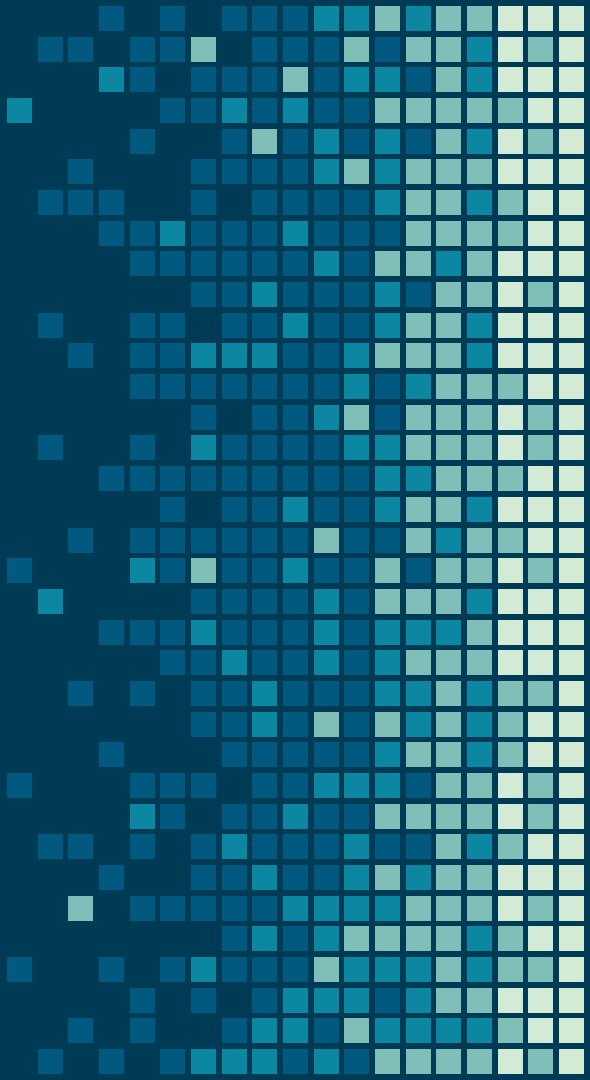
Mastering Data Engineering

Roman Golovnya
28 September 2024



About me

- Data Engineer in SUSE
- Over 10 years working experience in data related jobs
- Education: Finance, Data Analytics & Computer Science
- Ex kaggler
- Founder & organiser of DSEClub

What is Apache Airflow?

- Apache Airflow is an open-source platform for authoring, scheduling and monitoring data and computing workflows.

  DAG Directed Acyclic Graph – is a collection of all the tasks you want to run, organized in a way that reflects their relationships and dependencies.

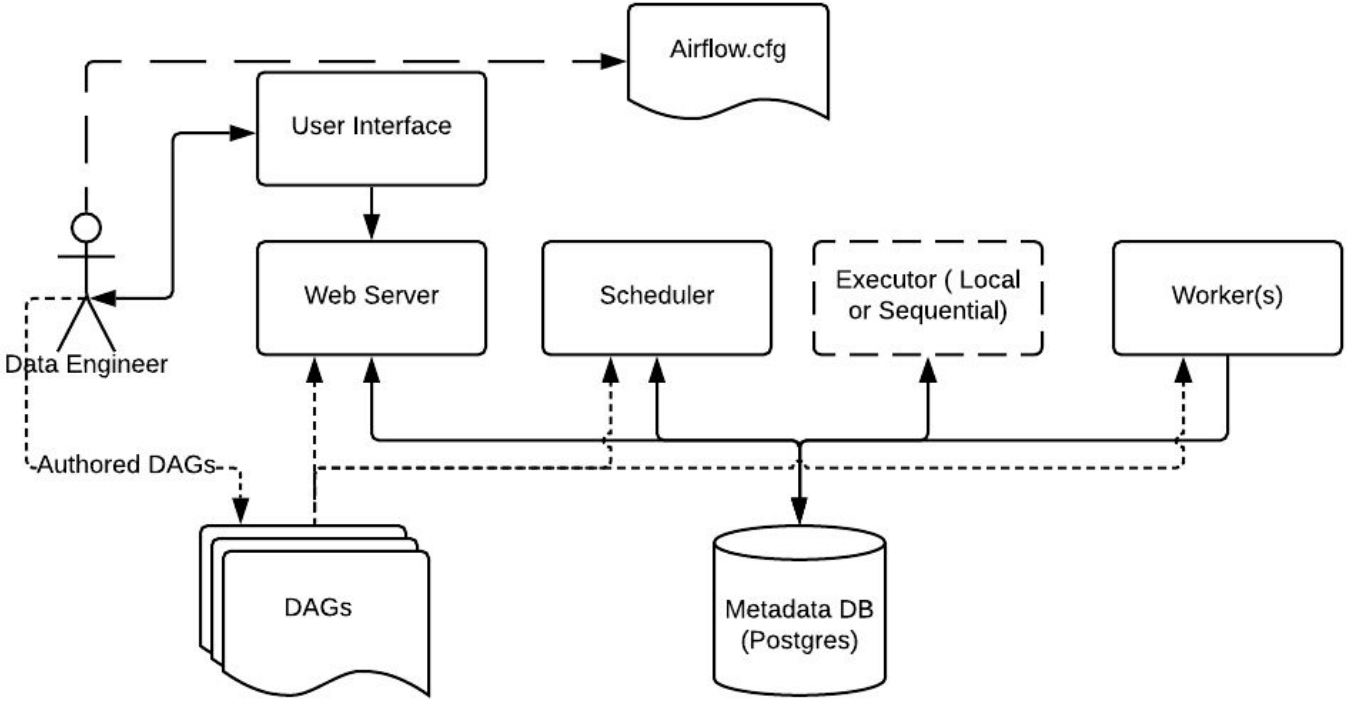
  Batch oriented data processing - scheduler cron

