

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

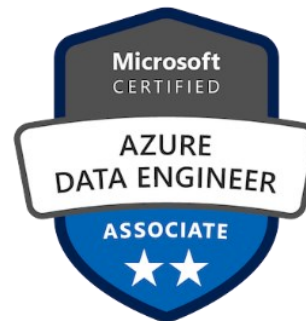
Introducing Apache Airflow Integration in ADF



# 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?

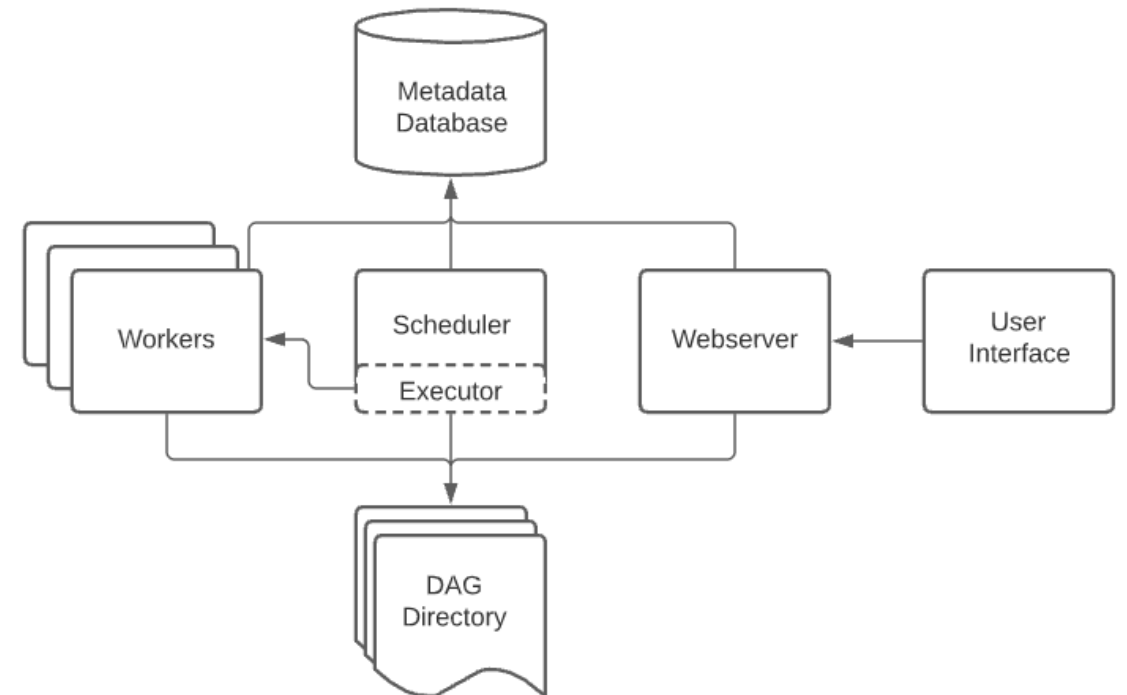
- Platform for authoring, scheduling & monitoring workflows
- Open-Source
- Widely used – 16 Mio downloads / month
- Code-centric – using Python language
- Execute nearly anything in any environment (using Python)
- Extensible by own operators, executors and libraries
- Scalable

