DEVOPS
W PROJEKCIE DATA

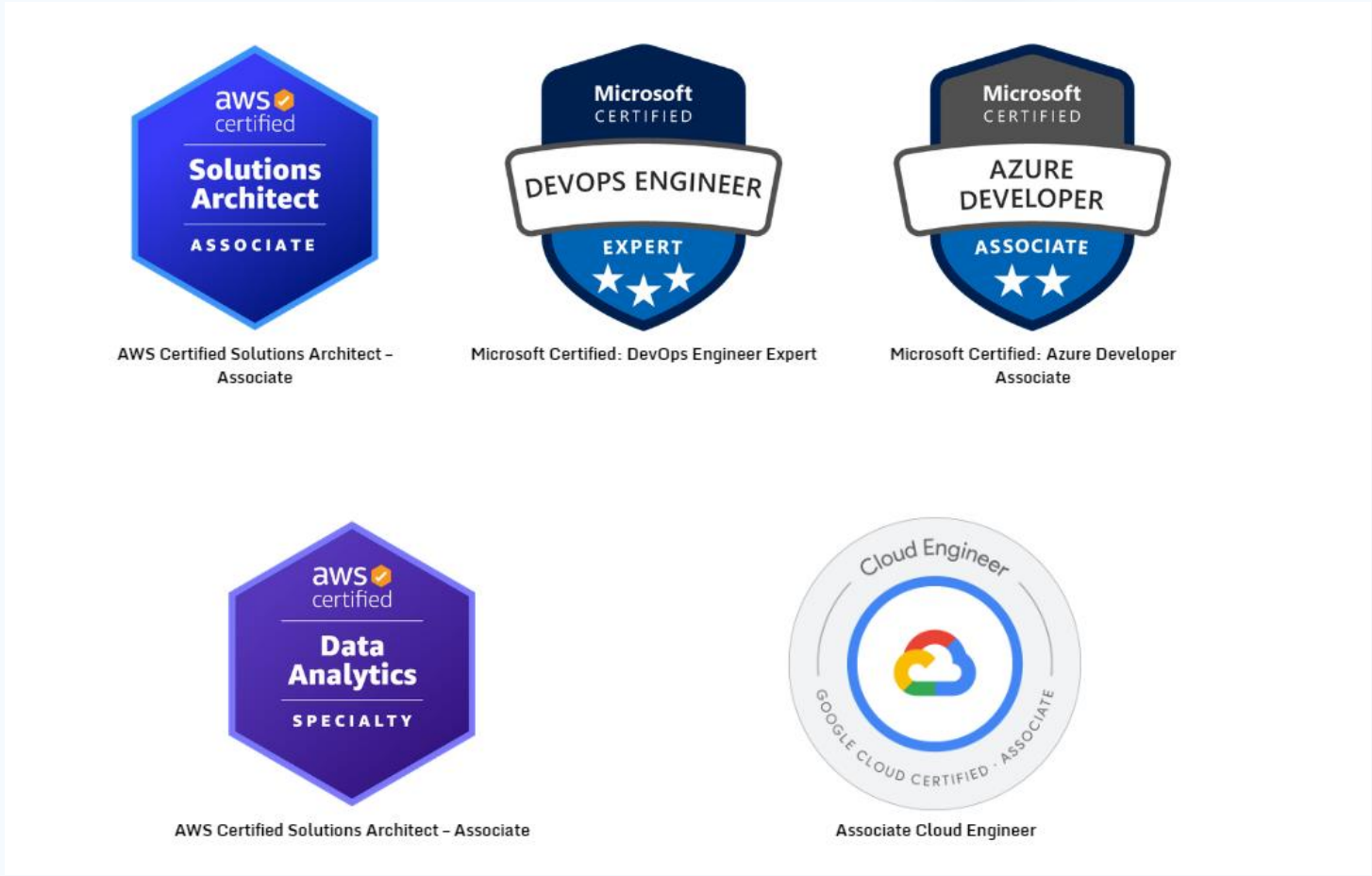
Kasper Kalfas

Cloud Architect

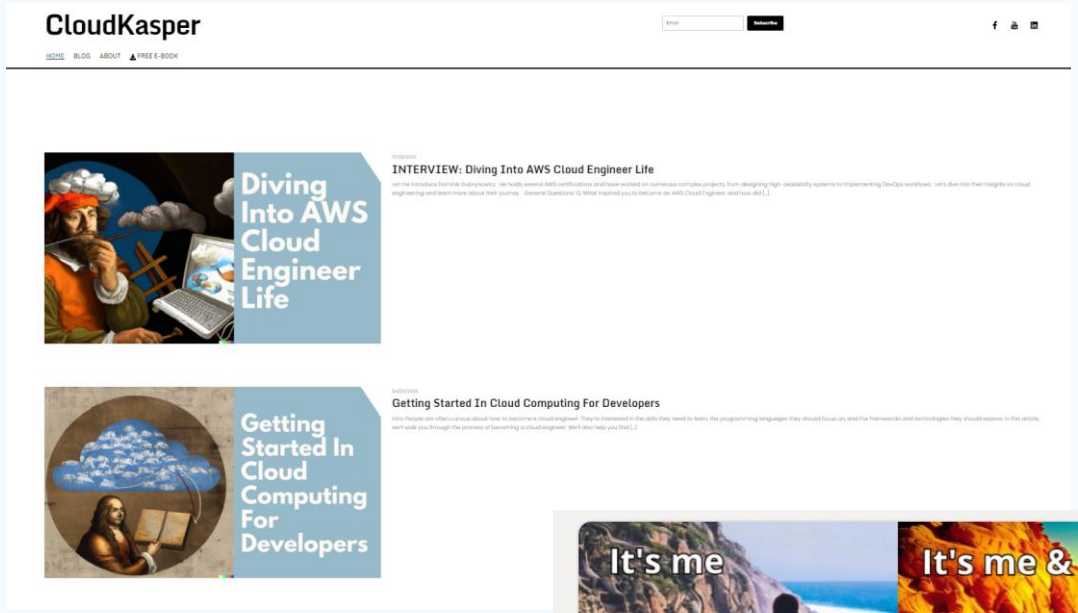
e: cloudkasperpro@gmail.com

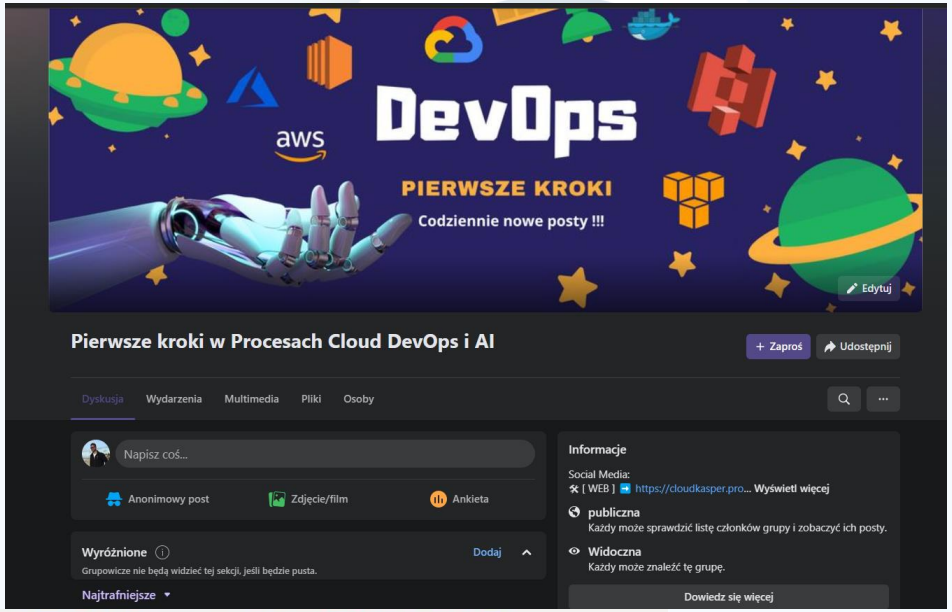
Chmurowisko

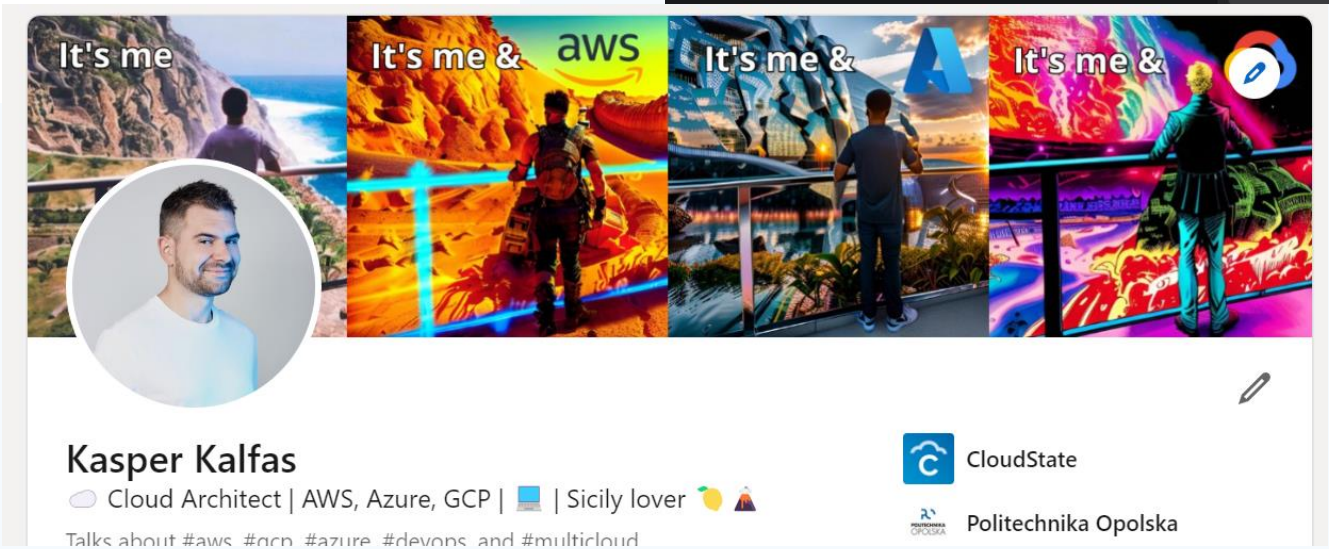
Moje ordery


aws certified
Solutions Architect
ASSOCIATE
AWS Certified Solutions Architect – Associate
Microsoft CERTIFIED
DEVOPS ENGINEER
EXPERT
Microsoft Certified: DevOps Engineer Expert
Microsoft CERTIFIED
AZURE DEVELOPER
ASSOCIATE
Microsoft Certified: Azure Developer Associate
aws certified
Data Analytics
SPECIALTY
AWS Certified Solutions Architect – Associate
Cloud Engineer
GOOGLE CLOUD CERTIFIED · ASSOCIATE
Associate Cloud Engineer

