

Hola

soy Nando Quintana

senior software engineer en *Heepsy*

Una historia basada en hechos real

dos personas emprendedoras

ipor cron!

```
23 59 * * * backups.sh  
59 23 * * * /home/backups/backups.sh  
32 95 * * * /bin/bash /home/backups/backups.sh
```

Busca una solución

pipelines de datos
scheduling tareas
visibilidad
alarmas



Apache Airflow





DAGs

Search:

		DAG	Schedule	Owner	Recent Tasks	Last Run	DAG Runs	Links
	On	hola_mundo_0	1:00:00	ExtrePython	<div><div>3</div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	2019-03-29 10:52	<div><div>2</div><div></div><div></div></div>	
	On	hola_mundo_1	1:00:00	ExtrePython	<div><div>3</div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	2019-03-29 10:52	<div><div>2</div><div></div><div></div></div>	
	On	hola_mundo_2	1:00:00	ExtrePython	<div><div>3</div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	2019-03-29 10:52	<div><div>2</div><div></div><div></div></div>	
	On	hola_mundo_3	1:00:00	ExtrePython	<div><div>3</div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	2019-03-29 10:52	<div><div>2</div><div></div><div></div></div>	

Showing 1 to 4 of 4 entries

[Hide Paused DAGs](#)



Airflow

DAGs

Security

Browse

Admin

Docs

About

2019-03-29, 12:55:50 UTC

Admin User

On DAG: hola_mundo_2

schedule: 1:00:00

Graph View

Tree View

Task Duration

Task Tries

Landing Times

Gantt

Details

Code

Refresh

Delete

SUCCESS

Base date:



2019-03-29 10:52:53+00

Number of runs:

25

Run:

scheduled__2019-03-29T10:52:52.093705+00:00

Layout:

Left->Right

Go

Search for...

