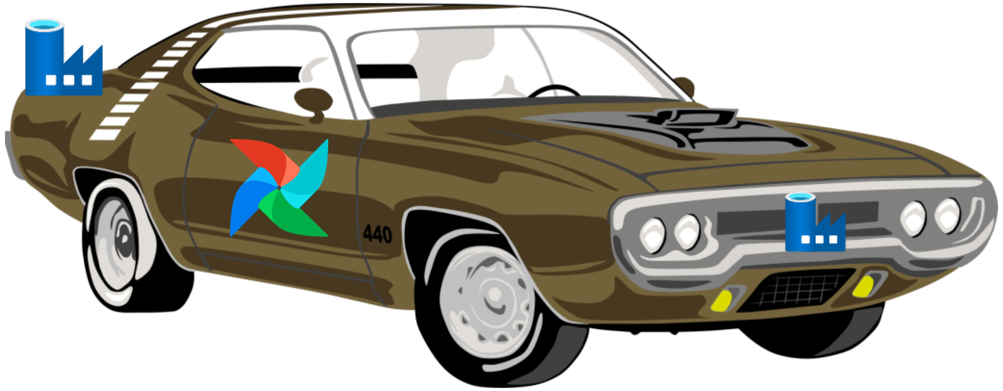
scieneers
DRIVEN BY DATA

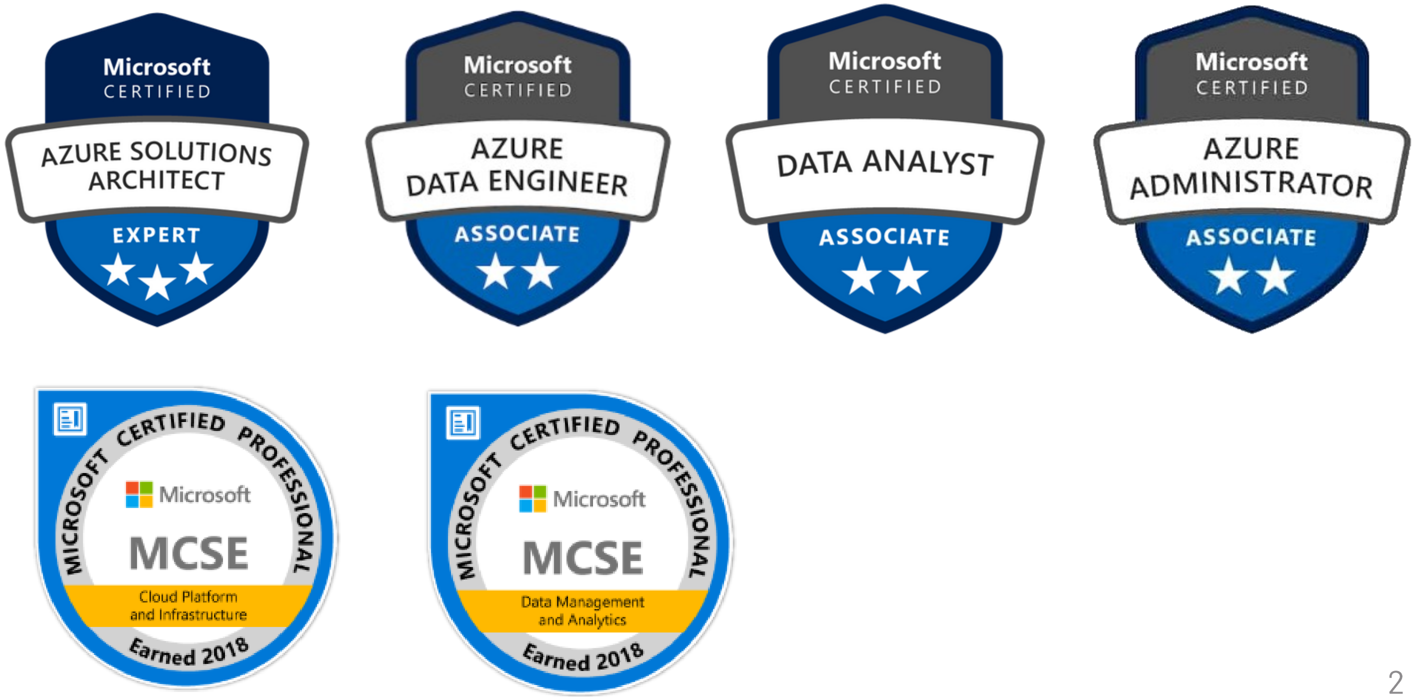
A new car is in town – Apache Airflow in der Azure Data Factory

Introducing Apache Airflow Integration in ADF



Stefan Kirner

- PASS Chapter Lead Karlsruhe ski@sqlpass.de & Beirat
- Director Business Intelligence scieneers GmbH
- Twitter: @KirnerKa



Agenda

- Apache Airflow in a nutshell
- Airflow concepts
- Airflow in Azure Data Factory
- Demo
- Comparison Airflow vs. ADF pipelines
- Best fit use cases for Airflow in ADF
- Bugs, Alternatives, Links

Apache Airflow in a nutshell

What is it and how does it work?

What is Airflow?

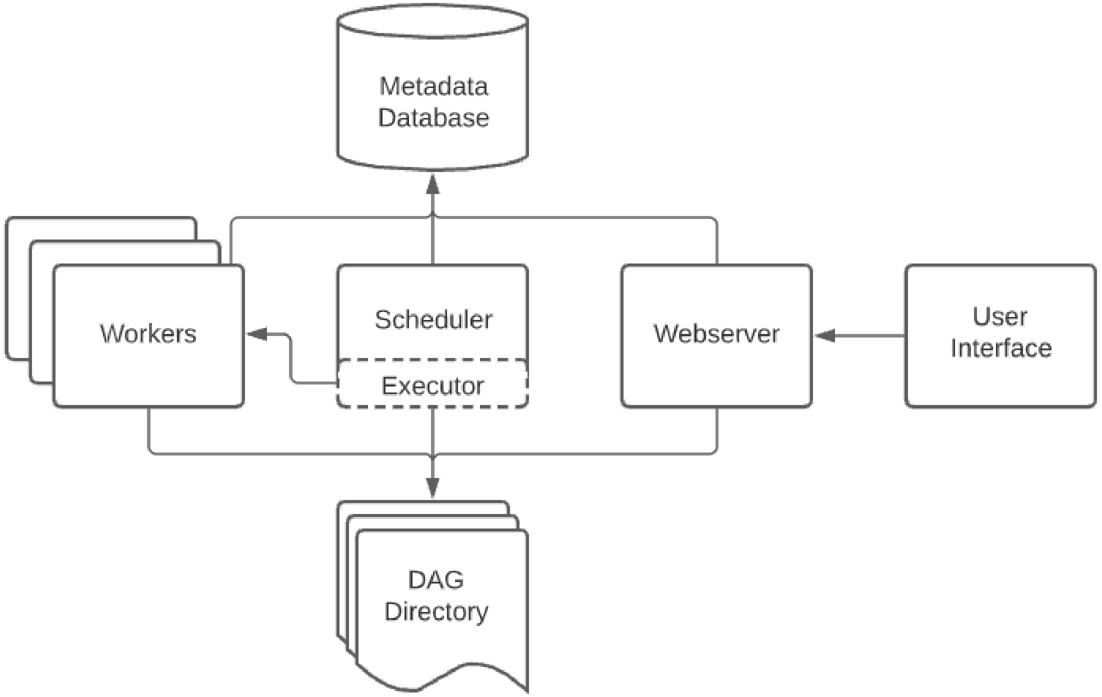
- Platform for authoring, scheduling & monitoring workflows
- Open-Source
- Widely used – 16 Mio downloads / month
- Code-centric – using Python language
- Execute nearly anything in any environment (using Python)
- Extensible by own operators, executors and libraries
- Scalable

What is Airflow **not**?

