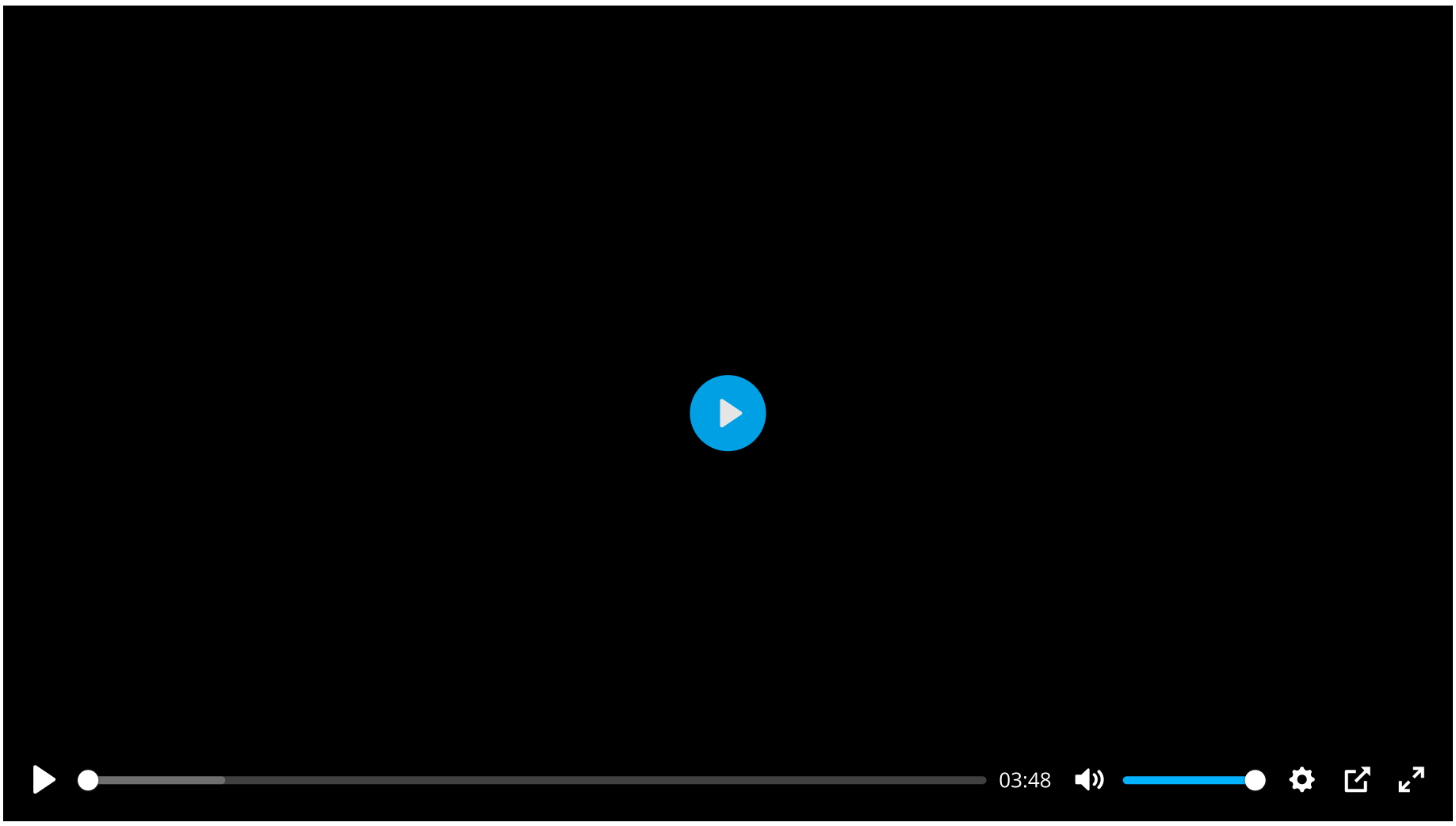
8.2 Airflow - знакомство и практические задачи



Презентация к уроку 8.2.3

Текстовая расшифровка видео:

APACHE AIRFLOW. АРХИТЕКТУРА

План:

- Архитектура Apache Airflow;
- 1 нода;
- Несколько нод;
- Механизм взаимодействия;
- Механизм работы распределенного инстанса Airflow.

