Microsoft Learn
Student Ambassadors

Orchestrate your data pipelines using Apache Airflow

Presented by Mahmoud Fettal

in /mahmoud-fettal

/mahmoudfettal

/MahmoudFettal

Where do we go now?

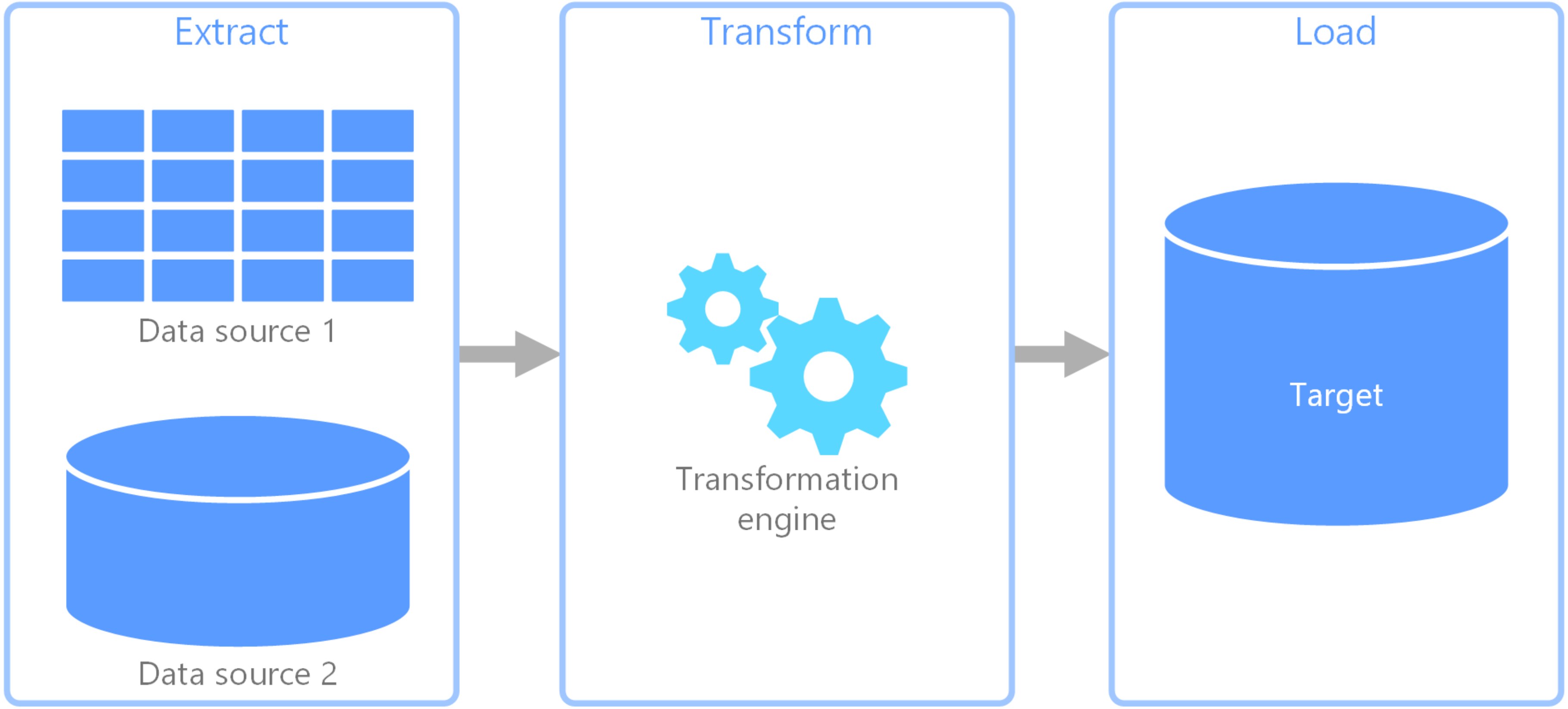
What is a data pipeline?

“Data pipelines generally consist of several tasks or actions that need to be executed to achieve the desired result”

Source: https://www.manning.com/books/data-pipelines-with-apache-airflow

What is a data pipeline?

What is a data pipeline?



Source: https://learn.microsoft.com/en-us/azure/architecture/data-guide/relational-data/etl?wt.mc_id=studentamb_123265

What are the challenges of data pipelines?

Monitoring

Scheduling

Linked tasks

Scalability

What is Apache Airflow?

airbnb

Apache Airflow

“Airflow is a batch workflow orchestration platform. The Airflow framework contains operators to connect with many technologies and is easily extensible to connect with a new technology. If your workflows have a clear start and end, and run at regular intervals, they can be programmed as an Airflow DAG.”

Source: https://airflow.apache.org/docs/apache-airflow/stable/index.html

What does Apache Airflow do/provide?

Schedule pipelines

Web interface for monitoring

Pipelines as code (mainly python)

Pre-built components

Easy integration with cloud services

Source: https://airflow.apache.org/docs/apache-airflow/stable/index.html

Why you probably don’t need Apache Airflow?

Tasks do not rely on each other -> Use CRON jobs + generate logs of the tasks

Streaming data -> Consider using Apache Kafka or Apache storm

Source: https://airflow.apache.org/docs/apache-airflow/stable/index.html

Alternatives to Apache Airflow

Azkaban

By

LinkedIn

conductor

By

NETFLIX

Luigi

By

Source: https://airflow.apache.org/docs/apache-airflow/stable/index.html

What is a DAG?