- No GUI-based ETL Tool
- No special support for ETL functionality
  - like parallel threaded bulk load, schema or surrogate key creation
- Not very lightweight (# of components to operate)

Airflow systems architecture overview

- **Scheduler**: triggering scheduled workflows, and submitting taks to the executor
- **Executor**: do the work – itself or push it to workers
- **Webserver**: user interface
- **DAG Directory**: folder of DAG files
- **metadata database**: storing state by the components



https://airflow.apache.org/docs/apache-airflow/stable/core-concepts/overview.html#

Airflow systems architecture detailed sample

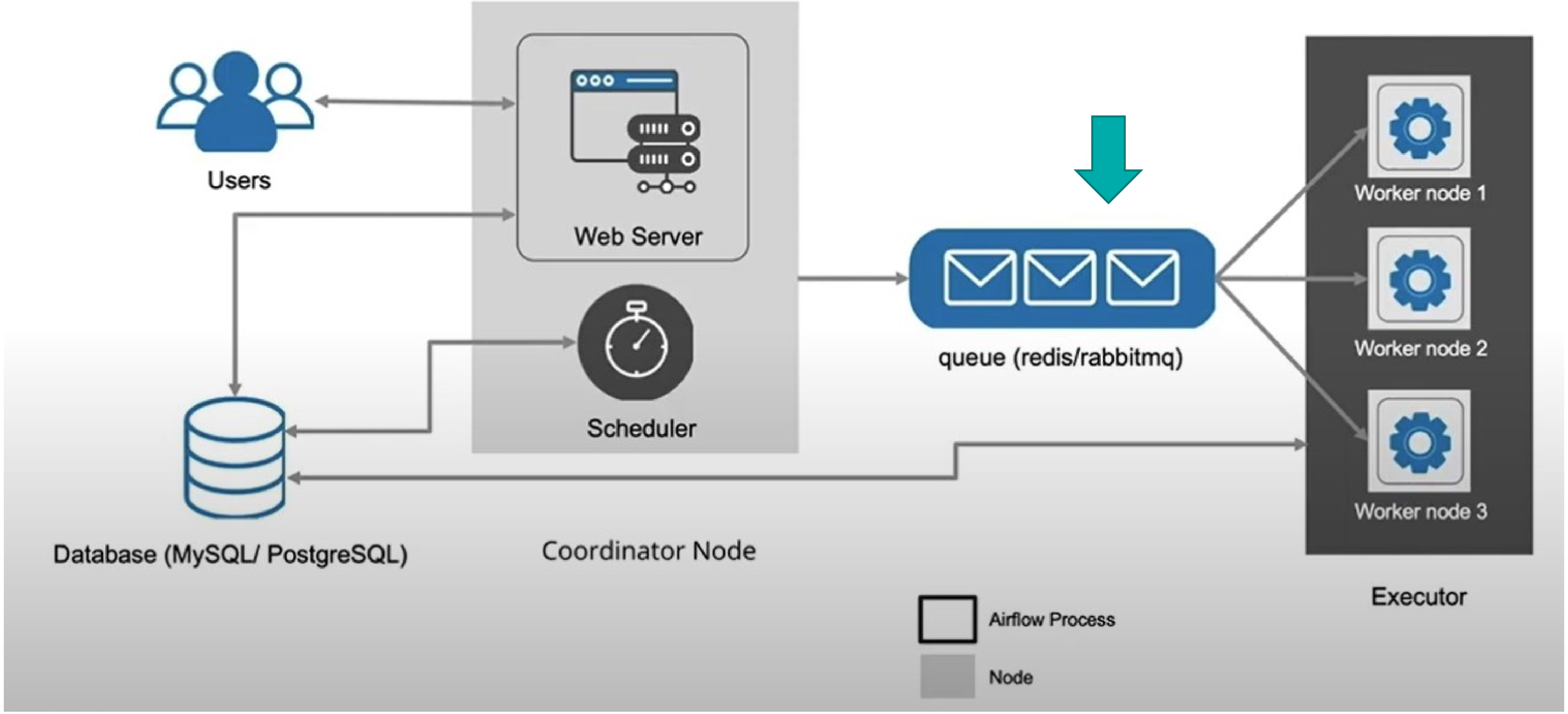
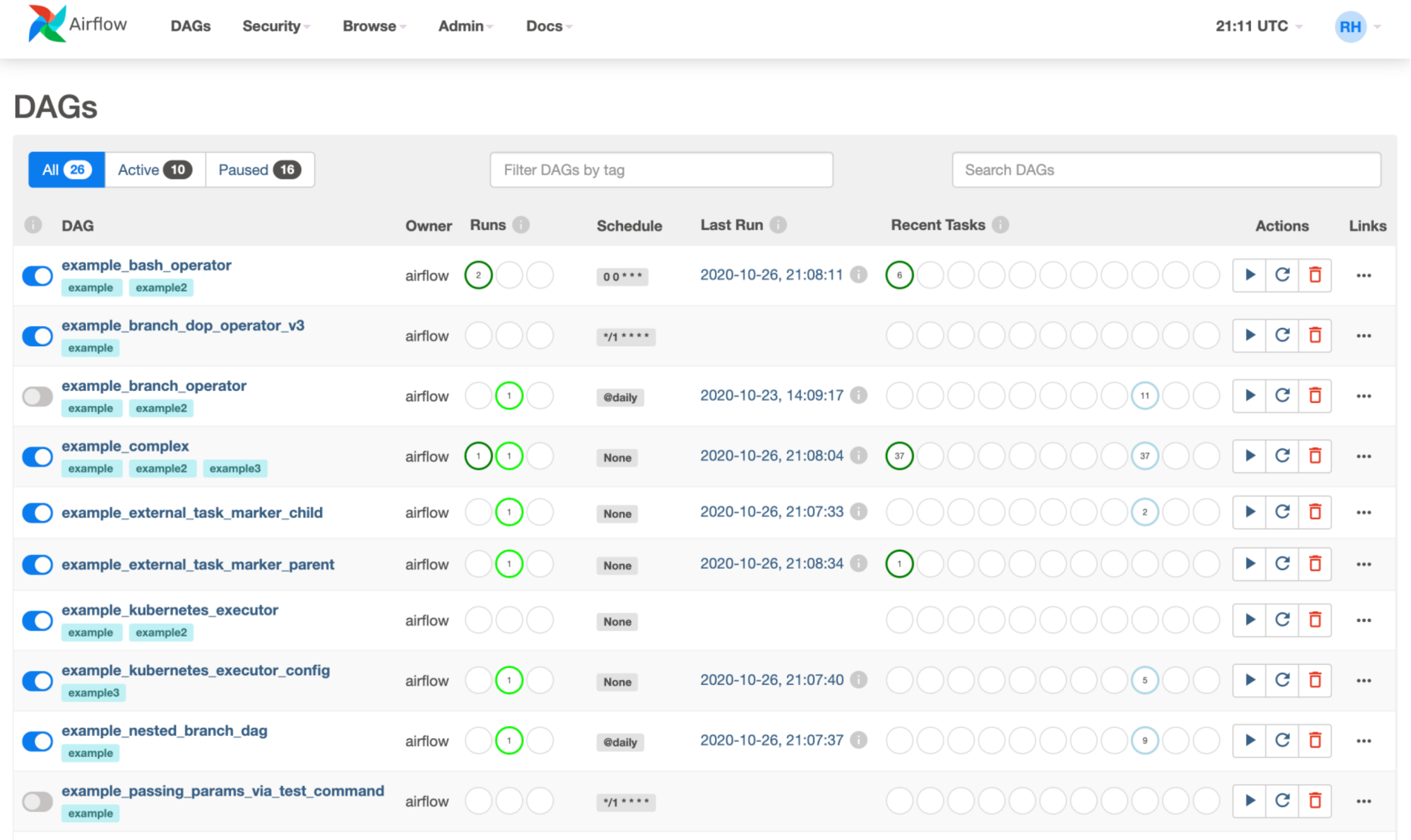


