Scaling Airflow

Kan Ouivirach

Kan Ouivirach, PhD

- Data Product Developer & Technical Coach, **ODDS**
- Community Organizer of **Data Council Bangkok**
- Facebook Page Admin of **Data Engineer Cafe**
- Contributor in **Apache Airflow**

- Ph.D. in Computer Science, Computer Vision & Machine Learning, **Asian Institute of Technology**
- ex-Software Engineer Manager, **Pronto Tools**

Course Module

No.	Date / Time	Module
1	Jan 20, 2024 / 9:00 - 12:00	Introduction to Data Pipeline and Apache Airflow
2	Jan 27, 2024 / 9:00 - 12:00	Implementing Advanced Concepts in Airflow
3	Feb 3, 2024 / 9:00 - 12:00	Building a Production Grade Data Pipeline
4	Feb 10, 2024 / 9:00 - 12:00	Maintaining and Monitoring Data Pipelines in Airflow
	Feb 17, 2024 / 9:00 - 12:00	– No Class –
5	Feb 24, 2024 / 9:00 - 12:00	**วันมาฆบูชา** & Ensuring Data Quality with Automated Data Pipelines (Online)
6	Mar 2, 2024 / 9:00 - 12:00	Introducing Analytics Engineering
7	Mar 9, 2024 / 9:00 - 12:00	Applying Analytics Engineering in Data Pipelines
8	**Mar 16, 2024 / 9:00 - 12:00**	**Scaling Airflow**

Agenda

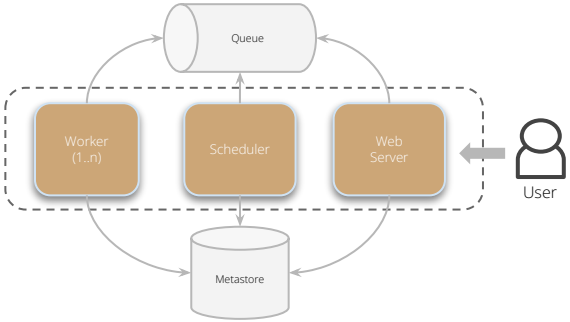
- Architecture of Airflow
- Scaling Airflow

