Ailio

DuckDB x Data Lakehouse

- Wie funktioniert ein DLH/DWH?
- Wozu Kataloge?
- Wodurch entstehen Kosten in der Visualisierung?
- Wie hilft DuckDB?

databricks

DuckDB

Streamlit

Das Bin Ich

Janis Gösser
CTO & Data Engineer @ Ailio
www.ailio.de

LinkedIn
https://de.linkedin.com/in/jgoesser

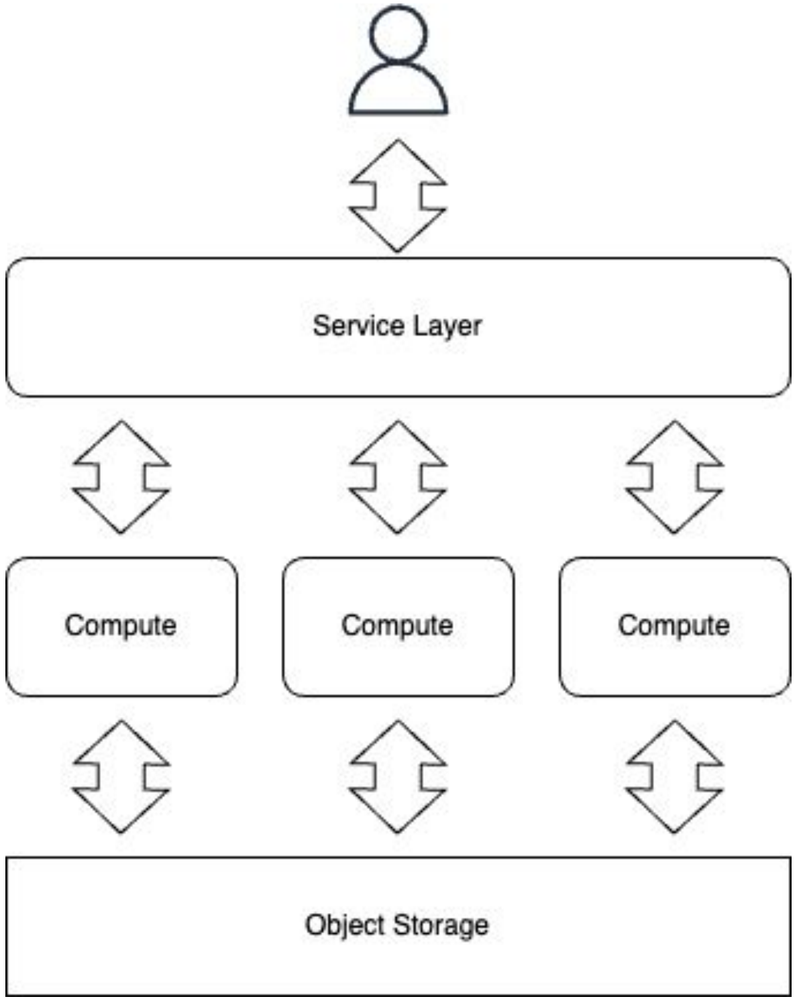
Ailionauten-Podcast
https://ailio.de/podcast/

Personal Website
https://jgoesser.me/

Multi-Layer -Strategie (klassisch)

DLH / DWHs haben typischerweise seperate Layer:

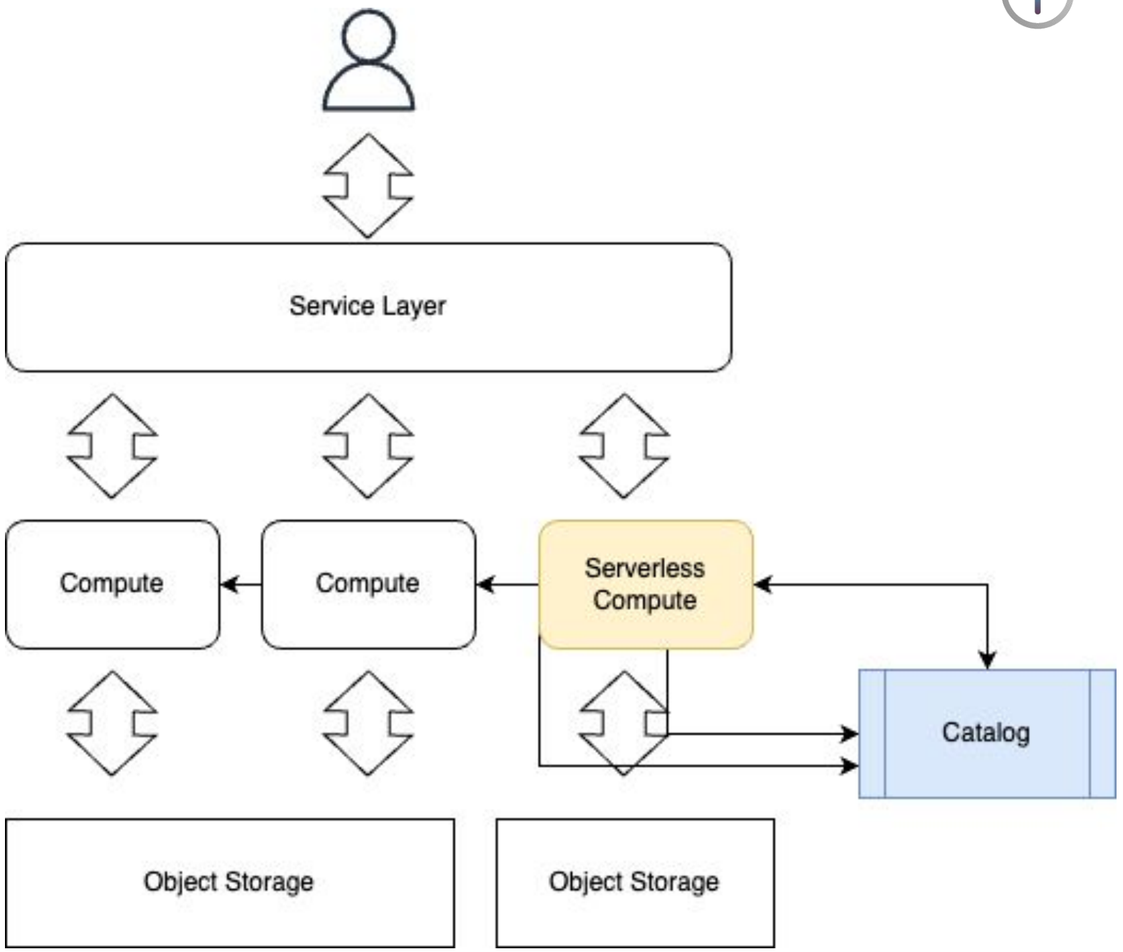
- **Service Layer** - mit dem Interagieren wir als User entweder via API, CLI oder UI
- **Compute Layer** - liefert beinahe beliebig skalierbare, parallelisierbare “Rechenpower” um Datenmengen zu verarbeiten
- **Storage** - meist ein ausgelagerter Object Storage (z.B. S3, Azure Blob Storage) auf den mit einem Open File Format (Hudi, Delta, Iceberg, CSV, Parquet, ...) geschrieben wird

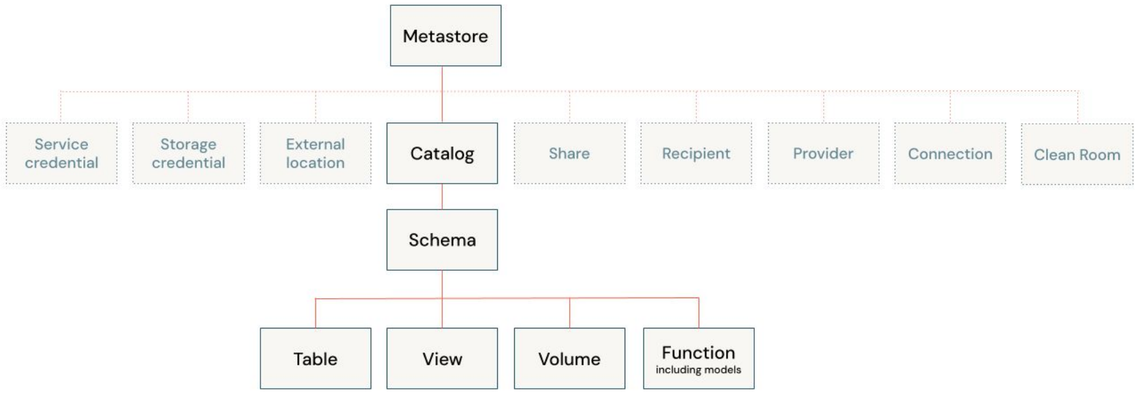


Multi-Layer -Strategie (moderner)

DLH / DWHs haben typischerweise seperate Layer:

- **Serverless Compute** - Vorteil: stark reduzierte Startup Zeiten, Nachteil: Daten verlassen den eigenen Compute
- **Catalog** - Behälter für Metainformationen von Tabellen und vielem mehr





Wie kann lesend auf Daten zugegriffen werden?