Diagram from Qubole Airflow 101 Architecture of Apache Airflow Link

How does this fancy user interface look?



DAGs

All 26 | Active 10 | Paused 16

Filter DAGs by tag

Search DAGs

DAG	Owner	Runs	Schedule	Last Run	Recent Tasks	Actions	Links
example_bash_operator example example2	airflow	2	0 0 * * *	2020-10-26, 21:08:11	6	▶ ↻ 🗑	…
example_branch_dop_operator_v3 example	airflow		*/1 * * * *			▶ ↻ 🗑	…
example_branch_operator example example2	airflow	1	@daily	2020-10-23, 14:09:17	11	▶ ↻ 🗑	…
example_complex example example2 example3	airflow	1 1	None	2020-10-26, 21:08:04	37 37	▶ ↻ 🗑	…
example_external_task_marker_child	airflow	1	None	2020-10-26, 21:07:33	2	▶ ↻ 🗑	…
example_external_task_marker_parent	airflow	1	None	2020-10-26, 21:08:34	1	▶ ↻ 🗑	…
example_kubernetes_executor example example2	airflow		None			▶ ↻ 🗑	…
example_kubernetes_executor_config example3	airflow	1	None	2020-10-26, 21:07:40	5	▶ ↻ 🗑	…
example_nested_branch_dag example	airflow	1	@daily	2020-10-26, 21:07:37	9	▶ ↻ 🗑	…
example_passing_params_via_test_command example	airflow		*/1 * * * *			▶ ↻ 🗑	…

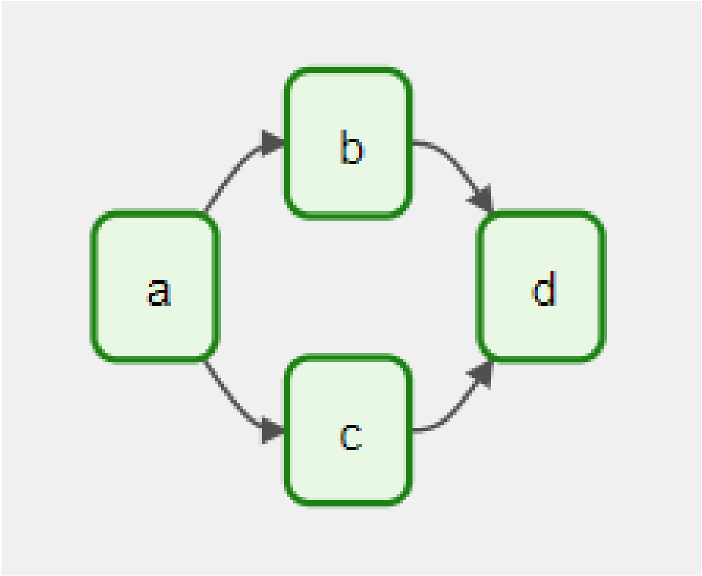
https://airflow.apache.org/docs/apache-airflow/stable/core-concepts/overview.html#user-interface

Airflow concepts

Building blocks

What are Directed Acyclic Graphs (DAGs)

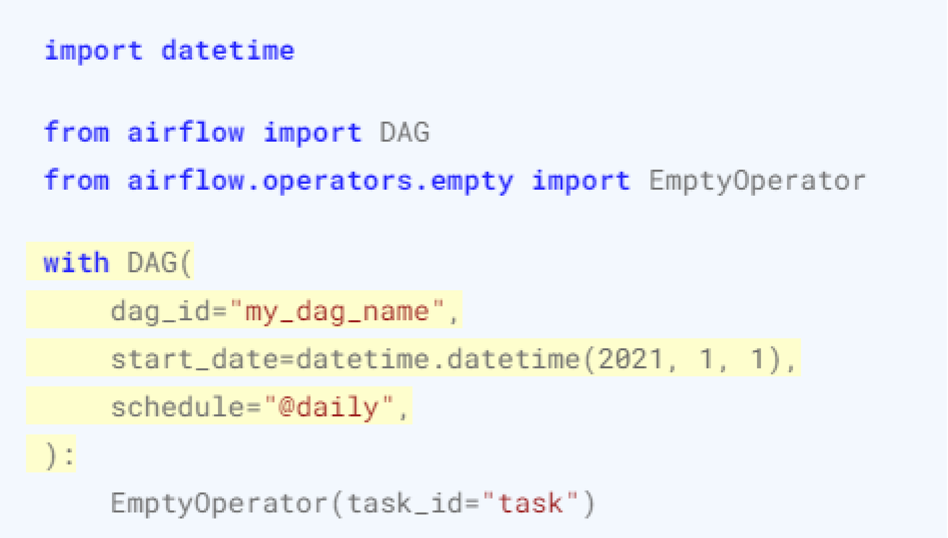
- Graph consists of edges and nodes
- Each edge point towards a node
- Directed means that no cycles formed by the edges



https://airflow.apache.org/docs/apache-airflow/stable/core-concepts/dags.html

DAGs in Airflow

- Collection of **tasks**
- Authored in Python
- Define running order and dependencies of tasks
- Scheduling and number of repeats defined
- DAGRun: instance of a DAG at runtime



```
import datetime

from airflow import DAG
from airflow.operators.empty import EmptyOperator

with DAG(
    dag_id="my_dag_name",
    start_date=datetime.datetime(2021, 1, 1),
    schedule="@daily",
):
    EmptyOperator(task_id="task")
```

