Data-Pipelines with airflow

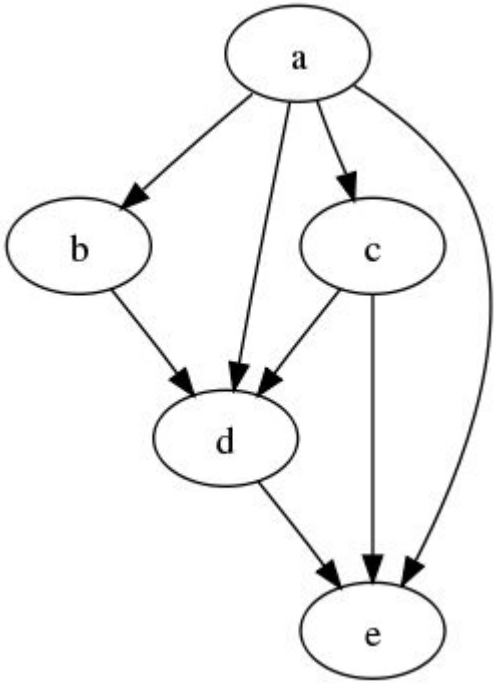
anup Sethuram

Agenda

- ➔ DE
- ➔ Airflow
- ➔ Workflow
- ➔ Tool Views
- ➔ Value
- ➔ Q?

Airflow

"Platform to programmatically Author, Schedule and Monitor workflows"

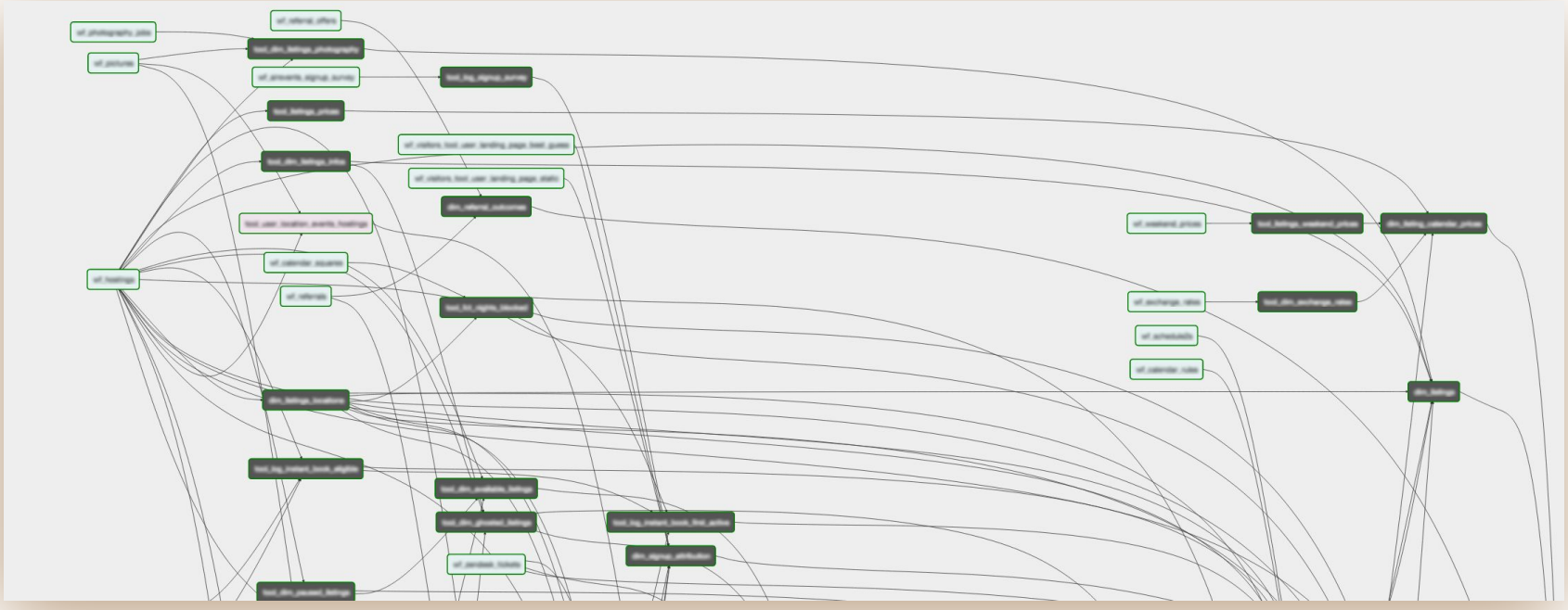


Philosophy: Configuration as code

Concepts: DAGs, Tasks, Operators and Runs.

