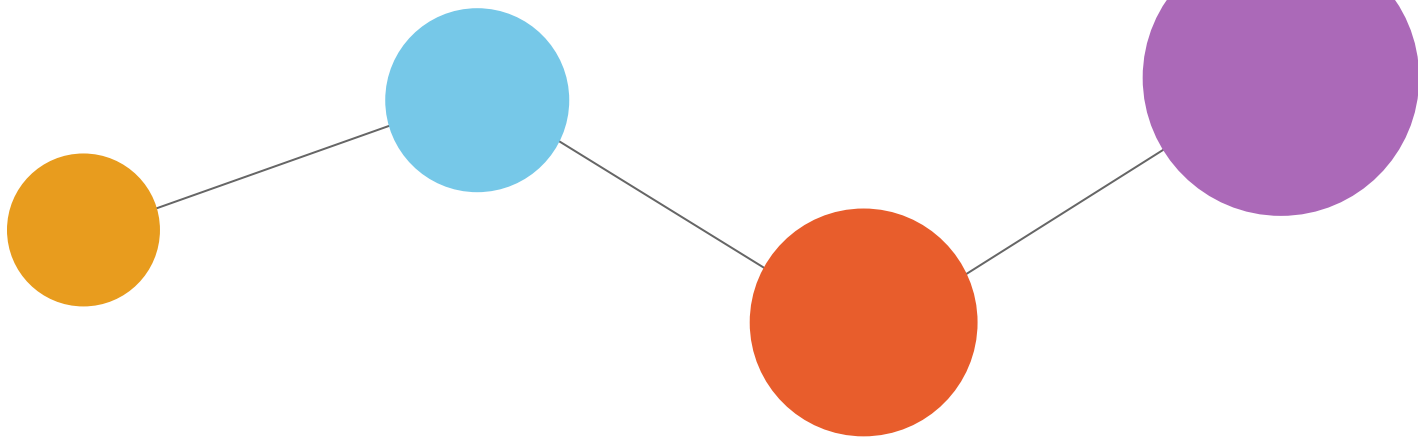


Orquestación de datos con Apache Airflow

Fiorella Piriz Sapio



Índice

Introducción

Apache Airflow

Caso Práctico

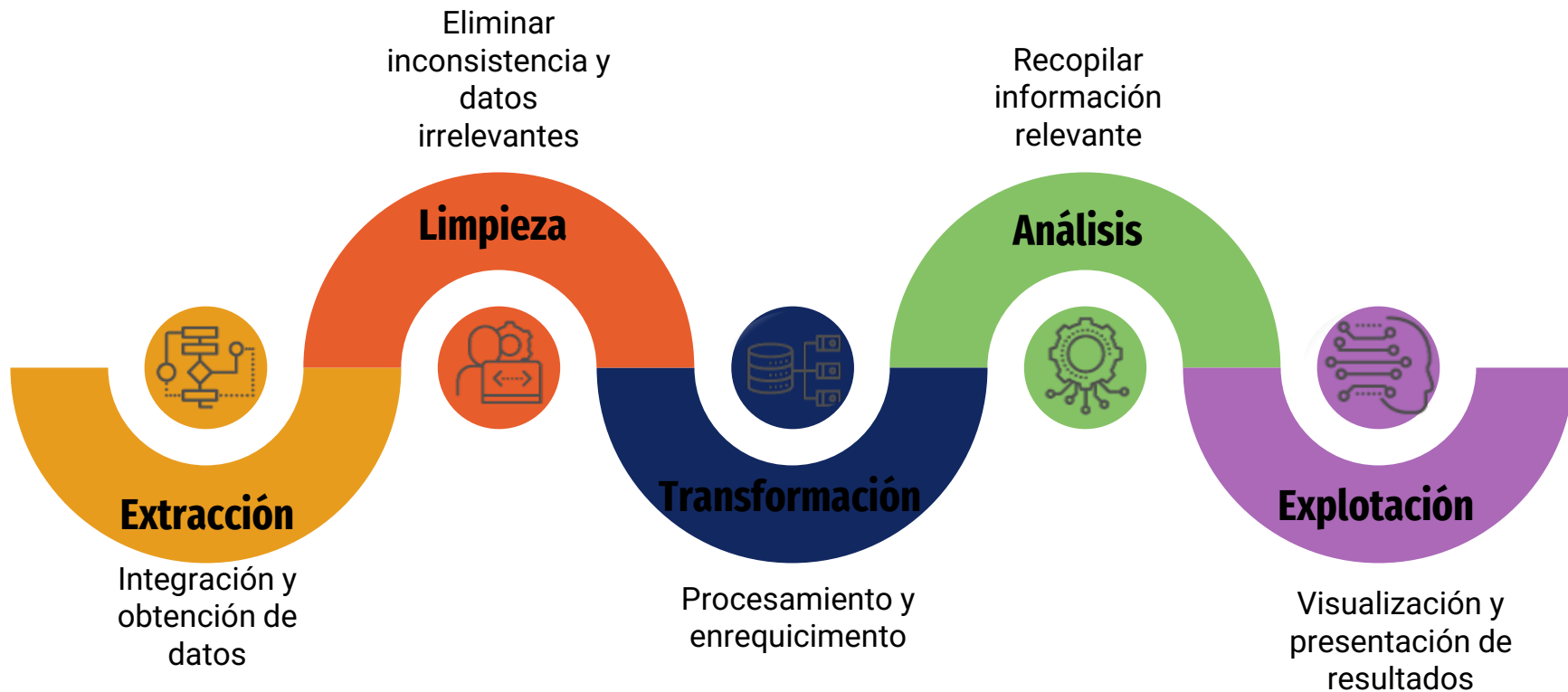
Siguientes Pasos

01



¿Qué es la orquestación de datos?

Orquestación de datos



02

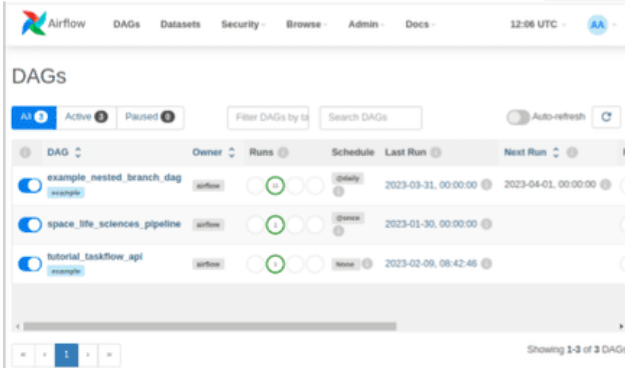
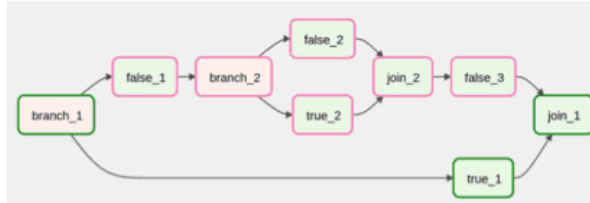


¿Qué es Apache Airflow?

Apache Airflow

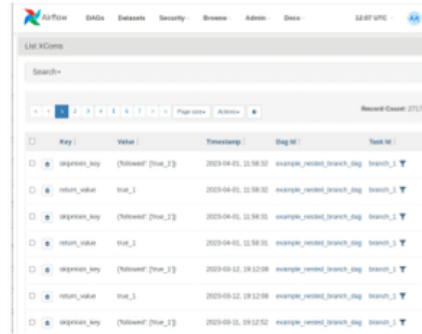


Apache
Airflow



A screenshot of the Apache Airflow web interface showing a list of DAGs. The interface includes a navigation bar with links for DAGs, Datasets, Security, Browse, Admin, and Docs. The main content area is titled 'DAGs' and shows a table of DAGs with columns for DAG name, Owner, Runs, Schedule, Last Run, and Next Run. The table lists three DAGs: 'example_nested_branch_dag', 'space_life_sciences_pipeline', and 'tutorial_taskflow_api'.

DAG	Owner	Runs	Schedule	Last Run	Next Run
example_nested_branch_dag	airflow	1	interval: 1d	2023-03-31, 00:00:00	2023-04-01, 00:00:00
space_life_sciences_pipeline	airflow	1	interval: 1d	2023-01-30, 00:00:00	
tutorial_taskflow_api	airflow	1	interval: 1d	2023-02-09, 08:42:46	



A screenshot of the Apache Airflow web interface showing a task log. The interface includes a navigation bar with links for DAGs, Datasets, Security, Browse, Admin, and Docs. The main content area is titled 'List XComs' and shows a table of XComs with columns for Key, Value, Timestamp, Dag ID, and Task ID. The table lists several XComs for the 'example_nested_branch_dag' DAG.

