

DEVOPS W PROJEKCIE DATA

+

•

○



Kasper Kalfas

Cloud Architect

e: cloudkasperpro@gmail.com

The logo for Chmurowisko features a stylized blue cloud icon above the word "Chmurowisko" in a bold, blue, sans-serif font.

Chmurowisko



Moje ordery



AWS Certified Solutions Architect - Associate



Microsoft Certified: DevOps Engineer Expert



Microsoft Certified: Azure Developer Associate



AWS Certified Solutions Architect - Associate



Associate Cloud Engineer

CloudKasper.pro

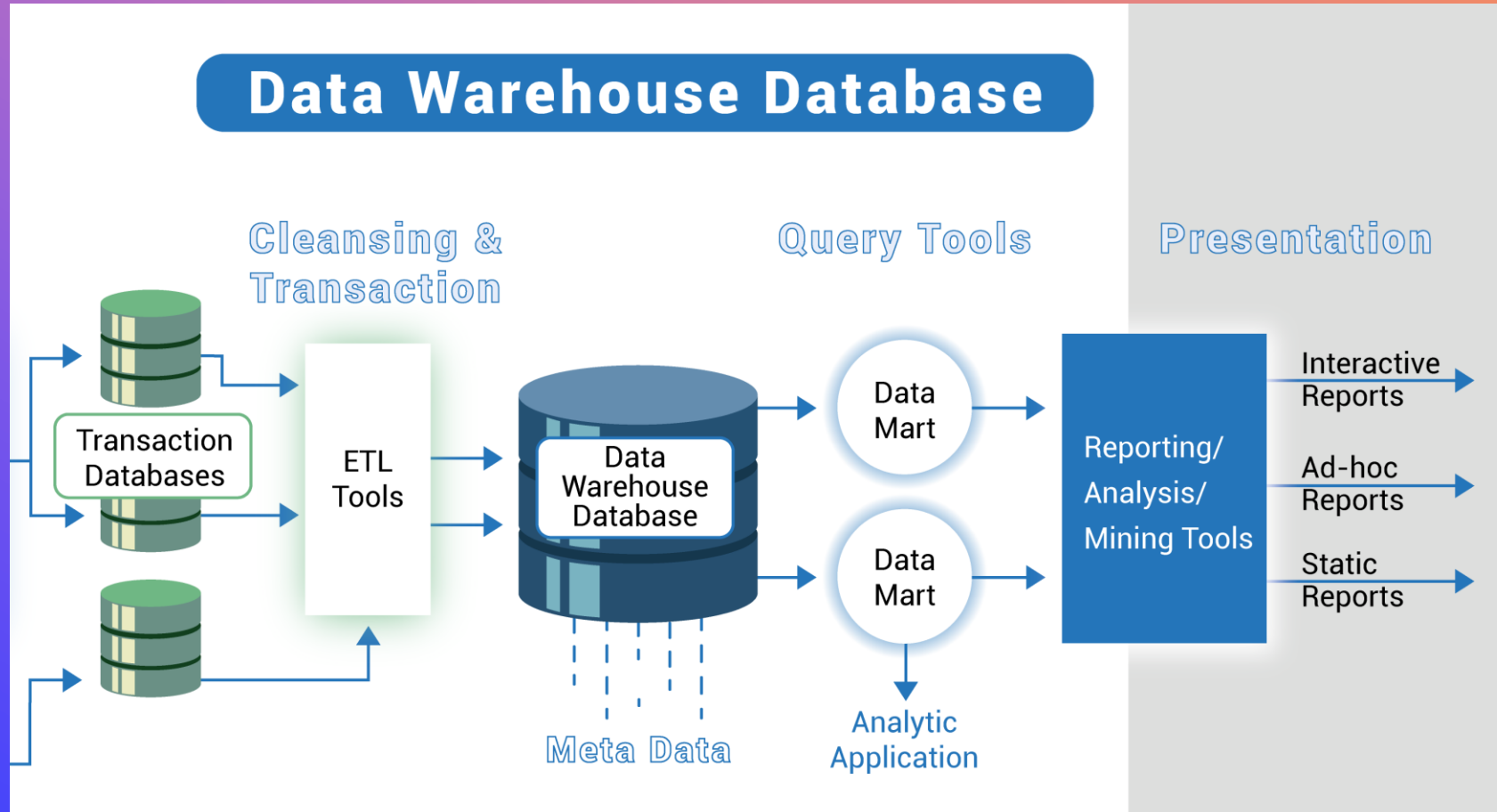
The screenshot shows the homepage of CloudKasper.pro. At the top, there is a navigation bar with links for HOME, BLOG, ABOUT, and a FREE E-BOOK. Below the navigation, there are two featured articles. The first article is titled "Diving Into AWS Cloud Engineer Life" and features an illustration of a man in a red hat and orange coat, reminiscent of Columbus, sitting at a desk with a laptop. The second article is titled "Getting Started In Cloud Computing For Developers" and features an illustration of a man in a blue coat, reminiscent of Leonardo da Vinci, sitting at a desk with a laptop. Both articles have a blue header with the title and a small image of the author.

The screenshot shows a Facebook group post for a group named "DevOps". The post features a vibrant, space-themed banner with the text "DevOps" in large white letters, "aws" in the Amazon logo, and "PIERWSZE KROKI" in yellow. Below the banner, it says "Codziennie nowe posty !!!". The post is titled "Pierwsze kroki w Procesach Cloud DevOps i AI" and has buttons for "Zaprosz" and "Udostępnij". Below the title, there are tabs for "Dyskusja", "Wydarzenia", "Multimedia", "Pliki", and "Osoby". The post content area shows a text input field "Napisz coś...", a "Zdjęcie/film" button, and an "Ankieta" button. On the right side, there is an "Informacje" section with social media links and group settings.

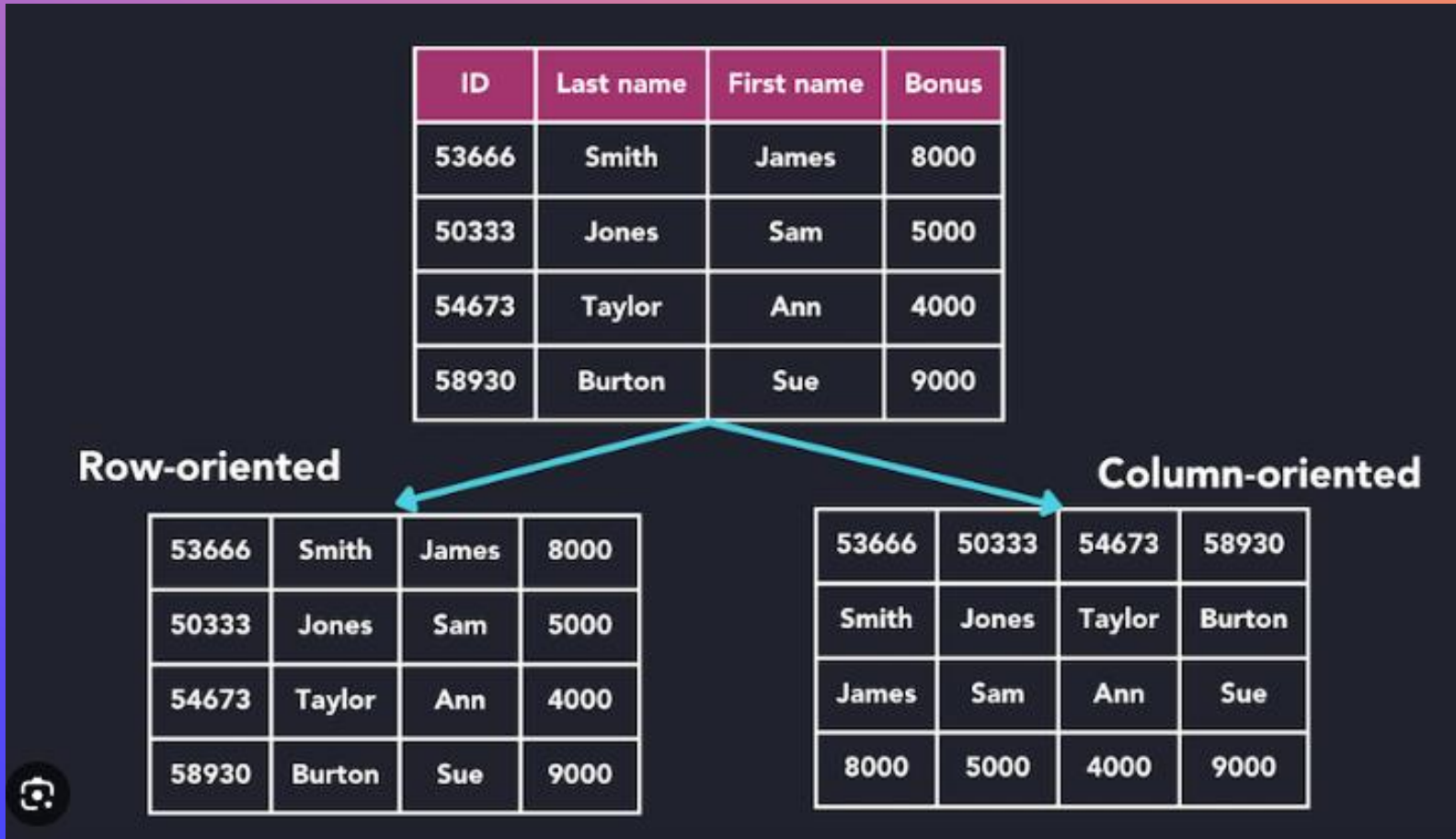
The screenshot shows the Facebook profile of Kasper Kalfas. The profile picture is a circular portrait of a man with a beard. The cover photo is a collage of four images, each with the text "It's me &". The images show a man standing on a balcony overlooking a beach, a man in a red jacket standing in a field of orange flowers, a man in a blue shirt standing on a balcony overlooking a city, and a man in a blue suit standing in a colorful, abstract environment. The profile bio reads "Kasper Kalfas" and "Cloud Architect | AWS, Azure, GCP | Sicily lover". Below the bio, there are tags for "CloudState" and "Politechnika Opolska". The profile also shows a list of interests, including "Cloud Architect", "AWS", "Azure", "GCP", "Sicily lover", and "Politechnika Opolska".