- Airflow was created in 2014 by Maxime Beauchemin at Airbnb.

Features of Apache Airflow

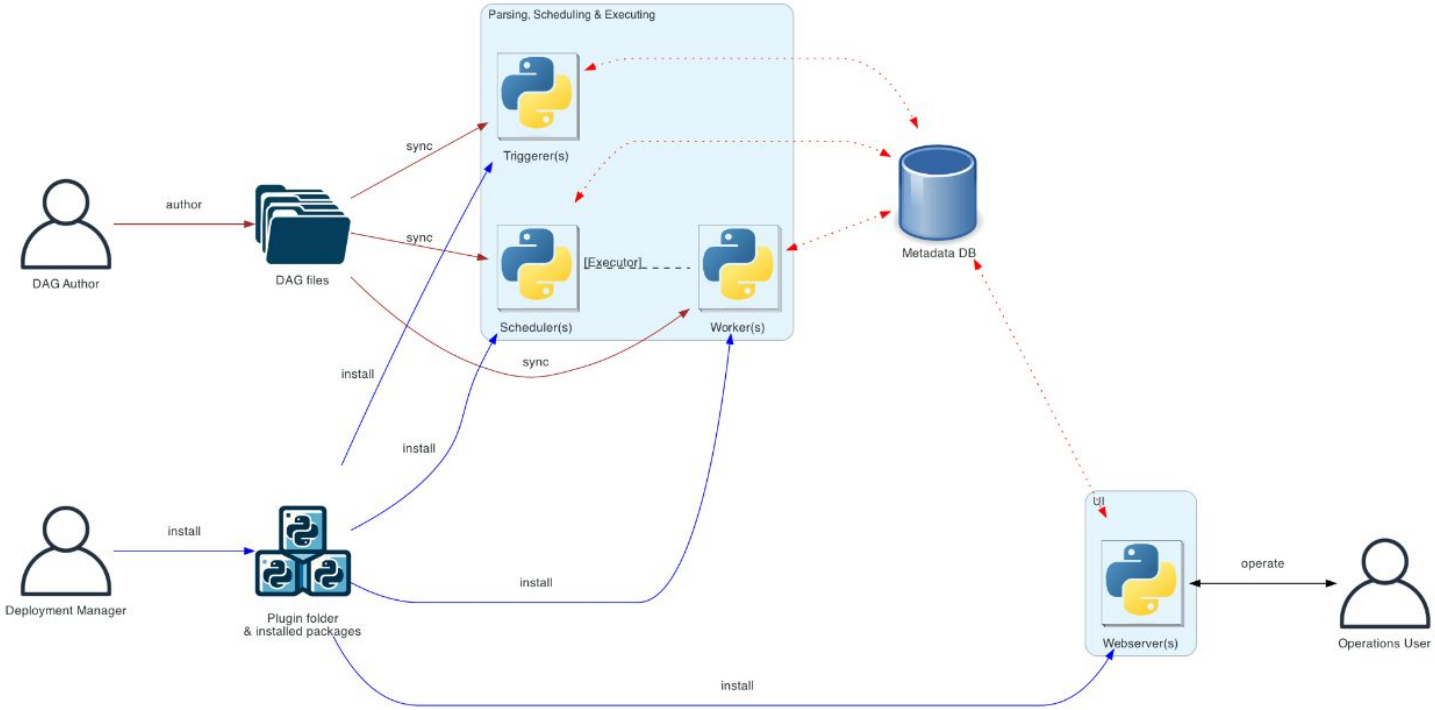
- Simple and friendly User Interface
- Open source
- Python code
- Robust Integrations
- Jinja templates
- Scalable
- Wide and active community

