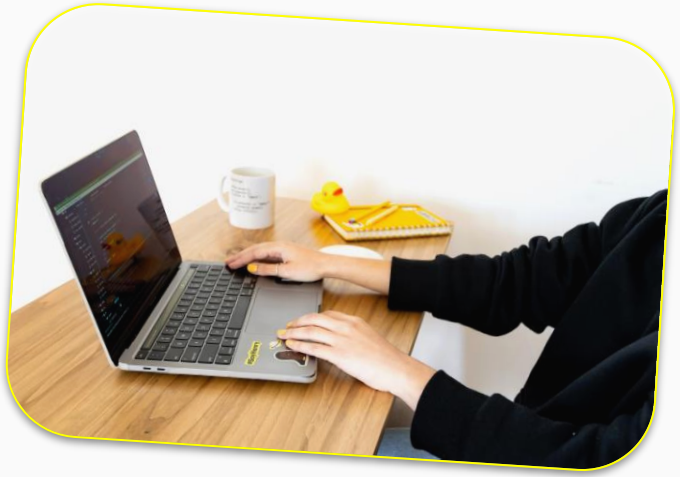
HENRY

Flujos de Trabajo

OBJETIVOS DE CLASE

- **Aplicar** los conceptos de Flujos de Trabajo y DAG
- **Entender** la Notación CRON



AGENDA

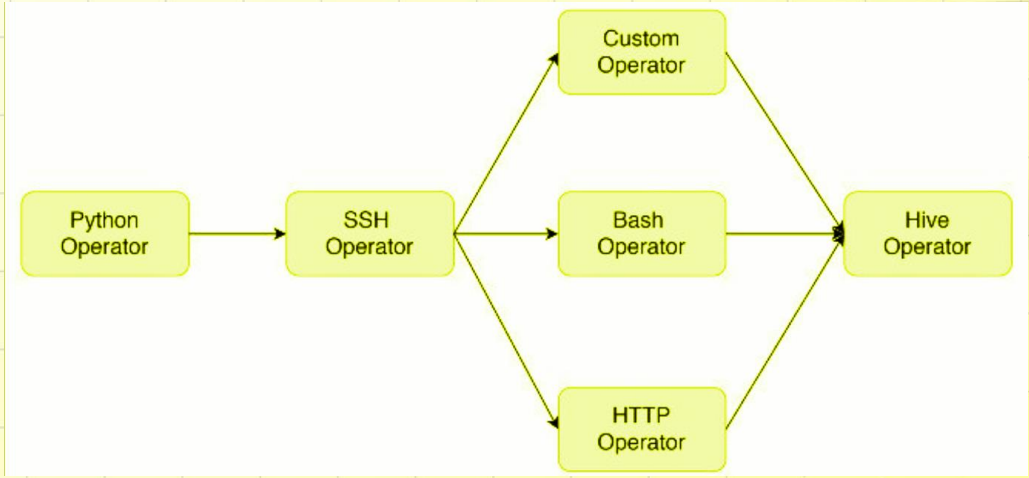
- DAG
- Notación CRON
- Oozie
- Nifi
- Airflow

Flujos de trabajo

Introducción

Flujos de trabajo

Un proyecto de **Big Data** implica realizar múltiples tareas en diferentes sistemas en un orden específico. Es por este motivo que existe la necesidad de contar con orquestadores de flujos de trabajo que permitan automatizar el **movimiento y la transformación de los datos.**

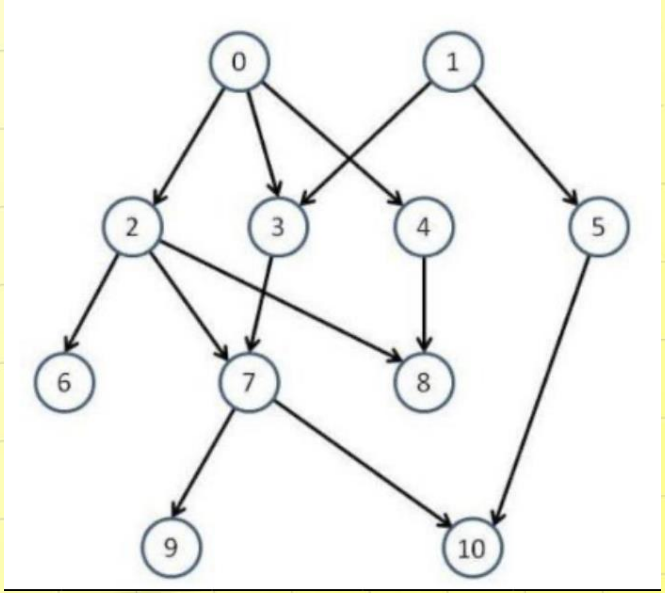


DAG

Directed Acyclic Graph

Directed Acyclic Graph Es una representación conceptual de una serie de actividades.