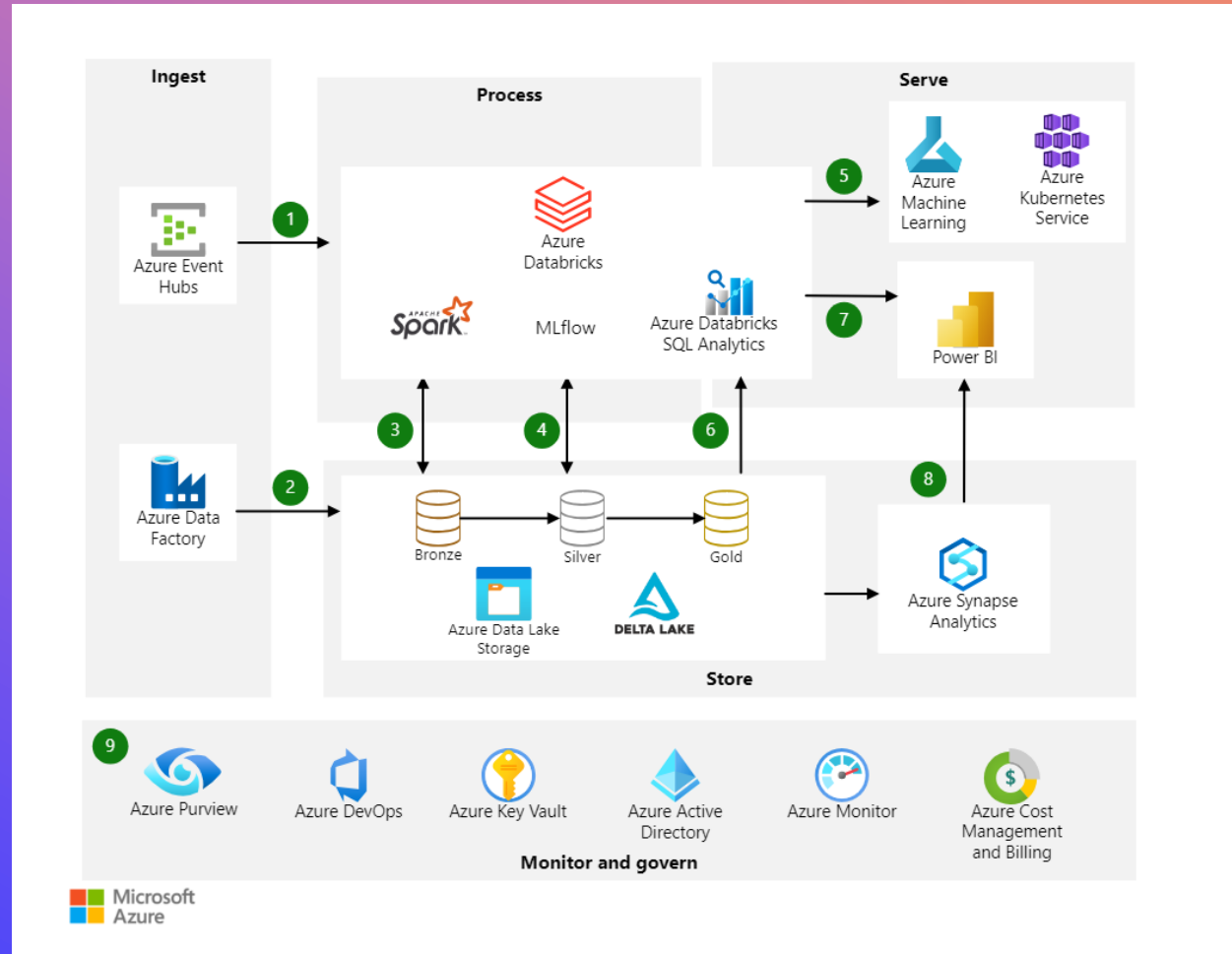
DATA PROJECT



ZORIENTOWANIE DANYCH



DATA PROJECT IN CLOUD



Przechowywanie danych

Azure Blob storage - \$21.00/msc



Azure SQL \$400.19/msc



Azure DB VM ok. 500\$/msc

Azure Blob Storage

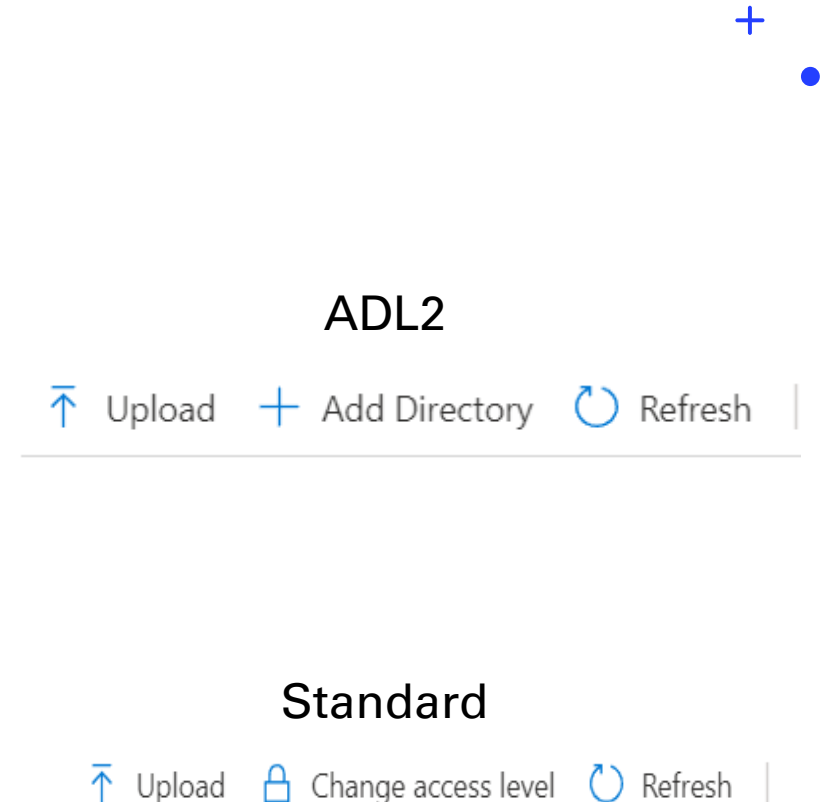
• **Definition:** Azure Blob Storage is a scalable, cloud-based object storage solution for unstructured data. It efficiently handles data of any type or size, from small documents to large media files.

• **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



Hot

Online

Suits frequently accessed data.

High storage costs, but low access costs.



Cool

Online

Suits infrequently accessed data.

Lower storage costs, but higher access costs.

Fee if deleted/moved tier earlier than 45 days.



Archive

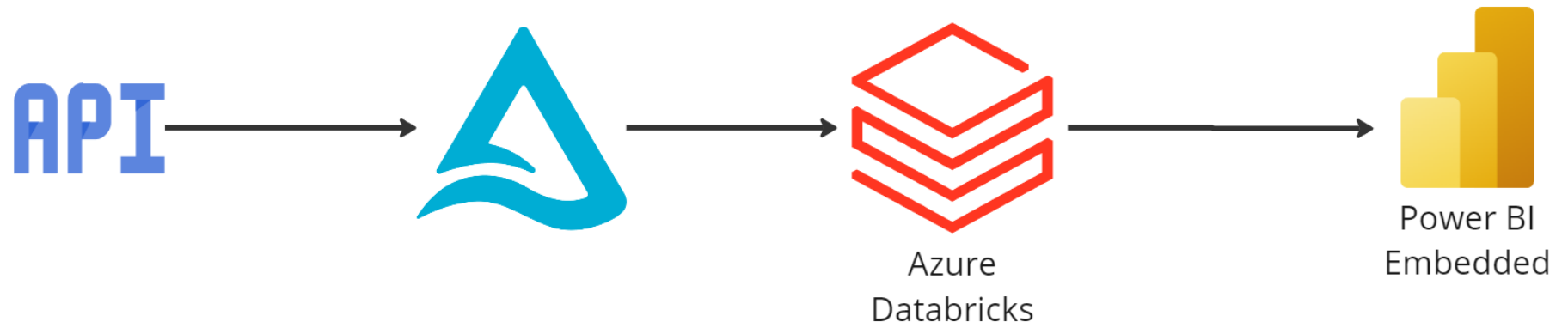
Offline

Suits rarely accessed data.