# What is Airflow **not**?

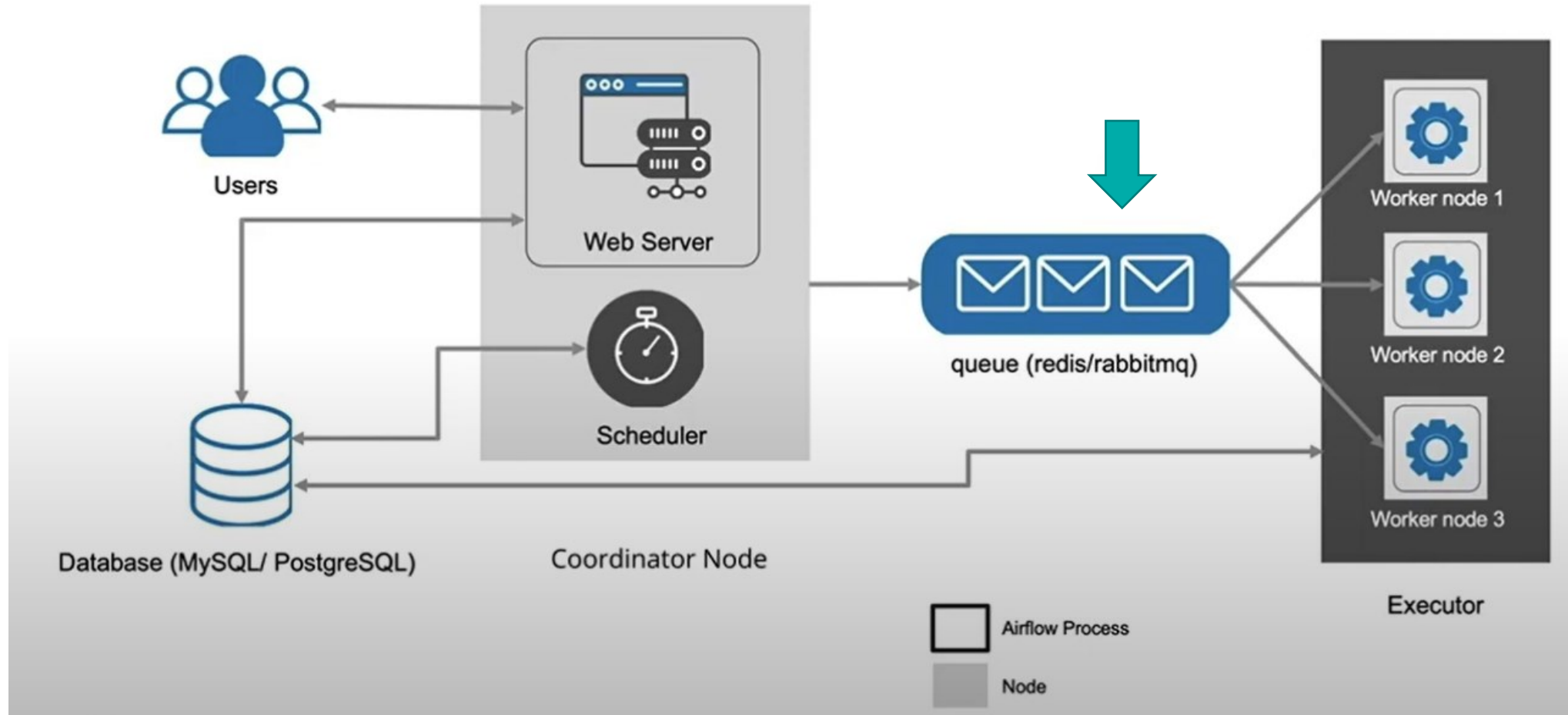
- No GUI-based ETL Tool
- No special support for ETL functionality
  - like parallel threaded bulk load, schema or surrogate key creation
- Not very lightweight (# of components to operate)

# Airflow systems architecture overview

- **Scheduler:** triggering scheduled workflows, and submitting tasks 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