Архитектура Apache Airflow

Airflow – open source-платформа для создания, планирования расписания и мониторинга рабочих процессов.

Airflow не отвечает за выполнение задач, этим занимаются другие виды софта, например, Python-интерпретатор или Bash.

Airflow DAG – это компонент Airflow, позволяющий отслеживать выполнение задач.

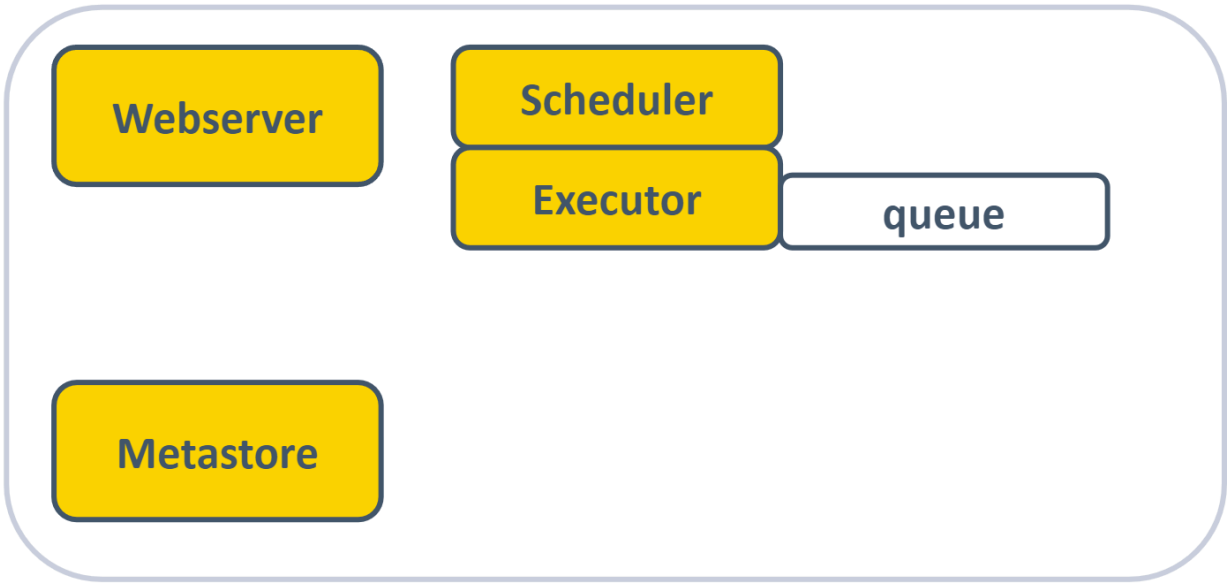
DAG – направленный ацикличный граф.

Граф состоит из операторов, где **Оператор** представляет собой **задачи, из которых состоит DAG**.

1 нода

Архитектура одной ноды подразумевает под собой хранение всех компонентов на одной фактической машине.

Рассмотрим схему:



Webserver – это Webqueue, позволяющий пользователю отслеживать выполнение тасков, задавать настройки и т.д.