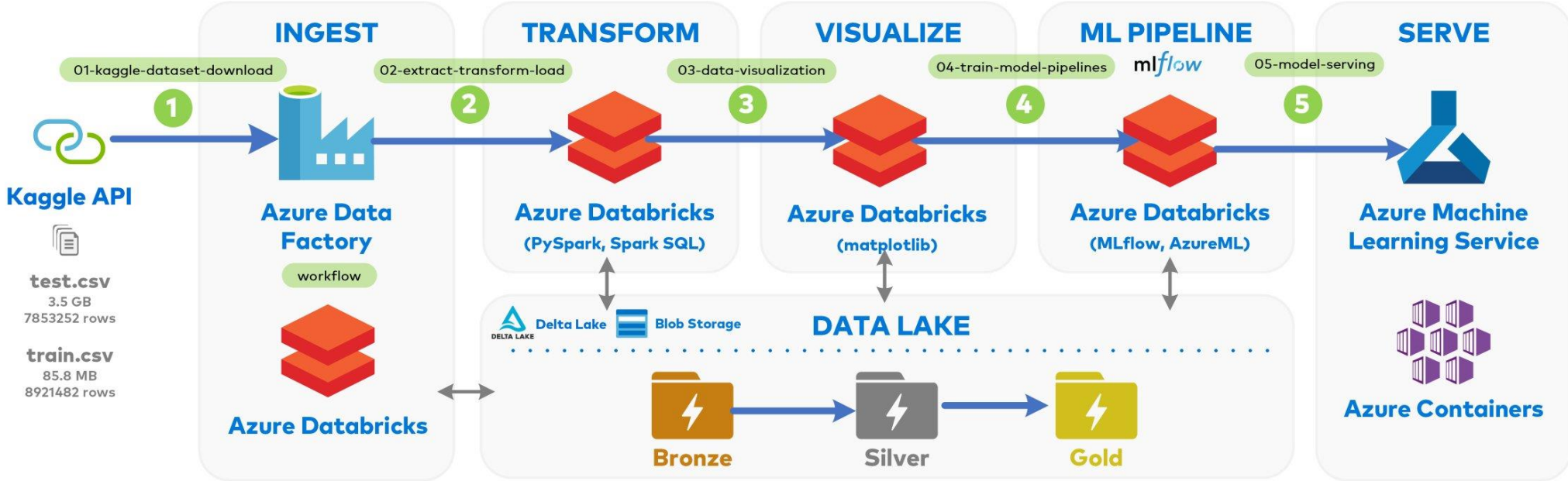
Lower storage costs, but higher access costs.

Fee if deleted/moved tier earlier than 180 days.

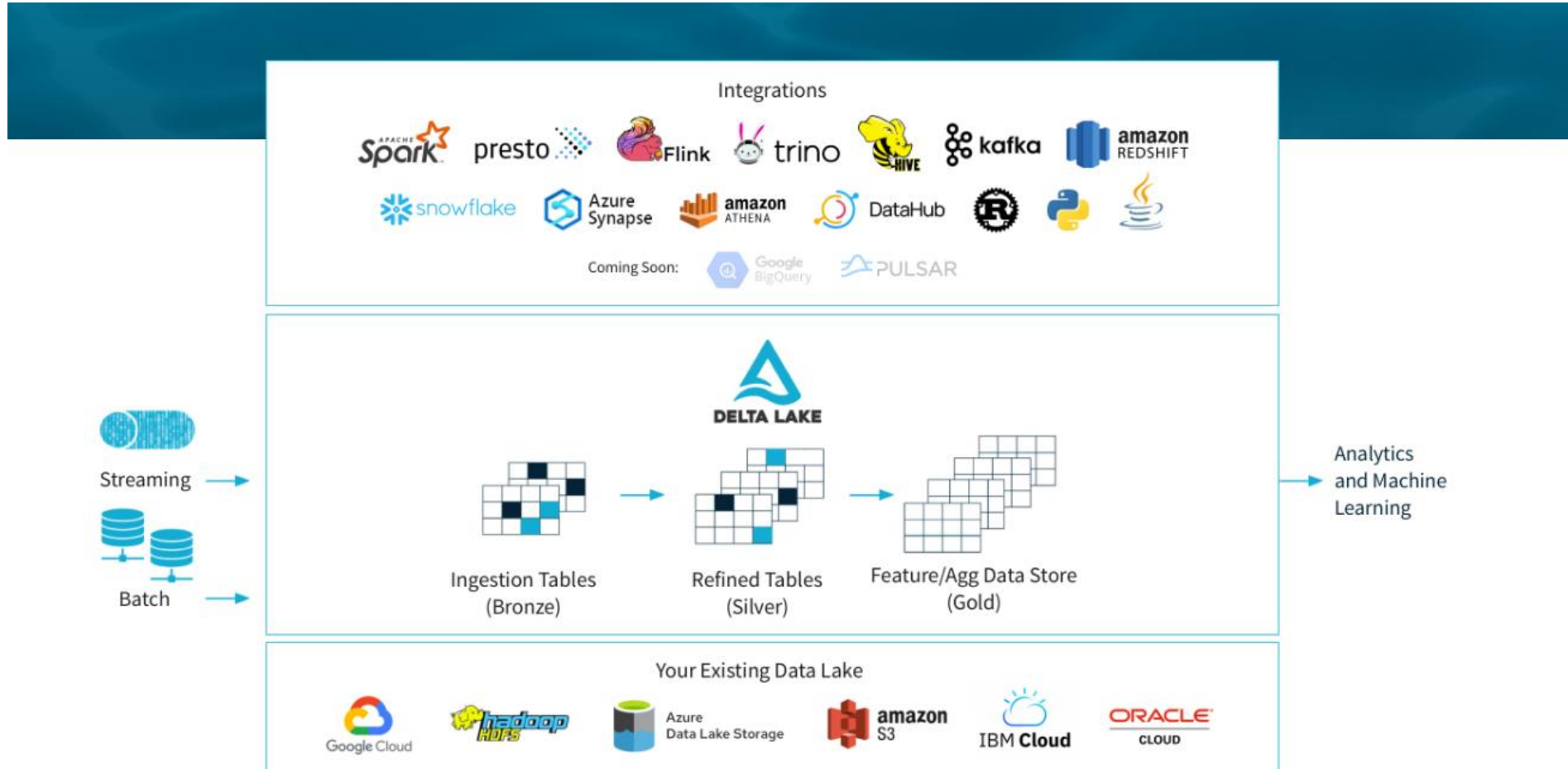
Delta Lake



Architecture




Delta Lake





Delta Lake


[Home](#) > [gsfdatalake](#) | [Containers](#) >

 **sensor-data** ...
Container


<<


 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](#)

Name	Modified
<input type="checkbox"/>  [-]	
<input type="checkbox"/>  _delta_log	
<input type="checkbox"/>  year=2018	
<input type="checkbox"/>  year=2019	
<input type="checkbox"/>  year=2020	
<input type="checkbox"/>  year=2021	
<input type="checkbox"/>  year=2022	
<input type="checkbox"/>  year=2023	

Delta Lake (parquet)

[↑ 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](#) / [year=2018](#) / [month=12](#)

Search blobs by prefix (case-sensitive)

Name	Modified	Access tier	Archive status
<input type="checkbox"/> 📁 [-]			
<input type="checkbox"/> 📄 part-00004-5f23dea4-1b2e-4e03-aa8b-b871887e4730.c000.snappy.parquet	4.10.2023, 14:43:42	Hot (Inferred)	
<input type="checkbox"/> 📄 part-00013-eafb7316-4efc-4f78-b8bf-3542f96d1975.c000.snappy.parquet	9.10.2023, 14:03:48	Hot (Inferred)	

Databricks

Microsoft Azure **databricks** ⌘ + 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

```
DROP TABLE wikicc
```