DummyOperator PythonOperator

success

running

failed

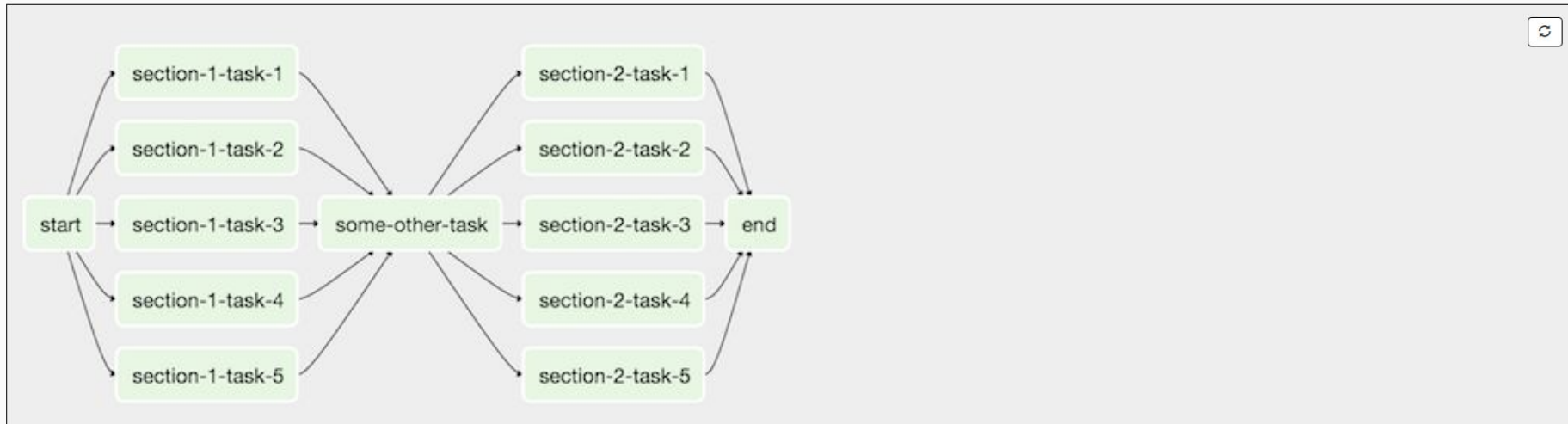
skipped

rescheduled

retry

queued

no status





Airflow

DAGs

Security

Browse

Admin

Docs

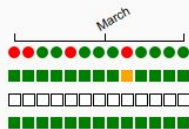
About

2019-03-29, 12:57:53 UTC

Admin User

On DAG: hola_mundo_2

schedule: 1:00:00

[Graph View](#) [Tree View](#) [Task Duration](#) [Task Tries](#) [Landing Times](#) [Gantt](#) [Details](#) [Code](#) [Refresh](#) [Delete](#)Base date: Number of runs: ☐ DummyOperator ☐ PythonOperator☒ success ☒ running ☒ failed ☒ skipped ☒ rescheduled ☒ retry ☒ queued ☒ no status



Airflow

DAGs

Security

Browse

Admin

Docs

About

2019-03-29, 13:03:00 UTC

Admin User

On DAG: hola_mundo_2

schedule: 1:00:0

[Graph View](#) [Tree View](#) [Task Duration](#) [Task Tries](#) [Landing Times](#) [Gantt](#) [Details](#) [Code](#) [Refresh](#) [Delete](#)Task Instance: hola_mundo [Task Instance Details](#) [Rendered Template](#) [Log](#) [XCom](#)

Log by attempts

1

```
*** Reading local file: /home/airflow/logs/hola_mundo_2/hola_mundo/2019-03-29T10:52:52.093705+00:00/1.log
[2019-03-29 11:54:06,664] {models.py:1359} INFO - Dependencies all met for <TaskInstance: hola_mundo_2.hola_mundo 2019-03-29T10:52:52.093705+00:00 [queued]>
[2019-03-29 11:54:06,692] {models.py:1359} INFO - Dependencies all met for <TaskInstance: hola_mundo_2.hola_mundo 2019-03-29T10:52:52.093705+00:00 [queued]>
[2019-03-29 11:54:06,692] {models.py:1571} INFO -
-----
Starting attempt 1 of 2
-----

[2019-03-29 11:54:06,733] {models.py:1593} INFO - Executing <Task(PythonOperator): hola_mundo> on 2019-03-29T10:52:52.093705+00:00
[2019-03-29 11:54:06,733] {base_task_runner.py:118} INFO - Running: ['bash', '-c', 'airflow run hola_mundo_2 hola_mundo 2019-03-29T10:52:52.093705+00:00 --job_id 901 --raw -sd DAGS_FOLDER/hola_mundo_2.py --cfg_path /tmp/tmp3v7g80/c']
[2019-03-29 11:54:08,028] {base_task_runner.py:101} INFO - Job 901: Subtask hola_mundo [2019-03-29 11:54:08,016] {settings.py:174} INFO - settings.configure_orm(): Using pool settings. pool_size=5, pool_recycle=1800, pid=6757
[2019-03-29 11:54:09,189] {base_task_runner.py:101} INFO - Job 901: Subtask hola_mundo [2019-03-29 11:54:09,179] {default_celery.py:90} WARNING - You have configured a result_backend of redis://:extrep@localhost:6379/0, it is highly recommended to use an alternative result_backend (i.e. a database).
[2019-03-29 11:54:09,189] {base_task_runner.py:101} INFO - Job 901: Subtask hola_mundo [2019-03-29 11:54:09,181] {__init__.py:51} INFO - Using executor CeleryExecutor
[2019-03-29 11:54:09,799] {base_task_runner.py:101} INFO - Job 901: Subtask hola_mundo [2019-03-29 11:54:09,777] {models.py:273} INFO - Filling up the DagBag from /home/airflow/dags/hola_mundo_2.py
[2019-03-29 11:54:09,855] {base_task_runner.py:101} INFO - Job 901: Subtask hola_mundo [2019-03-29 11:54:09,855] {cli.py:520} INFO - Running <TaskInstance: hola_mundo_2.hola_mundo 2019-03-29T10:52:52.093705+00:00 [running]> on host airflow
[2019-03-29 11:54:09,911] {python_operator.py:95} INFO - Exporting the following env vars:
```