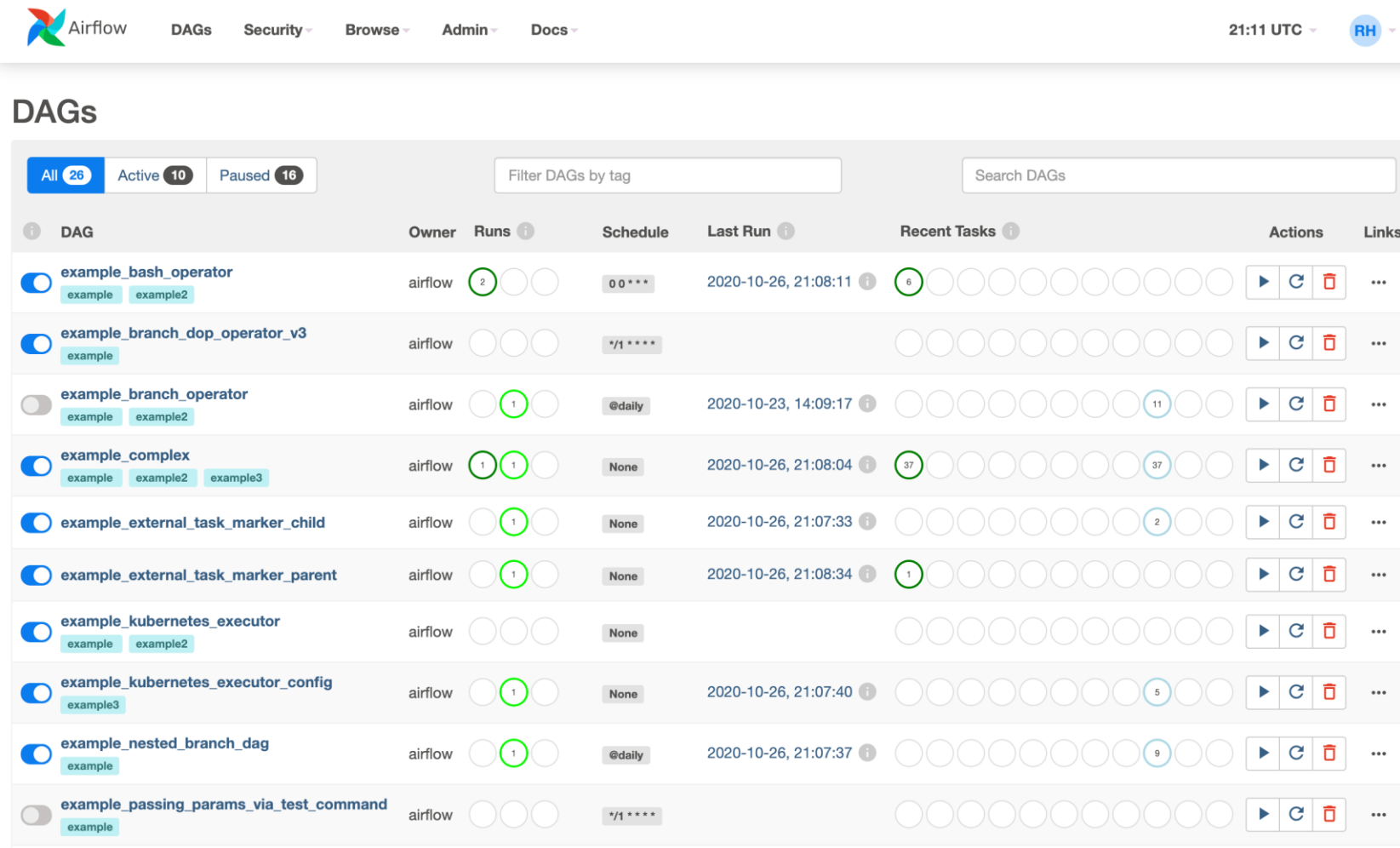


# Airflow systems architecture detailed sample





# How does this fancy user interface look?



The screenshot shows the Apache Airflow web interface. At the top, there is a navigation bar with the Airflow logo and menu items: DAGs, Security, Browse, Admin, and Docs. On the right, it displays the current time as 21:11 UTC and the user as RH.

The main section is titled "DAGs" and features a filter bar with buttons for "All 26", "Active 10", and "Paused 16". There are also input fields for "Filter DAGs by tag" and "Search DAGs".

Below the filter bar is a table listing various DAGs. Each row includes a toggle switch, the DAG name, its owner, a "Runs" column with a progress indicator, the "Schedule", the "Last Run" time, a "Recent Tasks" column with a progress indicator, and "Actions" and "Links" columns.