SQL

 Copy

```
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

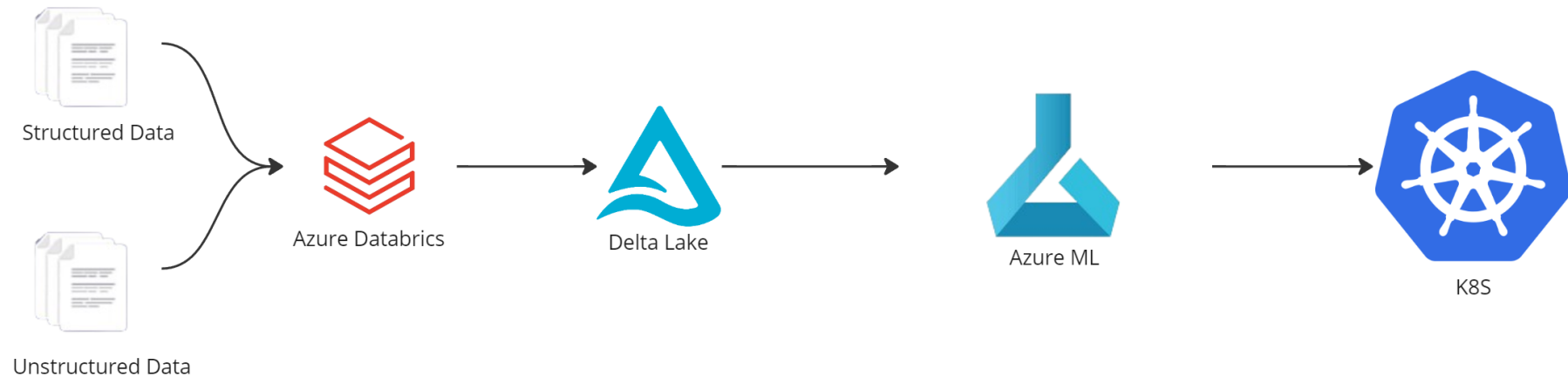
Apache Spark to silnik do przetwarzania danych. Zawiera całą masę bibliotek, których można używać do przetwarzania danych w klastrze komputerów.

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

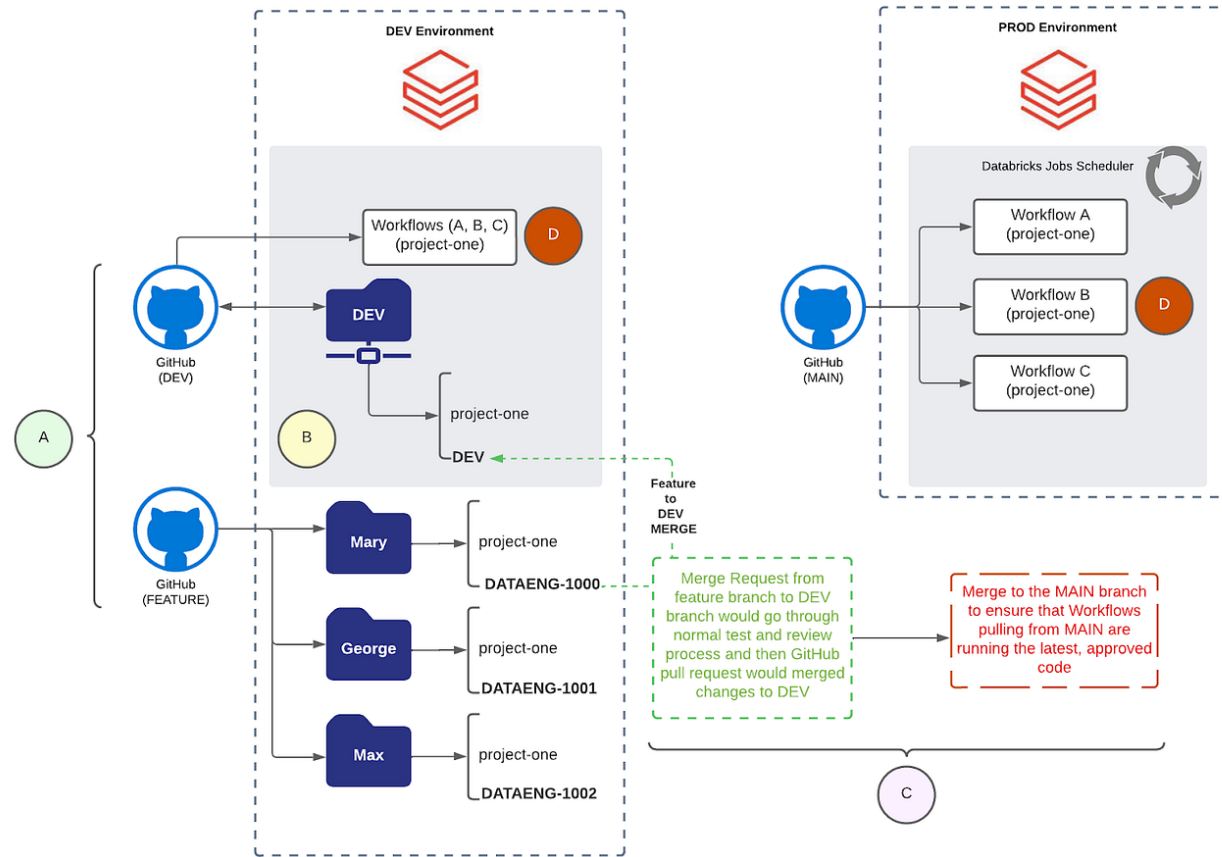
Obsługuje wiele języków programowania (**Python, SQL, Scala i R**). Możesz rozwiązywać problemy na wiele sposobów, używając SQL, transmisji danych („data streaming”) i uczenia maszynowego.



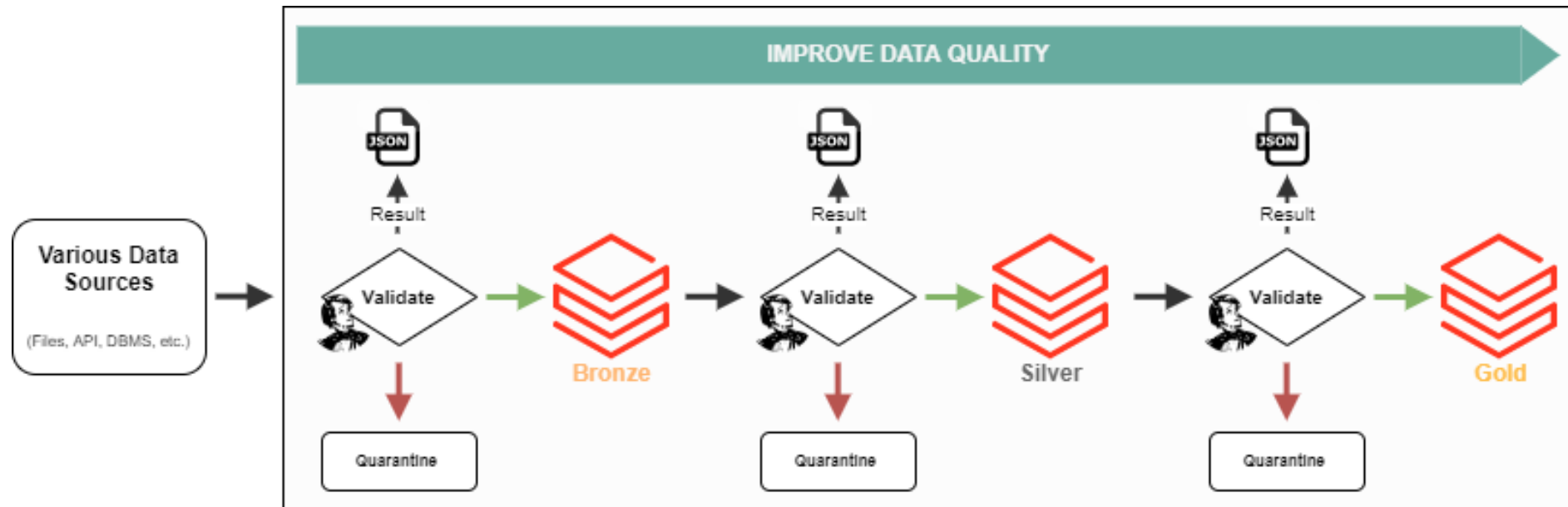
Developing process of ETL



Developing process of ETL



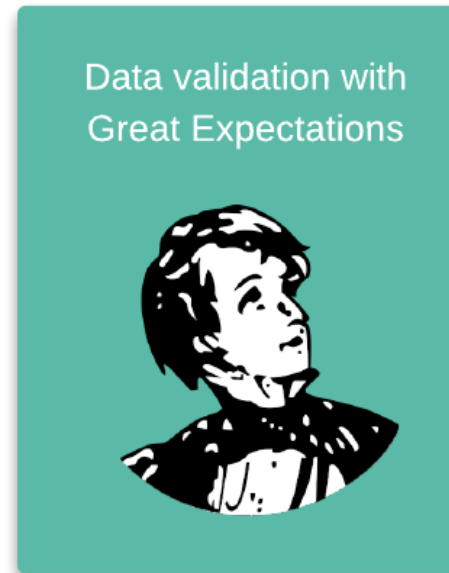
Jakość i niezawodność danych



Jakość i niezawodność danych



Your data assets:
database tables, flat
files, dataframes...



High quality data in
your data products



Data documentation
& data quality reports



Logging & alerting

Jakość i niezawodność danych

< re GREAT_EXPECTATIONS SAMPLE GREAT_EXPECTATIONS Updated 2 months ago View history Invite

great expectations

Data Docs autogenerated using Great Expectations.

Data Docs | local_site

Validation Results [Expectation Suites](#)

Search Clear Filters

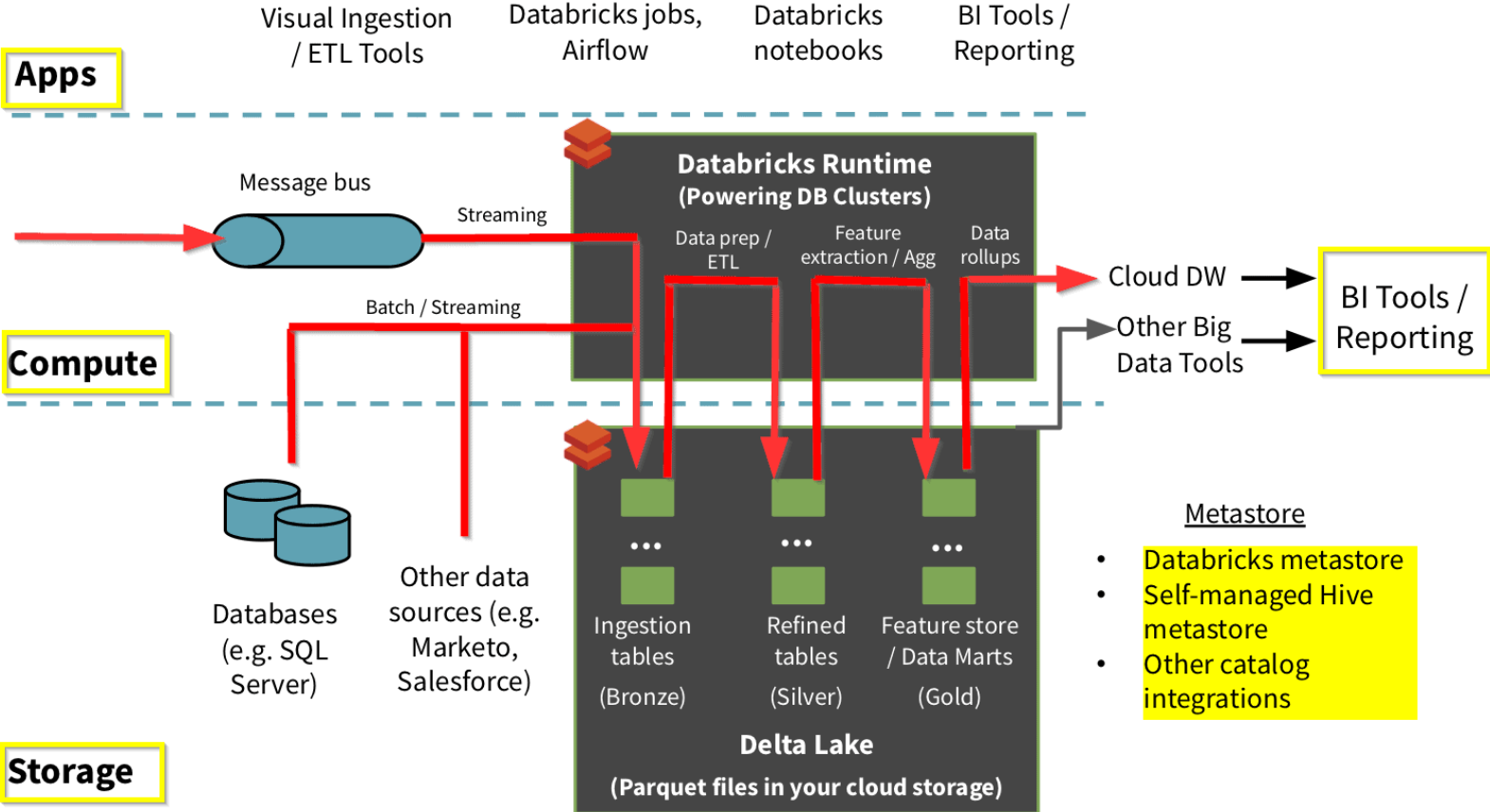
Status	Run Time	Run Name	Asset Name	Batch ID	Expectation Suite
✘	2022-08-10 19:35:56 CEST	20220810-173556-my-run-name-template		444fa93fe34e9e162c5f910bca5b5916	getting_started_expectation_suite_taxi
✔	2022-08-10 19:34:48 CEST	20220810-173448-my-run-name-template		3aa0a5a68a2bbe5abd3b08ea9739616c	getting_started_expectation_suite_taxi