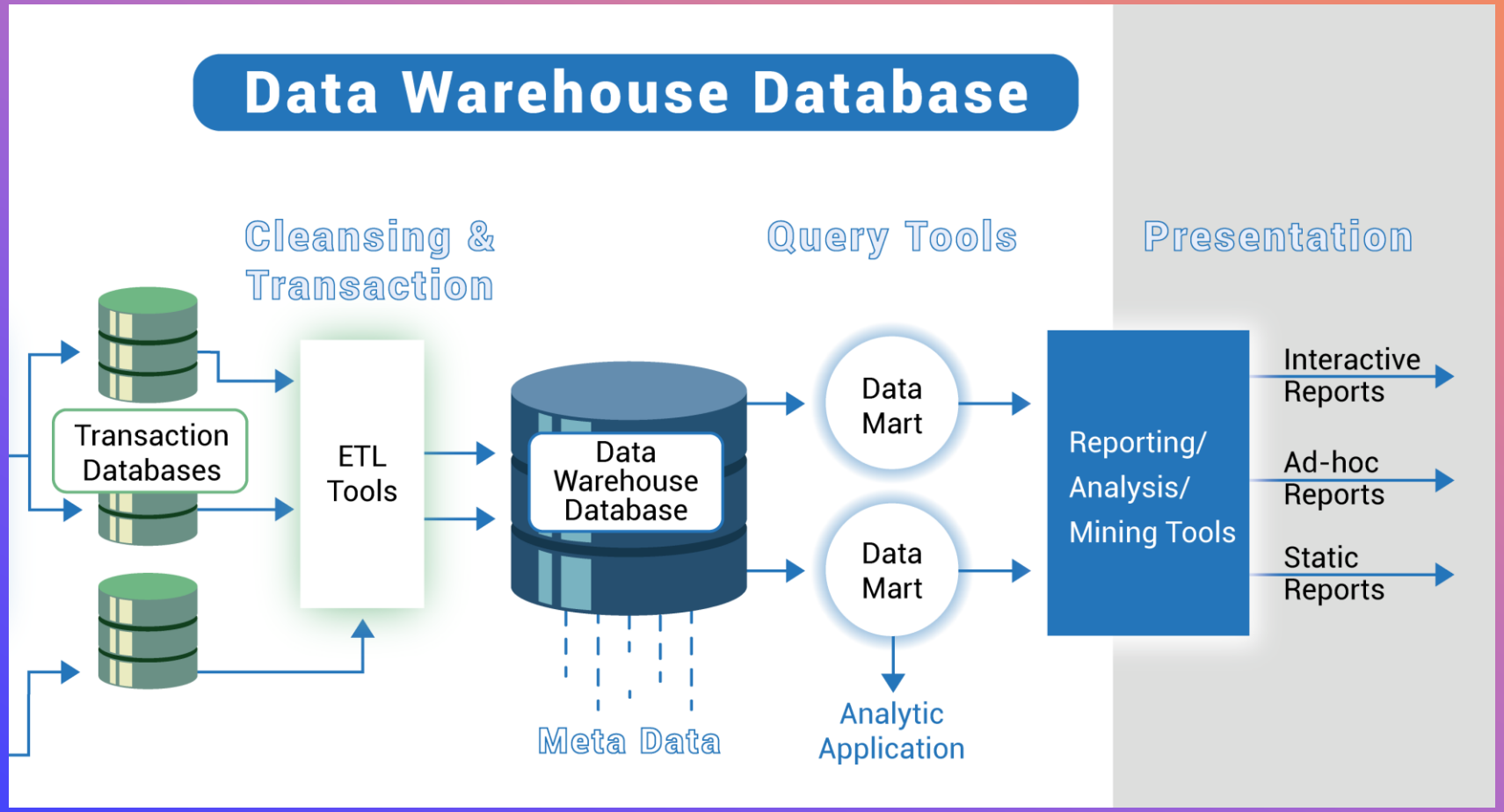
CloudKasper.pro



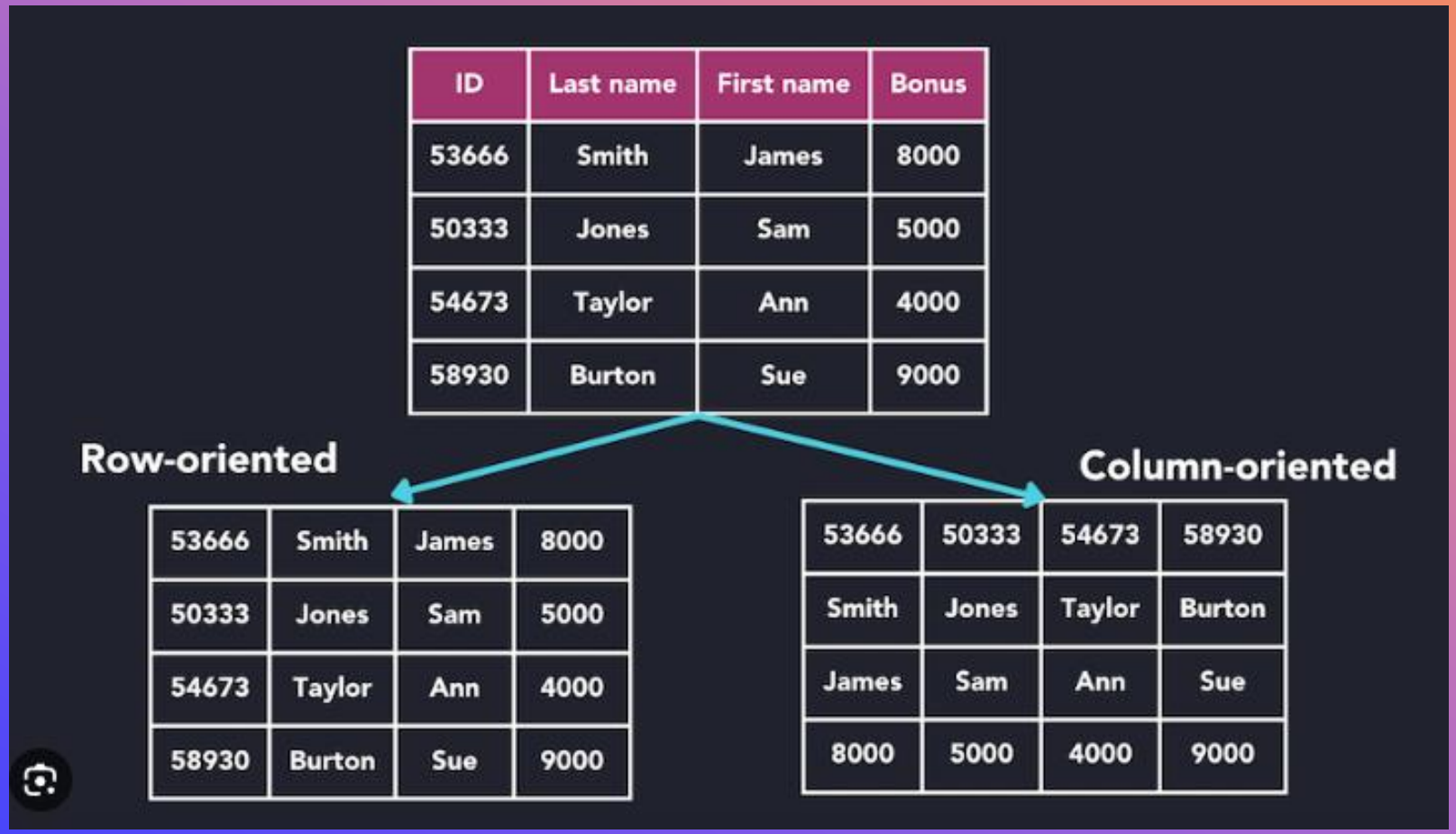




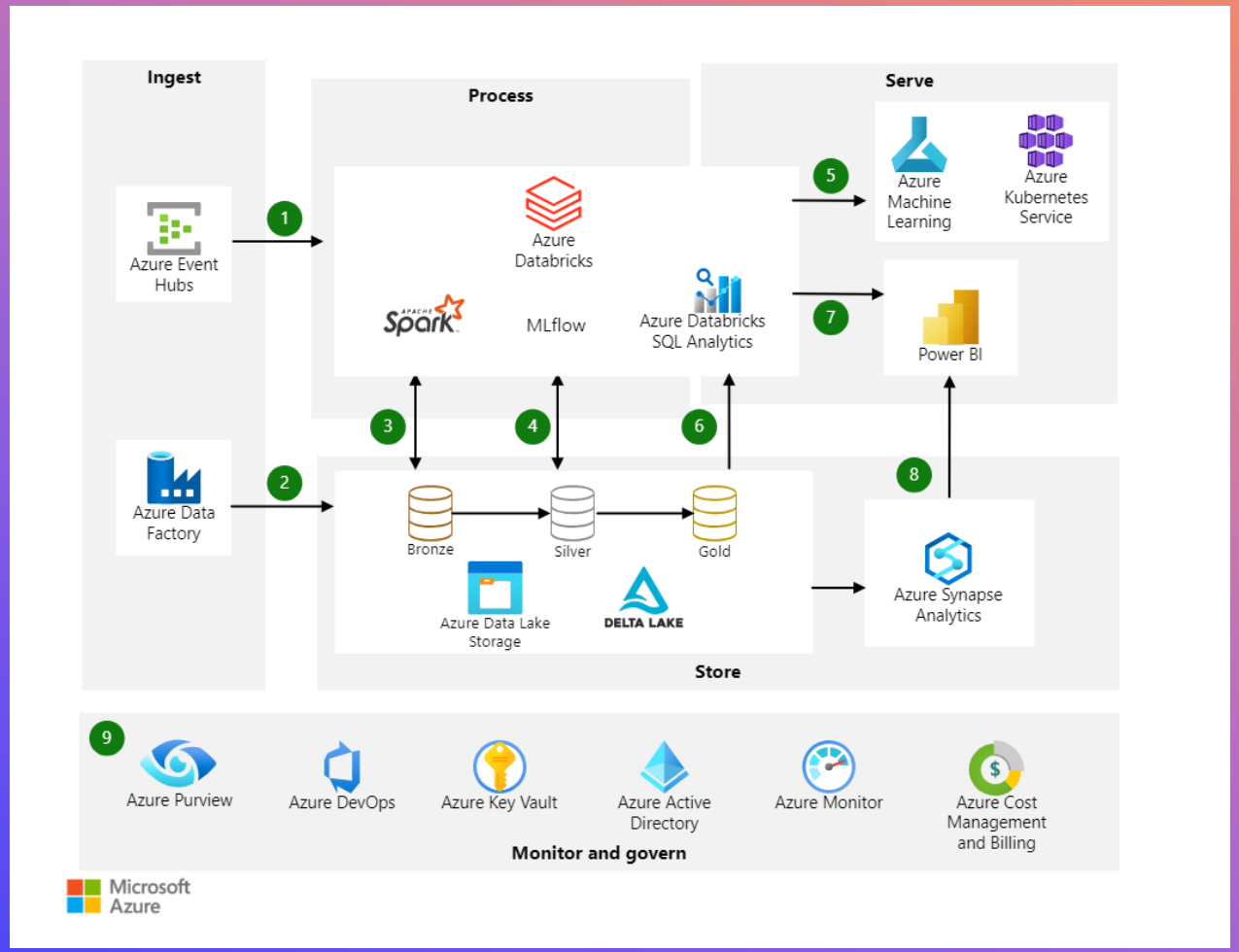
DATA PROJECT



ZORIENTOWANIE DANYCH



DATA PROJECT IN CLOUD



Przetrzymywanie danych

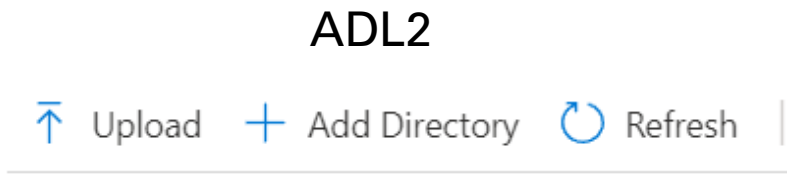
Azure Blob storage - $21.00/msc

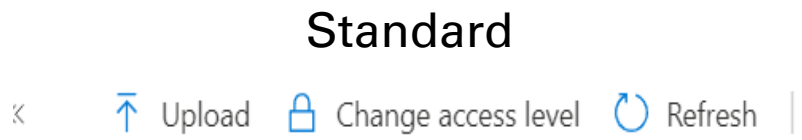
Azure SQL $400.19/msc

Azure DB VM ok. 500$/msc

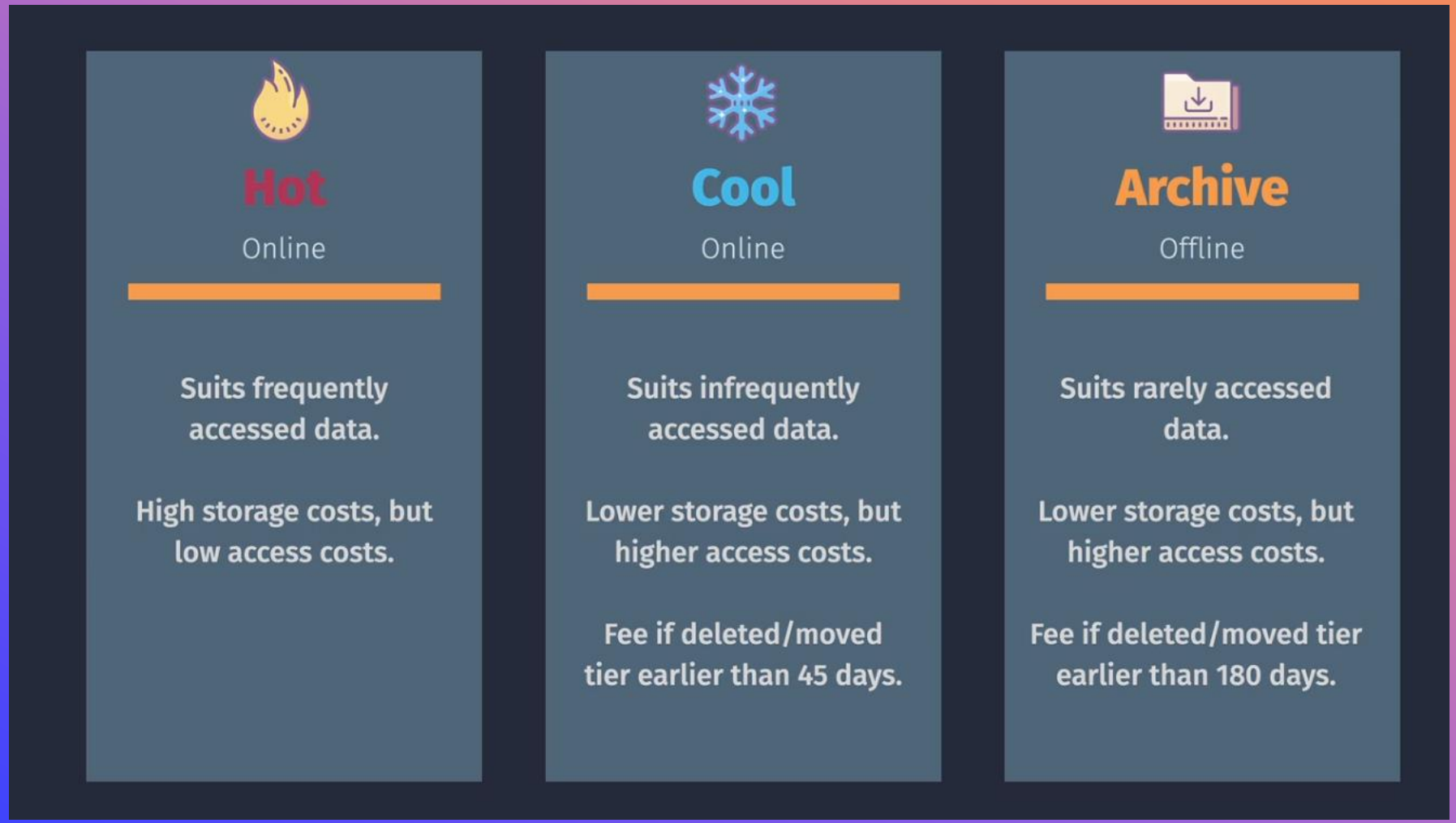
Azure Blob Storage

- **Definition**: Azure Blob Storage is a scalable, cloud-based object storage solution for unstructured data. It efficiently handles data of any type or size, from small documents to large media files.
- **Use Cases**:
  - Ideal for storing large media files, backup and restore, archive, data analytics, and much more.
- **Integration**:
  - Seamlessly integrates with Azure services and applications.