Workflow

Analytics Pipeline

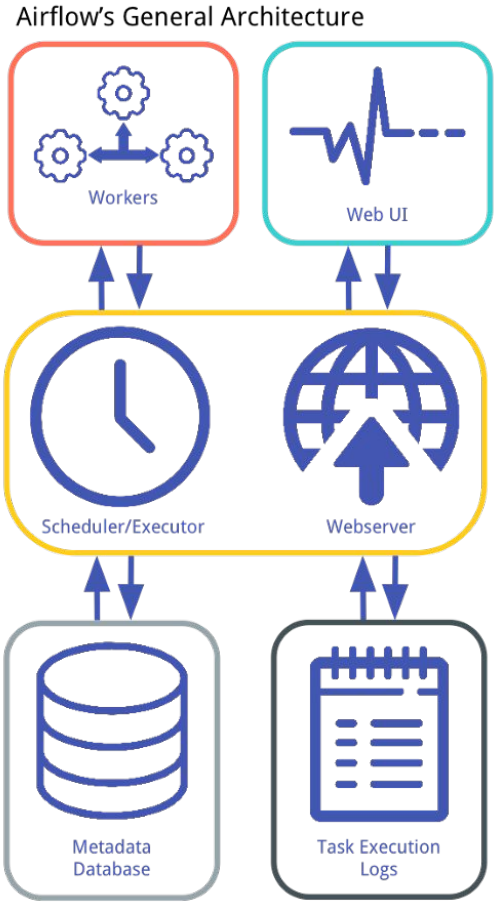


Workflow Code

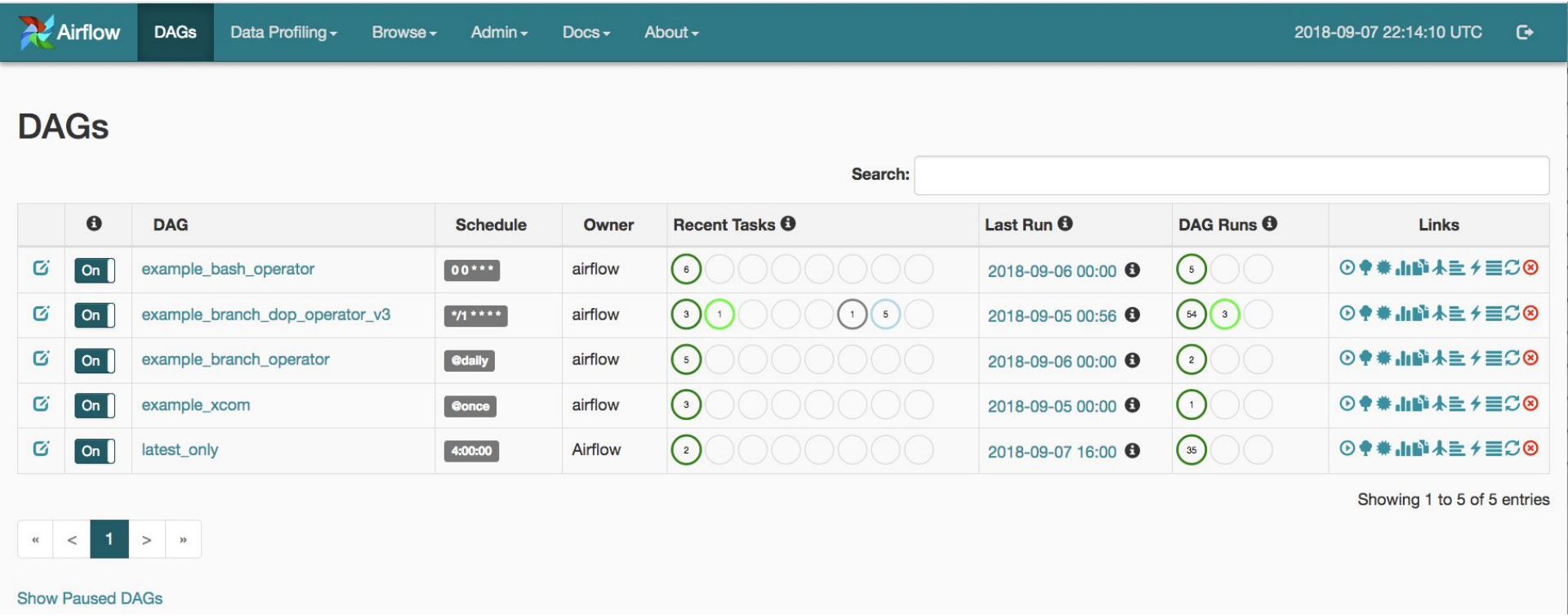
```python
from pprint import pprint

from airflow import DAG
from airflow.operators.bash_operator import BashOperator
from airflow.operators.python_operator import PythonOperator
from datetime import datetime, timedelta
from workflow.domain.alerts import slack


default_args = {
    'owner': 'airflow',
    'depends_on_past': False,
    'start_date': datetime(2019, 7, 1, 11, 0),
    'email': ['eanups@gmail.com'],
    'email_on_failure': False,
    'email_on_retry': False,
    'retries': 1,
    'retry_delay': timedelta(minutes=1),
    'on_failure_callback': slack.web_hook,

}

# DAG instantiation
dag = DAG(
    'sample_sc08', default_args=default_args, schedule_interval=timedelta(days=1))


# HELPER functions
```

```python
# SAMPLE TASKS
t1 = BashOperator(
    task_id='ingest_data',
    bash_command='date',
    dag=dag)

t2 = BashOperator(
    task_id='transform_data',
    bash_command='sleep 5',
    retries=3,
    dag=dag)

t3 = BashOperator(
    task_id='analyse_process',
    bash_command=templated_command,
    params={'my_param': 'Internal analytics'},
    dag=dag)

t4 = PythonOperator(
    task_id='report_data',
    python_callable=report,
    op_kwargs={'message': ' Report!'},
    dag=dag)


# CREATE Dependency graph
t1 >> t2 >> t3 >> t4
```