✔	2022-08-10 19:31:46 CEST	__none__		3aa0a5a68a2bbe5abd3b08ea9739616c	getting_started_expectation_suite_taxi
✔	2022-08-10 19:28:14 CEST	__none__		3aa0a5a68a2bbe5abd3b08ea9739616c	getting_started_expectation_suite_taxi

Showing 1 to 4 of 4 rows

Actions

Show Walkthrough

Jakość i niezawodność danych



Skalowalność

Compute > New compute > UI preview [Send feedback](#)

Cluster [✎](#)

Unrestricted [▼](#)

Multi node Single node

Access mode [?](#) Single user access [?](#)

Single user [▼](#)

Performance

Databricks runtime version [?](#)

Runtime: 13.3 LTS (Scala 2.12, Spark 3.4.1) [▼](#)

Use Photon Acceleration [?](#)

Worker type [?](#)

Min workers: Max workers: Spot instances [?](#)

General purpose

Standard_DS3_v2	14 GB Memory, 4 Cores
Standard_DS4_v2	28 GB Memory, 8 Cores
Standard_DS5_v2	56 GB Memory, 16 Cores
Standard_D4s_v3	16 GB Memory, 4 Cores
Standard_D8s_v3	32 GB Memory, 8 Cores

[46 more](#) [Add](#)

General purpose (HDD)

Summary

2-8 Workers 28-112 GB Memory
8-32 Cores

1 Driver 14 GB Memory, 4 Cores

Runtime 13.3.x-scala2.12

[Photon](#) [Standard_DS3_v2](#) [4-14 DBU/h](#)

Monitoring of ETL



Notifications ⓘ

No notifications

Edit notifications


Add email addresses or webhooks to notify when runs of this job begin, complete, or this job fails.

Monitoring of ETL



A job run has terminated with the error:

Message


 Task failed. This caused all downstream tasks to get skipped.

Run details

Workspace

Job

Job run

Status	 Failed
Started at	2023-08-22 04:53:51 UTC
Duration	6m 32s