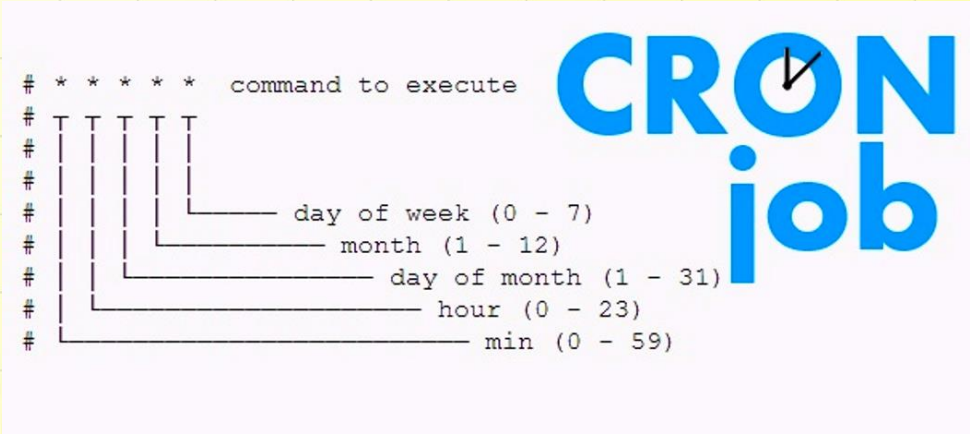
- Dirigido: cada relación entre nodos tiene un sentido único
- Acíclico: no hay ningún camino que nos permita volver al nodo inicial



Notación CRON

Notación CRON

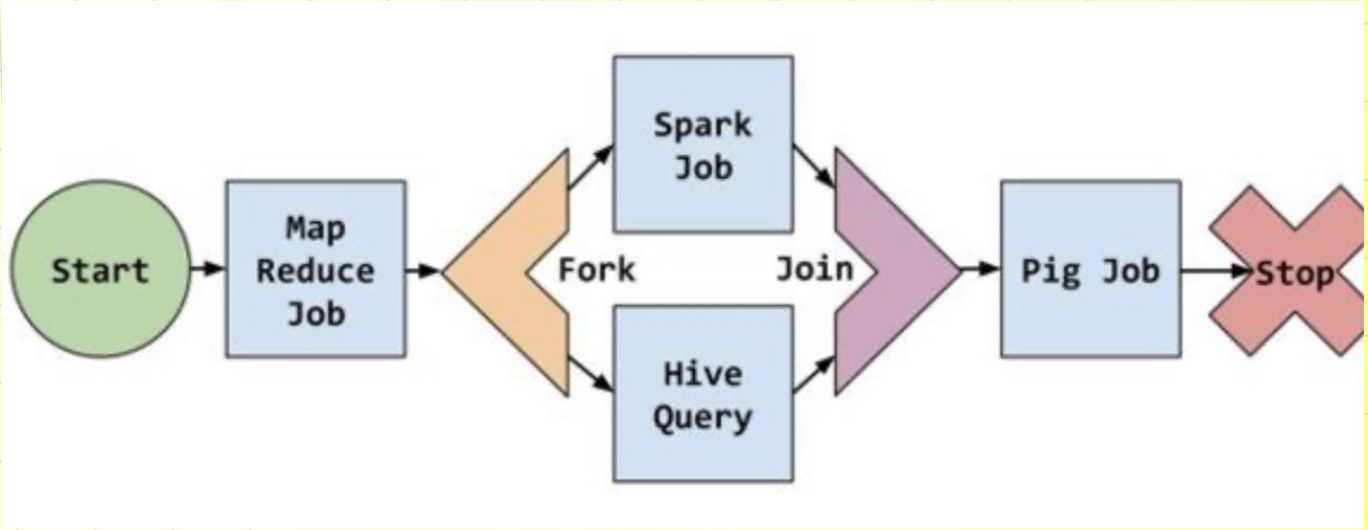
En el sistema operativo Unix, cron es un administrador regular de procesos en segundo plano (demonio) que ejecuta procesos o guiones a intervalos regulares (por ejemplo, cada minuto, día, semana o mes). Los procesos que deben ejecutarse y la hora en la que deben hacerlo se especifican en el fichero crontab. Como usuario podemos agregar comandos o scripts con tareas a cron para automatizar algunos procesos.