https://airflow.apache.org/docs/apache-airflow/stable/core-concepts/dags.html

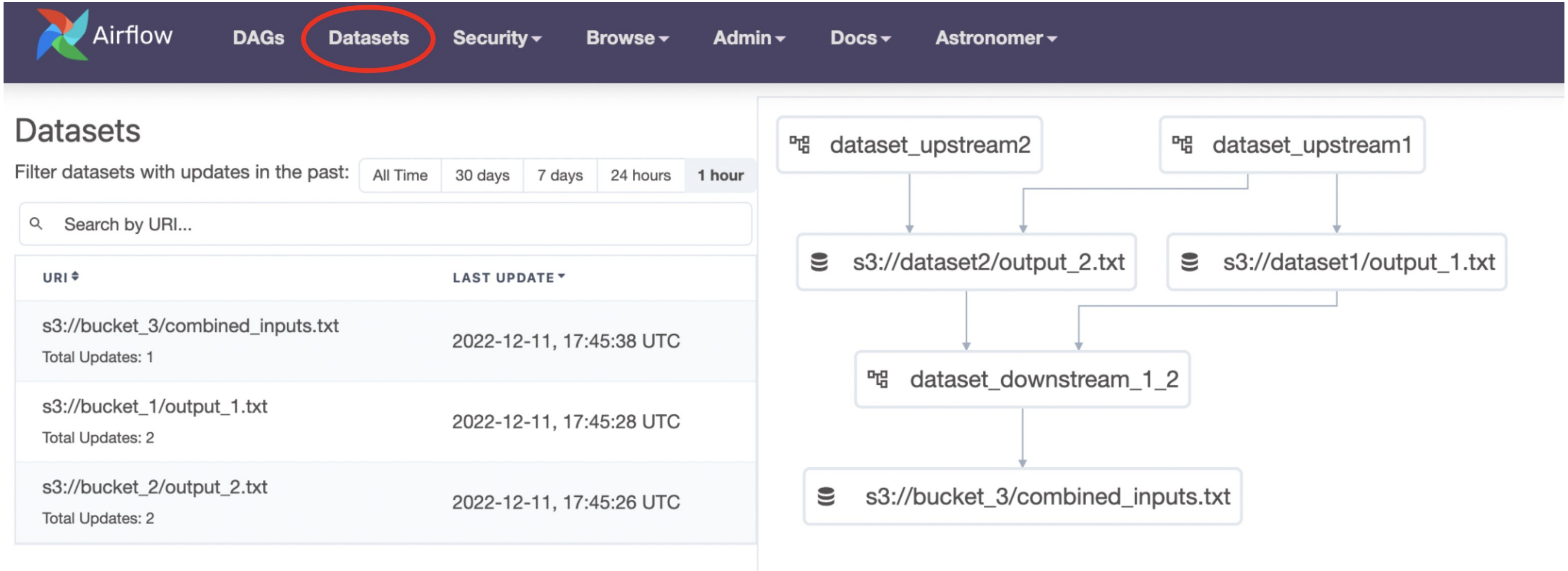
Tasks / Workloads

- Operators
  - Template for a predefined task like Bash, Python, Email or Database functionality
  - E.g. MsSqlOperator, S3FileTransformOperator…
- Sensors
  - Special operator type waiting for events like a new file occurs
- TaskFlow
  - Packaged custom python functions

Scheduling & Dependencies

- Time based schedules
- Sensor (event) based schedules
- Any kind of dependencies of tasks possible
- Grouping of tasks
- Branching
- Schedules on update of datasets

Datasets



Datasets

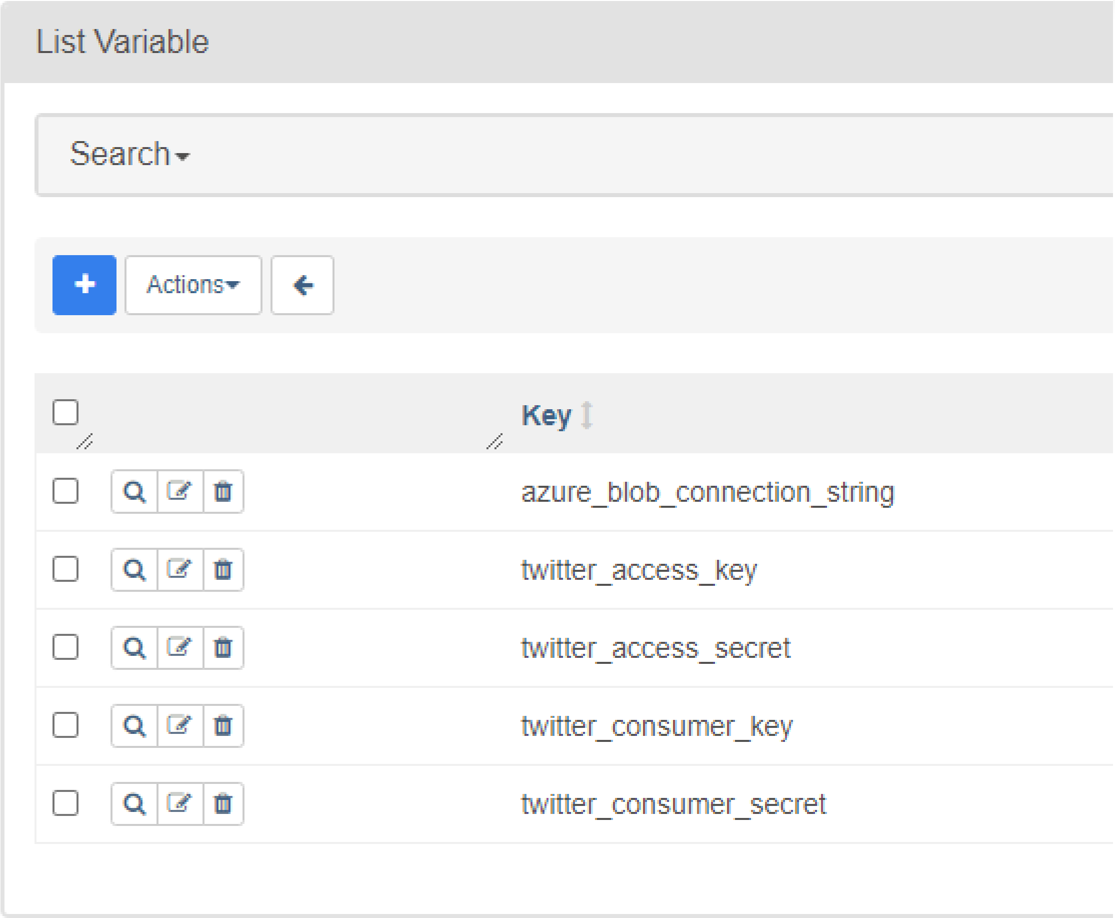
Filter datasets with updates in the past: All Time | 30 days | 7 days | 24 hours | **1 hour**

Search by URI...