Architecture of Airflow

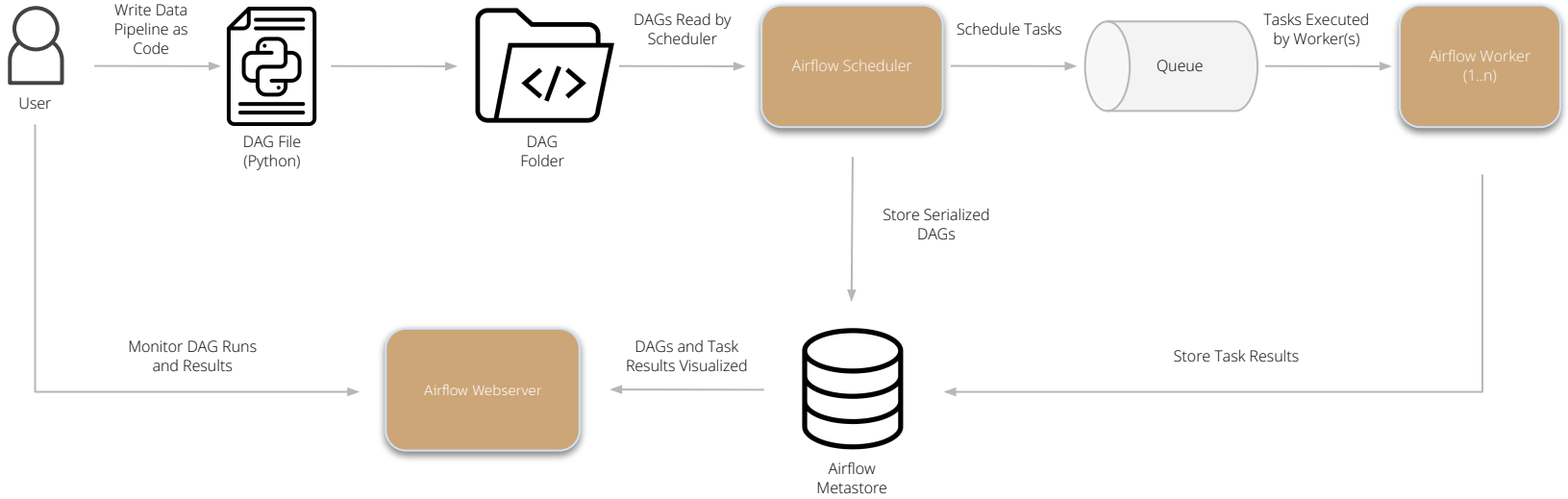
Architecture of Airflow

There are 5 main components

- **Web Server**: It provides a control dashboard for users
- **Scheduler**: It orchestrates execution of jobs on a trigger or schedule
- **Worker**: It executes the operations defined in a DAG
- **Queue**: It holds the state of running DAGs and tasks
- **Metastore**: It contains the status of the DAG runs and task instances