Oozie

Oozie

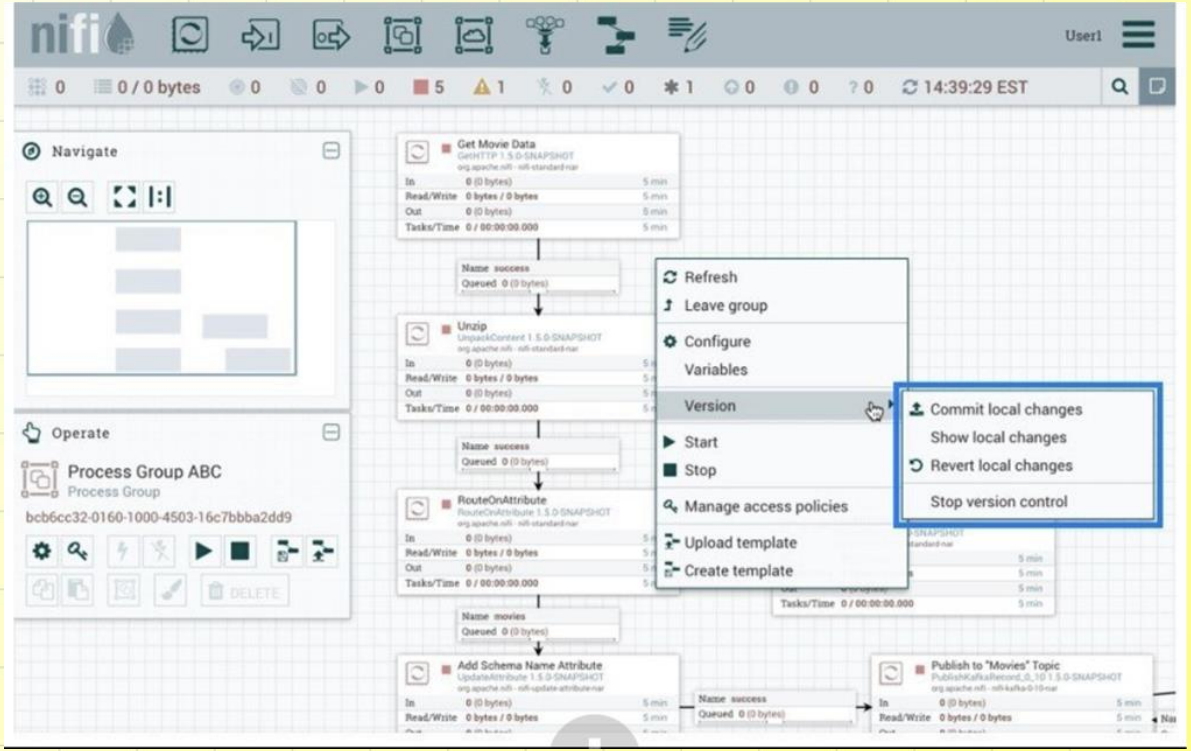
- Es un sistema de programación de workflows incluido en distribuciones de Hadoop.
- Los flujos de trabajo en Oozie están definidos como una colección de tareas representadas en un DAG.
- Acciones soportadas: MapReduce, Shell, Pig, Hive, Spark, Java, entre otros.



Nifi

Nifi

- Es una herramienta desarrollada por la NSA que permite automatizar flujos de datos entre sistemas.
- Posee una interfaz web que permite crear flujos sin necesidad de escribir código.
- Brinda funcionalidades de seguridad, monitoreo y linaje de datos en movimiento.