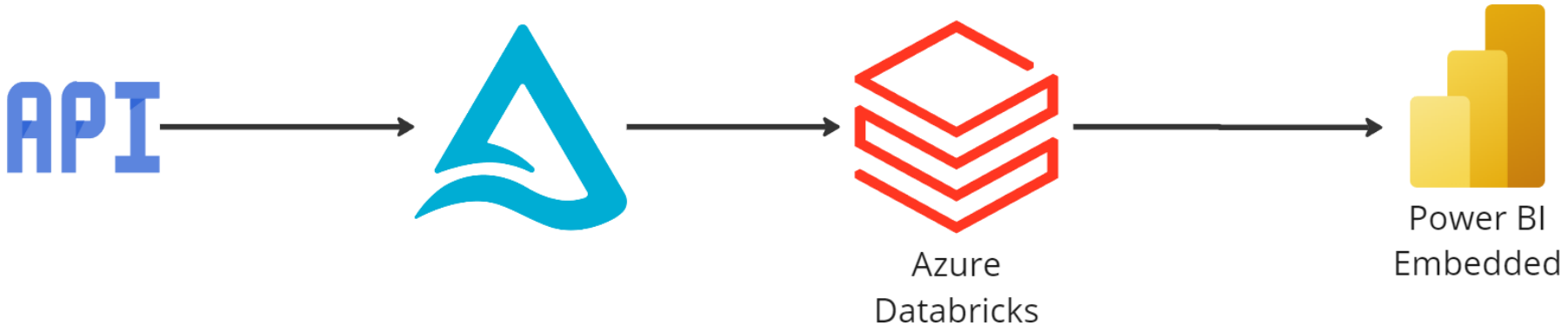




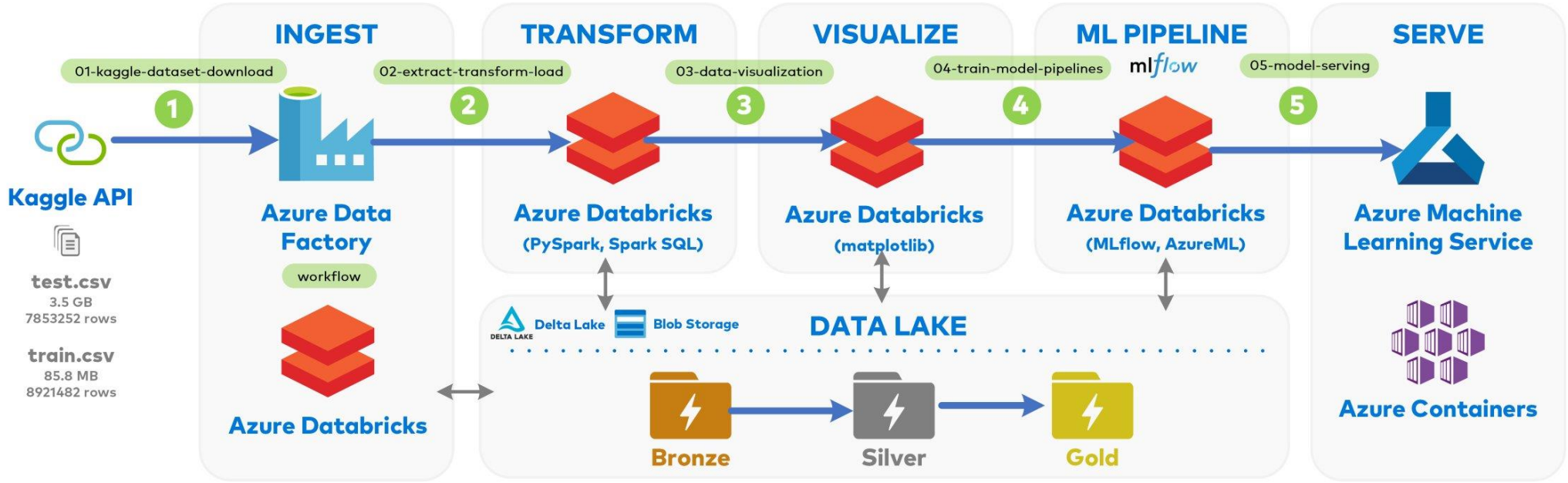
ACCESS TIERS



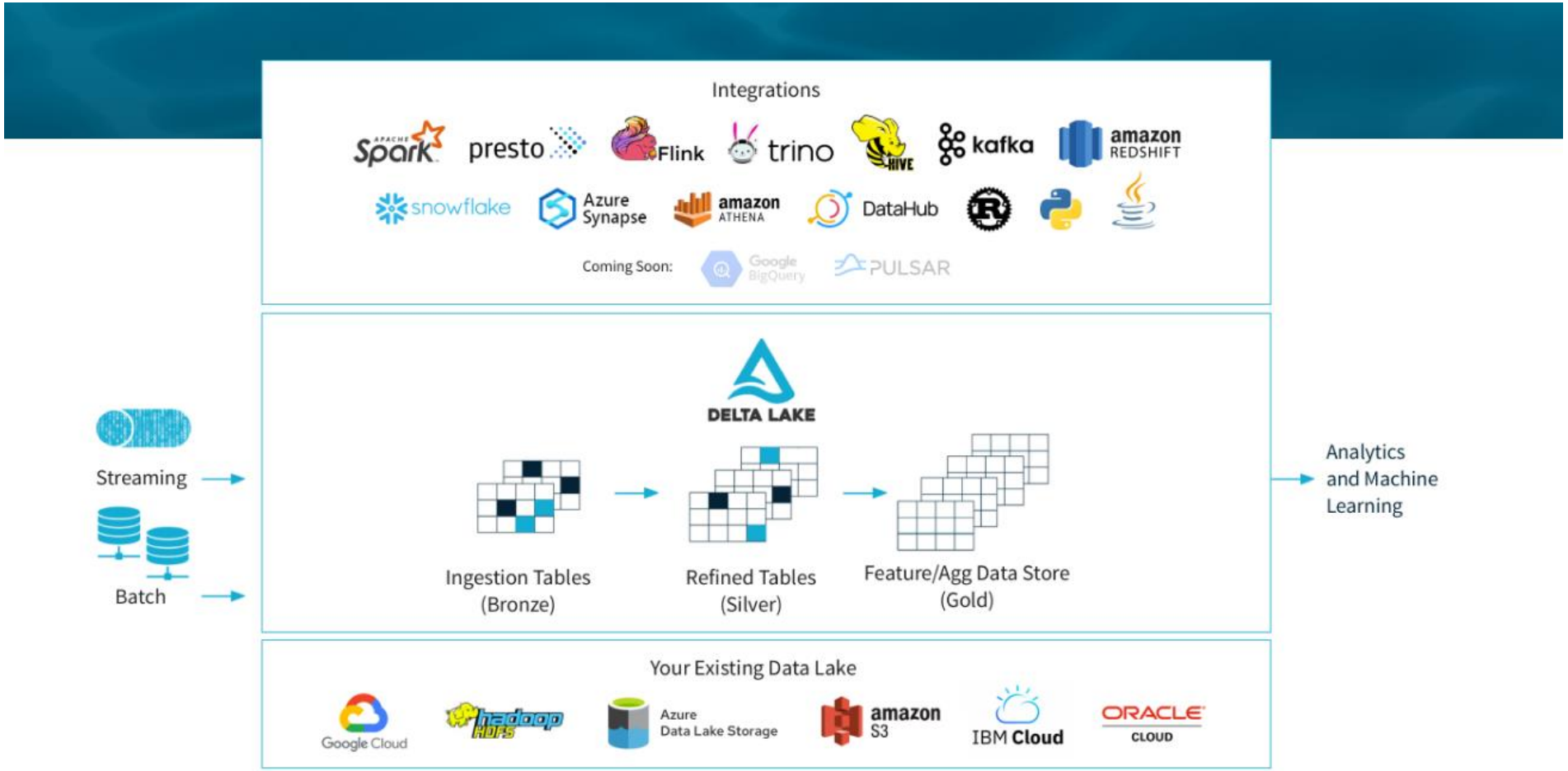
Delta Lake



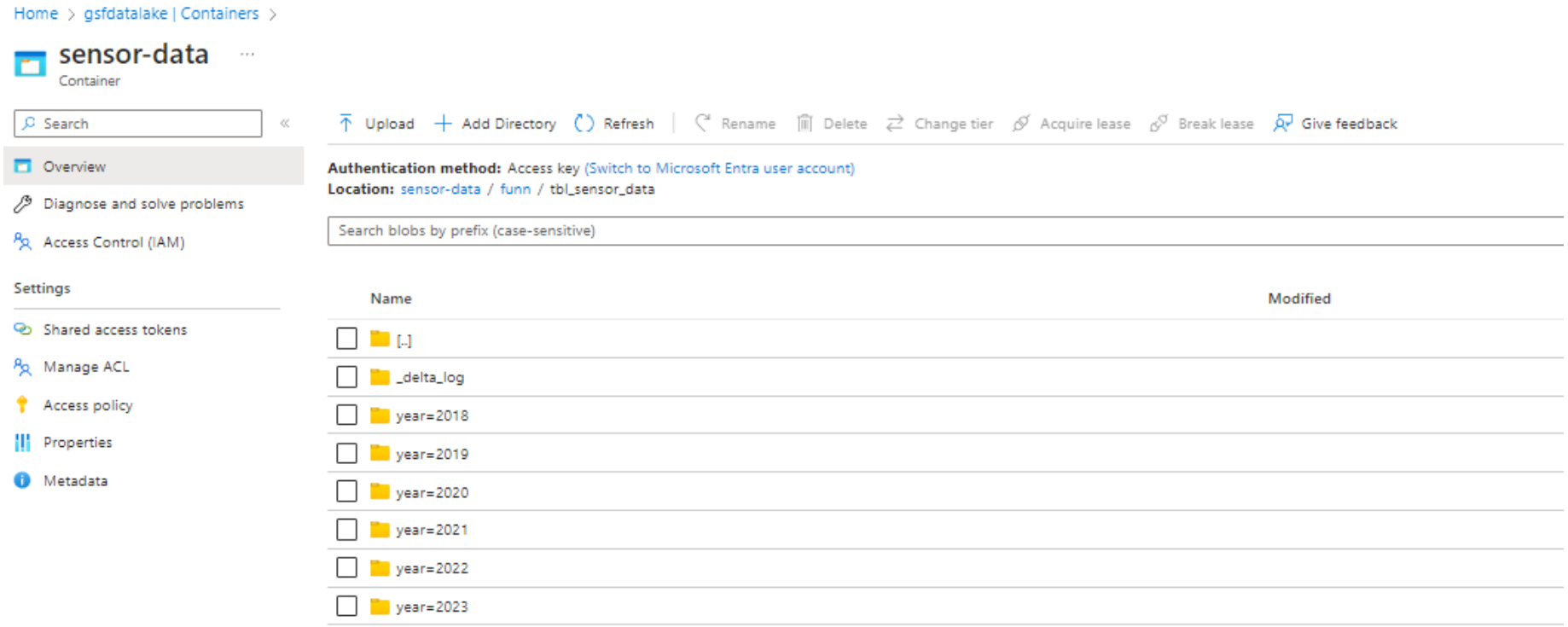
Architecture



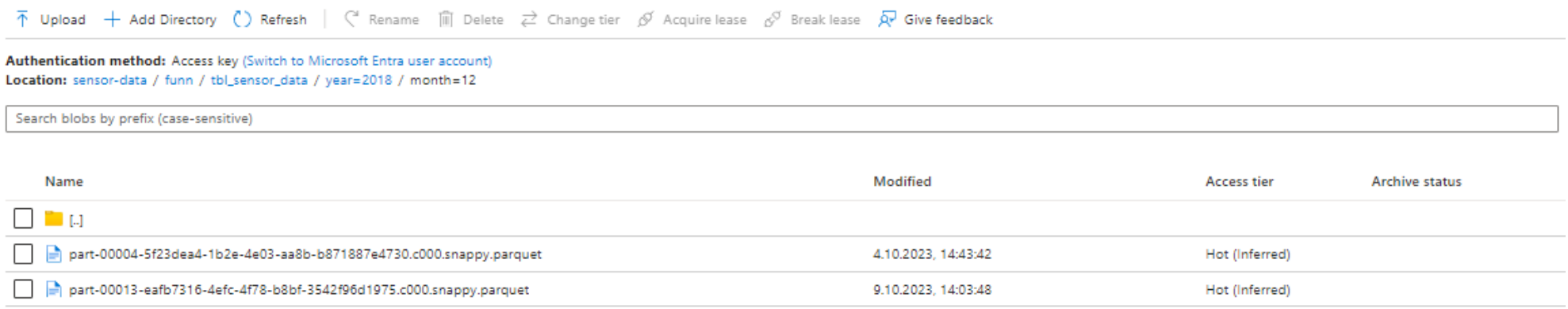
Delta Lake


Integrations
Spark
presto
Flink
trino
HIVE
kafka
amazon REDSHIFT
snowflake
Azure Synapse
amazon ATHENA
DataHub
Coming Soon:
Google BigQuery
PULSAR
Streaming
Batch
DELTA LAKE
Ingestion Tables (Bronze)
Refined Tables (Silver)
Feature/Agg Data Store (Gold)
Analytics and Machine Learning
Your Existing Data Lake
Google Cloud
hadoop HDFS
Azure Data Lake Storage
amazon S3
IBM Cloud
ORACLE CLOUD

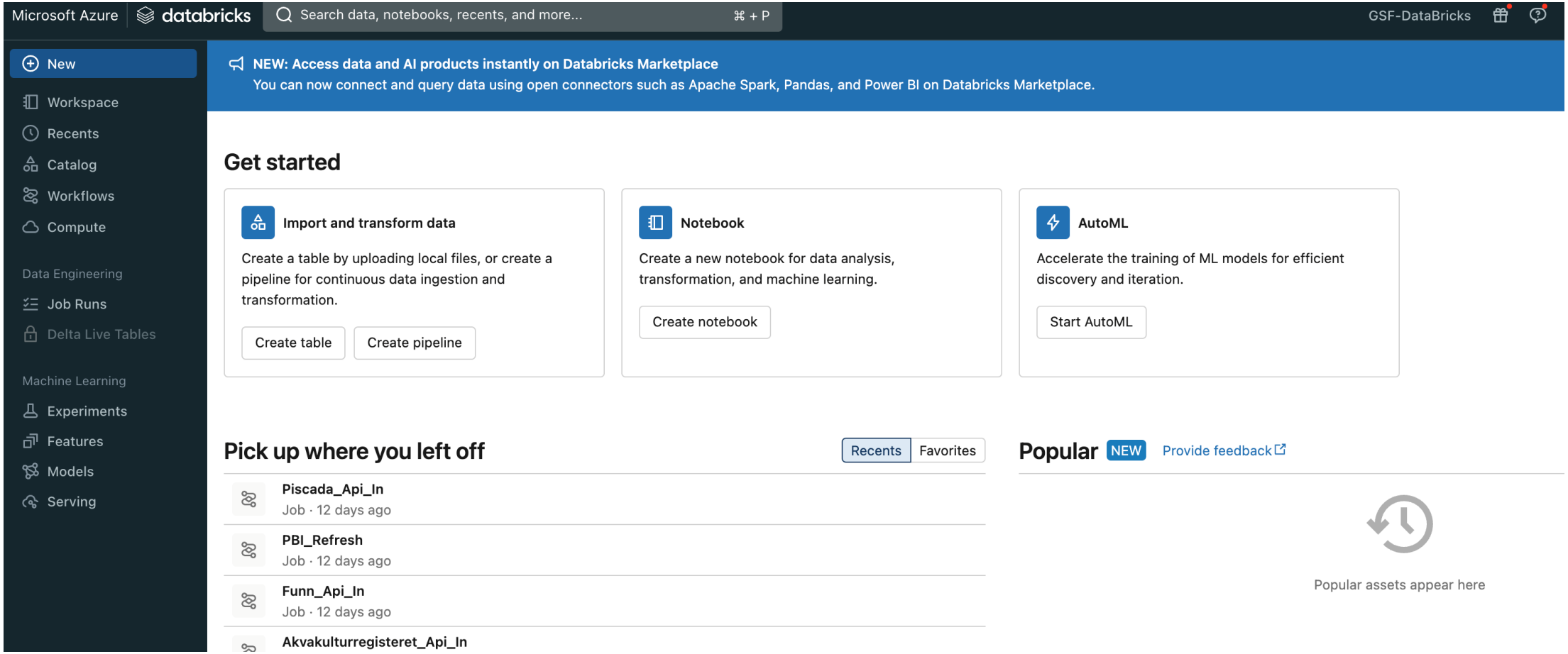
Delta Lake


Home > gsfdatalake | Containers >
sensor-data
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Give feedback
Authentication method: Access key (Switch to Microsoft Entra user account)
Location: sensor-data / funn / tbl_sensor_data
Search blobs by prefix (case-sensitive)
Name
Modified
[..]
_delta_log
year=2018
year=2019
year=2020
year=2021
year=2022
year=2023

Delta Lake (parquet)



Databricks


Microsoft Azure
databricks
Search data, notebooks, recents, and more...
⌘ + P
GSF-DataBricks
New
Workspace
Recents
Catalog
Workflows
Compute
Data Engineering
Job Runs
Delta Live Tables
Machine Learning
Experiments
Features
Models
Serving
NEW: Access data and AI products instantly on Databricks Marketplace
You can now connect and query data using open connectors such as Apache Spark, Pandas, and Power BI on Databricks Marketplace.
Get started
Import and transform data
Create a table by uploading local files, or create a pipeline for continuous data ingestion and transformation.
Create table
Create pipeline
Notebook
Create a new notebook for data analysis, transformation, and machine learning.
Create notebook
AutoML
Accelerate the training of ML models for efficient discovery and iteration.
Start AutoML
Pick up where you left off
Recents
Favorites
Piscada_Api_In
Job · 12 days ago
PBI_Refresh
Job · 12 days ago
Funn_Api_In
Job · 12 days ago
Akvakulturregisteret_Api_In
Popular
NEW
Provide feedback
Popular assets appear here

Databricks

1. **Połączenie najlepszych cech magazynu danych i hurtowni danych**: Architektura Lake House łączy najlepsze elementy magazynu danych i hurtowni danych, umożliwiając redukcję kosztów i dostarczanie różnych przypadków użycia związanych z uczeniem maszynowym i analizą danych.