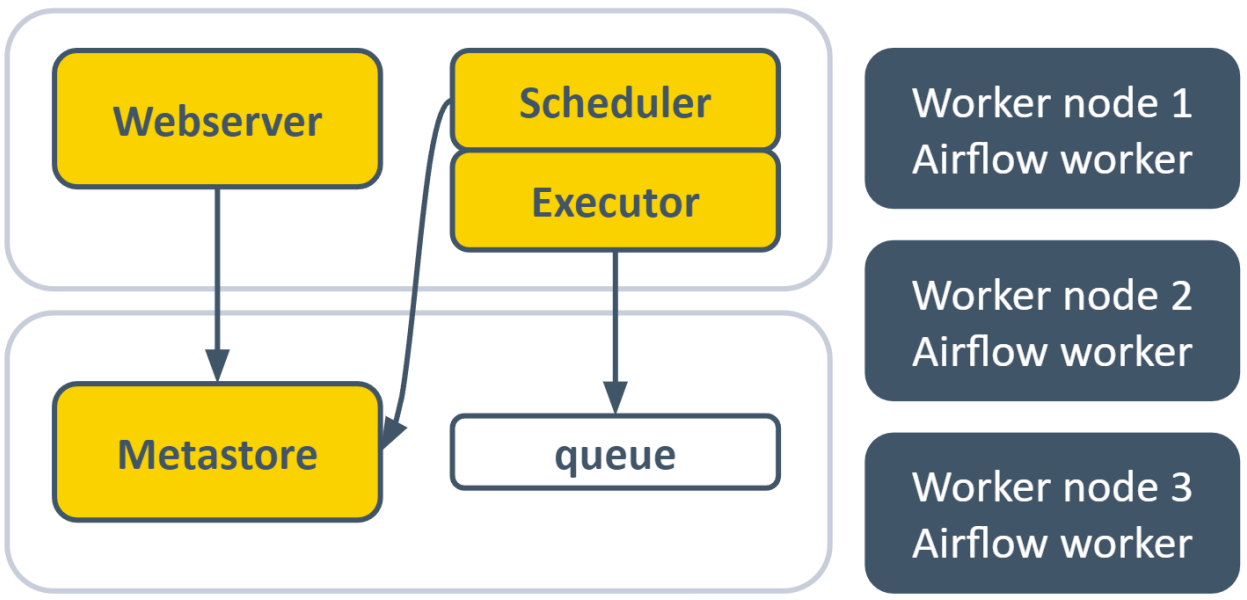
Metastore – это база данных, в которой хранятся все данные по инстансу (пользователи, права и т.д.).

Scheduler – это планировщик задач, который отправляет задачи на исполнение и следит за их выполнением. Он хранит в себе расписание и интервалы, с которыми нужно выполнять задачи.

Executor отвечает за выполнение задач, за их распределение в очереди. Из очередей задачи отправляются в воркеры.

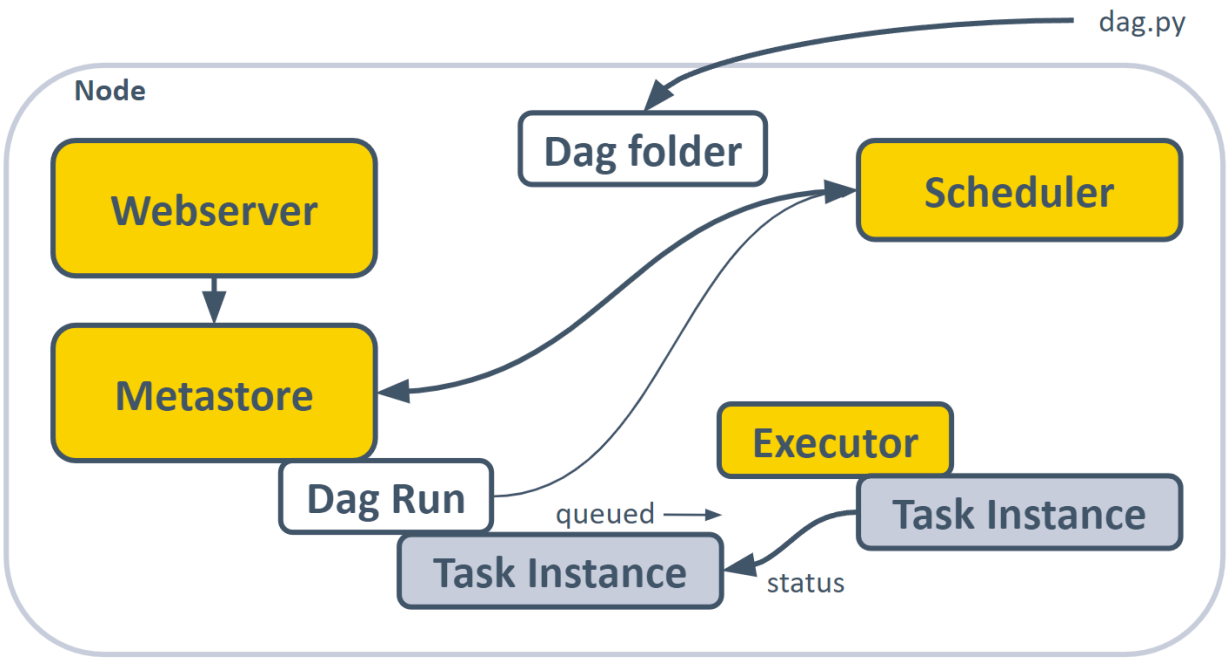
Несколько нод

В архитектуре, когда у нас несколько нод (несколько фактических машин), мы можем разнести компоненты на разные машины. Для воркеров используют более мощные машины с бóльшим количеством памяти:



Механизм взаимодействия

Как это работает:



- Сначала мы кладем в Dag folder файл Python – **dag**.