Key	Value	Timestamp	Dag ID	Task ID
skip_this_key	['true_2']	2023-04-01, 11:58:32	example_nested_branch_dag	branch_1
return_value	true_1	2023-04-01, 11:58:32	example_nested_branch_dag	branch_1
skip_this_key	['true_2']	2023-04-01, 11:58:32	example_nested_branch_dag	branch_1
return_value	true_1	2023-04-01, 11:58:32	example_nested_branch_dag	branch_1
skip_this_key	['true_2']	2023-04-01, 11:58:32	example_nested_branch_dag	branch_1
return_value	true_1	2023-04-01, 11:58:32	example_nested_branch_dag	branch_1
skip_this_key	['true_2']	2023-04-01, 11:58:32	example_nested_branch_dag	branch_1
return_value	true_1	2023-04-01, 11:58:32	example_nested_branch_dag	branch_1

Opensource

Fácil de usar

Basada en Python

Múltiples conectores

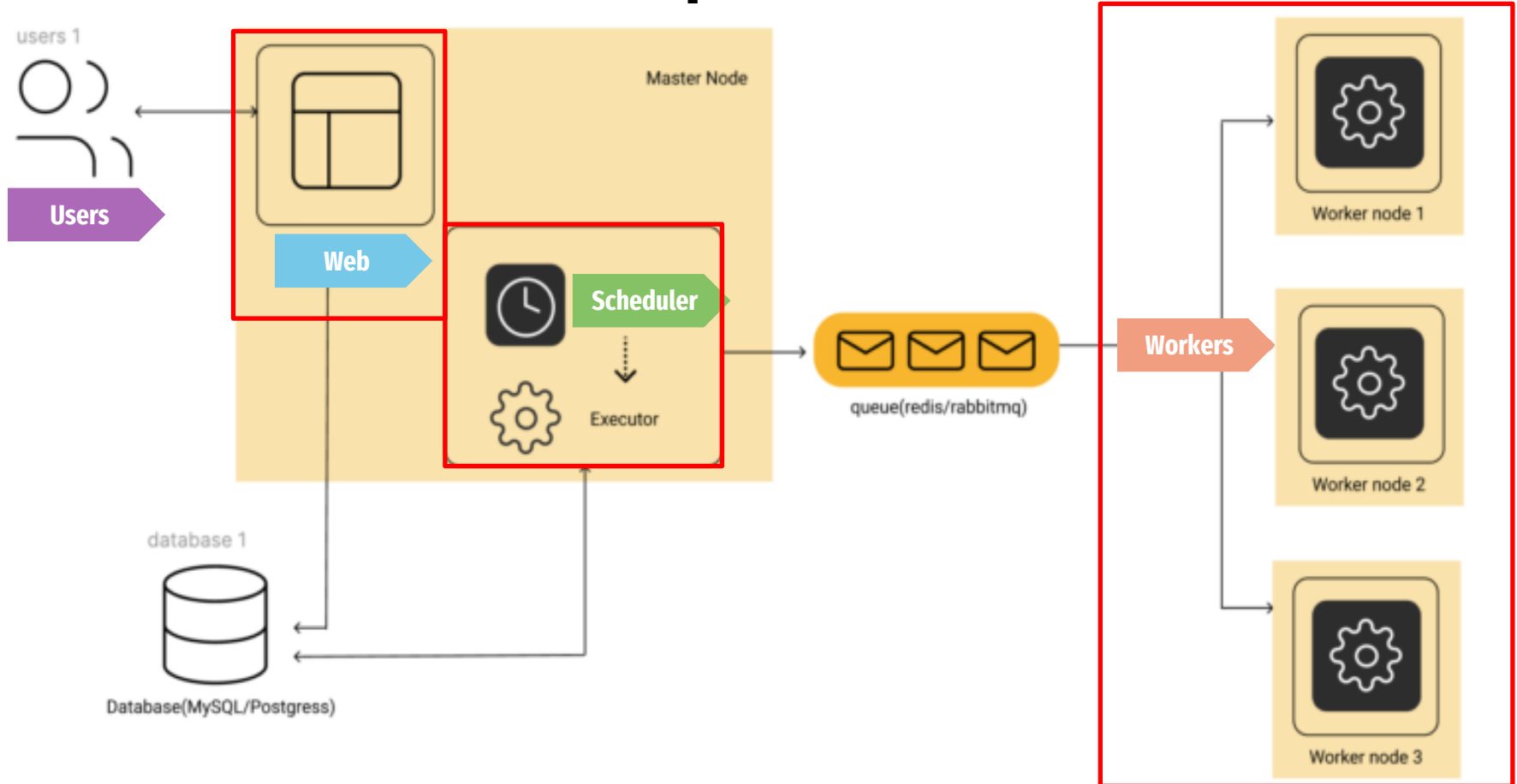
VS

Oozie

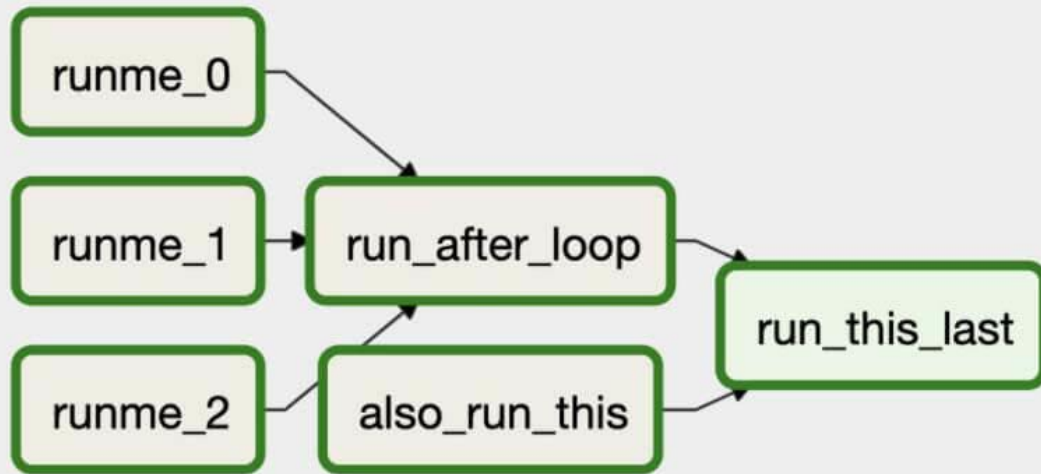
Azkabam

Luigi

Componentes



DAGS



03



Instalación y preparación del entorno

Instalación de Airflow en Linux



01

Instalar Apache Airflow

`pip install "apache-airflow[gcp]"`



02

Configuración

`nano /root/airflow/airflow.cfg`



03

Comprobar la versión