2. **Szybkość i niezawodność w budowaniu i wdrażaniu modeli AI**: Dzięki unikalnemu podejściu opartemu na danych, Databricks umożliwia szybkie i niezawodne budowanie oraz wdrażanie modeli sztucznej inteligencji. Można łatwo eksplorować dane, trenować modele i wdrażać je w środowisku produkcyjnym.

3. **Optymalizacja składowania i przetwarzania danych**: Databricks oferuje zoptymalizowane narzędzia do składowania i przetwarzania danych. Można łatwo zarządzać dużymi zbiorami danych, wykorzystując technologie takie jak Delta Lake, które zapewniają wysoką wydajność i niezawodność.

4. **Elastyczność i skalowalność**: Platforma Databricks jest elastyczna i skalowalna, co oznacza, że można dostosować jej moc obliczeniową do zmieniających się potrzeb. Można łatwo zwiększać lub zmniejszać zasoby w zależności od wymagań projektu.

Hive

- **Przetwarzanie dużych zbiorów danych:** Dzięki niemu można przetwarzać ogromne ilości danych, co jest szczególnie przydatne dla organizacji i osób pracujących z dużymi zbiorami danych.
- **Zapytania w stylu SQL:** Apache Hive umożliwia wykonywanie zapytań przy użyciu HiveQL, czyli języka zapytań w stylu SQL. Dzięki temu można łatwo formułować zapytania i analizować dane w sposób zrozumiały dla osób z doświadczeniem w SQL.

SQL Copy

```sql
DROP TABLE wikicc
```

SQL Copy

```sql
CREATE EXTERNAL TABLE `wikicc`(
  `country` string,
  `count` int)
ROW FORMAT SERDE
  'org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe'
STORED AS INPUTFORMAT
  'org.apache.hadoop.mapred.TextInputFormat'
OUTPUTFORMAT
  'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
LOCATION
  '<path-to-table>'
```

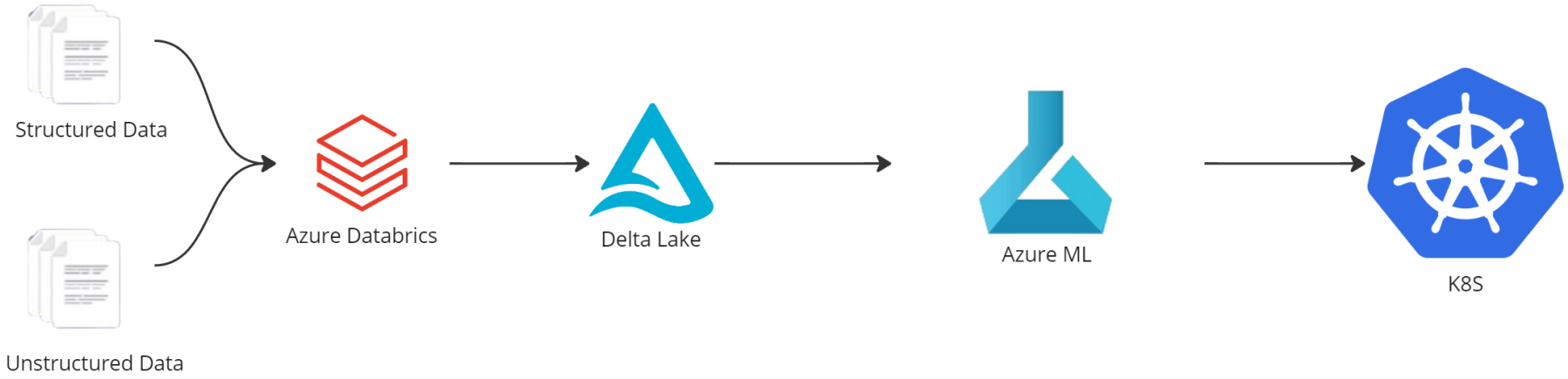
Apache Spark

Apache Spark to silnik do przetwarzania danych. Zawiera całą masę bibliotek, których można używać do przetwarzania danych w klastrze komputerów.

Najważniejszą korzyścią jest możliwość równoległego przetwarzania danych. Obecnie jest jednym z najpopularniejszych narzędzi do Big Data.

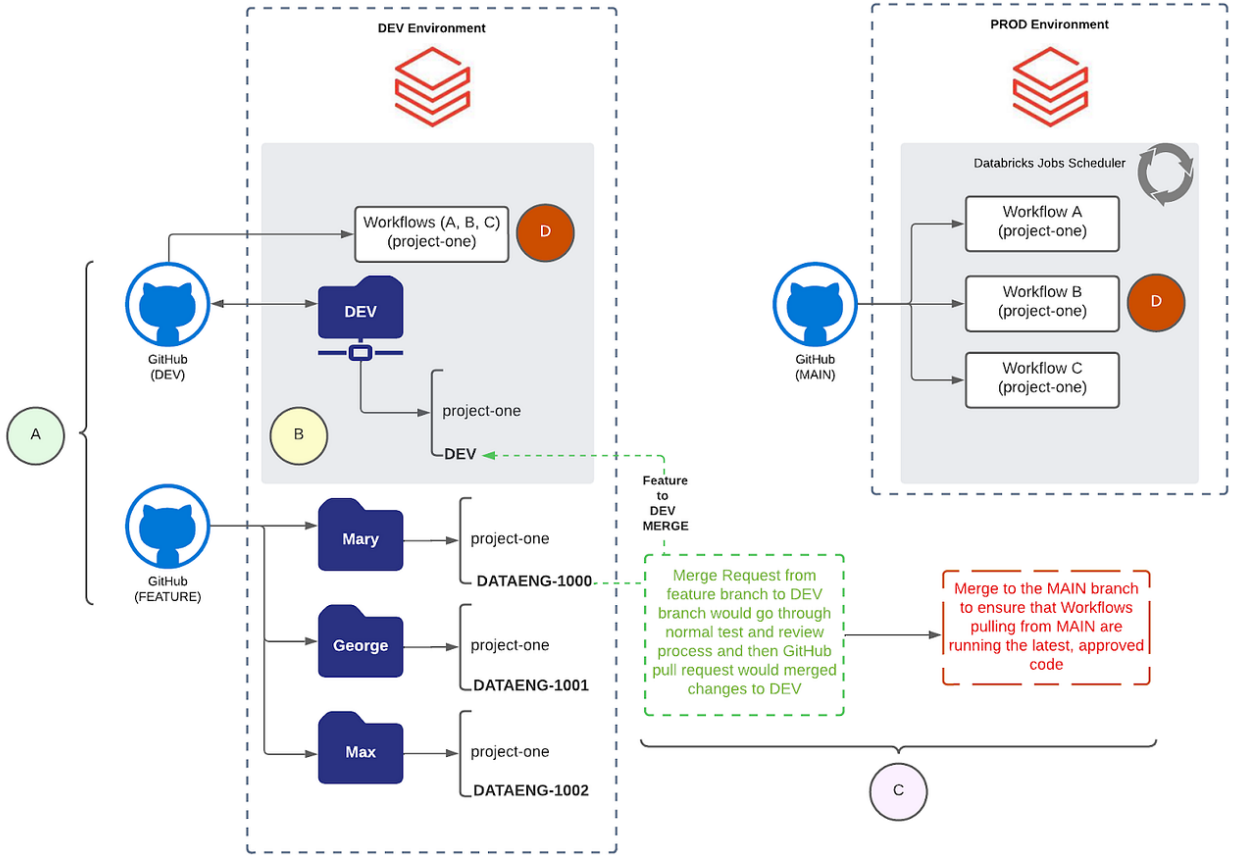
Obsługuje wiele języków programowania (**Python, SQL, Scala i R**). Możesz rozwiązywać problemy na wiele sposobów, używając SQL, transmisji danych („data streaming”) i uczenia maszynowego.