Heepsy



Airflow

DAGs

Security

Browse

Admin

Docs

About

2019-03-29, 13:05:36 UTC

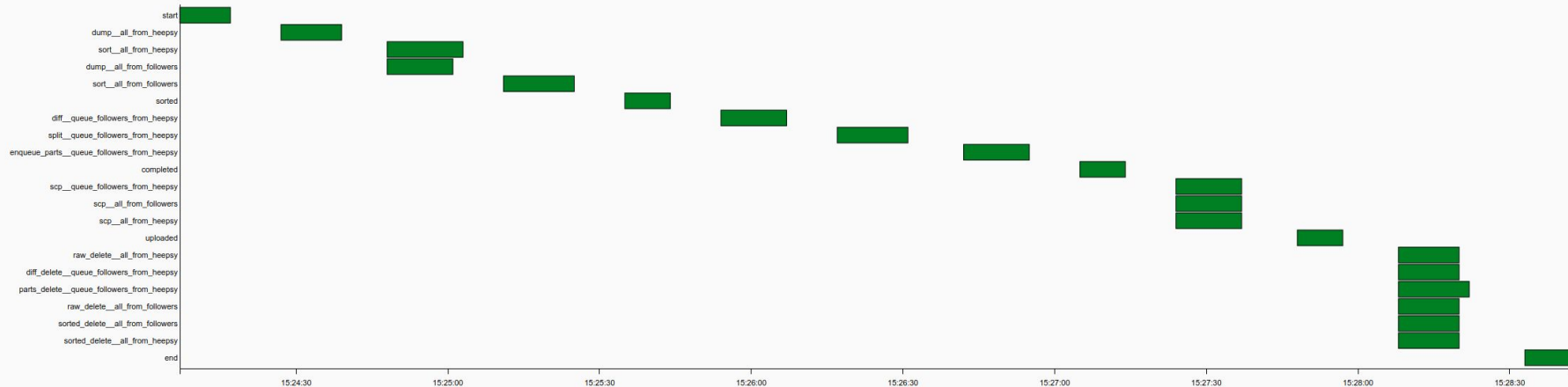
Admin User

On DAG: hola_mundo_2

schedule: 1:00:00

[Graph View](#) [Tree View](#) [Task Duration](#) [Task Tries](#) [Landing Times](#) [Gantt](#) [Details](#) [Code](#) [Refresh](#) [Delete](#)

Base date: 2019-03-29 10:52:53+00 Number of runs: 25 Run: scheduled__2019-03-29T10:52:52.093705+00:00 Go



DAGs

```
dag = DAG(  
    ...  
    start_date=datetime(2019, 1, 14),  
    end_date=datetime(2019, 1, 20)  
    schedule_interval='@daily')
```

TASKs

```
...
```

```
task1 = BashOperator(..., dag)
```

```
task2 = SSHOperator(..., dag)
```

```
for i in [1, 2]:
```

```
    task = BashOperator(..., dag)
```

```
...
```

```
start >> [task1, task2] >> task4 >> end
```

```
task3 >> task4
```

OPERATORS

```
command = """  
    DATASET="{{dag.dag_id}}-{{task.task_id}}-{{run_id}}"  
    comm -2 -3 {{params.file_a}} {{params.file_b}} > $DATASET.csv;  
    echo "$DATASET"  
    """
```

```
t1 = BashOperator(...,  
    params={'file_a': 'a.csv',  
            'file_b': 'b.csv'}  
    command,  
    dag)
```

HOOKs

```
query = "SELECT * from table"
hook = PostgresHook(postgres_conn_id="postgres_cachedb")
results = hook.get_records(query)

for result in results:
    print(result)
```