DAG	Owner	Runs	Schedule	Last Run	Recent Tasks	Actions	Links
<input checked="" type="checkbox"/> example_bash_operator <small>example example2</small>	airflow	2	0 0 ***	2020-10-26, 21:08:11	6	[▶] [↺] [🗑️]	...
<input checked="" type="checkbox"/> example_branch_dop_operator_v3 <small>example</small>	airflow		* / 1 * * * *			[▶] [↺] [🗑️]	...
<input type="checkbox"/> example_branch_operator <small>example example2</small>	airflow	1	@daily	2020-10-23, 14:09:17	11	[▶] [↺] [🗑️]	...
<input checked="" type="checkbox"/> example_complex <small>example example2 example3</small>	airflow	1 1	None	2020-10-26, 21:08:04	37	[▶] [↺] [🗑️]	...
<input checked="" type="checkbox"/> example_external_task_marker_child	airflow	1	None	2020-10-26, 21:07:33	2	[▶] [↺] [🗑️]	...
<input checked="" type="checkbox"/> example_external_task_marker_parent	airflow	1	None	2020-10-26, 21:08:34	1	[▶] [↺] [🗑️]	...
<input checked="" type="checkbox"/> example_kubernetes_executor <small>example example2</small>	airflow		None			[▶] [↺] [🗑️]	...
<input checked="" type="checkbox"/> example_kubernetes_executor_config <small>example3</small>	airflow	1	None	2020-10-26, 21:07:40	5	[▶] [↺] [🗑️]	...
<input checked="" type="checkbox"/> example_nested_branch_dag <small>example</small>	airflow	1	@daily	2020-10-26, 21:07:37	9	[▶] [↺] [🗑️]	...
<input type="checkbox"/> example_passing_params_via_test_command <small>example</small>	airflow		* / 1 * * * *			[▶] [↺] [🗑️]	...

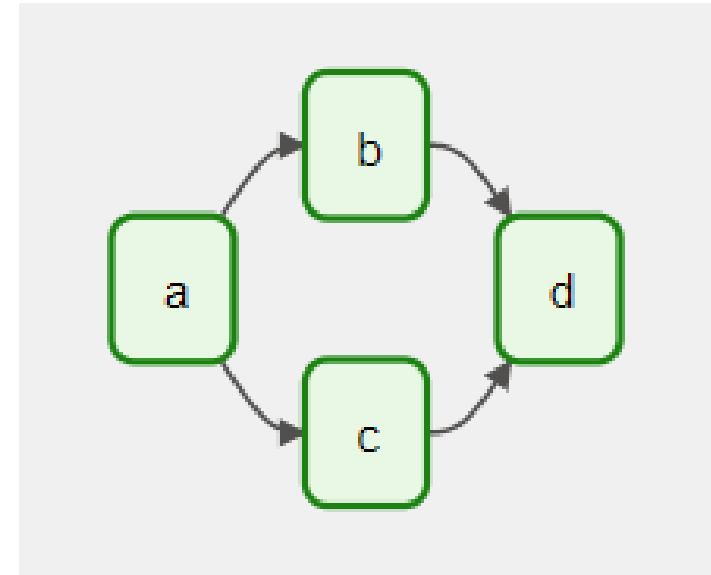


# Airflow concepts

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

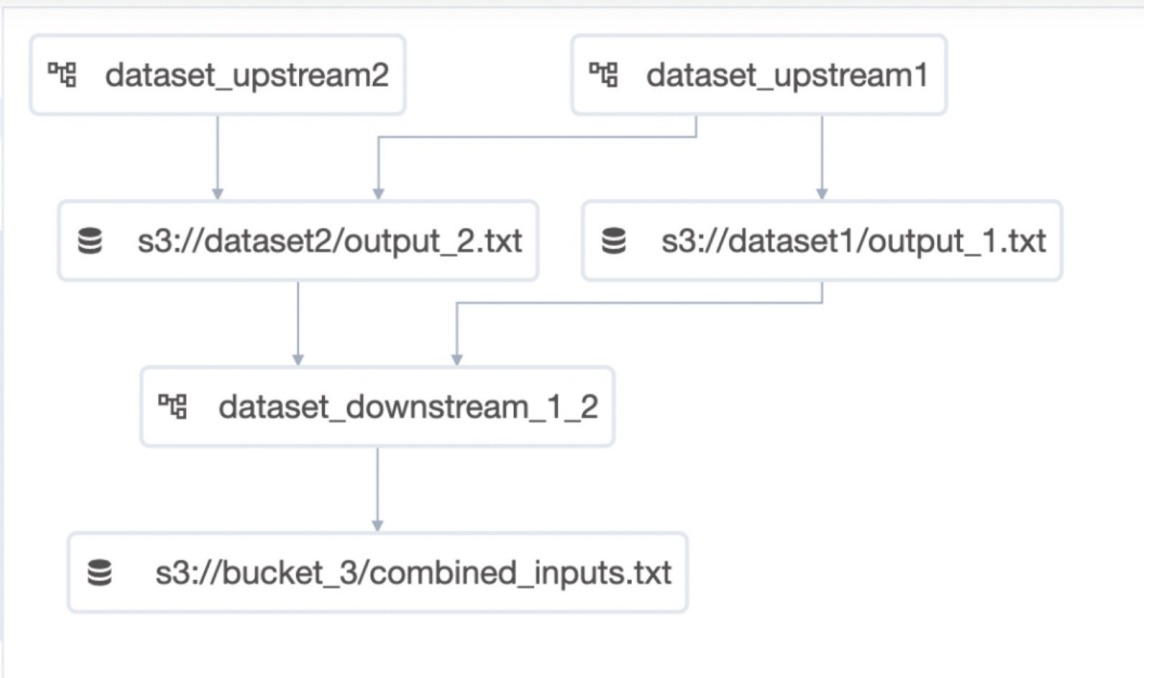
Airflow **DAGs** **Datasets** Security ▾ Browse ▾ Admin ▾ Docs ▾ Astronomer ▾