Developing process of ETL

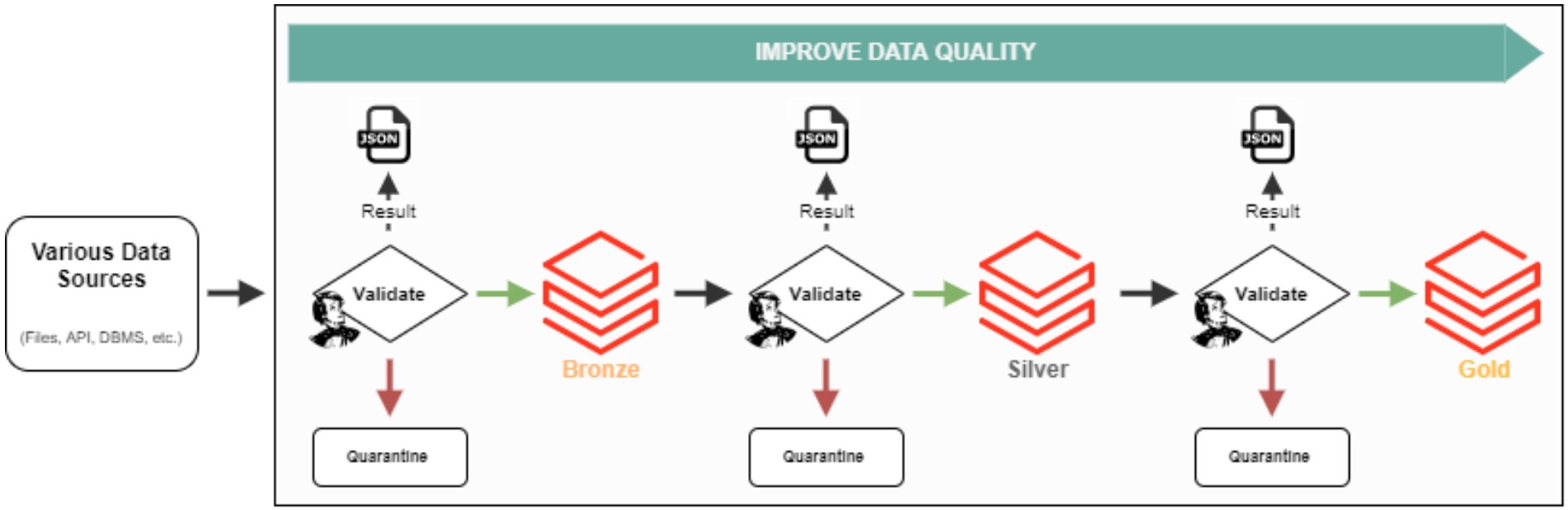


Developing process of ETL

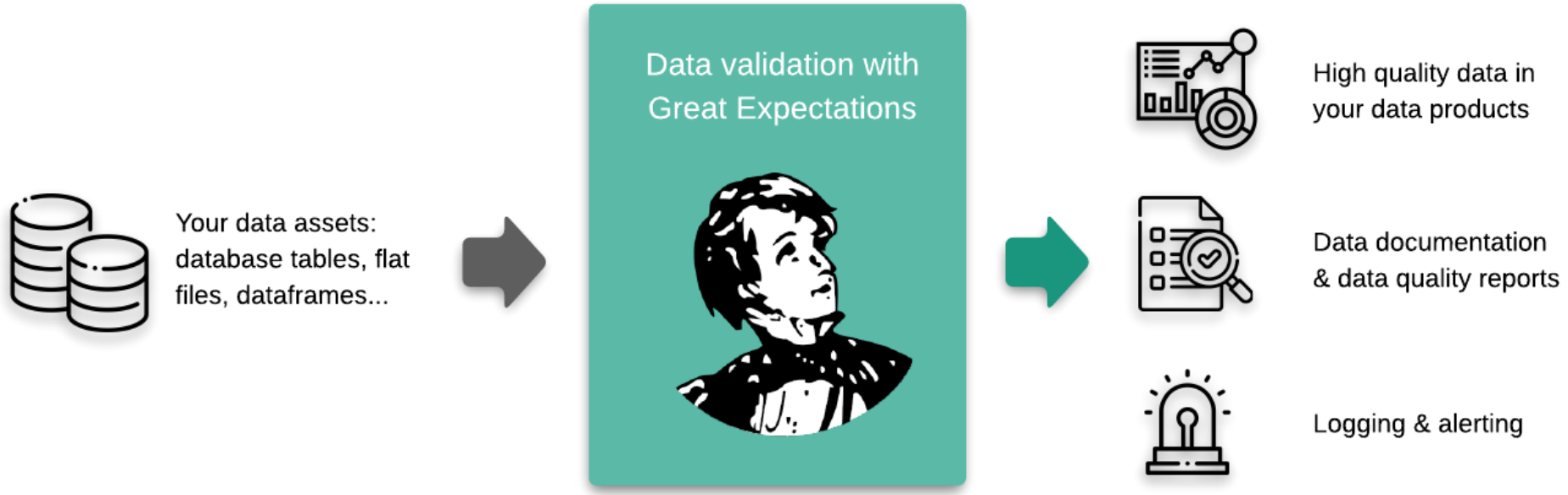
Developing process of ETL



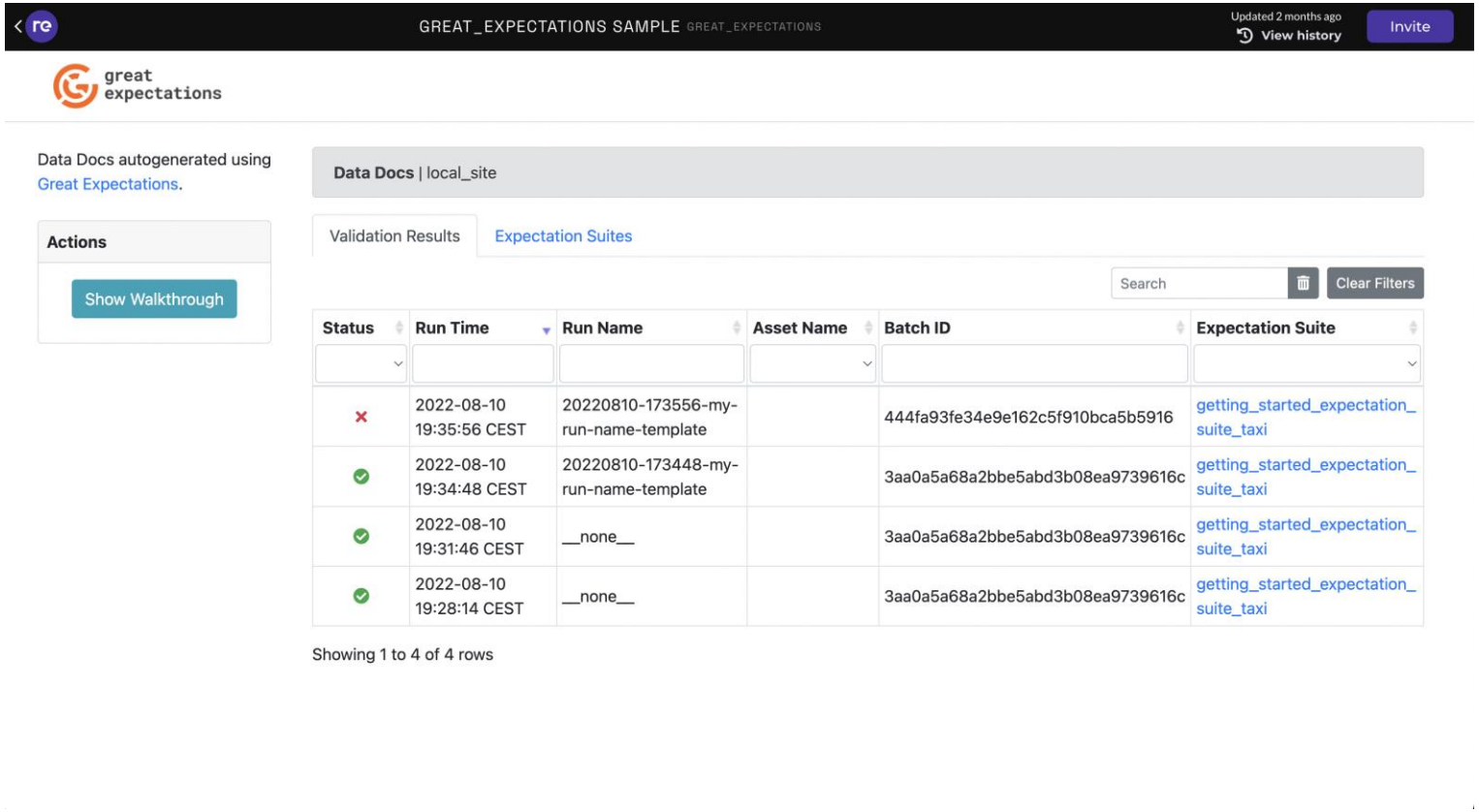
Jakość i niezawodność danych



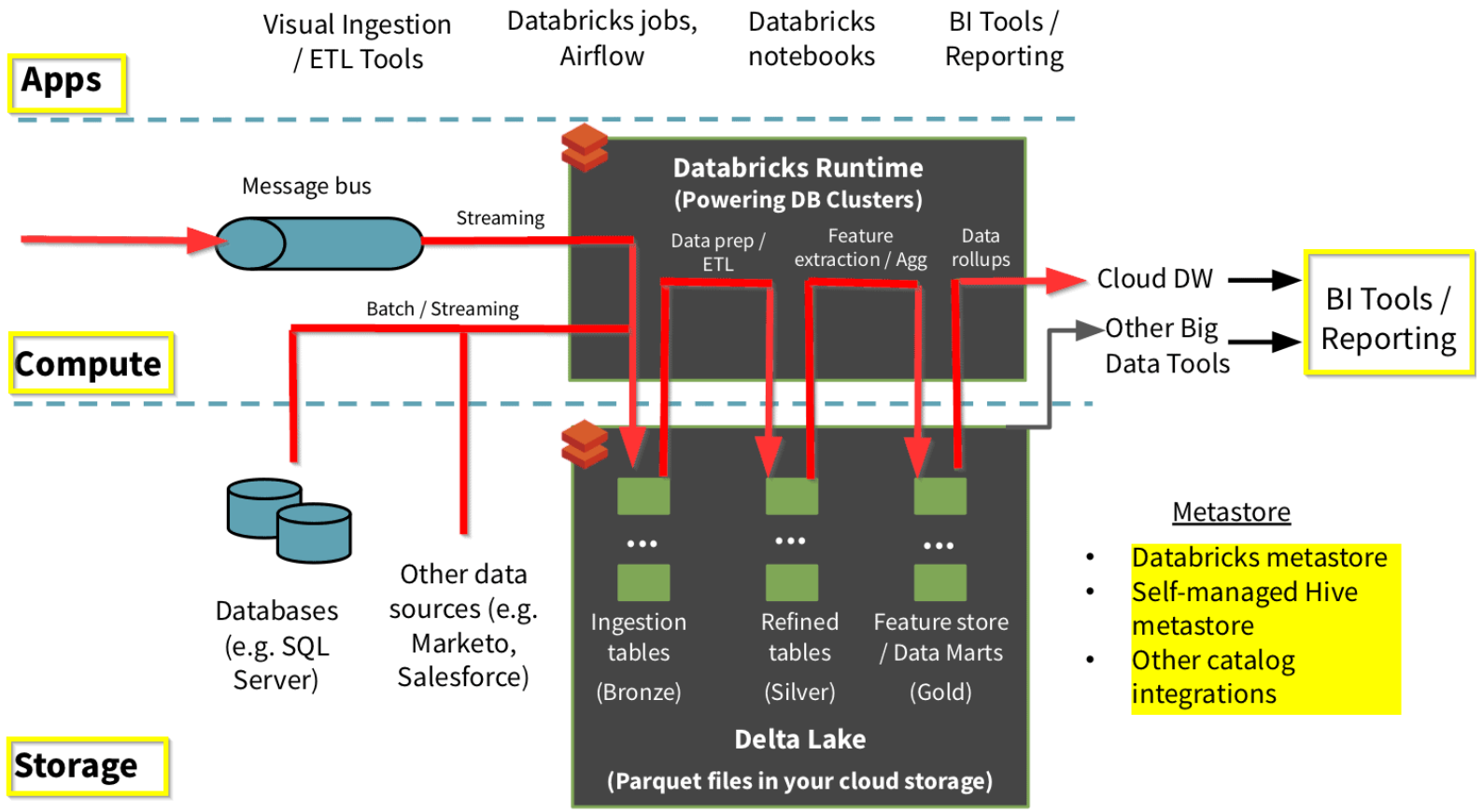
Jakość i niezawodność danych



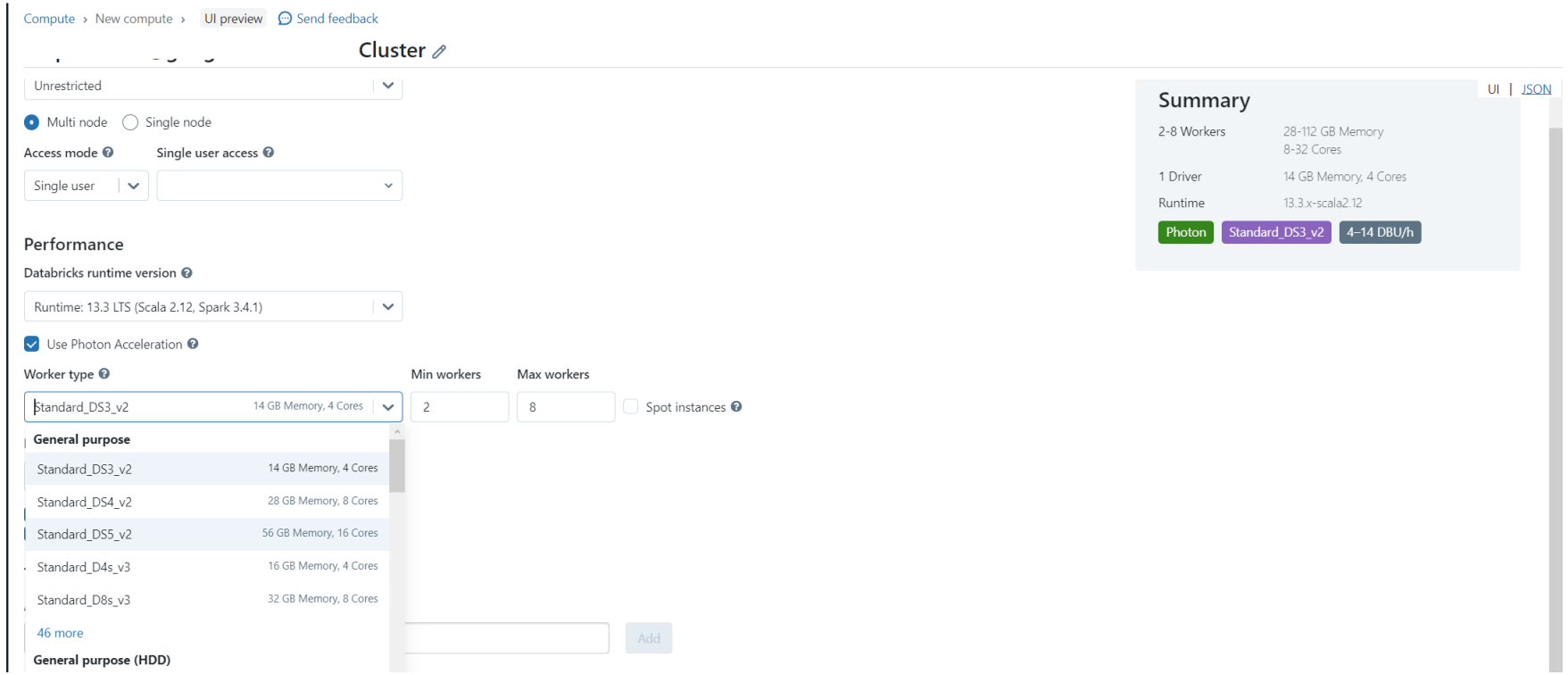
Jakość i niezawodność danych



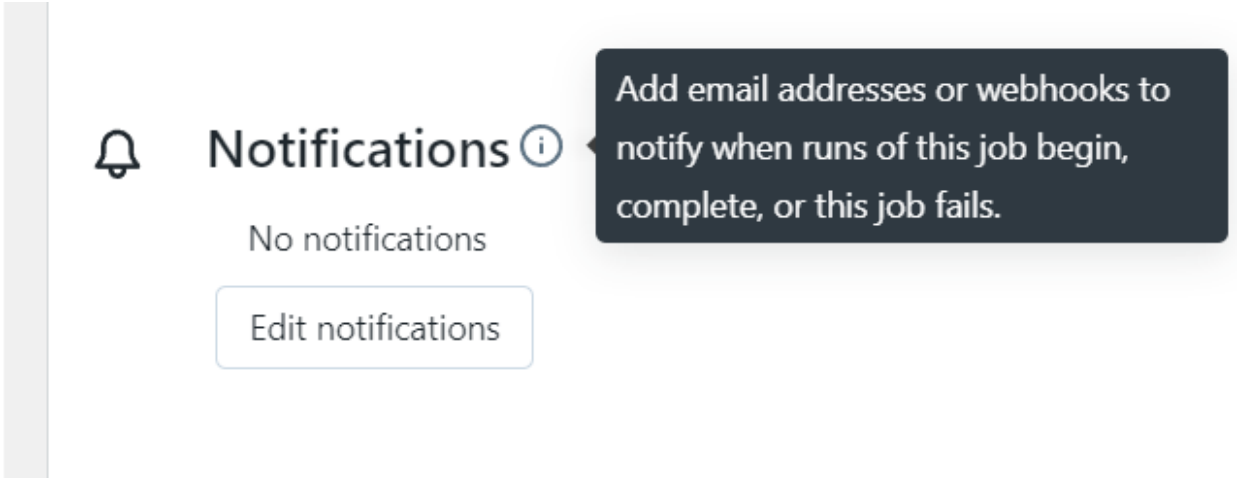
Jakość i niezawodność danych



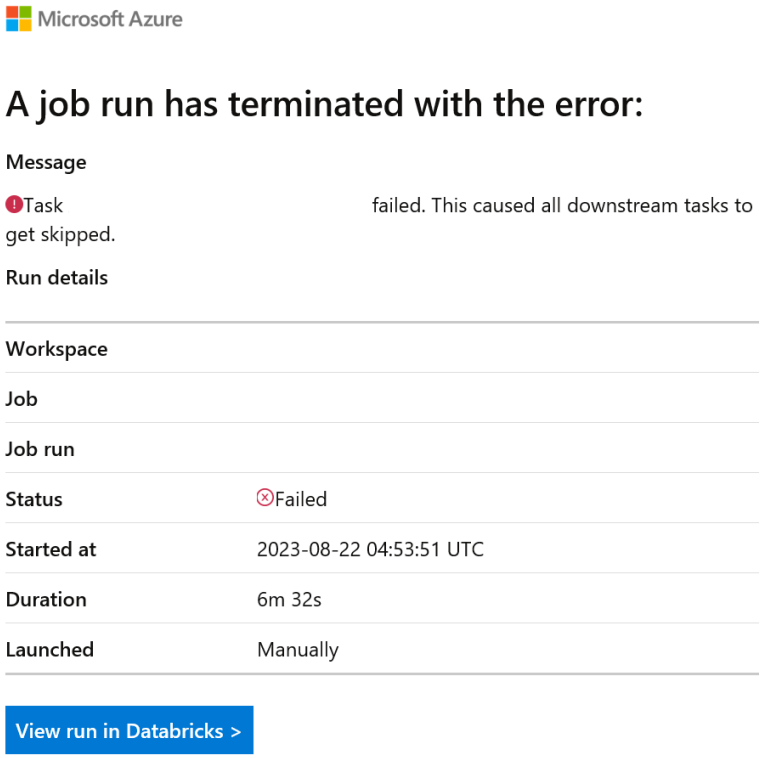
Skalowalność



Monitoring of ETL



Monitoring of ETL



Main | Search | Views | Aggregate Graphs | Compare Hosts | Events | Reports | Automatic Rotation | Live Dashboard | Cubism | Mobile

cluster Report at Mon, 18 Sep 2023 14:00:02 +0000 Get Fresh Data

Last hour | 2hr | 4hr | day | week | month | year | job or from [ ] to [ ] Go Clear