URI	LAST UPDATE
s3://bucket_3/combined_inputs.txt Total Updates: 1	2022-12-11, 17:45:38 UTC
s3://bucket_1/output_1.txt Total Updates: 2	2022-12-11, 17:45:28 UTC
s3://bucket_2/output_2.txt Total Updates: 2	2022-12-11, 17:45:26 UTC

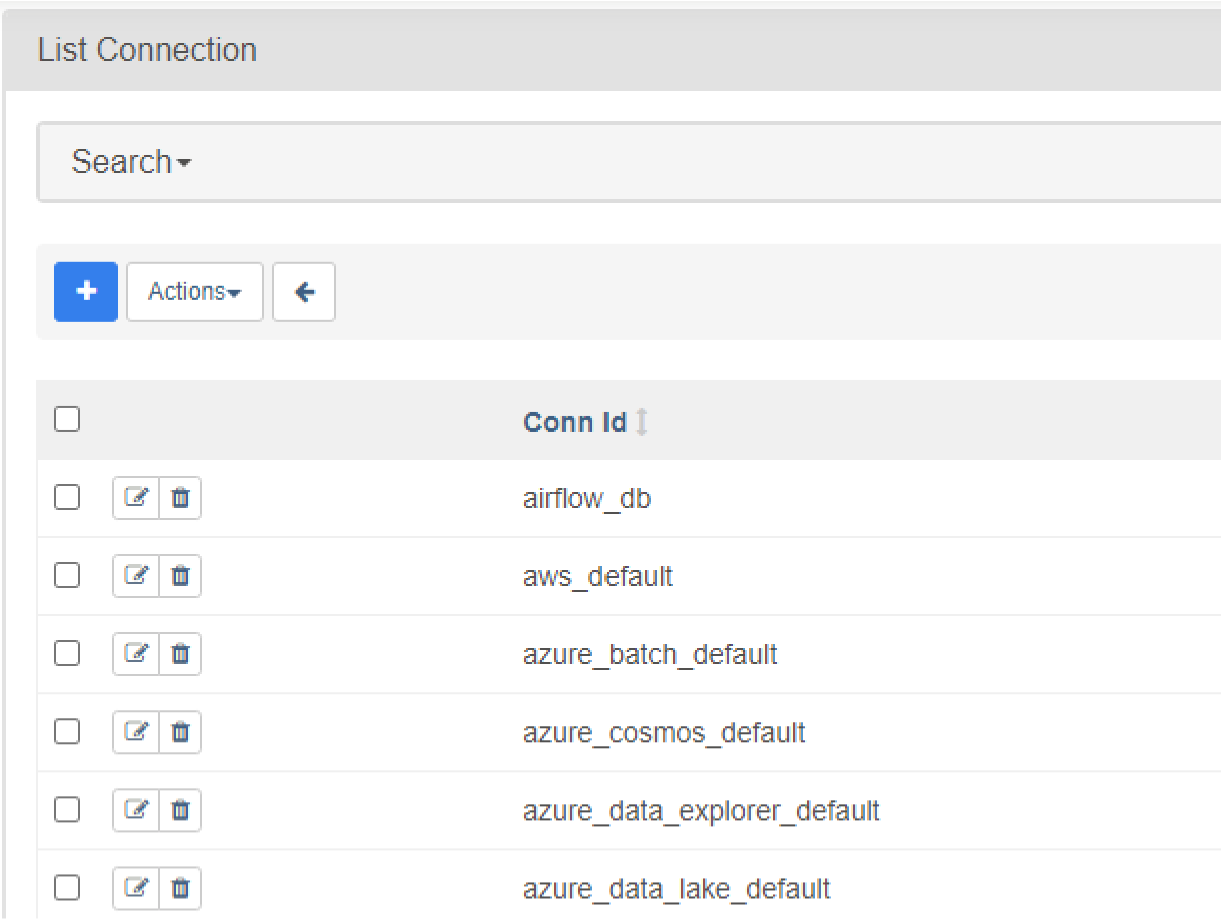
https://docs.astronomer.io/learn/airflow-datasets#:~:text=You%20can%20define%20datasets%20in,Uniform%20Resource%20Identifier%20(URI).

Secure Strings & Connections



List Variable

□		Key
□		azure_blob_connection_string
□		twitter_access_key
□		twitter_access_secret
□		twitter_consumer_key
□		twitter_consumer_secret



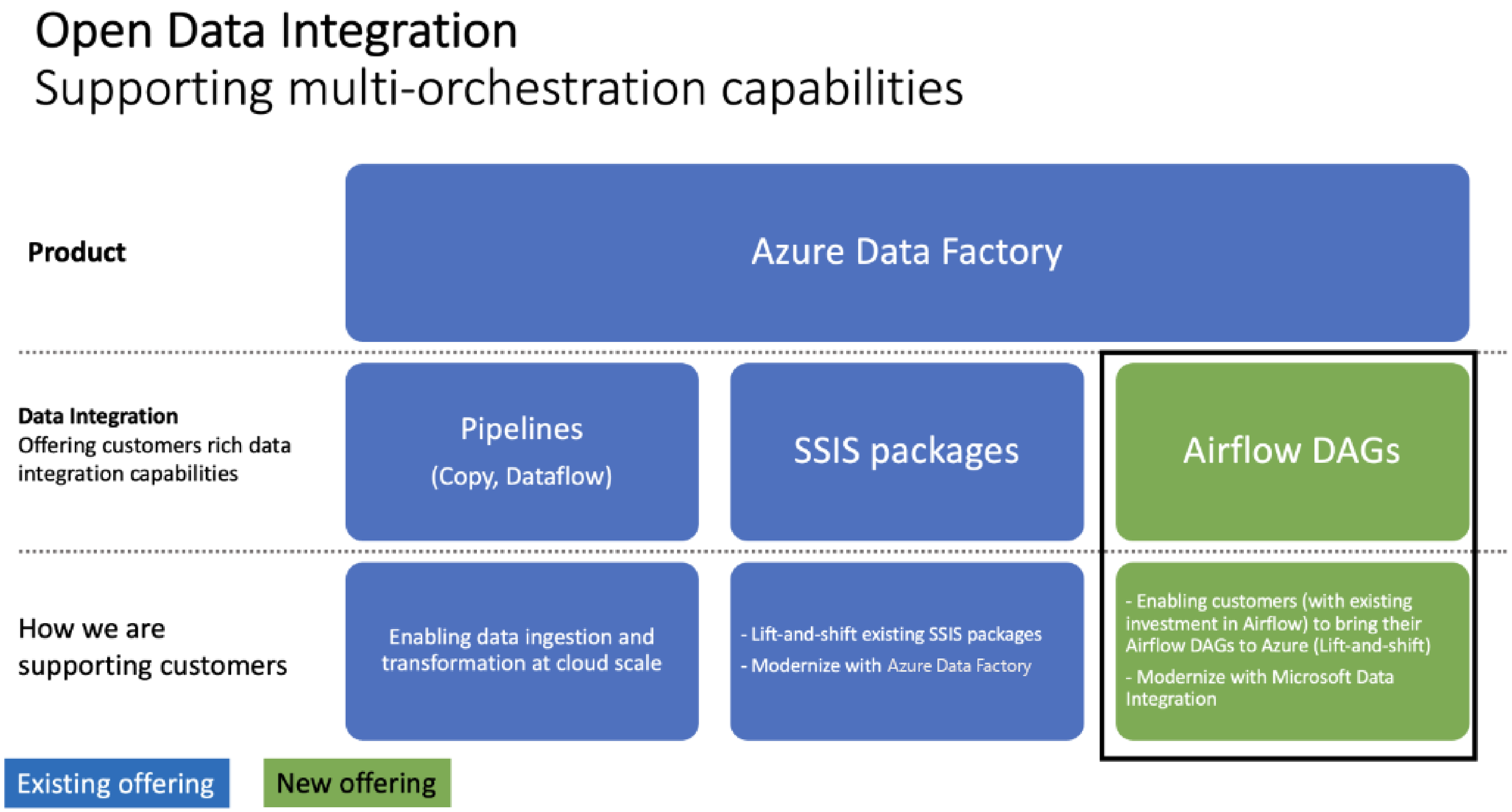
List Connection

□		Conn Id
□		airflow_db
□		aws_default
□		azure_batch_default
□		azure_cosmos_default
□		azure_data_explorer_default
□		azure_data_lake_default

Airflow in der Azure Data Factory

How does this fit in?