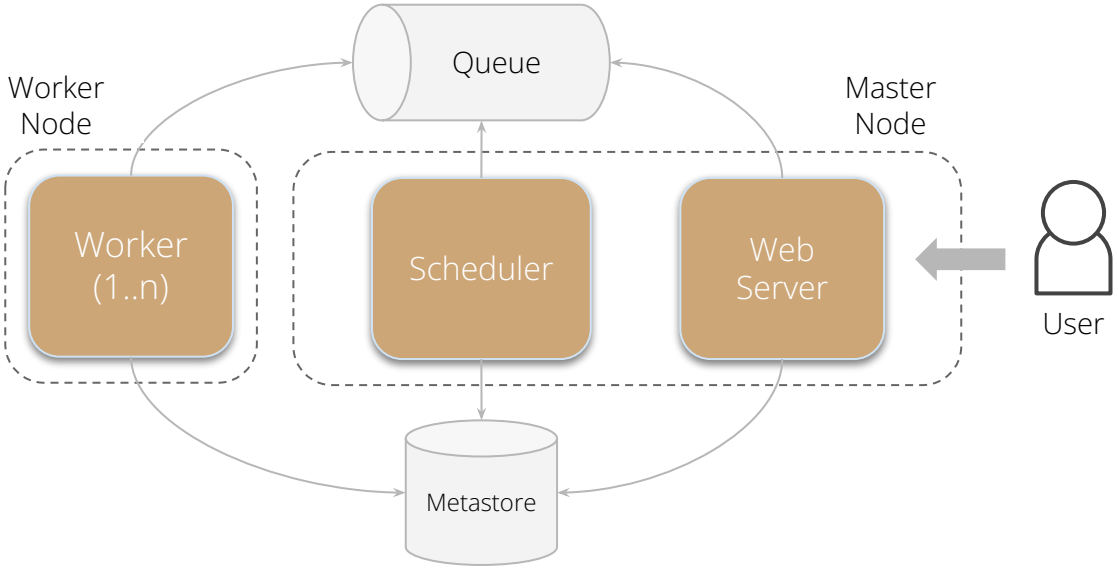


Overview of Process in Airflow

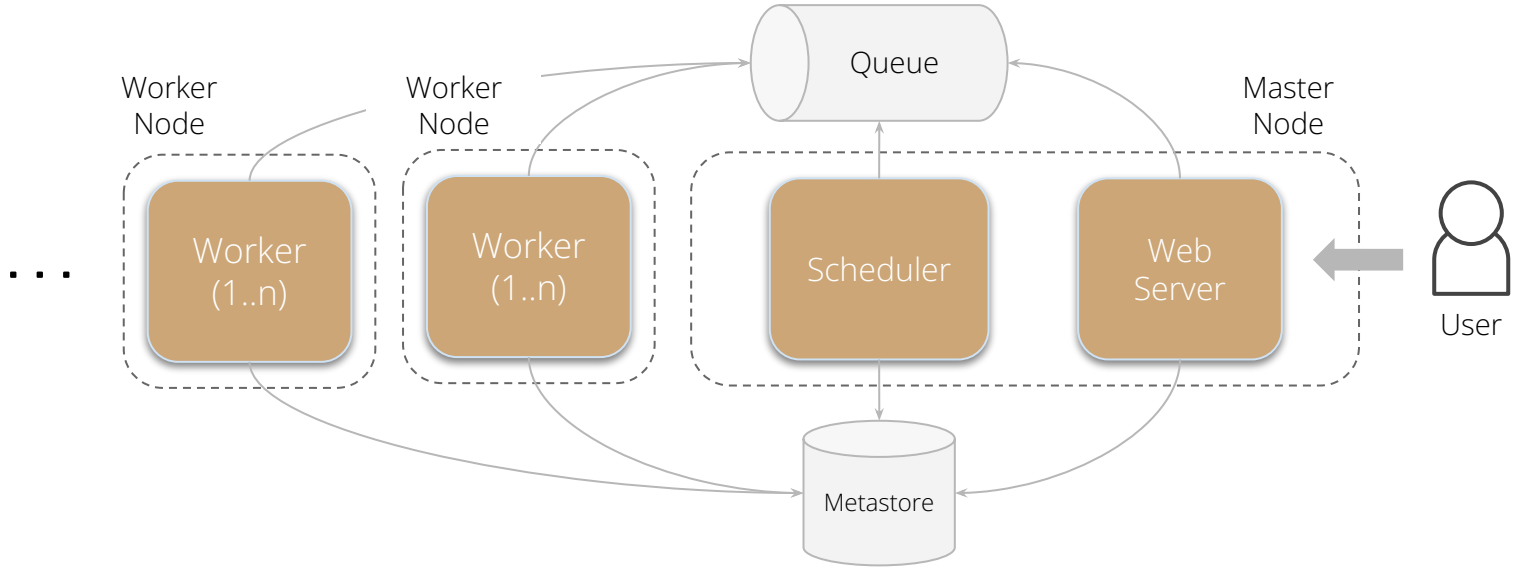


Scaling Airflow

Scaling Airflow



Scaling Airflow



Executors

Executors are the mechanism that runs the task instances

There are 3 main types:

1. Sequential Executor
2. Local Executor
3. Celery Executor

Other executors: Debug, Dask, Kubernetes, and CeleryKubernetes

Sequential Executor

Sequential Executor runs a single task instance at a time with no parallelism functionality

Pros:

- Helpful for debugging

Cons:

- No parallelism
- Not scalable at all
- Single point of failure

Local Executor

Local Executor completes tasks in parallel that run on a single machine

Pros:

- Easy to set up
- Cheap
- Still offering parallelism

Cons:

- Not scalable
- Single point of failure

Celery Executor

Celery Executor is for horizontal scaling

Pros:

- High availability
- Built for horizontal scaling

Cons:

- More expensive
- More effort to set up
- Worker maintenance

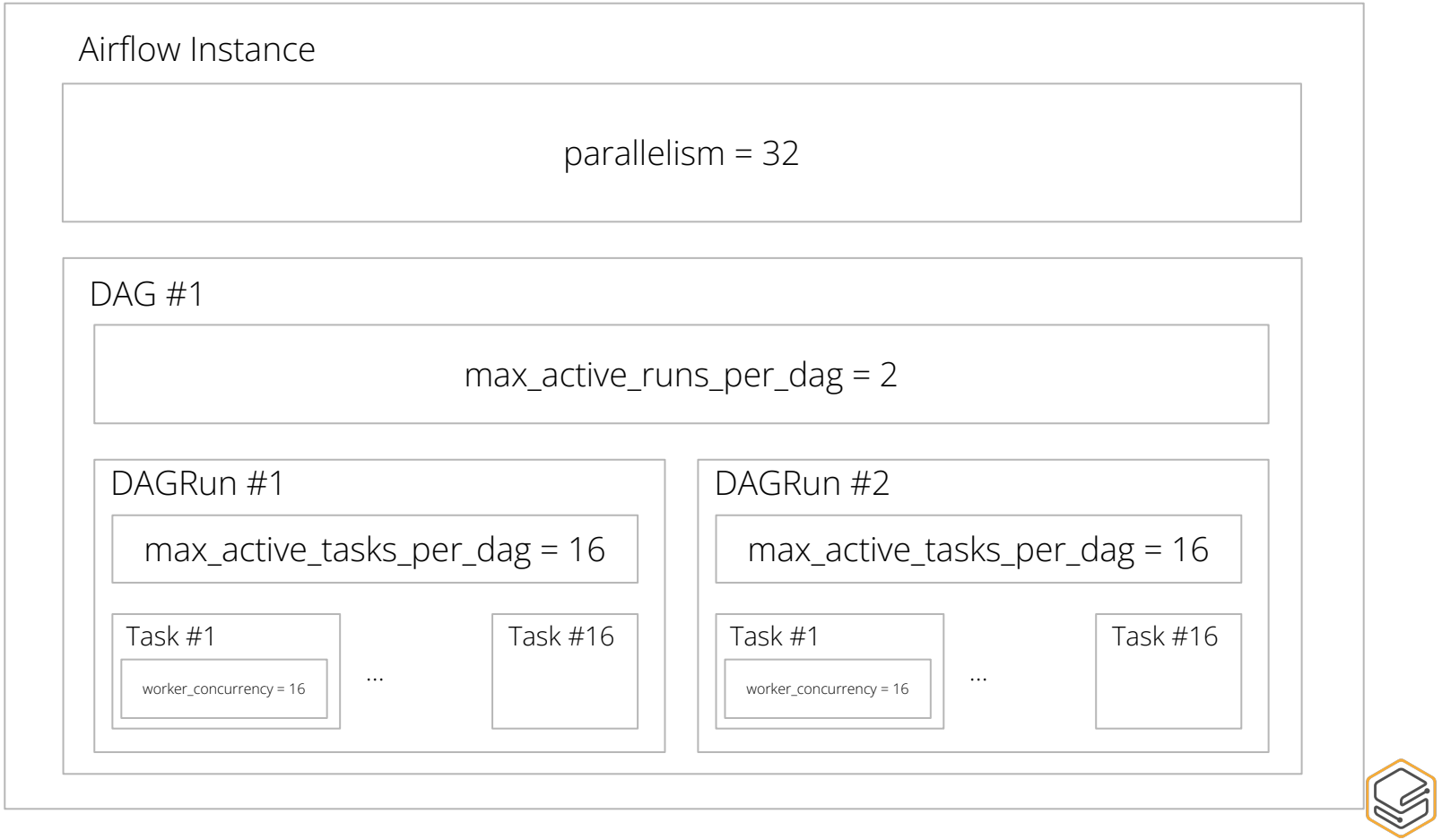
Which Executor is Right for You?

Executor	Distributed	Ease of Installation	Good Fit
SequentialExexcutor	No	Very Easy	Demoing or testing
LocalExexcutor	No	Easy	Single machine is enough
CeleryExexcutor	Yes	Moderate	Scale out over multiple machines
KubernetesExexcutor	Yes	Complex	Familiar with Kubernetes and prepare a containerized setup

Parallelism

We consider the following configuration variables

- `parallelism`
- `max_active_runs_per_dag`
- `max_active_tasks_per_dag`
- `worker_concurrency`


Airflow Instance
parallelism = 32
DAG #1
max_active_runs_per_dag = 2
DAGRun #1
max_active_tasks_per_dag = 16
Task #1
worker_concurrency = 16
...
Task #16
DAGRun #2
max_active_tasks_per_dag = 16
Task #1
worker_concurrency = 16
...
Task #16

Workshop

Scaling Airflow

Let’s try to scale Airflow with CeleryExecutor

Recap

Recap

- Studied the architecture of Airflow
- Learned how to scale Airflow with CeleryExecutor

Capstone Project

Capstone Project

Objective: Build a data pipeline to extract Reddit data from r/dataengineering for a trend analysis in Data Engineering

Technology stack should have

1. Apache Airflow
2. PostgreSQL or BigQuery
3. dbt (optional)

The data pipeline should be able to

- Extract the Reddit data from r/dataengineering daily
- Transform the data into the proper format
- Ensure the data quality

Extracting Reddit Data

- PRAW: The Python Reddit API Wrapper to extract the Reddit data
- See this doc to study how to obtain a Subreddit

Business Questions

- What are the total number of authors and posts in this subreddit?
- What is the average score?
- How many posts are published per day?
- What is the average number of comments per day?
- Is there any interesting trend at this moment?

Example of a Dashboard

Expectation

- Be able to extract the Reddit data
- Prepare the SQL for data transformation to answer the business questions mentioned in the previous slide
- The data model must be easy to show on a dashboard
- [Bonus] Use dbt to create your data model and schedule it in Airflow