Timezone: Server

Physical View

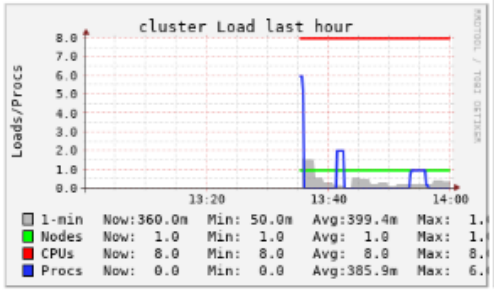
Grid > **cluster** > --Choose a Node

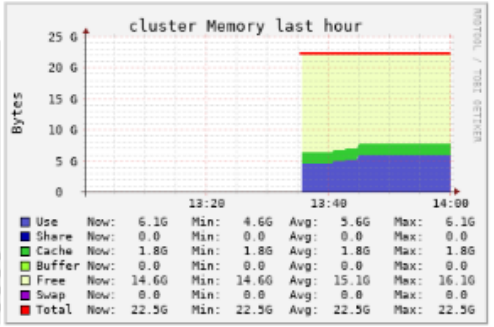
Overview of cluster @ 2023-09-18 14:00

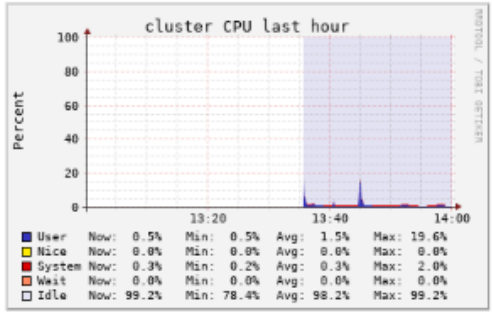
CPUs Total: **8**
Hosts up: **1**
Hosts down: **0**

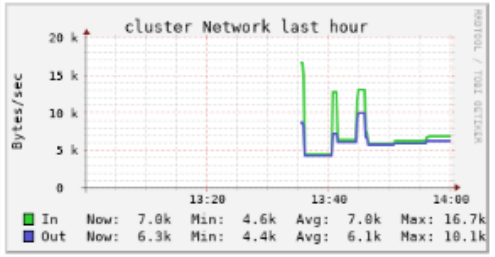
Current Load Avg (15, 5, 1m):
5%, 3%, 4%
Avg Utilization (last hour):
0%

Server Load Distribution









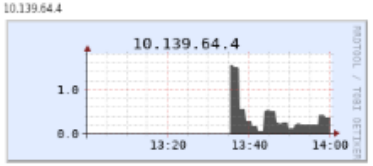
Stacked Graph - load_one

cluster **load_one** last **hour** sorted **by name**

Metric: load_one

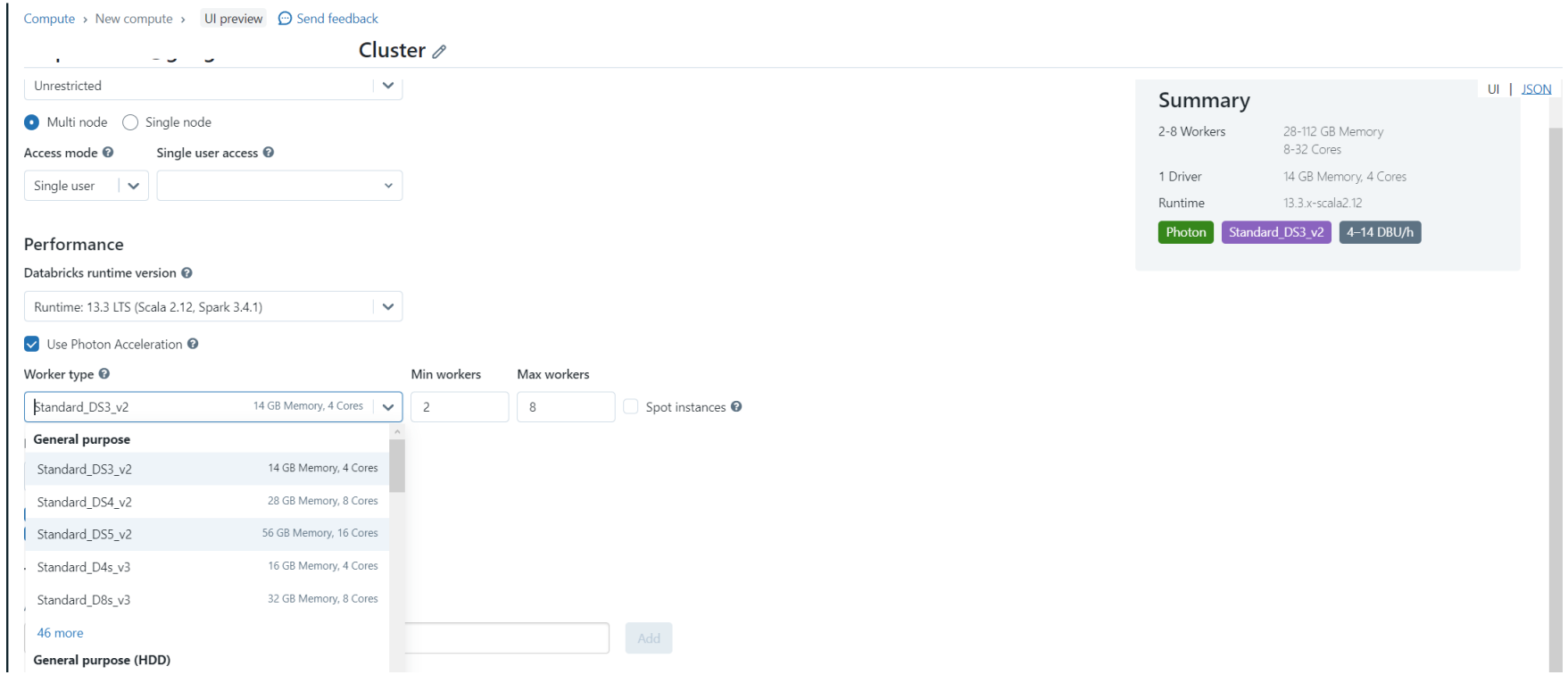
Show Hosts Scaled: Auto | Same | None Size: small Columns: 4 (0 = metric + reports)

Show only nodes matching [ ] Filter Max graphs to show: all Sorted: ascending | descending | by name



(Nodes colored by 1-minute load) | Legend

Skalowalność



Cena

DSv2 series

Instance	vCPU(s)	RAM	DBU Count	DBU Price	Pay As You Go Total Price	1 Year Reserved VM (% Savings) Total Price	3 Year Reserved VM (% Savings) Total Price	Spot (% Savings) Total Price
DS3 v2	4	14.00 GiB	0.75	$0.113/hour	$0.464/hour	$0.271/hour ~42% savings	$0.220/hour ~53% savings	$0.150/hour ~68% savings
DS4 v2	8	28.00 GiB	1.50	$0.225/hour	$0.927/hour	$0.542/hour ~42% savings	$0.439/hour ~53% savings	$0.300/hour ~68% savings
DS5 v2	16	56.00 GiB	3.00	$0.45/hour	$1.855/hour	$1.084/hour ~42% savings	$0.877/hour ~53% savings	$0.601/hour ~68% savings

Azure ML Studio

- **User Interface:** Azure ML Studio offers a user-friendly, drag-and-drop interface, making it accessible for beginners and those who prefer a visual approach to building machine learning models.

- **Ease of Use:** It simplifies the process of data science by abstracting complex coding requirements, allowing users to focus on the model design.

- **Scalability:** While it provides adequate scalability, it may not be as robust as Databricks when handling extremely large datasets or highly complex models.

- **Data Processing:** Azure ML Studio supports various data processing and transformation activities, but with a focus on simplicity rather than extensive customization.

- **Model Training:** Offers a wide range of pre-built algorithms and the ability to import custom models, though it might be less flexible compared to Databricks for advanced users.

Azure DataStudio

•**User Interface:** Databricks primarily uses a notebook-based interface, which is favored by users who are comfortable with coding and script-based model building.

•**Ease of Use:** More suited for users with a solid background in data science and programming. Offers more control and customization for complex tasks.

•**Scalability:** Excelling in scalability, Databricks is designed to handle large-scale data processing and complex machine learning tasks efficiently.

•**Data Processing:** Provides extensive support for data processing, with a focus on big data handling, using Spark and other big data technologies.

•**Model Training:** Allows more flexibility in model training, including the use of advanced machine learning and deep learning frameworks.

Gdzie hostowac?

- **Azure Kubernetes Service (AKS)**

Overview: Offers scalable and flexible deployment options for complex applications, using Kubernetes.

Best For: Large-scale, enterprise-grade ML deployments. Ideal if you need high scalability and robust orchestration for deploying complex models.

Gdzie hostowac?

- **Azure Functions**

Overview: A serverless compute service that lets you run event-triggered code without having to explicitly provision or manage infrastructure.

Best For: Lightweight models, especially in scenarios where you need to run small pieces of code in response to events.

Gdzie hostowac?

- **Azure Container Instances (ACI)**

Overview: Provides a simple way to deploy containers with a quick setup and without the need for orchestration.

Best For: Deploying models in containers for scenarios where you don't require the full capabilities of Kubernetes. Suitable for simple, small-scale deployments.

Gdzie hostowac?

- **Azure Web Apps**

Overview: Allows you to build and host web applications in the programming language of your choice without managing infrastructure.

Best For: ML models that are integrated into web applications, especially when you need easy scaling and integration with other Azure services.

Role w Projekcie

DevOps

DataScience

Cloud Architect

What is DevOps?

DevOps is anything that makes developing & releasing :

Fast

Automated

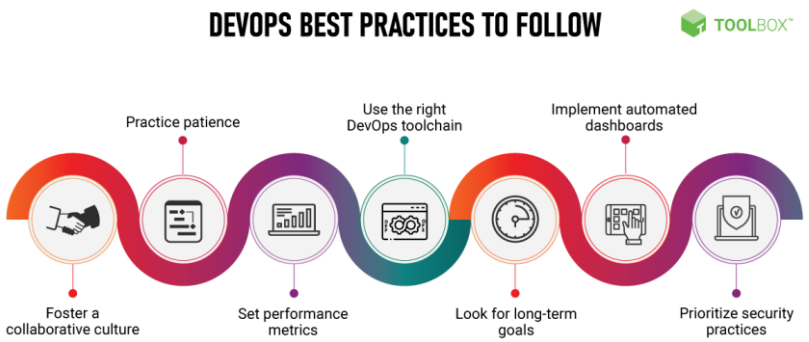
High-Quality

DevOps is combination of:

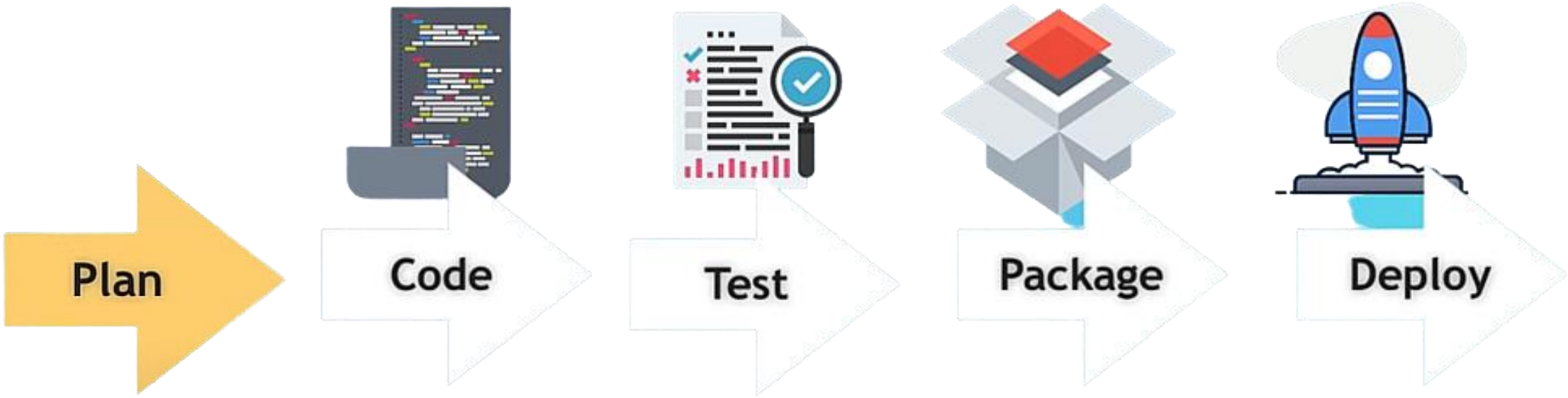
Tools

Concepts

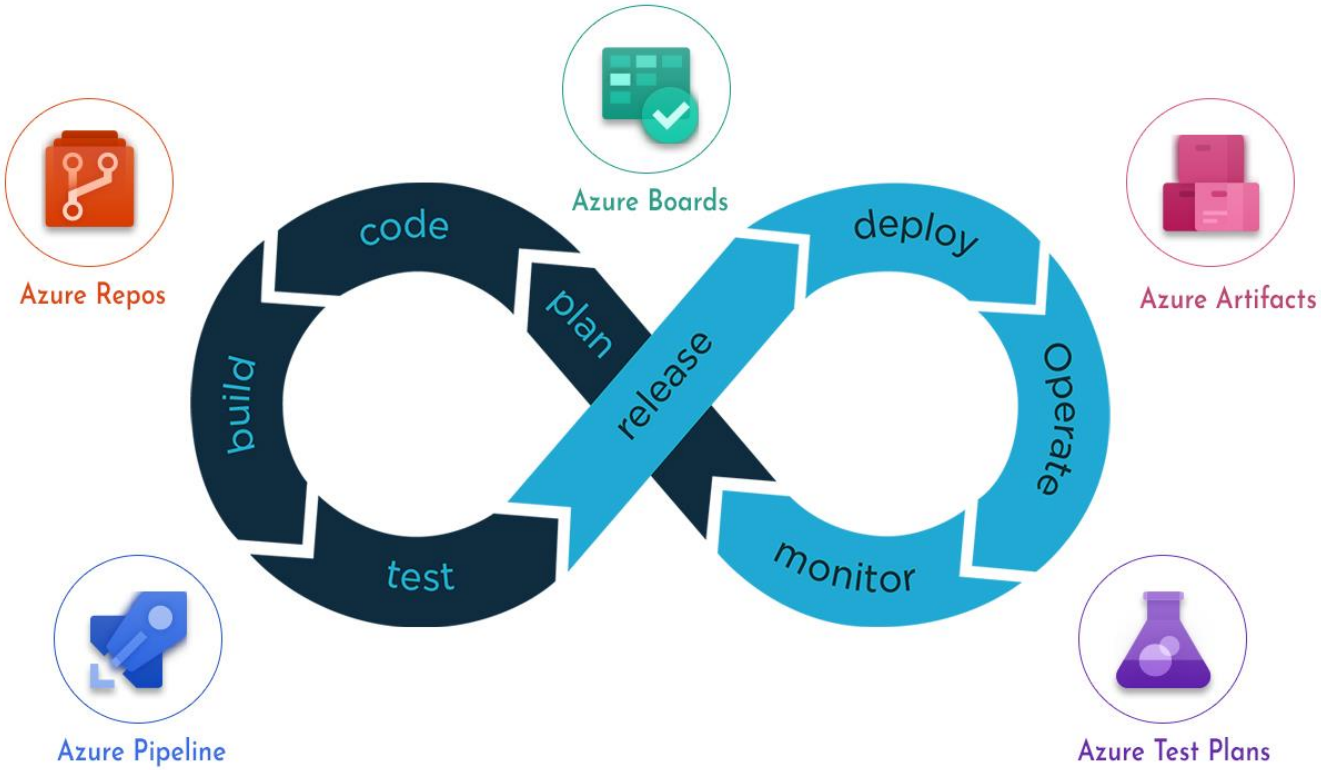
Practices



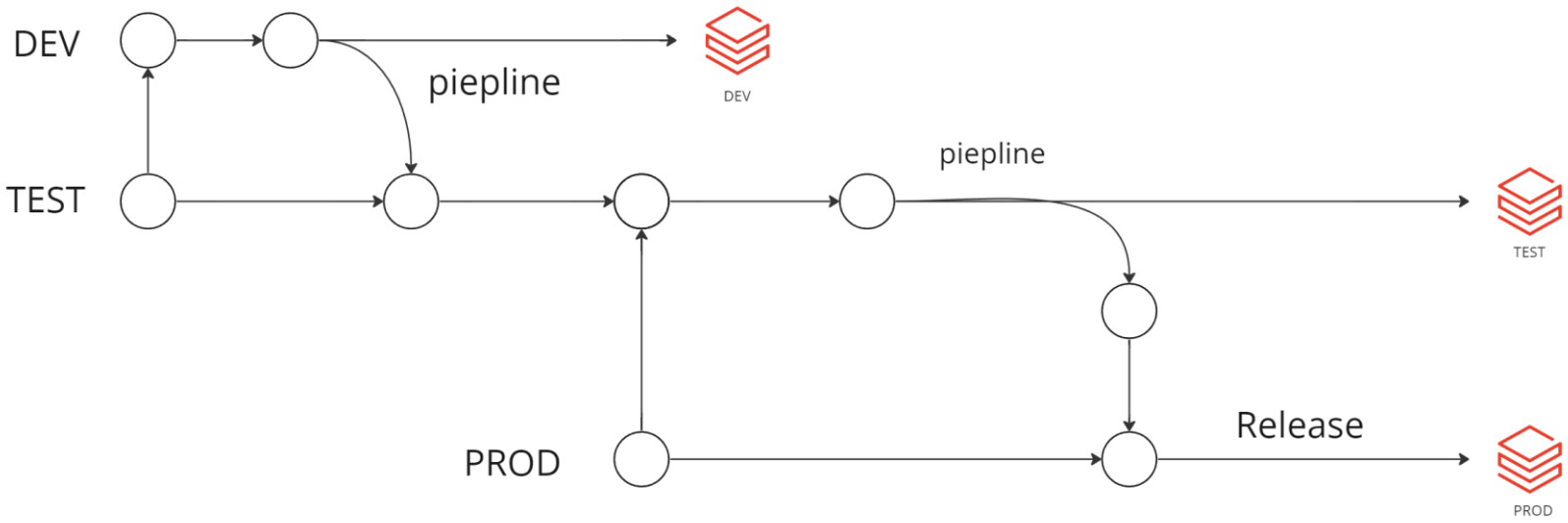
Developing process of ETL



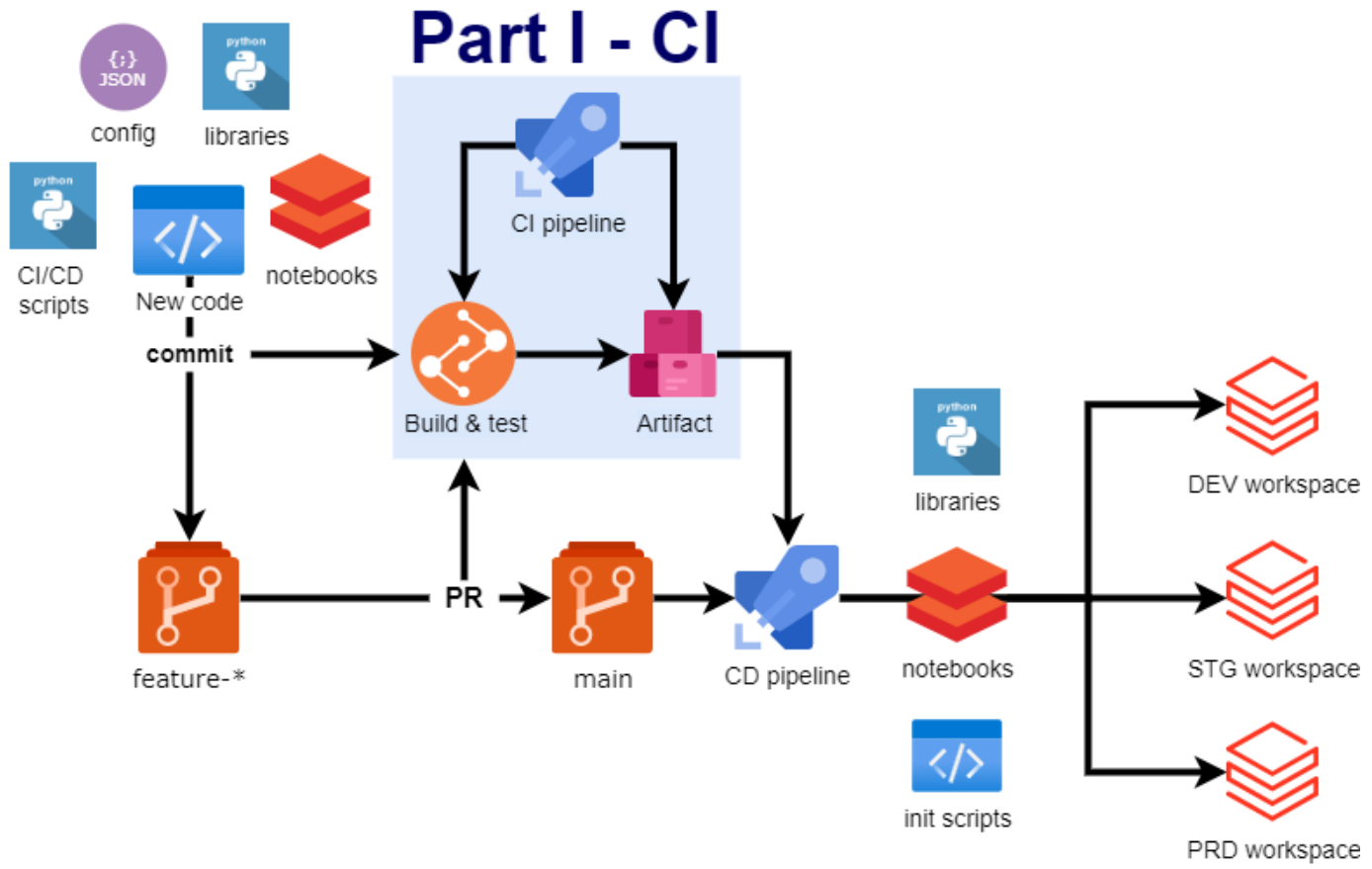
Developing process of ETL



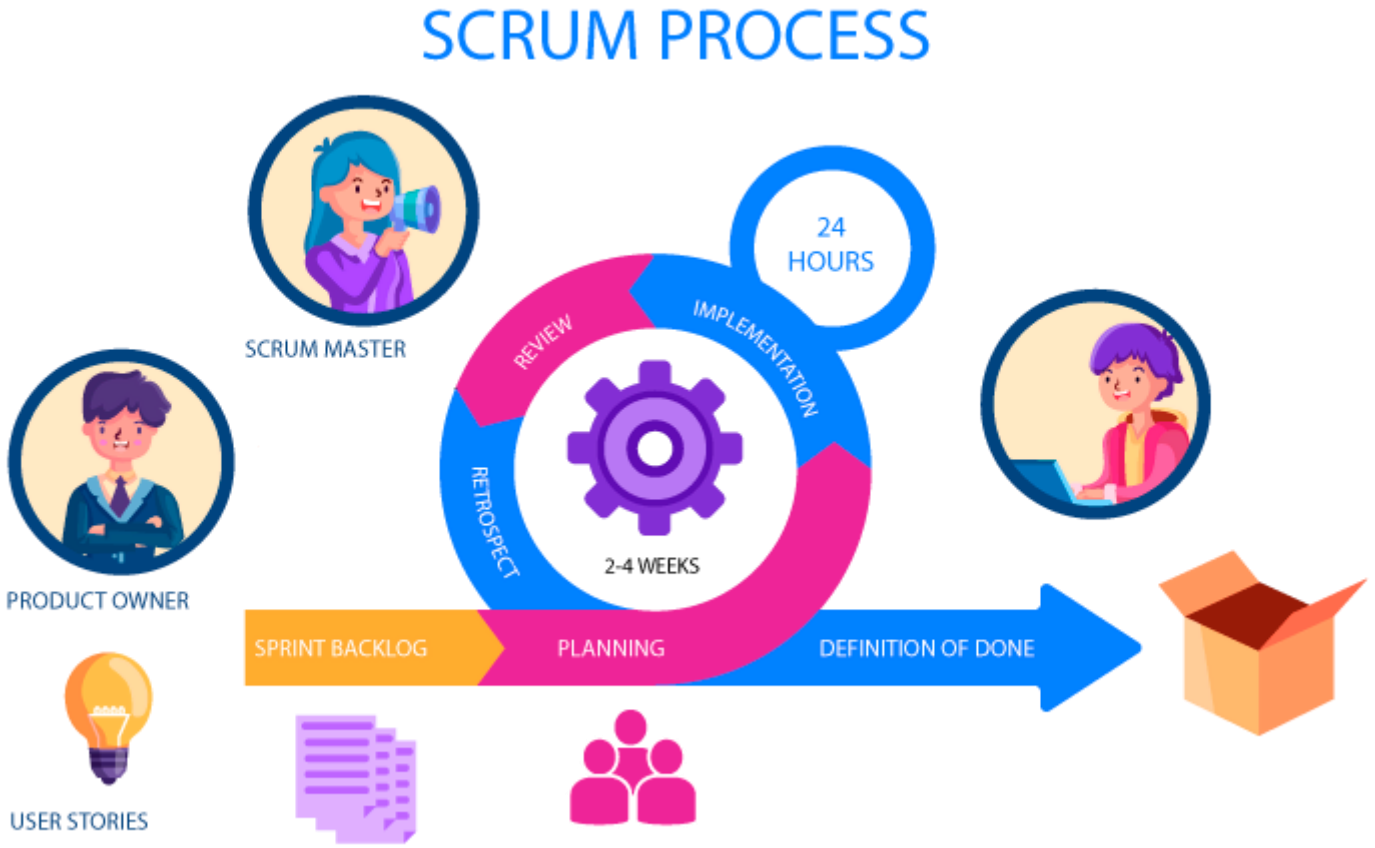
Developing process of ETL



Developing process of ETL

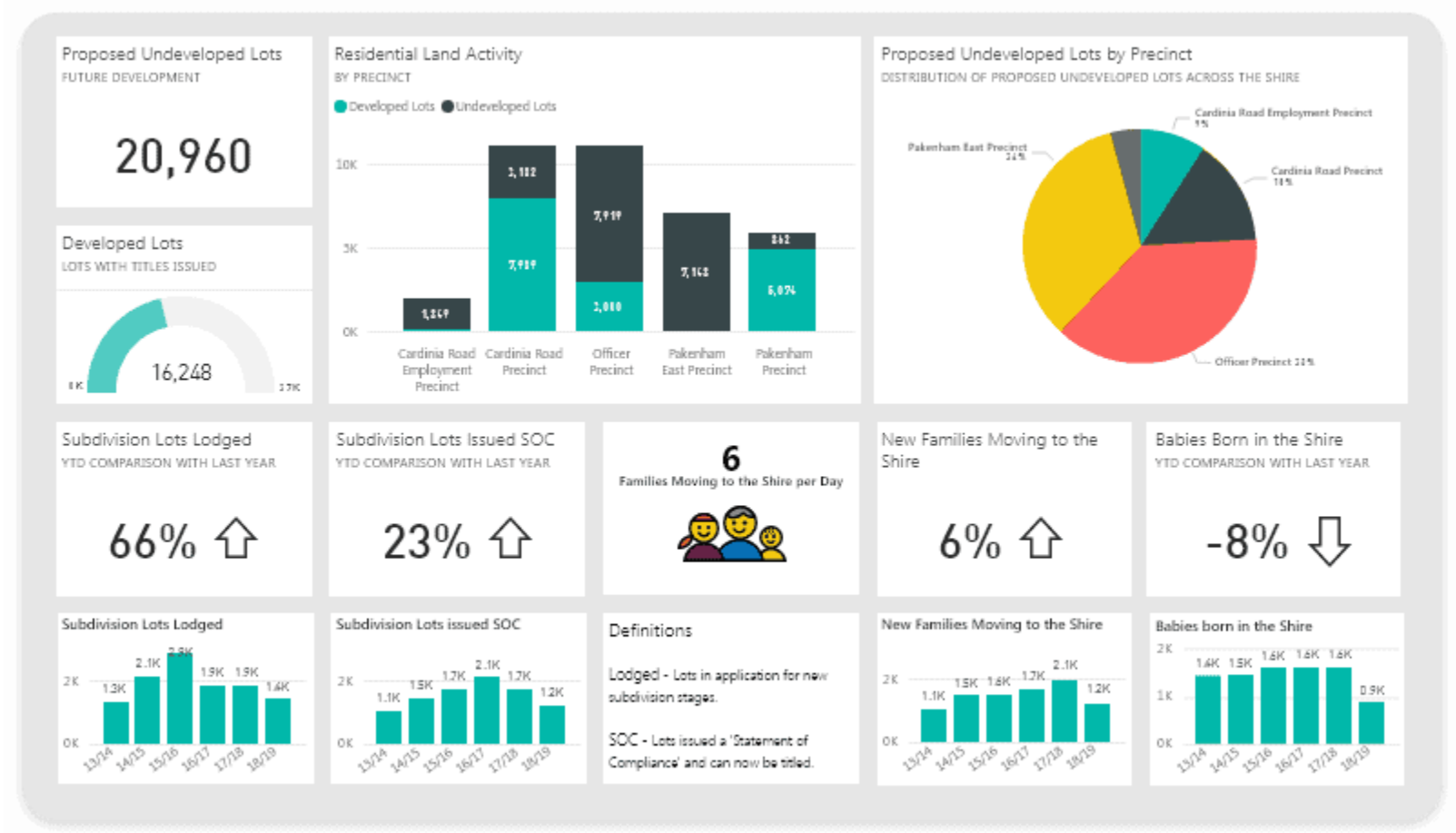


Kolaboracja między zespołami



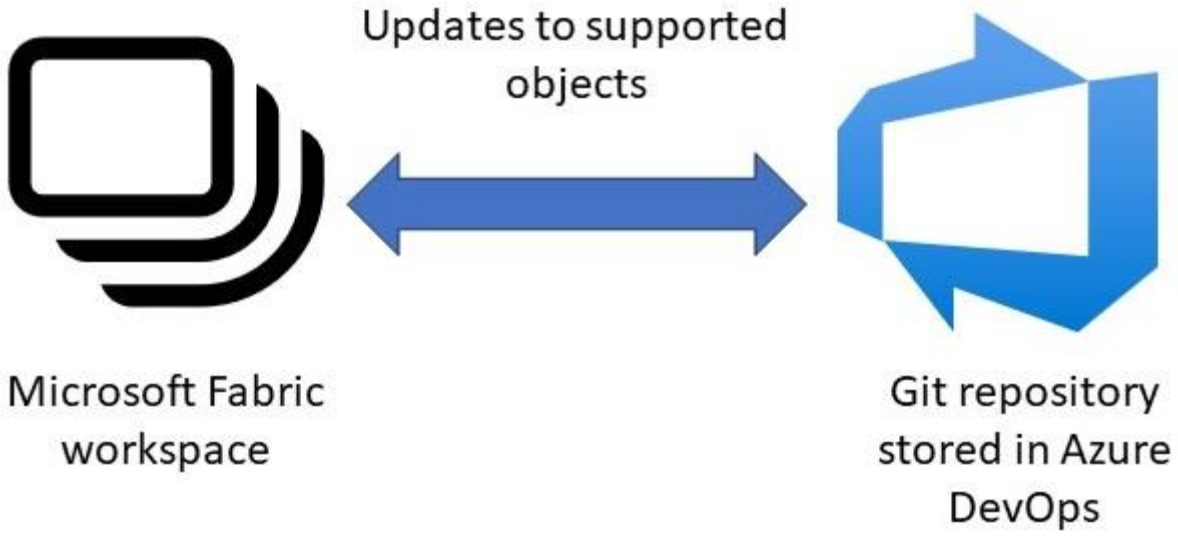
Szybsza adaptacja do zmian:

Szybsza adaptacja do zmian:

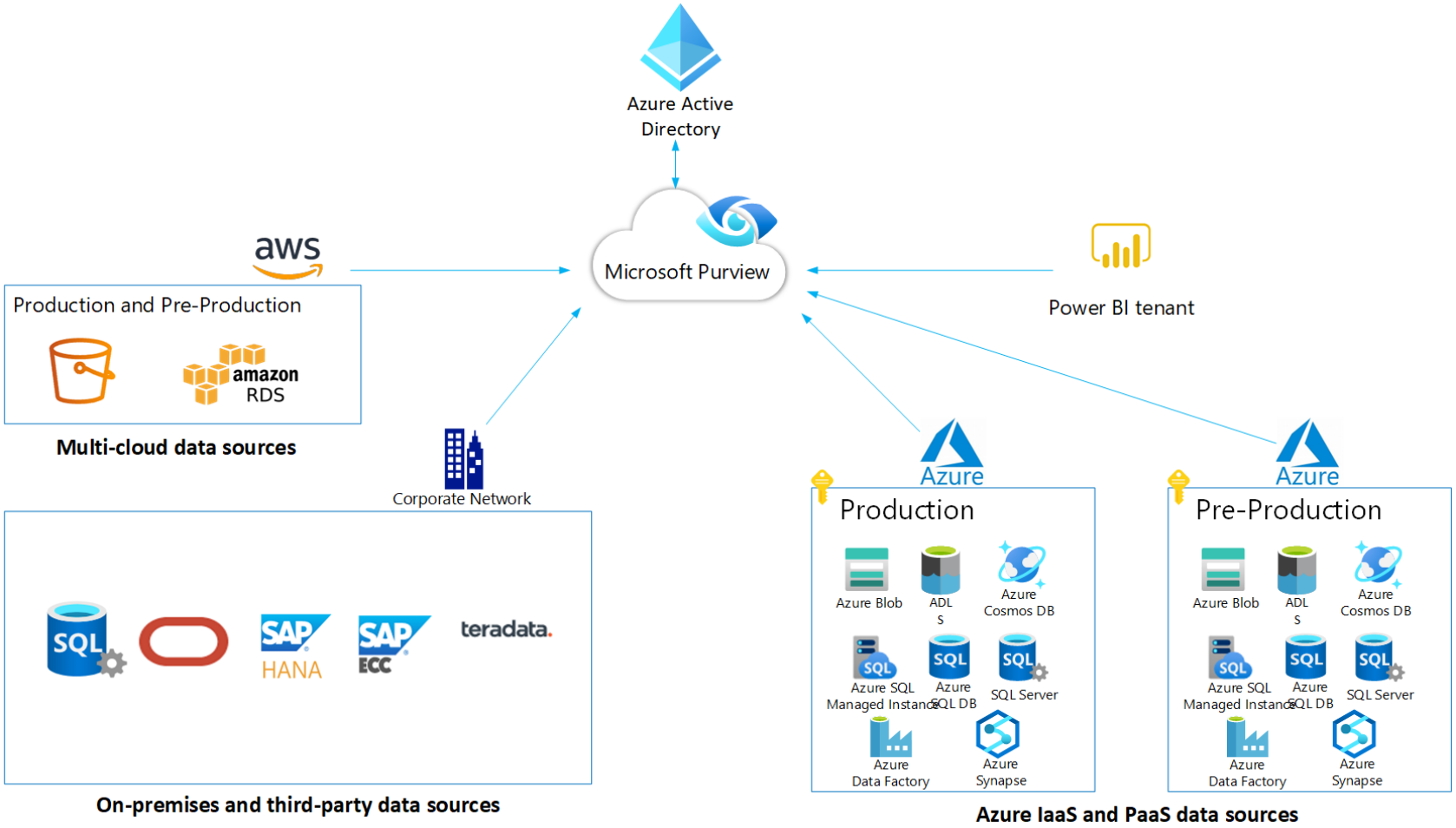


Szybsza adaptacja do zmian:

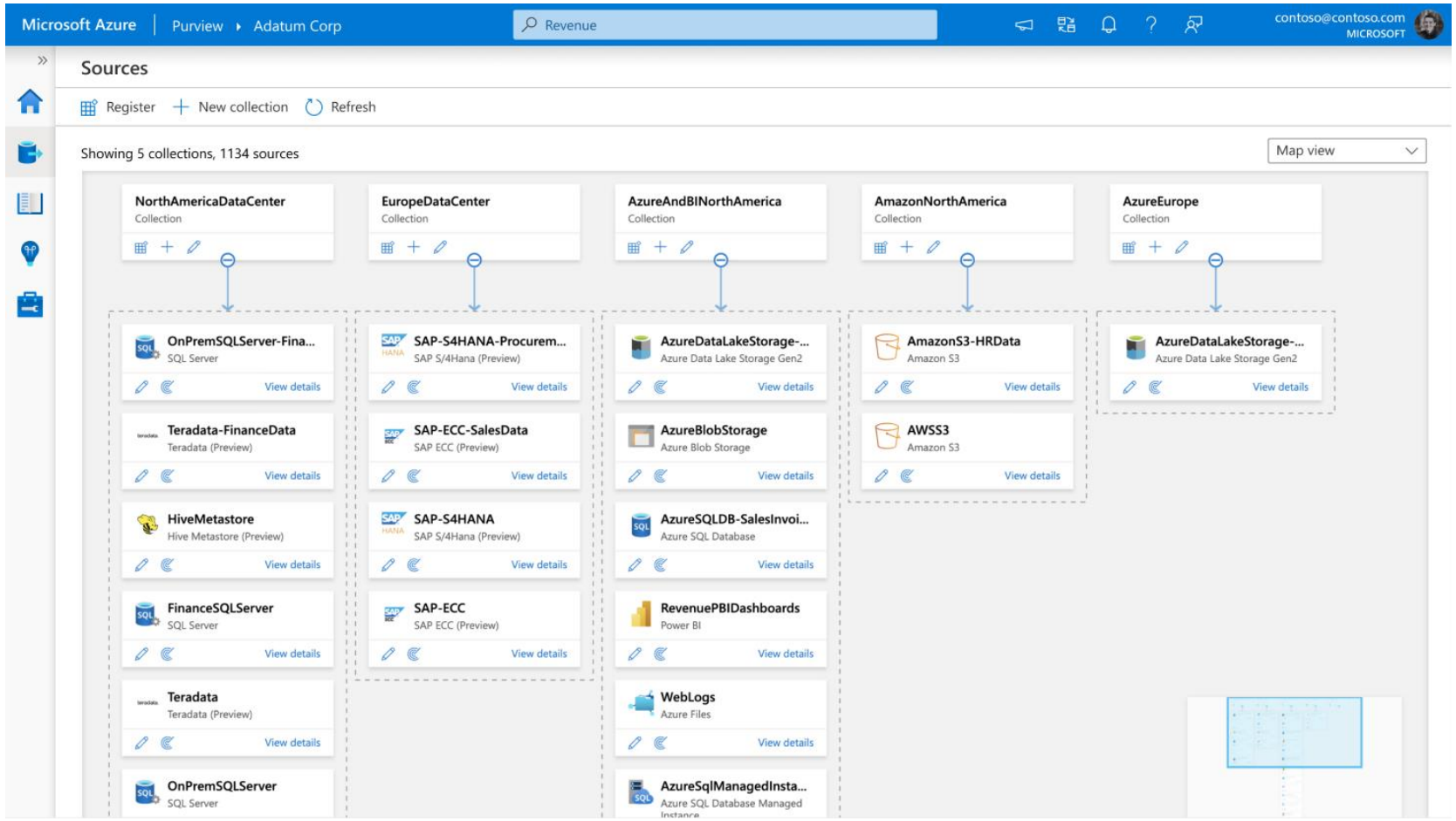
Szybsza adaptacja do zmian:



Azure Pureview



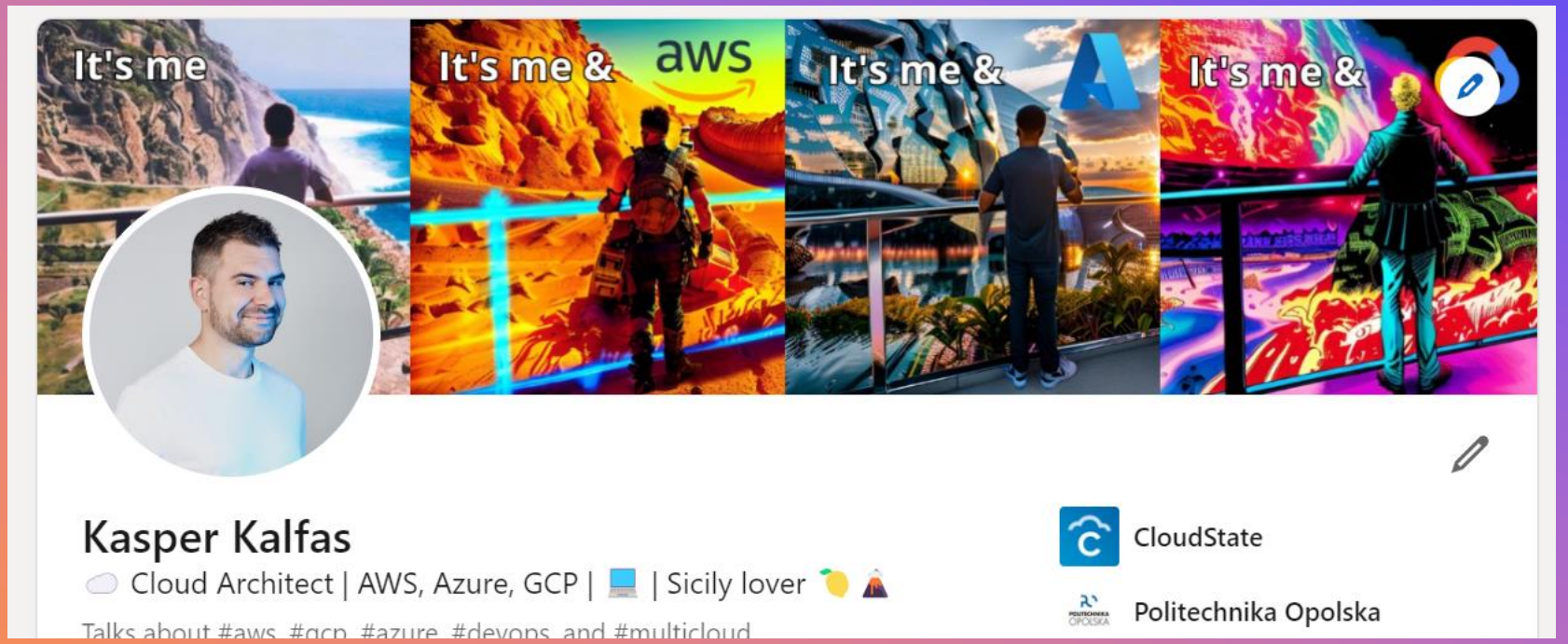
Azure Pureview



DevOps & Data

1. **Automatyzacja i ciągła integracja:**
2. **Kolaboracja między zespołami:**
3. **Jakość i niezawodność danych:**
4. **Skalowalność:**
5. **Szybsza adaptacja do zmian:**
6. **Bezpieczeństwo**

THANK YOU


It's me
It's me &
aws
It's me &
It's me &
Kasper Kalfas
Cloud Architect | AWS, Azure, GCP | | Sicily lover
Talks about #aws, #gcp, #azure, #devops, and #multicloud
CloudState
Politechnika Opolska

Q&A