Available runtimes in Azure Data Factory



Open Data Integration
Supporting multi-orchestration capabilities

Product	Azure Data Factory		
Data Integration Offering customers rich data integration capabilities	Pipelines (Copy, Dataflow)	SSIS packages	Airflow DAGs
How we are supporting customers	Enabling data ingestion and transformation at cloud scale	- Lift-and-shift existing SSIS packages - Modernize with Azure Data Factory	- Enabling customers (with existing investment in Airflow) to bring their Airflow DAGs to Azure (Lift-and-shift) - Modernize with Microsoft Data Integration

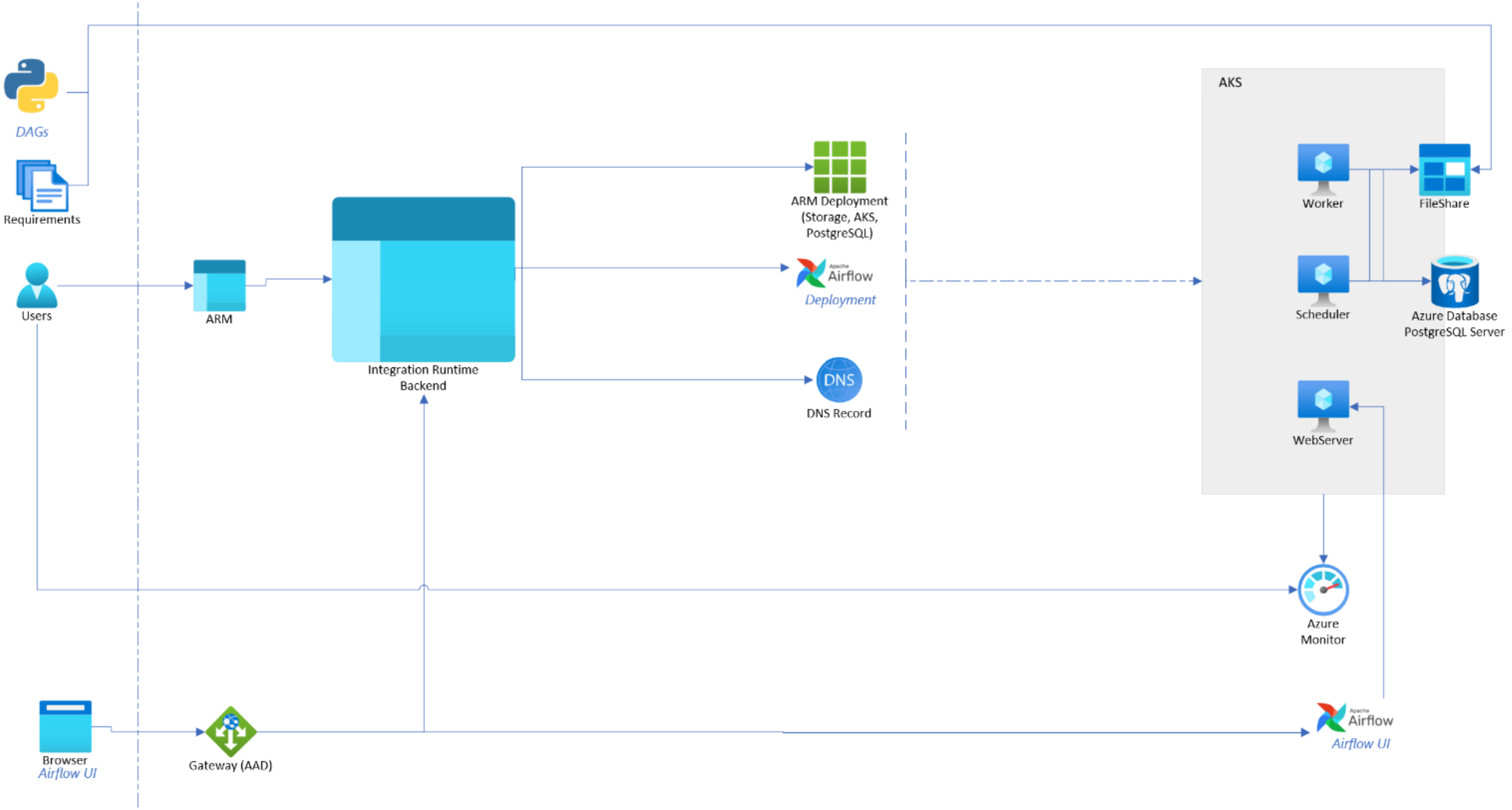
🟦 Existing offering 🟩 New offering

https://learn.microsoft.com/en-us/azure/data-factory/concept-managed-airflow

Features of Airflow in Azure Data Factory

- System managed for you by MS
- Fast and simple deployment (of airflow environment itself)
- Azure Active Directory integration
- Metadata encryption
- Azure Monitoring and alerting
- Managed Virtual Network integration (not yet)