SENSORs

```
...
sensor1 = SqlSensor(
    task_id='sensor1',
    sql='SELECT COUNT(*) from table WHERE ...'
    self.conn_id="postgres_cachedb",
    da=dag)
```


```
sensor1 >> task3
```


XCOM


```
task1 = PythonOperator(  
    task_id='task1',  
    python_callable=lambda: return "hello",  
    dag=dag)
```


```
task2 = BashOperator(  
    task_id='task2',  
    bash_command="echo '{{task_instance.xcom_pull(task_ids='task1')}}'; echo 'goodbye'",  
    xcom_push=True,  
    dag=dag)
```



DAG Run


 DAGs

 Security ▾

 Browse ▾

 Admin ▾

 Docs ▾

 About ▾

2019-03-29, 13:10:26 UTC

Admin User ▾

List Dag Run

Search ▾

Actions ▾

⬅

Record Count: 2

<input type="checkbox"/>	State	Dag Id	Execution Date	Run Id	External Trigger
<input type="checkbox"/>	success	hola_mundo_0	03-29T10:52:51.108366+00:00	scheduled__2019-03-29T10:52:51.108366+00:00	False
<input type="checkbox"/>	success	hola_mundo_0	03-29T09:52:51.108366+00:00	scheduled__2019-03-29T09:52:51.108366+00:00	False


execution_date

start_date


end_date


schedule_interval


Task instance


Airflow


DAGs

Security

Browse

Admin

Docs

About

2019-03-29, 14:01:47 UTC

Admin User

List Task Instance

Search

Actions

Record Count: 6

	State	Dag Id	Task Id	Execution Date	Operator	Start Date	End Date	Duration	Job Id	Hostname	Unixname	Priority Weight	Queue	Queued Dttm	Try Number
<input type="checkbox"/>	success	hola_mundo_0	start	03-29T10:52:51.108366+00:00	DummyOperator	03-29T11:53:54.771163+00:00	03-29T11:53:59.098826+00:00	0:00:04.327663	889	airflow	airflow	3	default	03-29T11:52:55.054307+00:00	2
<input type="checkbox"/>	success	hola_mundo_0	start	03-29T09:52:51.108366+00:00	DummyOperator	03-29T11:53:55.004002+00:00	03-29T11:53:59.600783+00:00	0:00:04.596781	891	airflow	airflow	3	default	03-29T11:52:53.066659+00:00	2
<input type="checkbox"/>	success	hola_mundo_0	hola_mundo	03-29T10:52:51.108366+00:00	BashOperator	03-29T11:54:06.323260+00:00	03-29T11:54:09.421493+00:00	0:00:03.098233	898	airflow	airflow	2	default	03-29T11:54:01.232046+00:00	2
<input type="checkbox"/>	success	hola_mundo_0	hola_mundo	03-29T09:52:51.108366+00:00	BashOperator	03-29T11:54:06.775117+00:00	03-29T11:54:10.237414+00:00	0:00:03.462297	902	airflow	airflow	2	default	03-29T11:54:01.232073+00:00	2
<input type="checkbox"/>	success	hola_mundo_0	end	03-29T09:52:51.108366+00:00	DummyOperator	03-29T11:54:16.448955+00:00	03-29T11:54:18.438957+00:00	0:00:01.990002	906	airflow	airflow	1	default	03-29T11:54:11.279498+00:00	2

execution_date

upstream

downstream

trigger_rule (all_success, all_done, etc.)

Arquitectura



```
~# airflow webserver  
~# airflow scheduler  
~# airflow worker  
~# airflow worker  
~# airflow worker  
~# airflow flower
```

```
~# service postgresql start  
~# service redis-server start
```

¿ejemplos?

Otros operadores

SlackOperator

ShortCircuitOperator

TriggerDagRunOperator

<https://github.com/apache/airflow/tree/master/airflow/operators>

BaseOperator

Otros sensores

HttpSensor

SqlSensor

ExternalTaskSensor

TimeSensor

TimeDeltaSensor

<https://github.com/apache/airflow/tree/master/airflow/sensors>

BaseSensor