## 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



# Secure Strings & Connections

List Variable

Search ▾

+ Actions ▾ ←

<input type="checkbox"/>	Key ↑
<input type="checkbox"/>	azure_blob_connection_string
<input type="checkbox"/>	twitter_access_key
<input type="checkbox"/>	twitter_access_secret
<input type="checkbox"/>	twitter_consumer_key
<input type="checkbox"/>	twitter_consumer_secret

List Connection

Search ▾

+ Actions ▾ ←

<input type="checkbox"/>	Conn Id ↑
<input type="checkbox"/>	airflow_db
<input type="checkbox"/>	aws_default
<input type="checkbox"/>	azure_batch_default
<input type="checkbox"/>	azure_cosmos_default
<input type="checkbox"/>	azure_data_explorer_default
<input type="checkbox"/>	azure_data_lake_default



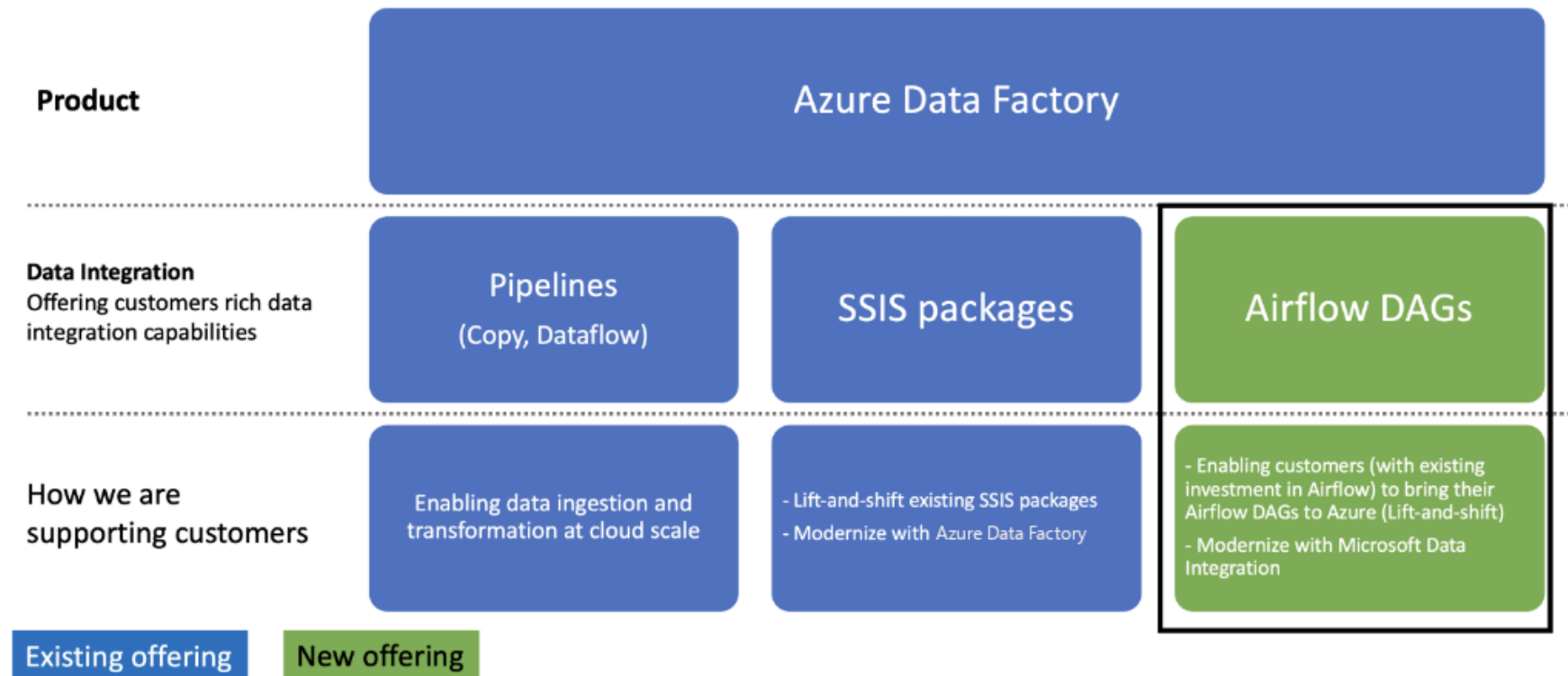
# Airflow in der Azure Data Factory

How does this fit in?



# Available runtimes in Azure Data Factory

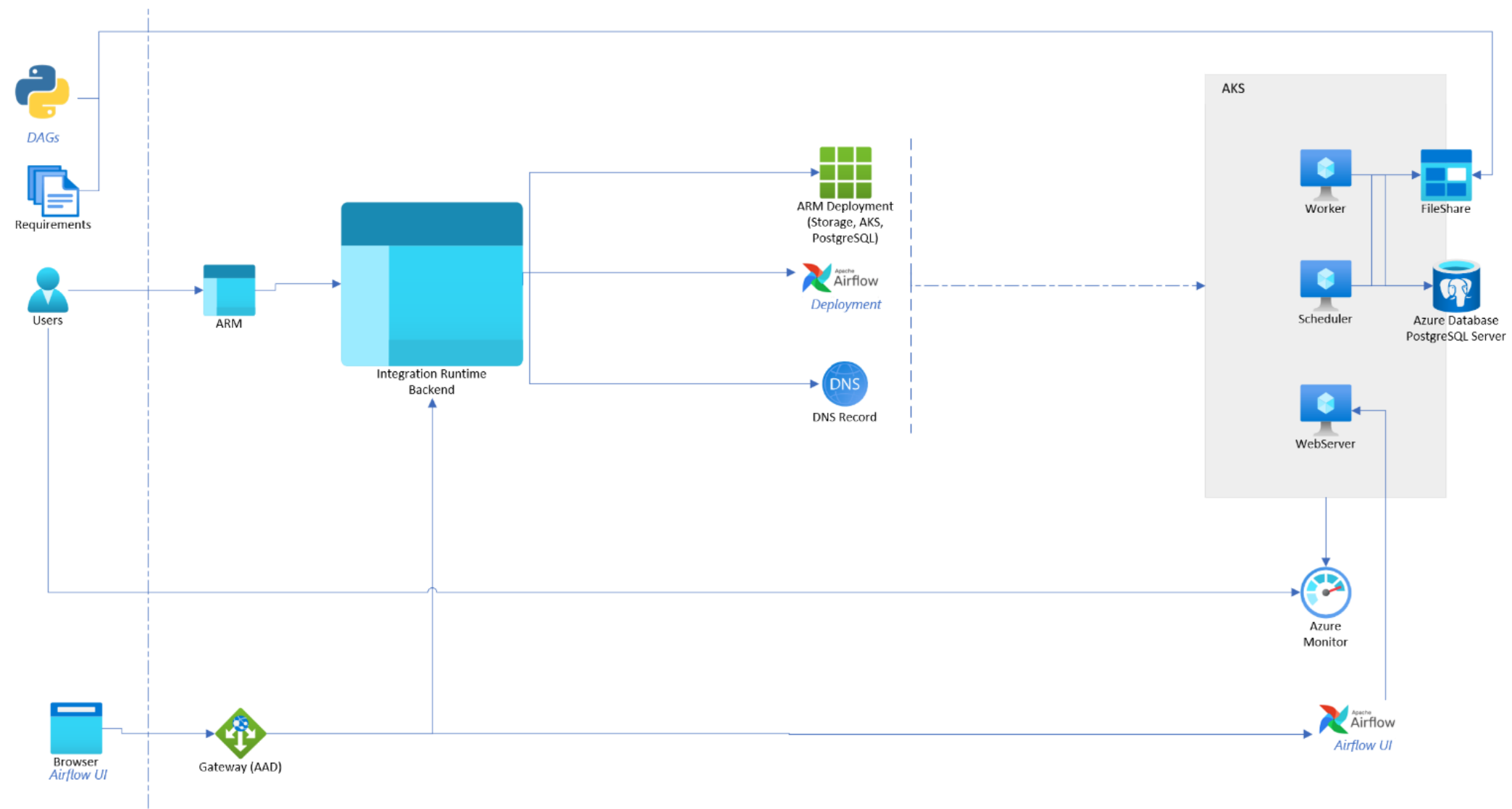
Open Data Integration  
Supporting multi-orchestration capabilities