`airflow version`



04

Inicializar la base de datos

`airflow db init`



05

Crear usuario

`airflow users create`



06

Arrancar webserver

`airflow webserver`



07

Arrancar scheduler

`airflow scheduler`

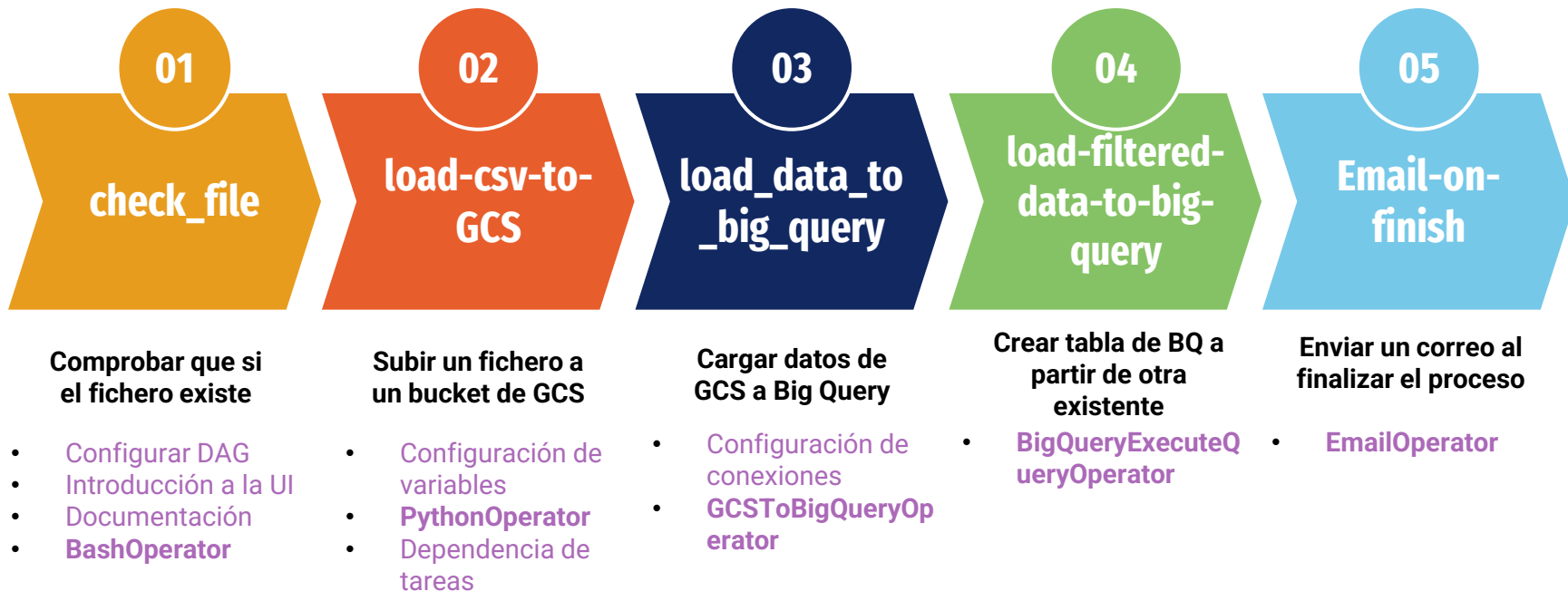
04



Caso práctico

Caso Práctico Encuestas de Barcelona

ETL

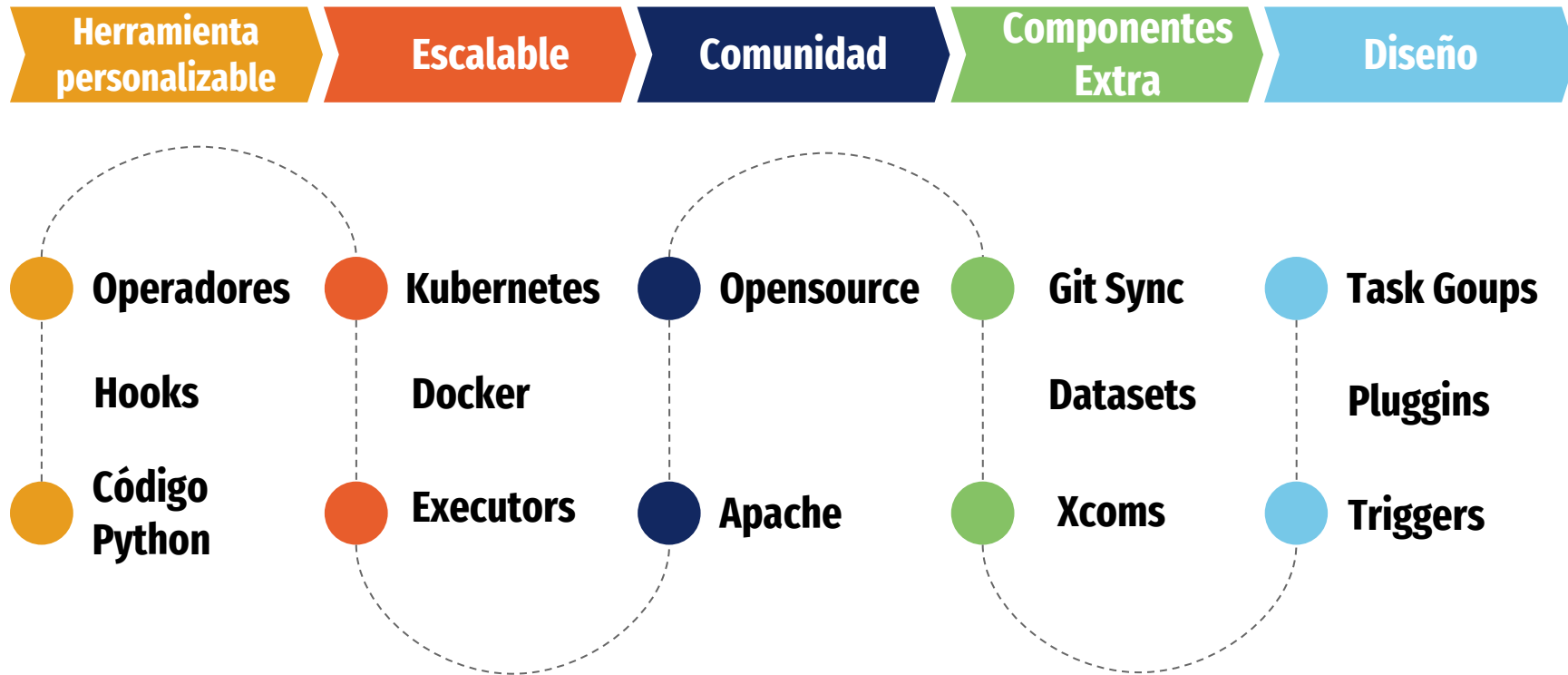


05



Siguientes pasos

¿Qué hemos aprendido?



Fin 

Muchas gracias

Fiorella Piriz Sapio

fpa@nextret.net

<https://www.linkedin.com/in/fiorella-piriz-sapio-74569a188/>