Launched	Manually

[View run in Databricks >](#)

cluster Report at Mon, 18 Sep 2023 14:00:02 +0000 [Get Fresh Data](#)

Last or from to

Timezone:

[Physical View](#)

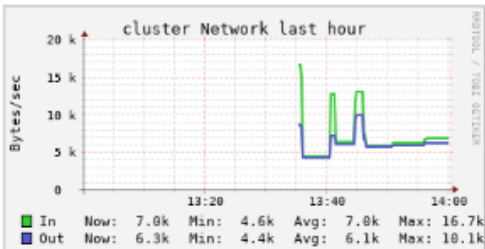
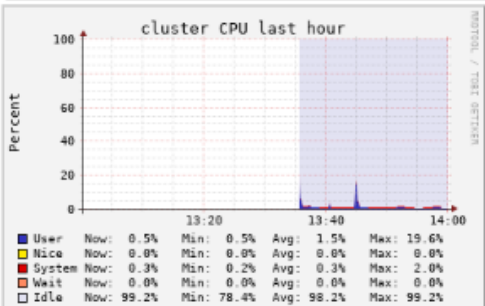
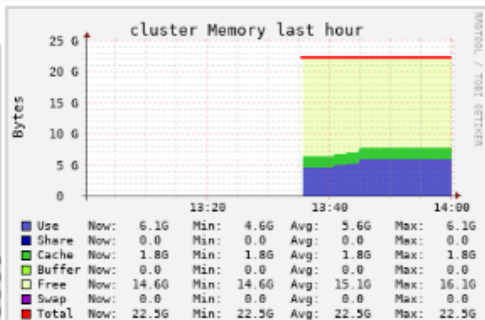
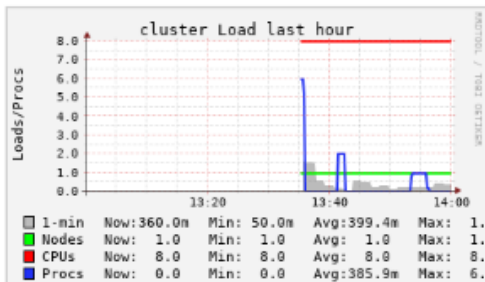
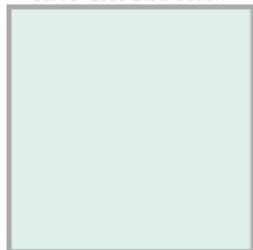
[Grid](#) > [cluster](#) >

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

CPU's 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



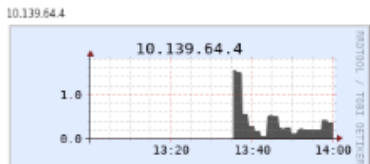
Stacked Graph - load_one

cluster **load_one** last hour sorted by name

Metric:

Show Hosts Scaled: Size: Columns: (0 = metric + reports)

Show only nodes matching Max graphs to show: Sorted:



(Nodes colored by 1-minute load) | [Legend](#)

Skalowalność

Compute > New compute > UI preview [Send feedback](#)

Cluster [✎](#)

Unrestricted [▼](#)

Multi node Single node

Access mode [?](#) Single user access [?](#)

Single user [▼](#)

Performance

Databricks runtime version [?](#)

Runtime: 13.3 LTS (Scala 2.12, Spark 3.4.1) [▼](#)

Use Photon Acceleration [?](#)

Worker type [?](#)

Min workers: Max workers: Spot instances [?](#)

General purpose

Standard_DS3_v2	14 GB Memory, 4 Cores
Standard_DS4_v2	28 GB Memory, 8 Cores
Standard_DS5_v2	56 GB Memory, 16 Cores
Standard_D4s_v3	16 GB Memory, 4 Cores
Standard_D8s_v3	32 GB Memory, 8 Cores

[46 more](#) [Add](#)

General purpose (HDD)

Summary

UI | [JSON](#)

2-8 Workers 28-112 GB Memory
8-32 Cores

1 Driver 14 GB Memory, 4 Cores

Runtime 13.3.x-scala2.12