# 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



# Adding provider packages

„requirements.txt“ in ADF GUI

Update 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

# Application Lifecycle Management

- Development with VS.Code etc.
- Deployment upload to Azure Blob Store
- Import to filesystem of airflow runtime (GUI only yet)
- Not overwrite – delete and new (but keeping metadata)
- Refresh airflow GUI and check for errors
- Difficult to automate deployment at the moment

# Limitations (06 / 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

# Costs (Spring 2023)

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

Additional node	Worker vCPU	Price per Hour
Small (D2v4)	2	<b>\$0.055</b>
Large (D4v4)	4	<b>\$0.22</b>

Small  
per Day: ~12\$ Month: ~353\$

Large  
per Day: ~24\$ Month: ~730\$

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



A group of people in a meeting, with one man pointing towards a screen. The man has curly brown hair and is wearing a dark blue jacket. He is pointing with his right hand towards the left side of the frame. In the background, there are three other people: a man with a beard and short dark hair, a man with short brown hair, and a woman with long blonde hair and glasses. They are all looking towards the left. The background is a blurred office setting with a white door and a wall with vertical slats.

# 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
Application lifecycle management not good (06/23)	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 announced 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

The image features two yellow directional signs mounted on a single metal post. The top sign is a right-pointing arrow, and the bottom sign is a left-pointing arrow. The background is a clear blue sky with scattered white clouds, and a green field is visible at the bottom. The text "Known bugs, alternatives, links" is overlaid in white on the top sign.

Known bugs, alternatives,  
links

# Bugs & fails of current state

- No full access to managed resources reduces usability
- Publishing DAGs is cumbersome and manual
- Deletion of DAG only works in ADF GUI, does not work in airflow part of GUI
- Auto-Refresh not working in airflow GUI
- Connections – error when creating connections
- MS Documentation insufficient
- Airflow runtime cannot be paused
- Using Airflow version 2.4.3 from November 22 (2.6.1 is current)

# Alternatives to run airflow on Azure

- Using managed service from Astronomer on your Azure Tenant  
<https://www.astronomer.io/>
- 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/DLBY8xfhlsQ?feature=share>
- Airflow 101 Turtorial
  - [https://www.youtube.com/watch?v=4\\_lfm4PNRyg&list=PLY-S9rU4aY6Y39ITqY6WVN-eph3QvzWw4&index=1](https://www.youtube.com/watch?v=4_lfm4PNRyg&list=PLY-S9rU4aY6Y39ITqY6WVN-eph3QvzWw4&index=1)
- Brian Cafferky – Reasons for not using Airflow
  - [https://www.youtube.com/watch?v=YQ056EKzCyw&list=PL7\\_h0bRfL52pygj88FC1laf9F1q7FWnZM](https://www.youtube.com/watch?v=YQ056EKzCyw&list=PL7_h0bRfL52pygj88FC1laf9F1q7FWnZM)



Questions?