What is Apache Airflow Architecture



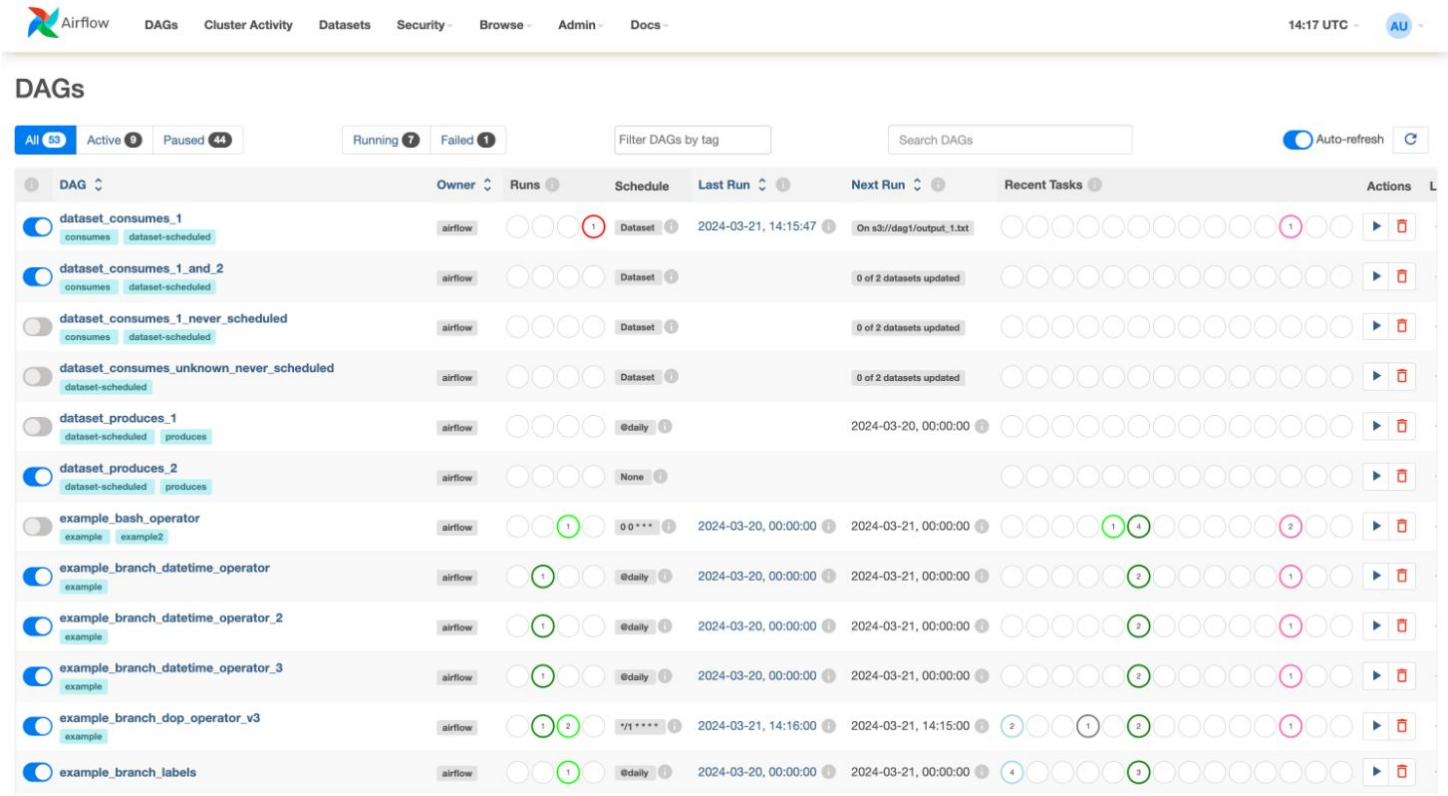
https://airflow.apache.org/docs/apache-airflow/2.0.1/concepts.html

Distributed Airflow architecture



https://airflow.apache.org/docs/apache-airflow/2.10.2/core-concepts

Airflow DAG and UI



https://airflow.apache.org/docs/apache-airflow/2.10.2/core-concepts

Airflow Components:

- **DAG**: It is the Directed Acyclic Graph – a collection of the tasks that you want to run which is organized and shows the relationship between different tasks.
- **Task** is the basic unit of execution in Airflow. Operators, Sensors,Taskflowdecorator
- **Web Server:** It is the user interface built on the Flask. It allows us to monitor the status of the DAGs and trigger them.
- **Metadata Database**: Airflow stores the status of all the tasks in a database Postgres and do all read/write operations of a workflow.
- **Scheduler**: is responsible for scheduling the execution of DAGs. It retrieves and updates the status of the task in the database.
- **Executor:** is process by which task instances are run
- **Worker:** a separate instance which job run specific task

Competitors of Apache Airflow:

- **Dagster** https://github.com/dagster-io/dagster
- **Prefect** https://github.com/PrefectHQ/prefect

Airflow learning resources:

- https://airflow.apache.org/docs/apache-airflow/stable/tutorial/index.html
- https://academy.astronomer.io/path/airflow-101
- https://www.astronomer.io/docs/learn/intro-to-airflow

https://theaisummer.com/apache-airflow-tutorial/

Contact me:

roman.golovnya@gmail.com

https://www.linkedin.com/in/romangolovnya

Via meetup.com/messages

https://github.com/dseclub

https://www.kaggle.com/rgolovnya

Thank you!

Any Questions?