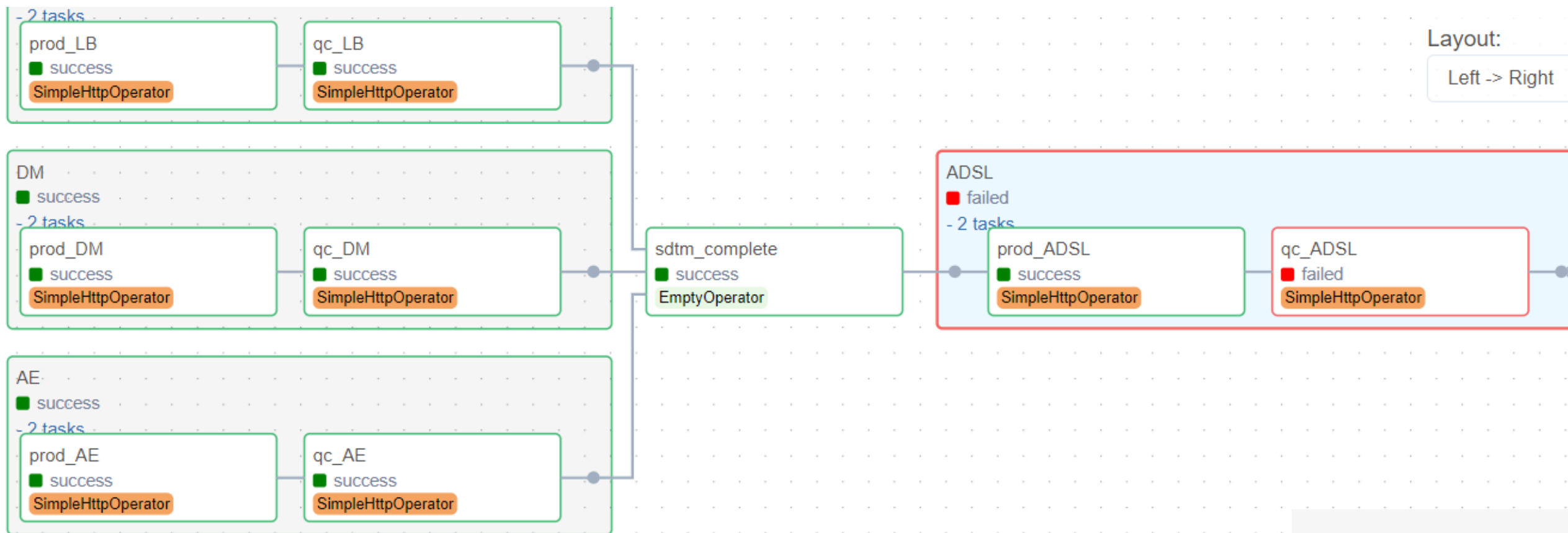
Recursive

Only Failed

Affected Tasks: 20

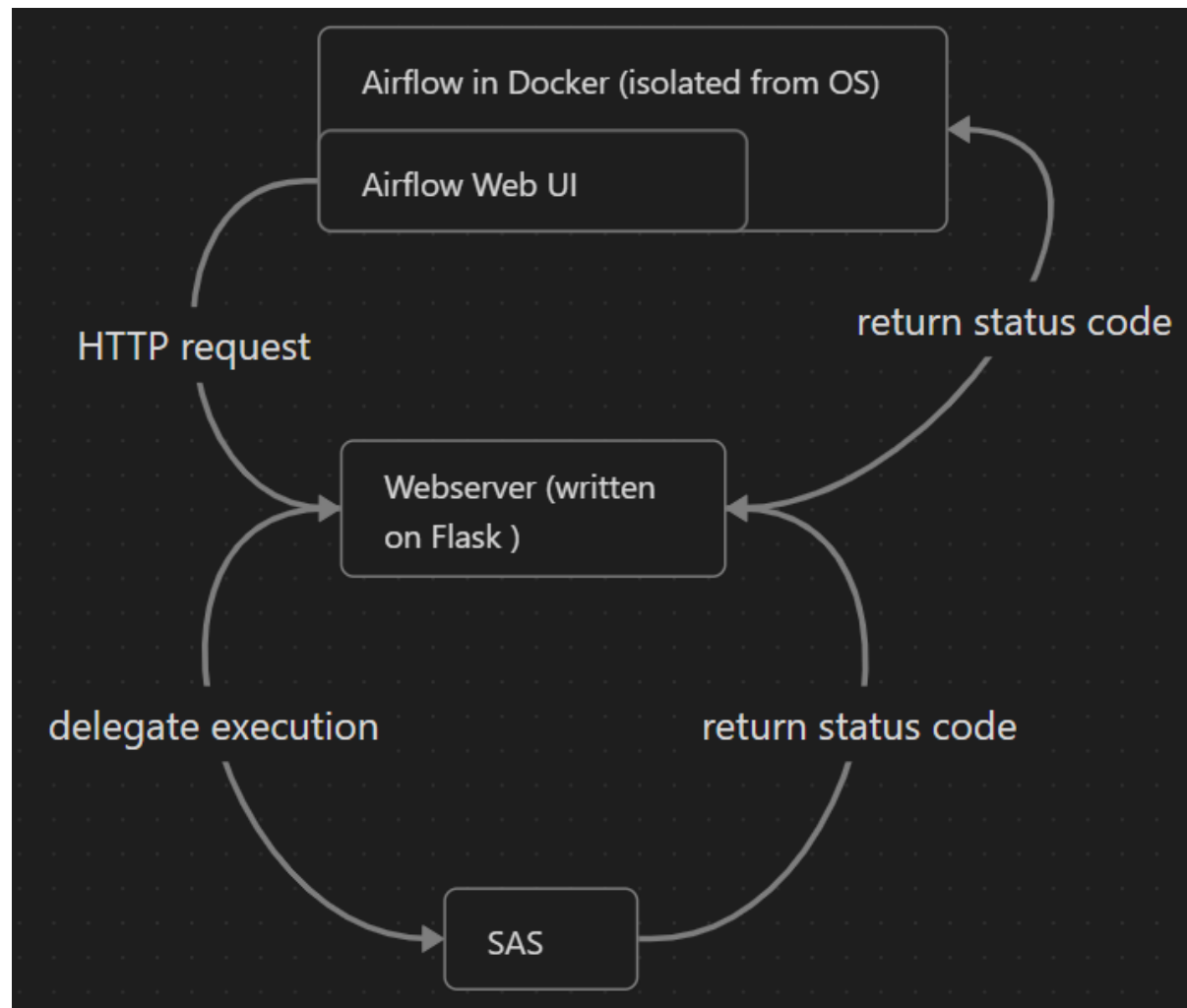
Cancel

Clear

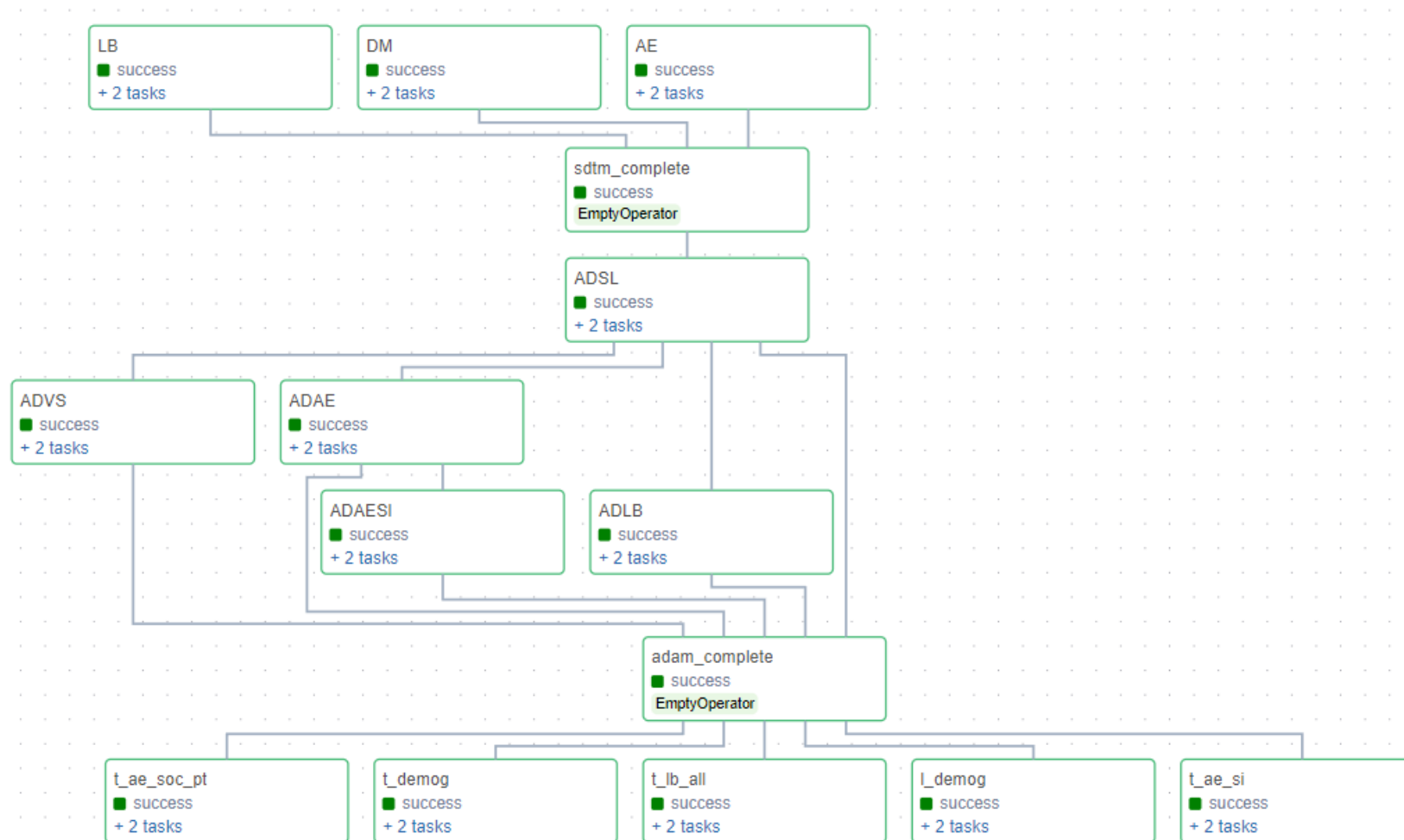


Our architecture

- Airflow runs in Docker;
- Airflow sends HTTP request to local webserver;
- Webserver is local program and can call host machine's programs (SAS, Powershell, etc.)



Clinical Trials Workflow



Let us know with Airflow

- `AIRFLOW__SMTP__SMTP_HOST: smtp.gmail.com`
- `AIRFLOW__SMTP__SMTP_STARTTLS : True`
- `AIRFLOW__SMTP__SMTP_SSL : False`
- `AIRFLOW__SMTP__SMTP_USER : #sender email address`
- `AIRFLOW__SMTP__SMTP_PASSWORD: #password (do not save password in config in production env)`
- `AIRFLOW__SMTP__SMTP_PORT: 587`
- `AIRFLOW__SMTP__SMTP_MAIL_FROM: #sender email address`

DAGs default arguments:

```
default_args = {  
    'email': ['receiver1@gmail.com', 'receiver2@yourdomain.com'],  
    'email_on_failure': True  
}
```

You shall not pass!