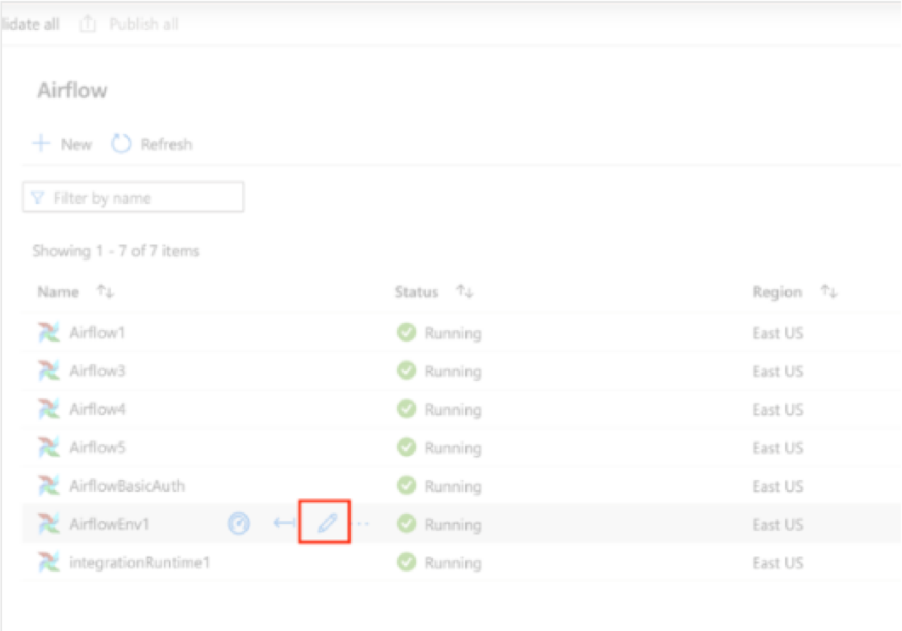
Systems architecture on Azure



https://learn.microsoft.com/en-us/azure/data-factory/concept-managed-airflow

Adding provider packages

„requirements.txt“ in ADF GUI



lidate all   Publish all

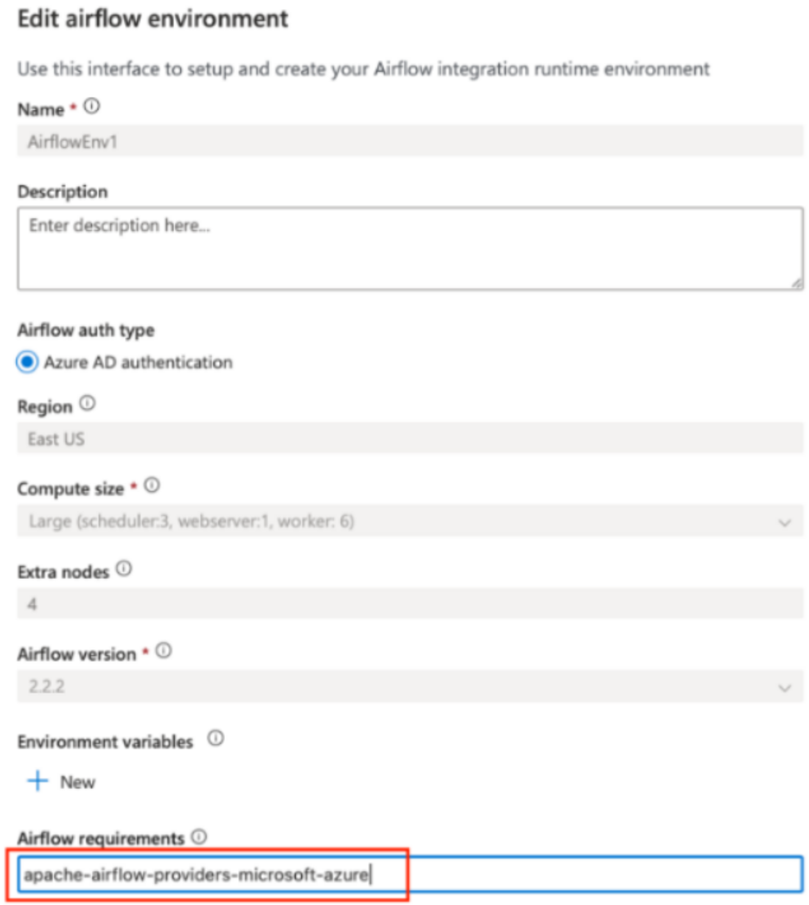
Airflow

+ New   Refresh

Filter by name

Showing 1 - 7 of 7 items

Name	Status	Region
Airflow1	Running	East US
Airflow3	Running	East US
Airflow4	Running	East US
Airflow5	Running	East US
AirflowBasicAuth	Running	East US
AirflowEnv1	Running	East US
integrationRuntime1	Running	East US



Edit airflow environment

Use this interface to setup and create your Airflow integration runtime environment

Name * : AirflowEnv1

Description: Enter description here...

Airflow auth type: ● Azure AD authentication

Region: East US

Compute size * : Large (scheduler:3, webserver:1, worker: 6)

Extra nodes: 4

Airflow version * : 2.2.2

Environment variables: + New

Airflow requirements: apache-airflow-providers-microsoft-azure

https://learn.microsoft.com/en-us/azure/data-factory/concept-managed-airflow

Application Lifecycle Management

- Manual:
  - Development with VS.Code etc.
  - Deployment upload to Azure Blob Store
  - Import to filesystem of airflow runtime (GUI only yet)
  - Refresh airflow GUI and check for errors
  - Difficult to automate deployment at the moment
- Git Hub Repo Sync (new)
  - Enable git sync feature while creating IR
  - Files will be updated
  - Or use Rest API for airflow in ADF
  - Should work w/o local dev environment (?)

https://learn.microsoft.com/en-us/azure/data-factory/airflow-sync-github-repository

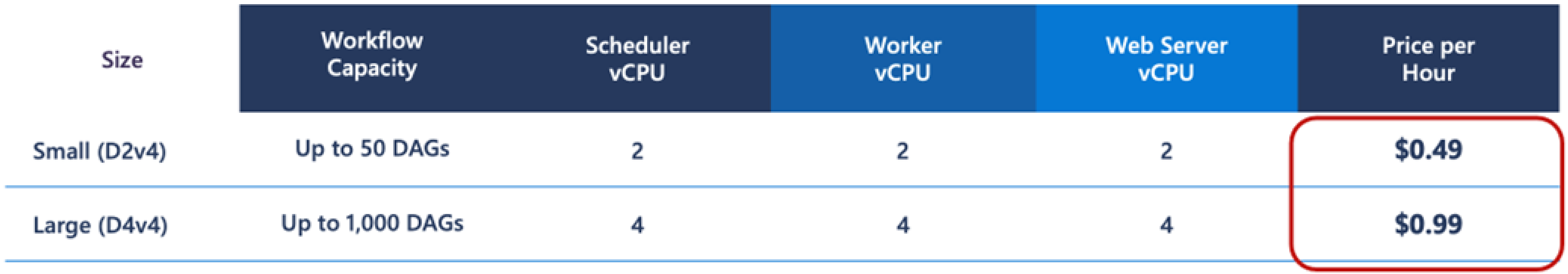
https://learn.microsoft.com/en-us/azure/data-factory/rest-apis-for-airflow-integrated-runtime

Limitations (10 / 2023)

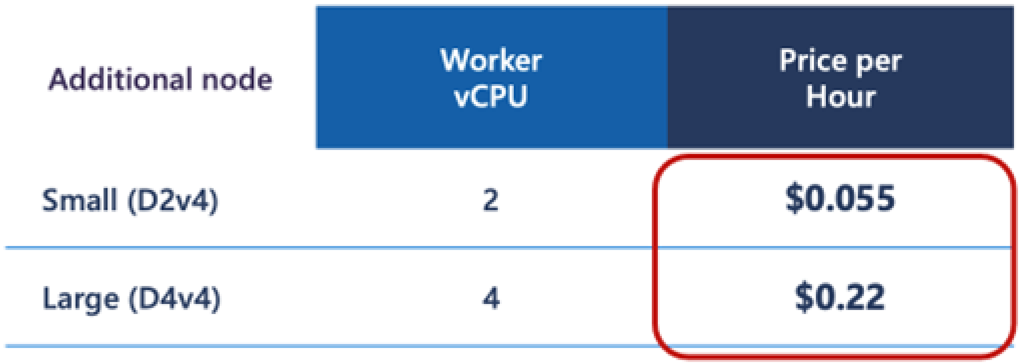
- Limited access to work folders
- Only limited regions supported
- Limited connectivity to on-premises data sources
  - (no SHIR support, VPN necessary for on-premises sources)
- Limitations for Blob Storage and ADLS support
  - Blob Storage behind Vnet not yet supported
  - Firewall whitelisting in other services possible using private IP address for airflow runtime

https://learn.microsoft.com/en-us/azure/data-factory/concept-managed-airflow

Costs (12/2023)



Size	Workflow Capacity	Scheduler vCPU	Worker vCPU	Web Server vCPU	Price per Hour
Small (D2v4)	Up to 50 DAGs	2	2	2	$0.49
Large (D4v4)	Up to 1,000 DAGs	4	4	4	$0.99