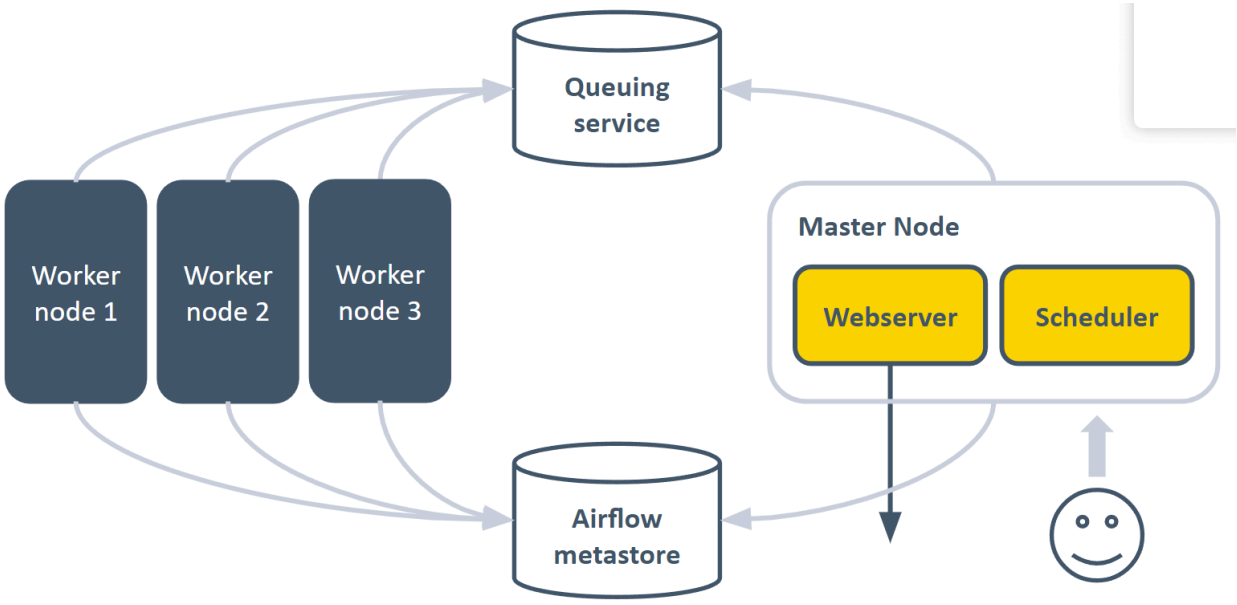
- Airflow, в конфигурации которого прописан директорий «Dag folder», смотрит туда. Если какие-то данные обновились, то Scheduler их подтягивает в Metastore.
- Metastore-база отправляет их на Webserver, где пользователь видит новый dag или изменения в уже имеющихся, а также генерирует объект «Dag Run» (это объект, который показывает, что Dag был успешно/неуспешно запущен).
- У Dag Run формируется Task Instance по каждому таску. Они закладываются в очередь Executor.
- Executor следит за выполнением на воркерах и отправляет статус выполнения, дополнительную информацию обратно в Metastore. Информация об успешности/неуспешности выполнения отправляется обратно в Scheduler, и таски закладываются на дальнейшее исполнение.
- Вся информация доступна пользователю на Webserver.

Механизм работы распределенного инстанса Airflow

Когда распределенные инстансы – Airflow, тип Executor'а должен быть Salary Executer.

Salary – это асинхронный планировщик задач, работающий с очередью. Из нее он автоматически подгружает задачи на воркеры.

Рассмотрим схему:



Здесь пользователь работает с Master Node, где находятся Scheduler и Webserver. С этой ноды пользователь получает данные в базе данных Airflow Metastore. Оттуда отправляются задачи на выполнение, а также происходит обмен данными по их выполнению.

Как вам урок?

Изучил, далее

Слёрм ©

+7 (495) 248-05-80

Лицензия №ДЛ-1368 от 22.08.2019

Политика конфиденциальности

Публичная оферта