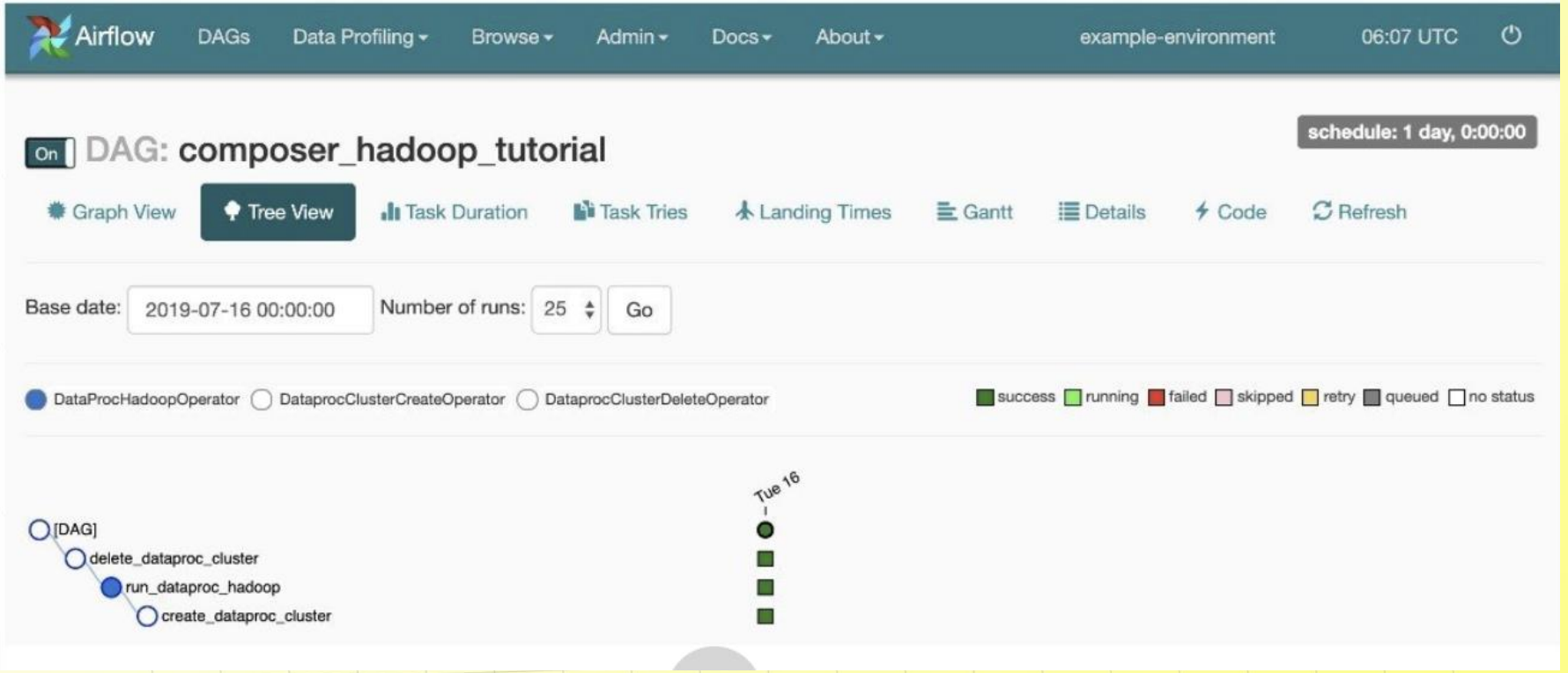


Airflow

¿QUÉ ES AirFlow?

- Es una plataforma de gestión de flujos de trabajo de código abierto desarrollada por Airbnb.
- Las tareas y dependencias se representan como DAG's definidos en scripts Python.
- Los DAG's pueden ser programados para ejecutarse en un horario predefinido o en función de la ocurrencia de eventos.

Apache
Airflow


Airflow
DAGs
Data Profiling
Browse
Admin
Docs
About
example-environment
06:07 UTC
On
DAG: composer_hadoop_tutorial
schedule: 1 day, 0:00:00
Graph View
Tree View
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Refresh
Base date:
2019-07-16 00:00:00
Number of runs:
25
Go
DataProcHadoopOperator
DataprocClusterCreateOperator
DataprocClusterDeleteOperator
success
running
failed
skipped
retry
queued
no status
Tue 16
[DAG]
delete_dataproc_cluster
run_dataproc_hadoop
create_dataproc_cluster

RESUMEN DE LA CLASE

- ✓ La importancia de los flujos de trabajo
- ✓ DAG
- ✓ Notación Cron
- ✓ Oozie
- ✓ Nifi
- ✓ Airflow

¿PREGUNTAS?

¡Muchas gracias!