Dag stands for **D**irect **A**cyclic **G**raph

A graph representation of the pipeline makes it easy to interpret.

The graphs are acyclic to avoid infinite loop.

Task 2 will never be able to execute, due to its dependency on task 3, which in turn depends on task 2.

Source: https://www.manning.com/books/data-pipelines-with-apache-airflow

Apache Airflow installation

Prerequisites

Starting with Airflow 2.3.0, Airflow is tested with:

- Python: 3.7, 3.8, 3.9, 3.10
- Databases:
  - PostgreSQL: 11, 12, 13, 14, 15
  - MySQL: 5.7, 8
  - SQLite: 3.15.0+
  - MSSQL(Experimental): 2017, 2019
- Kubernetes: 1.20.2, 1.21.1, 1.22.0, 1.23.0, 1.24.0

The minimum memory required we recommend Airflow to run with is 4GB, but the actual requirements depends wildly on the deployment options you have

Source: https://airflow.apache.org/docs/apache-airflow/stable/installation/prerequisites.html

Apache Airflow installation

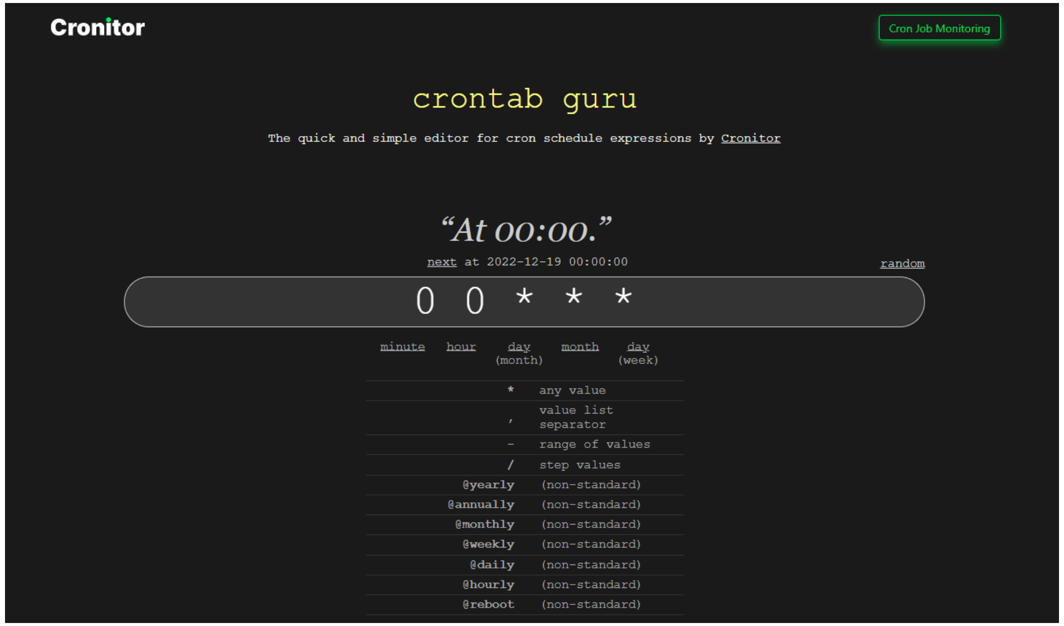
How to install Apache Airflow

- Simply use PyPI and use the command: *"pip install apache-airflow"*
- Using Docker production images.
- Using Airflow managed resources on different cloud providers.

Source: https://airflow.apache.org/docs/apache-airflow/stable/installation/index.html

One last thing before the demo

CRON expressions



crontab guru

The quick and simple editor for cron schedule expressions by Cronitor

“At 00:00.”

next at 2022-12-19 00:00:00

random

```
0 0 * * *
```

minute hour day (month) month day (week)

*	any value
,	value list separator
-	range of values
/	step values
@yearly	(non-standard)
@annually	(non-standard)
@monthly	(non-standard)
@weekly	(non-standard)
@daily	(non-standard)
@hourly	(non-standard)
@reboot	(non-standard)

Visit https://crontab.guru for more information

Source: https://airflow.apache.org/docs/apache-airflow/stable/dag-run.html#cron-presets

Let me risk a little more light.

DEMO TIME

Demo

What next?

- Apache Airflow in the clouds:
  https://azure.microsoft.com/fr-fr/blog/deploying-apache-airflow-in-azure-to-build-and-run-data-pipelines/
  https://aws.amazon.com/managed-workflows-for-apache-airflow/
  https://cloud.google.com/composer/docs/run-apache-airflow-dag

- Building custom components:
  https://airflow.apache.org/docs/apache-airflow-providers/index.html

- Apache Airflow best practices:
  https://airflow.apache.org/docs/apache-airflow/stable/best-practices.html

Source: https://airflow.apache.org/docs/apache-airflow/stable/dag-run.html#cron-presets

Thank you!

Microsoft Learn
Student Ambassadors

Orchestrate your data pipelines using Apache Airflow

Presented by Mahmoud Fettal

in /mahmoud-fettal

/mahmoudfettal