21:32 UTC

→ Log In

Sign In

Enter your login and password below:

Username:

 airflow

Password:



Sign In

Airflow roles

Role name	Intended users/usage	Default permissions
Admin	Only necessary when managing security permissions	All permissions
Public	Unauthenticated users	No permissions
Viewer	Read-only view of Airflow	Read access to DAGs
User	Useful if you want strict separation in your team between developers who can and cannot edit secrets (connections, variables, etc.). This role only grants permissions to create DAGs, not secrets.	Same as viewer but with edit permissions (clear, trigger, pause, etc.) on DAGs
Op	All permissions required for developing Airflow DAGs	Same as user but with additional permissions to view and edit connections, pools, variables, XComs, and configuration

User-defined roles

	First Name	Last Name	User Name	Email	Is Active?	Role
  	Peter	Parker	StudyBProgrammer	petersemail@email.com	True	[BaseUser, StudyBRole]
  	Bruce	Wayne	StudyAProgrammer	brucesemail@email.com	True	[BaseUser, StudyARole]
  	Airflow	Admin	airflow	airflowadmin@example.com	True	[Admin]

Peter Parker view


Airflow
Cluster Activity
Datasets
Browse
Admin
Docs
16:56 UTC
PP

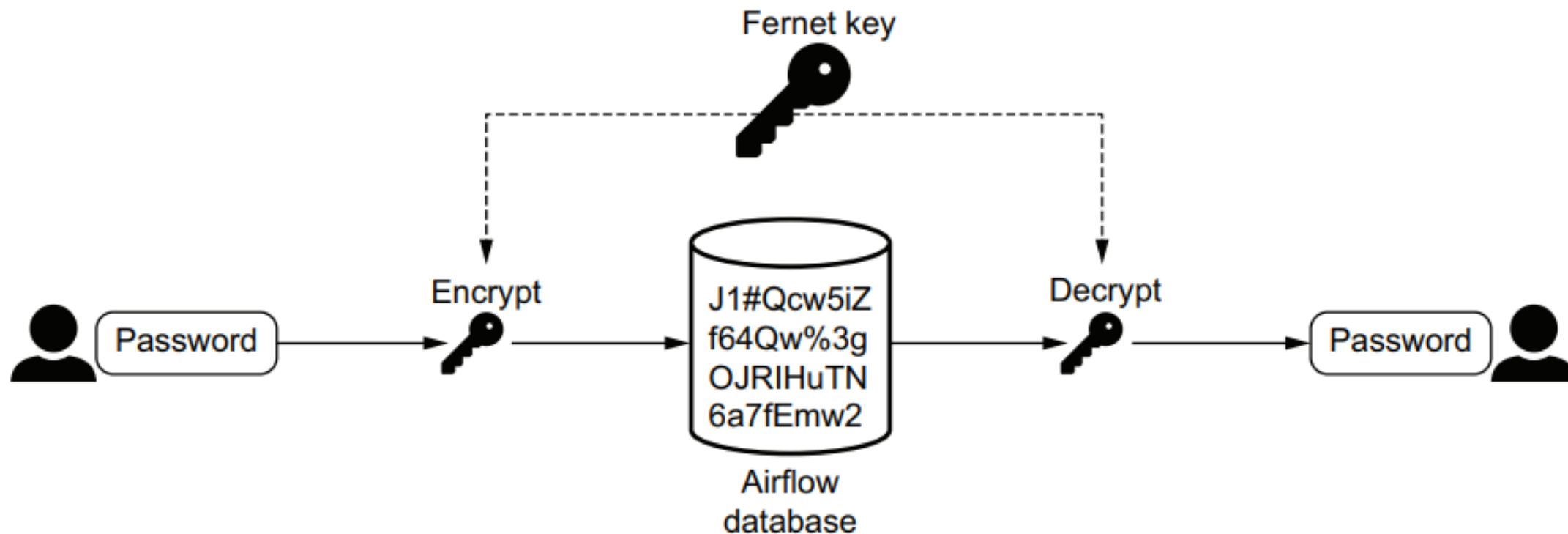
DAGs

All 1
Active 0
Paused 1
Running 0
Failed 0

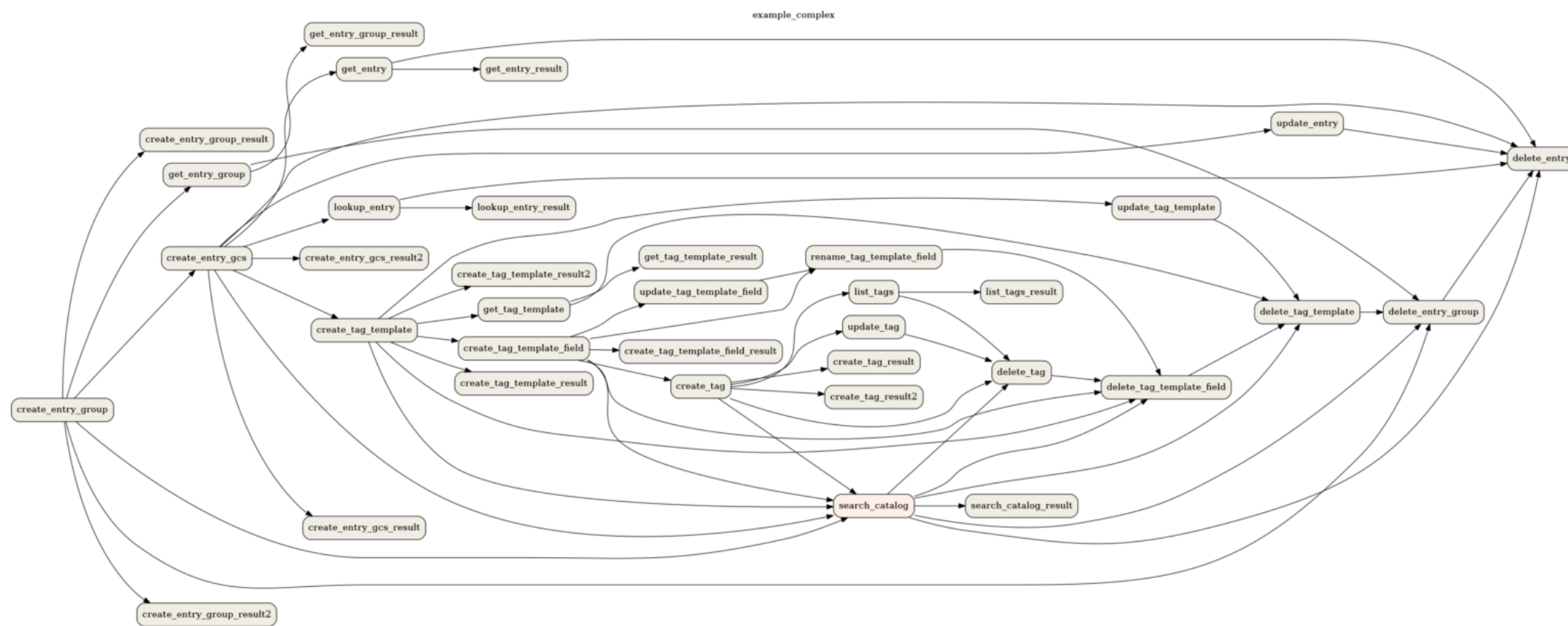
Auto-refresh
↺

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
☐ study_B_dag airflow		☐ ☐ ☐ ☐	None			☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐	▶ 🗑️	⋮

Metadata DB encryption



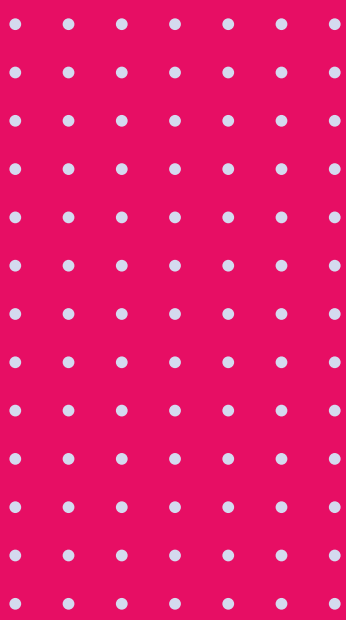
Airflow is more than that





Amazon SageMaker





Questions?