[Photon](#) [Standard_DS3_v2](#) [4-14 DBU/h](#)

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 hostować?

- **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 hostować?

- **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 hostować?



- **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 hostować?



- **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:

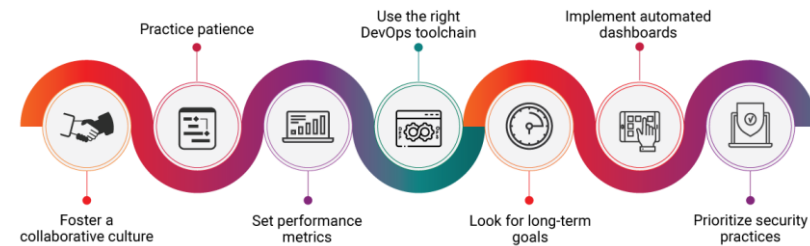
Tools

Concepts

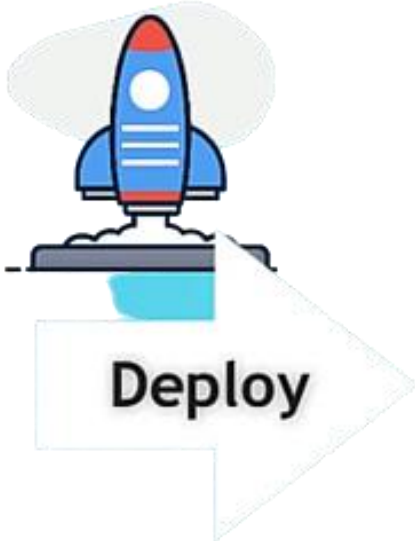
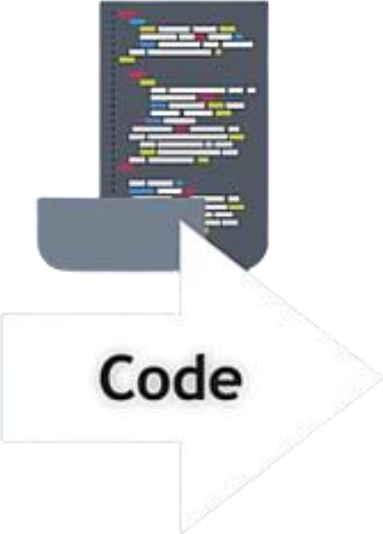
Practices



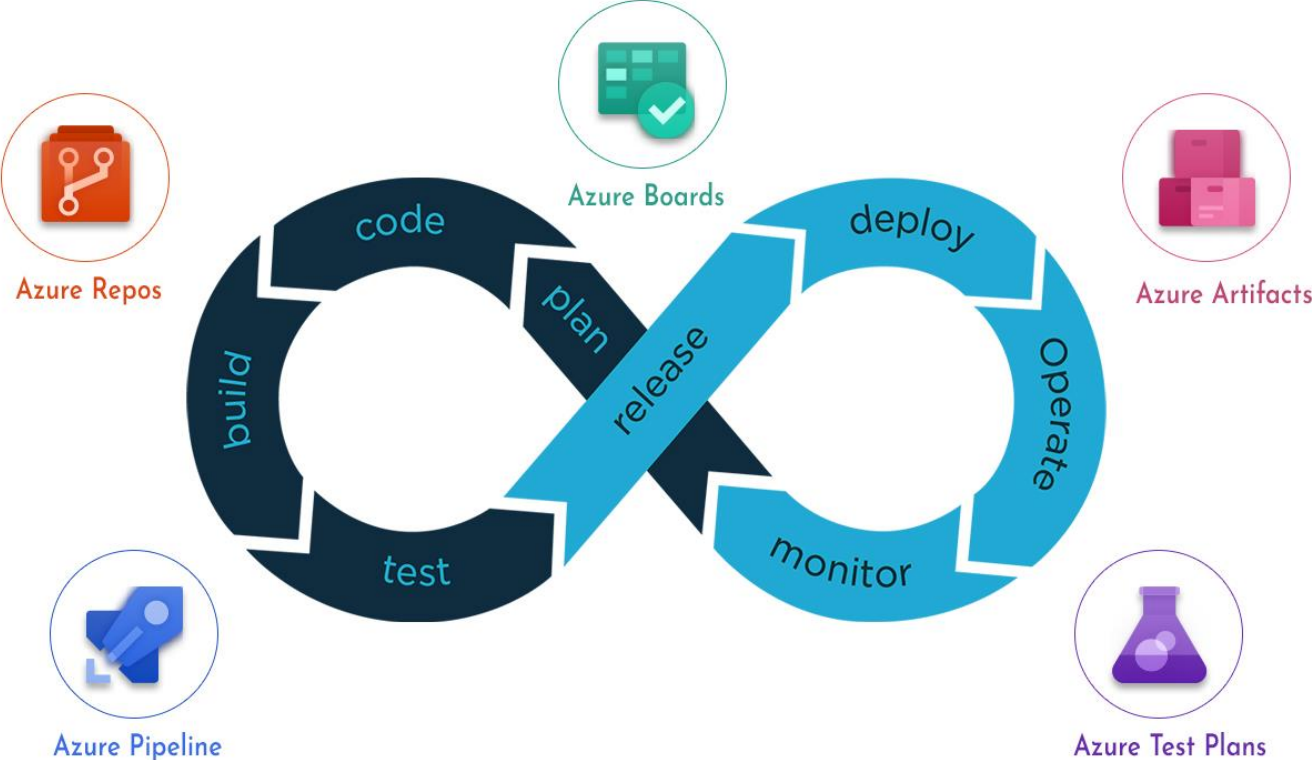
DEVOPS BEST PRACTICES TO FOLLOW



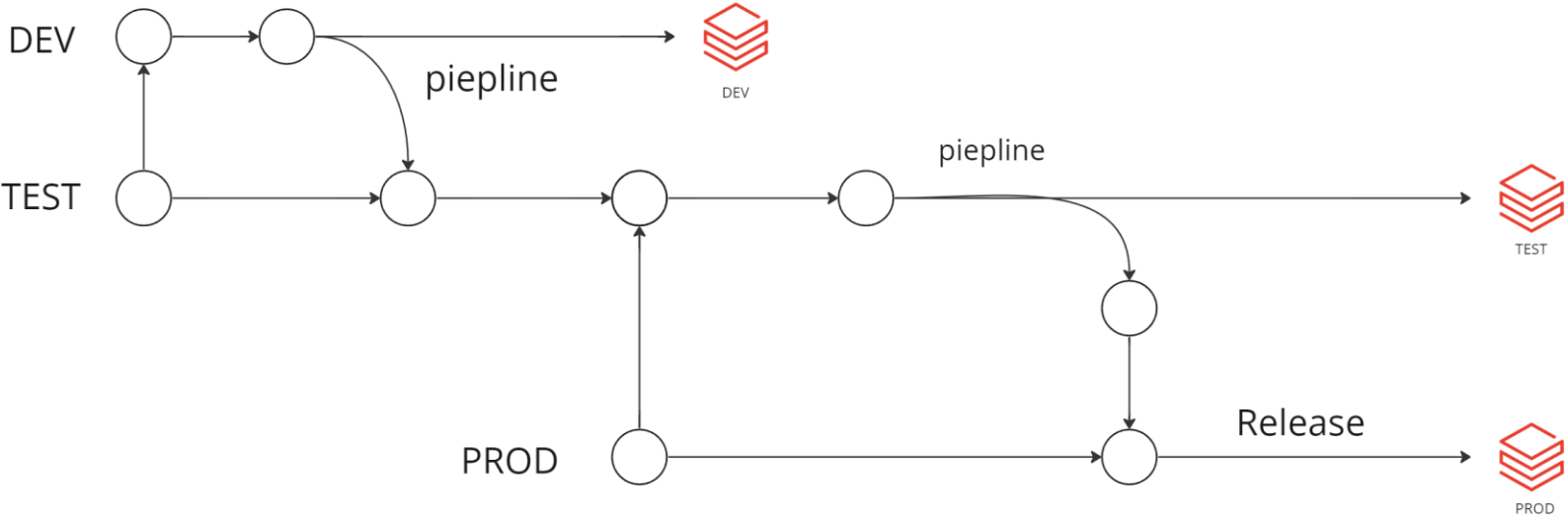
Developing process of ETL



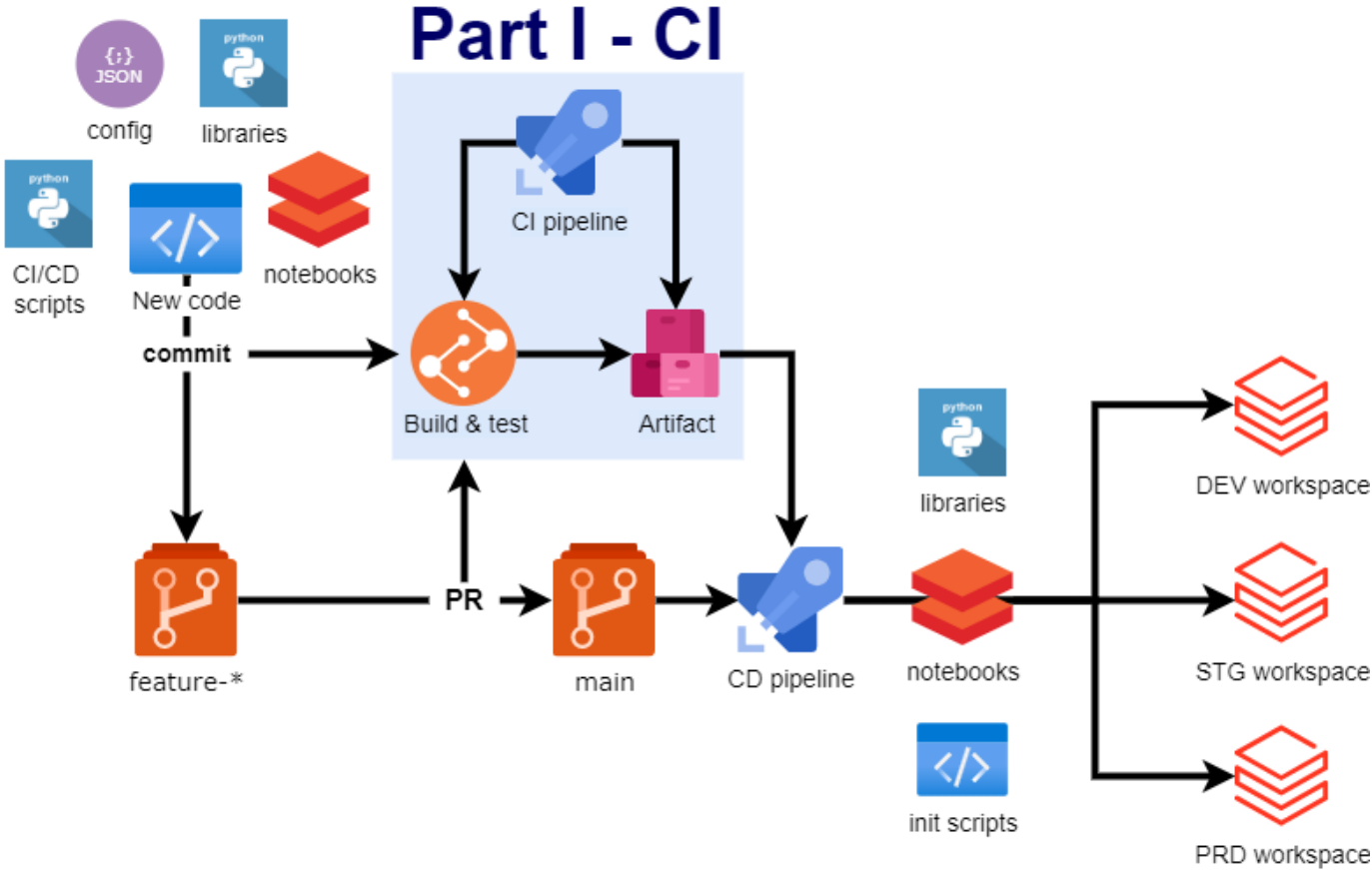
Developing process of ETL



Developing process of ETL



Developing process of ETL



Kolaboracja między zespołami

SCRUM PROCESS

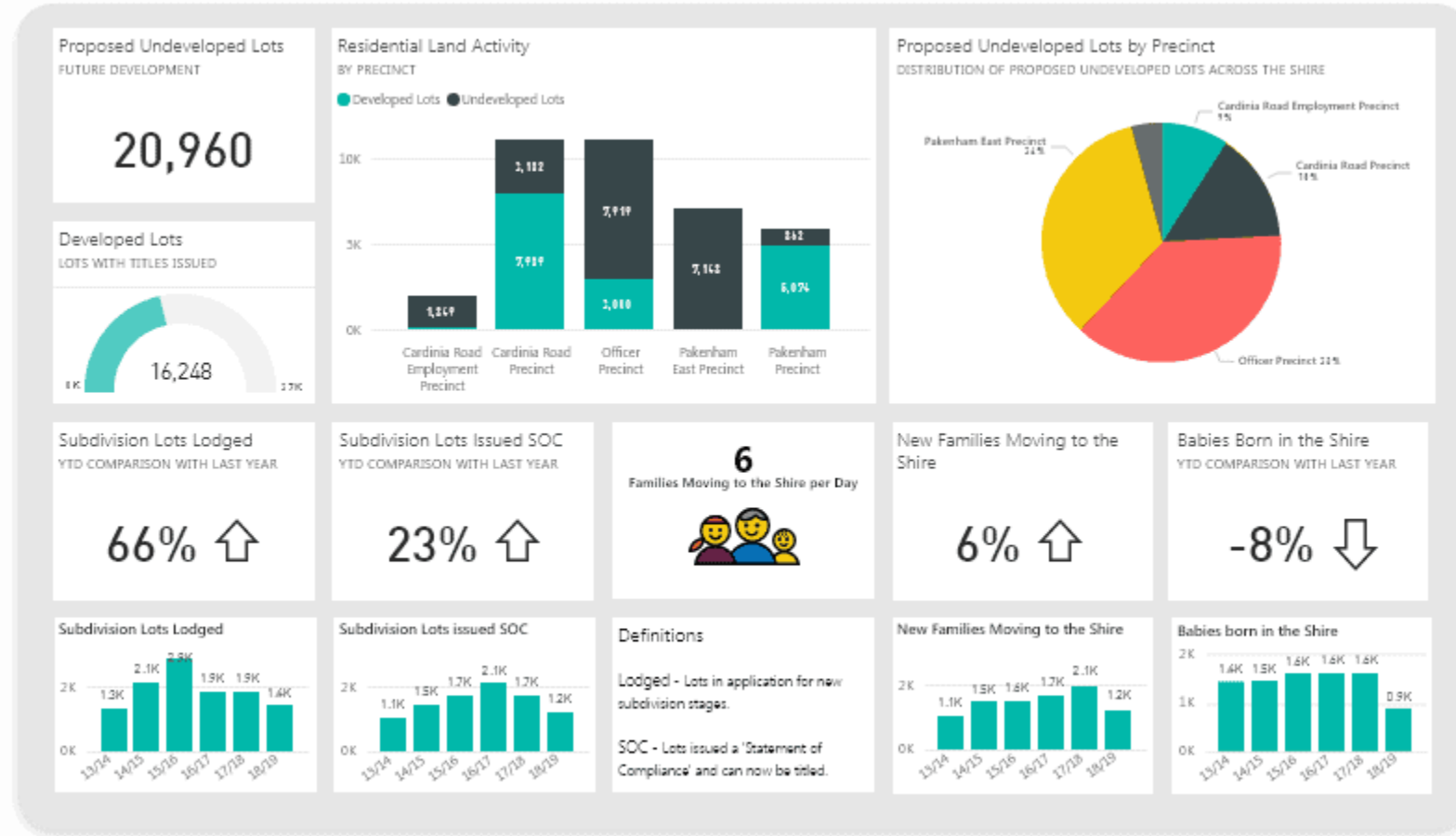


Szybsza adaptacja do zmian:

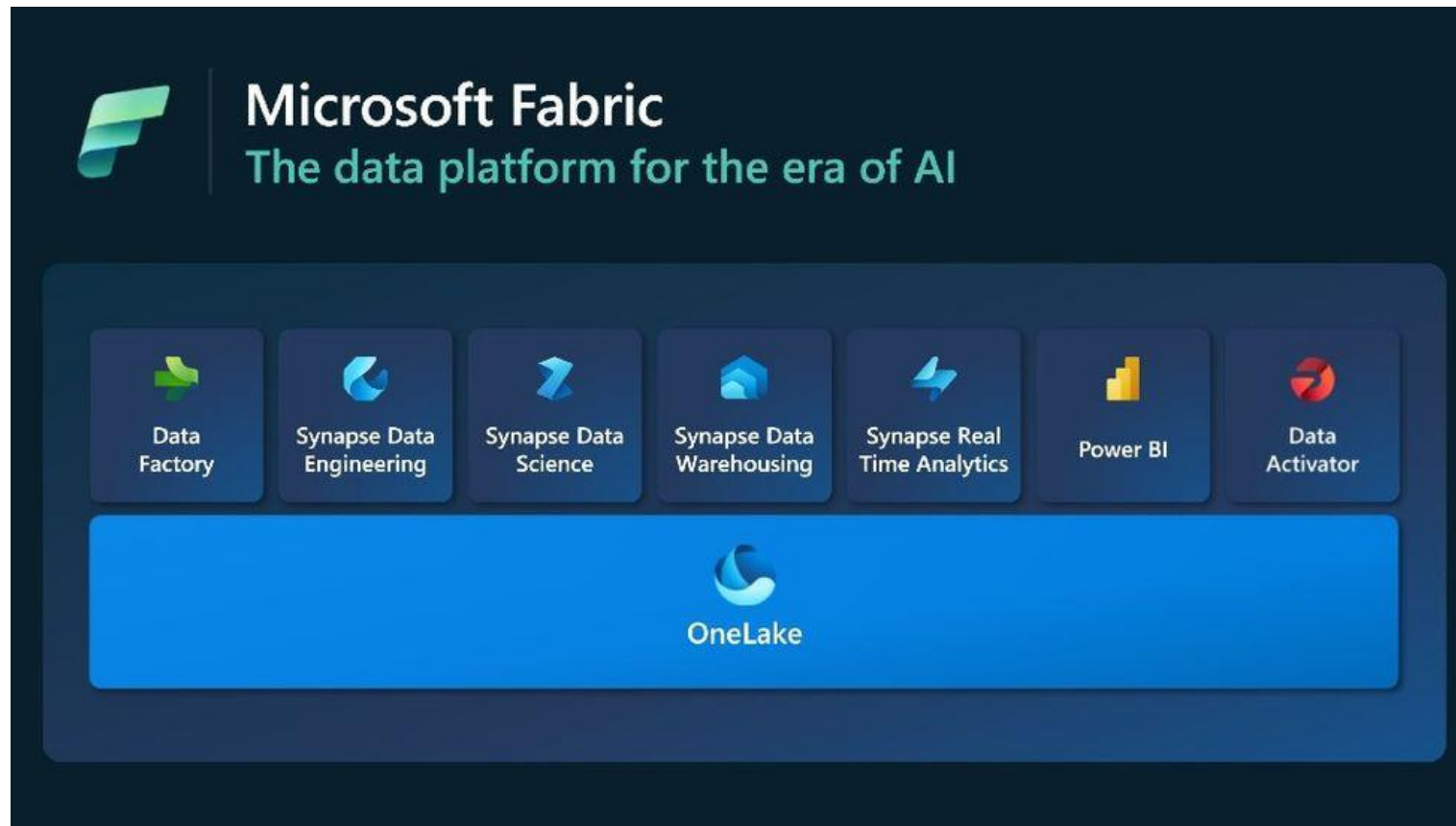


Power BI

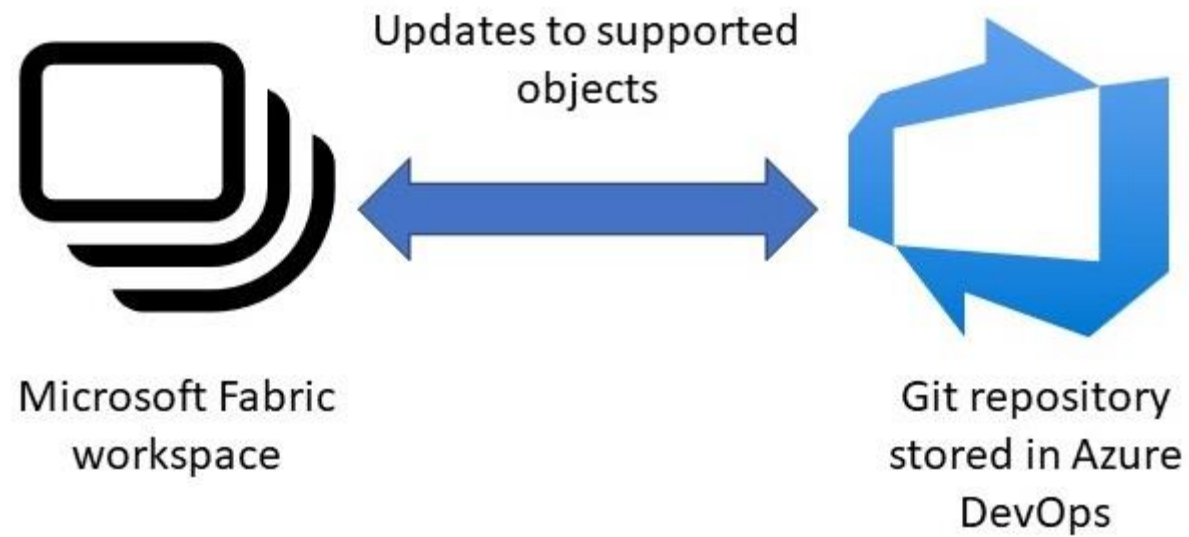
Szybsza adaptacja do zmian:



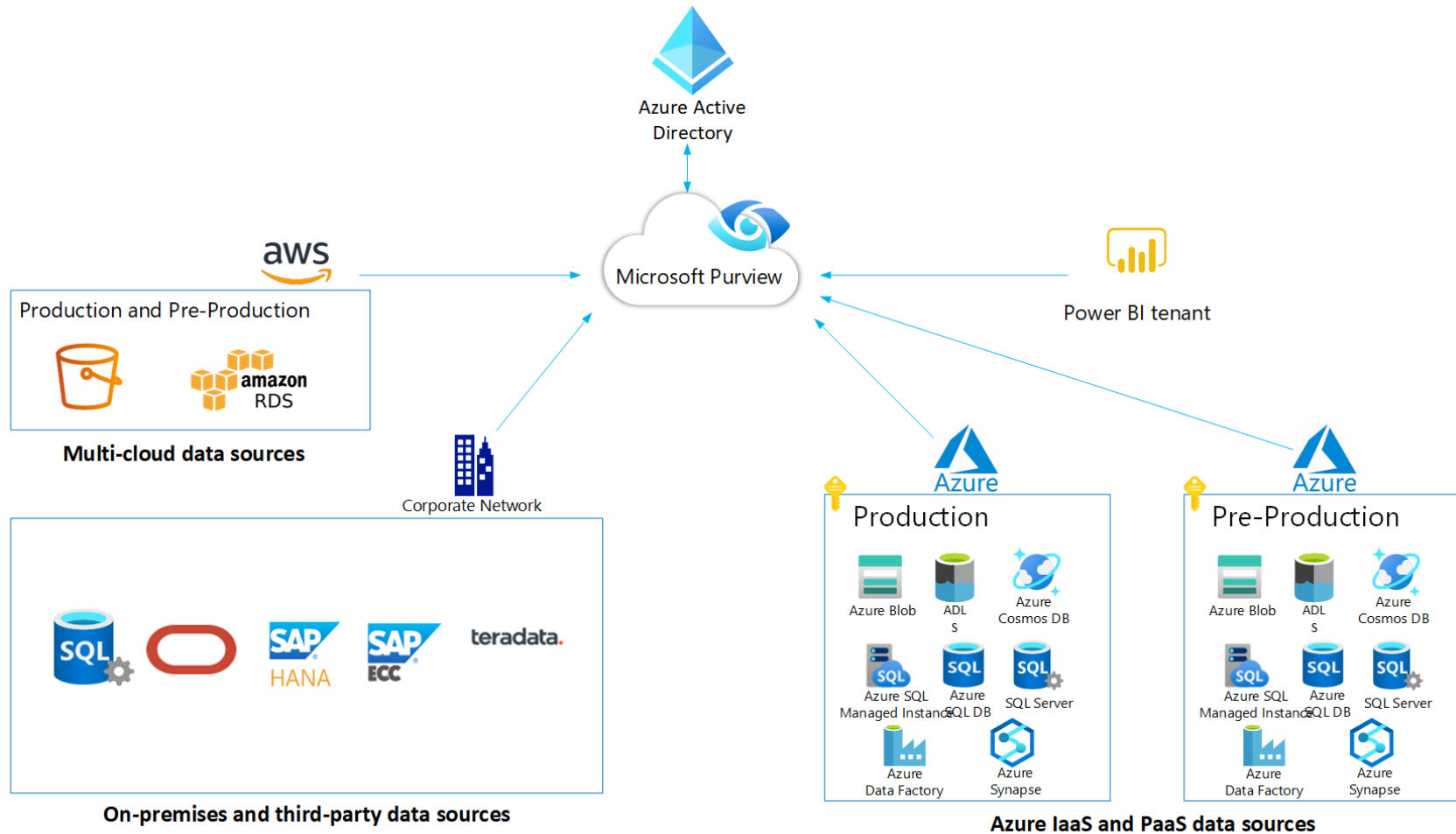
Szybsza adaptacja do zmian:



Szybsza adaptacja do zmian:



Azure Pureview



Azure Pureview

The screenshot displays the Microsoft Azure Purview interface for 'Adatum Corp'. The top navigation bar includes the Microsoft Azure logo, the user's email 'contoso@contoso.com', and the Microsoft logo. The main header shows 'Purview > Adatum Corp' and a search bar containing 'Revenue'. Below the header, the 'Sources' section is active, showing options to 'Register', '+ New collection', and 'Refresh'. A status bar indicates 'Showing 5 collections, 1134 sources' and a 'Map view' dropdown.