Option 1: Serverless -> bei sensiblen Daten nicht ratsam

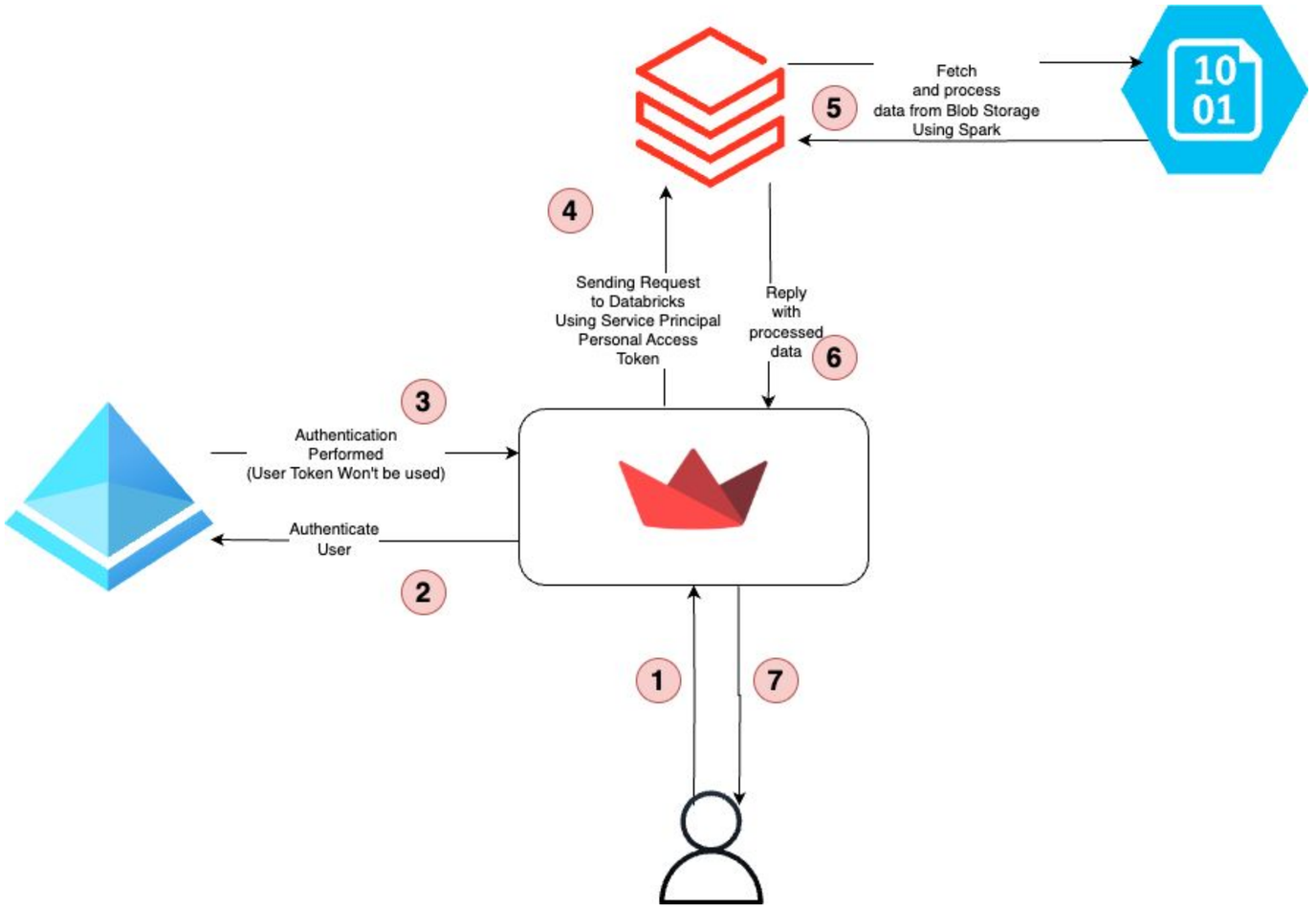
Option 2: Classic Compute -> hohe Startup Zeiten