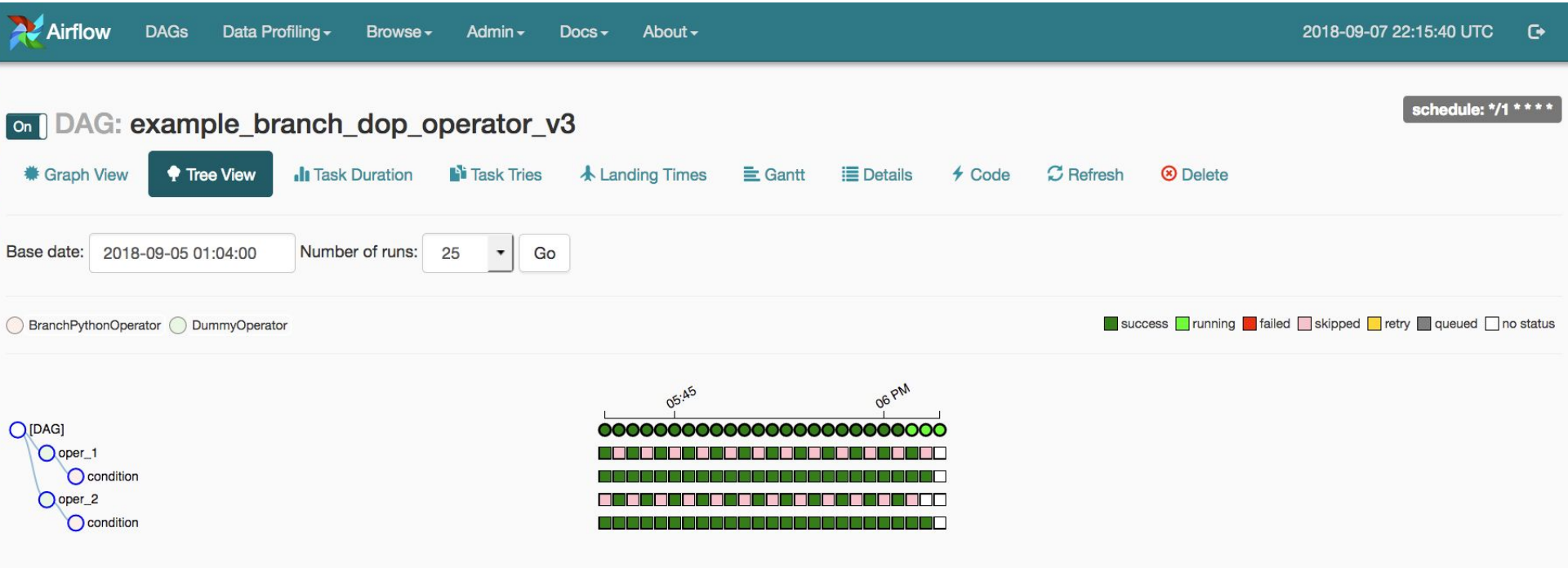
Composition



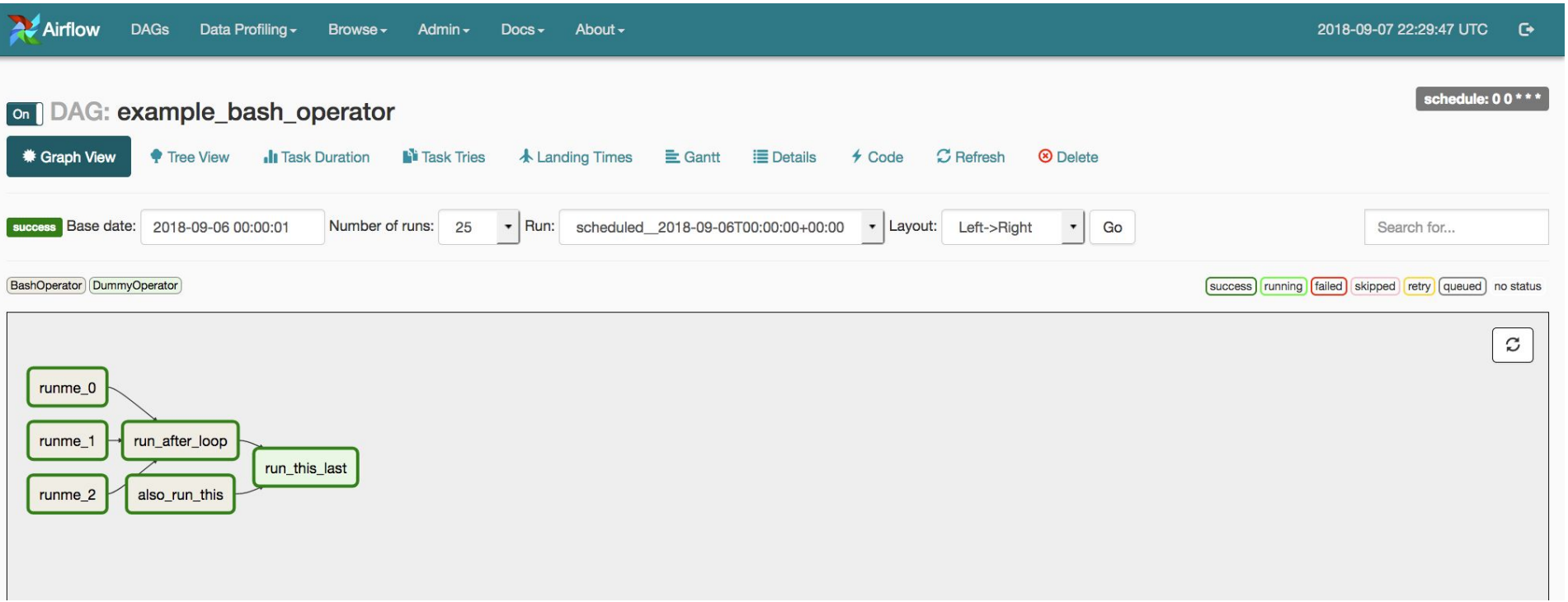
DAG view


Airflow
DAGs
Data Profiling
Browse
Admin
Docs
About
2018-09-07 22:14:10 UTC
DAGs
Search:
DAG
Schedule
Owner
Recent Tasks
Last Run
DAG Runs
Links
On
example_bash_operator
0 0 * * *
airflow
2018-09-06 00:00
On
example_branch_dop_operator_v3
*/1 * * * *
airflow
2018-09-05 00:56
On
example_branch_operator
@daily
airflow
2018-09-06 00:00
On
example_xcom
@once
airflow
2018-09-05 00:00
On
latest_only
4:00:00
Airflow
2018-09-07 16:00
Showing 1 to 5 of 5 entries
«
<
1
>
»
Show Paused DAGs

Tree view


Airflow
DAGs
Data Profiling
Browse
Admin
Docs
About
2018-09-07 22:15:40 UTC
On
DAG: example_branch_dop_operator_v3
schedule: */1 * * * *
Graph View
Tree View
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Refresh
Delete
Base date:
2018-09-05 01:04:00
Number of runs:
25
Go
BranchPythonOperator
DummyOperator
success
running
failed
skipped
retry
queued
no status
05:45
06 PM
[DAG]
oper_1
condition
oper_2
condition

Graph view


Airflow
DAGs
Data Profiling
Browse
Admin
Docs
About
2018-09-07 22:29:47 UTC
On
DAG: example_bash_operator
schedule: 0 0 * * *
Graph View
Tree View
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Refresh
Delete
success
Base date:
2018-09-06 00:00:01
Number of runs:
25
Run:
scheduled__2018-09-06T00:00:00+00:00
Layout:
Left->Right
Go
Search for...
BashOperator
DummyOperator
success
running
failed
skipped
retry
queued
no status
runme_0
runme_1
runme_2
run_after_loop
also_run_this
run_this_last

Value

- ➔ Visibility and Increased Data Governance
  - Replacement of CRON jobs for Batch processes
  - Proactive fixes along with Notifications.
  - Better Data Munging, Transformation & Reporting
  - Easier Backfill of historic data.
  - Data Correction.

- ➔ Automation and Future Opportunities
  - Adopt airflow for ML and AI pipelines
  - Modularize and simplify Data Modelling.
  - Incorporate Quality checks and tests


Questions?

Thank you!