Additional node	Worker vCPU	Price per Hour
Small (D2v4)	2	$0.055
Large (D4v4)	4	$0.22

Small

per Day: ~12$ Month: ~353$

Large

per Day: ~24$ Month: ~730$

Instances **cannot** not be **paused**. Auto-scaling not available, yet.

https://learn.microsoft.com/en-us/azure/data-factory/airflow-pricing

Demo

Airflow in Azure Data Factory

Airflow vs.
ADF GUI

Comparing functionality of
Airflow and pipelines / triggers

Airflow vs. Azure Data Factory pipelines

Airflow in ADF	ADF pipelines & triggers
Focus on scheduling & orchestration	Focus on ETL/ELT
Python knowledge needed	Easy to learn and use
Github sync as new feature	built-in Git and CI/CD support
Billed by run time of environment – not pausable	Cost-effective – pay what you need
Many plug-and-play operators – adaptable	Built-in connectivity to 90 data sources
	Fast and secure gateway to on-prem data sources
No such certificates	Compliance: HIPAA, GDPR, ISO 27001, others
Scale out using external infrastructure (Databricks etc), auto-scale of airflow instance anounced for GA	Scale out included (copy, mapping data flows) and optional external infrastructure

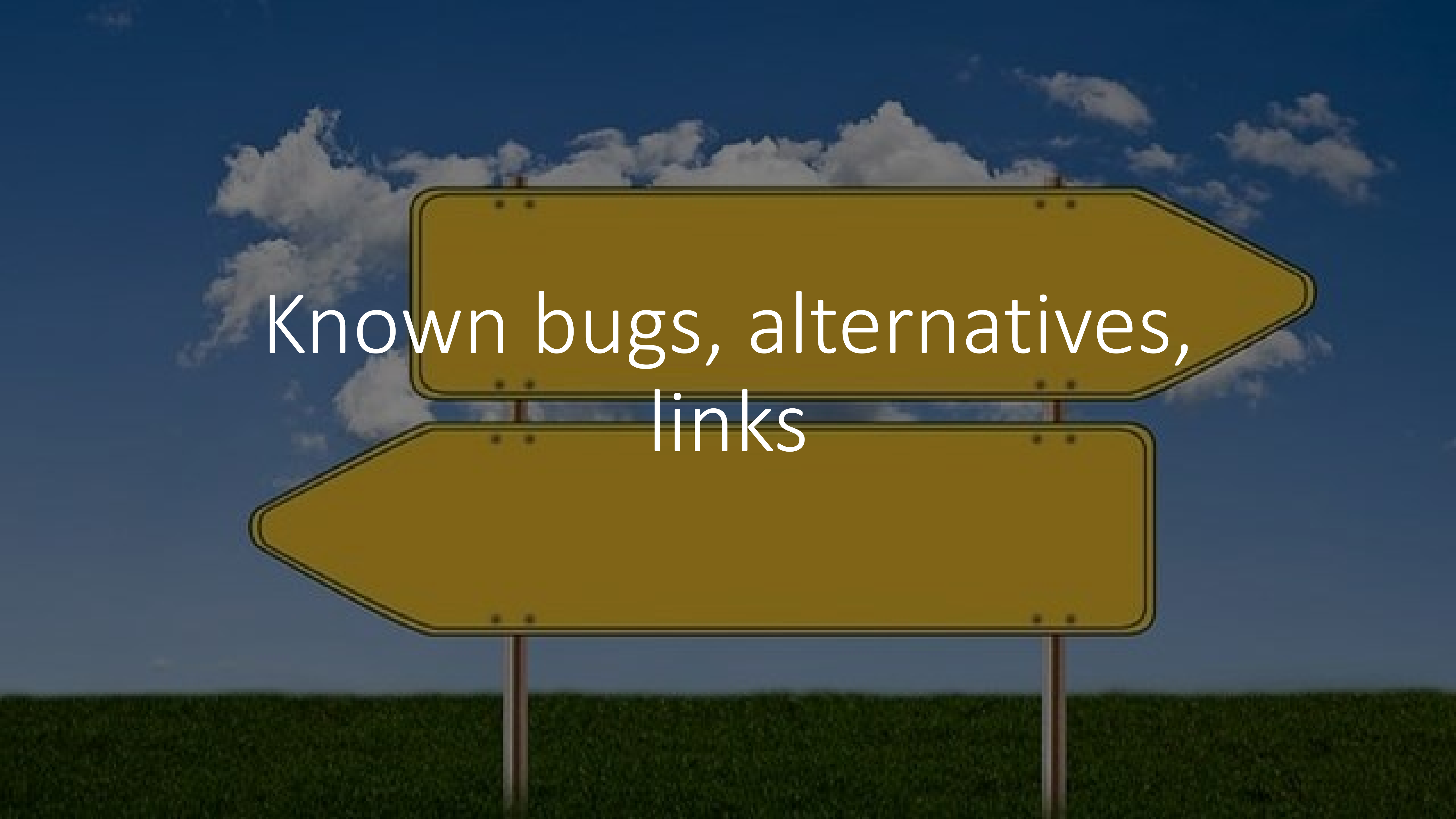
Use cases Airflow in ADF

Why should I use airflow in Azure Data Factory?

Best fit use cases for Airflow in ADF

- Lift & shift from existing on-premises environments
- Experienced team working completely in Python
  - Single stack
  - Data Science / Machine Learning projects
- No sufficient operations skills to run Airflow by yourself
- Very complicated dependencies of many processes
- Lots of testing of the scheduling and dependencies itself is necessary
- Using Airflow for enterprise orchestration which also calls ADF pipelines

Known bugs, alternatives, links



Bugs & fails of current state

- No full access to managed resources reduces usability
- Publishing DAGs manually is cumbersome (give github Sync a try)
- Deletion of DAG only works in ADF GUI, not works in airflow
- Auto-Refresh not working in airflow GUI
- Connections – error when creating connections
- MS Documentation insufficient
- Airflow runtime cannot be paused
- ...but development ongoing (gitsync, Azure Key Vault Integration...)

Alternatives to run airflow on Azure

- Using managed service from Astronomer on your Azure Tenant https://www.astronomer.io/
- Announced at Ignite 2023: Native Apache Airflow Service from Astronomer on Azure (Preview) – derzeit kostenlos nutzbar
- Run as Docker Images on Kubernetes cluster
- Run as VMs on Azure

Links which helped me to get in

- Microsofts Docs Airflow in ADF
  - https://learn.microsoft.com/en-us/azure/data-factory/concept-managed-airflow
- Apache Airflow Docs
  - https://airflow.apache.org/docs/apache-airflow/2.4.3/
- Microsoft Reactor
  - https://www.youtube.com/live/DLBY8xfhIsQ?feature=share
- Airflow 101 Turtorial
  - https://www.youtube.com/watch?v=4_Ifm4PNRyg&list=PLY-S9rU4aY6Y39lTqY6WVN-eph3QvzWw4&index=1
- Brian Cafferky – Reasons for not using Airflow
  - https://www.youtube.com/watch?v=YQ056EKzCyw&list=PL7_h0bRfL52pygj88FC1laf9F1q7FWnZM

Questions?

Thanks for your attention, I appreciate your feedback!