Option 3: Daten direkt vom Object Storage lesen -> umgeht Data Governance

Option 4: DuckDB 👑

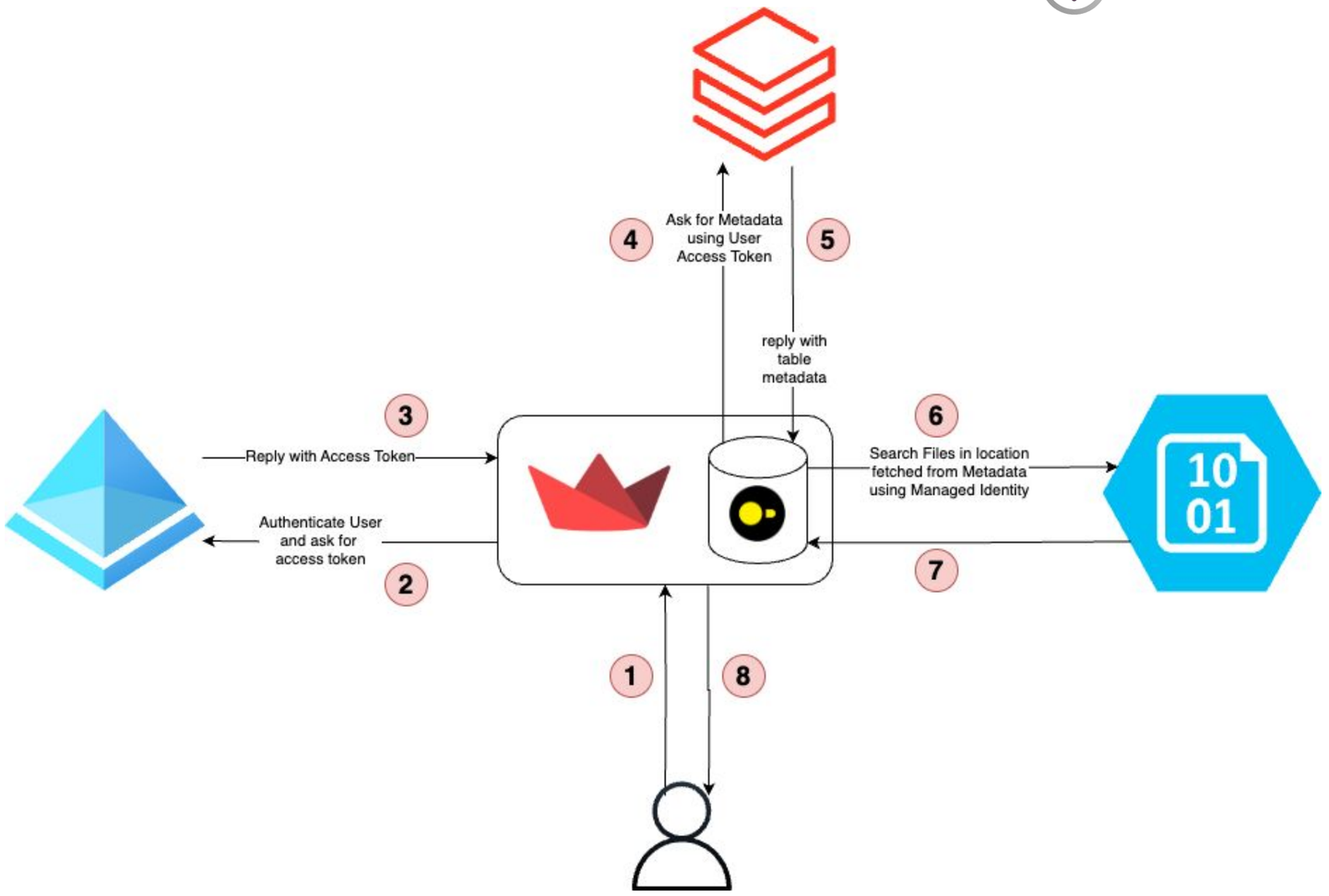
Visualisierung Option 1 / 2

- **SSO** – zur Dashboard Authentifizierung des Users nach Stramlit genutzt.
- **Technischer User** – in dieser Lösung verwendet die Streamlit Anwendung einen technischen User um Daten von DBX zu bekommen –> ***Governance geht womöglich partiell verloren***
- **Compute Endpunkt** – Wahlweise wird ein Cloud Compute oder Serverless Compute bereitgestellt. Der Datenzugriff und die Datenverarbeitung wird auf Databricks erledigt –> ***möglicherweise kostenintensiv***



Visualisierung Option 4

- **SSO** – zur Dashboard Authentifizierung des Users nach Stramlit genutzt.
- **UC Integration** –der Streamlit gibt den User an DuckDB (Azure Plugin) SSO authentifizierte User schickt Abfragen gegen den Unity Catalog (UC Plugin) –> ***Governance bleibt bestehen***
- **Blob Storage Zugriff** – dbx antwortet mit Metadaten. Diese Werden genutzt um mit DuckDB & SQL Daten beim Blob Storage zu lesen –> ***Compute Kosten werden gespart***



Fragen?