The interface is organized into five collection boxes, each with a grid icon, a plus sign, a pencil icon, and a minus sign:

- NorthAmericaDataCenter** Collection: Contains 'OnPremSQLServer-Fina...' (SQL Server), 'Teradata-FinanceData' (Teradata (Preview)), 'HiveMetastore' (Hive Metastore (Preview)), 'FinanceSQLServer' (SQL Server), 'Teradata' (Teradata (Preview)), and 'OnPremSQLServer' (SQL Server).
- EuropeDataCenter** Collection: Contains 'SAP-S4HANA-Procurem...' (SAP S/4Hana (Preview)), 'SAP-ECC-SalesData' (SAP ECC (Preview)), 'SAP-S4HANA' (SAP S/4Hana (Preview)), and 'SAP-ECC' (SAP ECC (Preview)).
- AzureAndBINorthAmerica** Collection: Contains 'AzureDataLakeStorage-...' (Azure Data Lake Storage Gen2), 'AzureBlobStorage' (Azure Blob Storage), 'AzureSQLDB-SalesInvoi...' (Azure SQL Database), 'RevenuePBIDashboards' (Power BI), 'WebLogs' (Azure Files), and 'AzureSqlManagedInsta...' (Azure SQL Database Managed Instance).
- AmazonNorthAmerica** Collection: Contains 'AmazonS3-HRData' (Amazon S3) and 'AWSS3' (Amazon S3).
- AzureEurope** Collection: Contains 'AzureDataLakeStorage-...' (Azure Data Lake Storage Gen2).

Each source card includes a 'View details' link. A dashed box highlights the 'AmazonNorthAmerica' and 'AzureEurope' collections and their associated sources. A small code editor window is visible in the bottom right corner.

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

+

o

.

Q_NA