ISR
A CENIT COMPANY

WHITEPAPER

DAS DATA LAKEHOUSE

Ein Anwendungsfall in Databricks und SAP Datasphere

ISR Information Products AG

INHALTSVERZEICHNIS

1 | EINLEITUNG
DAS DATA-LAKEHOUSE-KONZEPT

Hintergrund und Relevanz

Im Zeitalter datengetriebener Geschäftsentscheidungen stehen Unternehmen vor der Herausforderung, große Mengen an Daten effizient zu verwalten und zu analysieren. Das Data-Lakehouse-Konzept verspricht die besten Eigenschaften von Data Lakes und Data Warehouses zu vereinen, um diesen Herausforderungen zu begegnen.

Ein klassisches Data Warehouse bietet hohe Performance und Datenqualität durch die Integration von Speicher- und Rechenressourcen, ist jedoch weniger flexibel bei der Aufnahme semi- und unstrukturierter Daten und verursacht hohe Kosten durch permanente Rechenressourcen. Ein Data Lake ermöglicht die kostengünstige, flexible Speicherung großer Datenmengen in Rohform, aber ohne die strukturierte Organisation und hohe Datenqualität eines Data Warehouses. Data Lakes sind daher insbesondere für Anwendungsfälle wie maschinelles Lernen und prädiktive Analysen geeignet.

Das Data-Lakehouse-Konzept vereint die Flexibilität eines Data Lakes mit der strukturierten Organisation und hohen Datenqualität eines Data Warehouses und zielt so darauf ab, datengetriebene Entscheidungsprozesse zu vereinfachen, die Flexibilität zu erhöhen und die Kosten zu senken. Durch die Trennung von Speicher- und Rechenressourcen werden Letztere nur bei Bedarf genutzt. Open-Table-Formate gewährleisten die Datenkonsistenz und -qualität. Funktionen wie ACID-kompatible Transaktionen, Versio-nierung, Schema-Enforcement und -Evolution sowie die Unterstützung für Batch- und Streaming-Verarbeitung bieten eine robuste Datenverwaltung. Die Medaillenstruktur (Bronze, Silver und Gold Layer) sorgt für eine klare Datenorganisation, die den Anforderungen sowohl von Data Scientists als auch von Data-Warehouse-Experten gerecht wird.

Das Data-Lakehouse-Konzept hat in den letzten Jahren erheblich an Popularität gewonnen und wird als zukunftsweisender Ansatz betrachtet, um große und komplexe Datenmengen effizient zu verwalten und zu analysieren. In diesem Whitepaper möchten wir Ihnen anhand eines anschaulichen Anwendungsbeispiels zeigen, wie das Data-Lakehouse-Konzept in Databricks und SAP Datasphere praktisch umgesetzt werden kann.

Zielsetzung und Methodik

Das Ziel dieses Anwendungsfalls ist es, die Daten effizient zu speichern, zu transformieren und zu analysieren. Wir werden die Daten im mehrschichtigen Ansatz basierend auf der Medaillenstruktur (s. Abb. 1) verarbeiten:

- **Bronze Layer:** Speicherung der Rohdaten in ihrer ursprünglichen Form.
- **Staging/Delta Layer:** Konvertierung und erste Transformation der Daten in das Delta-Lake-Format.
- **Silver Layer:** Anwendung der Data Vault 2.0 Modellierung zur weiteren Datenverarbeitung.
- **Gold Layer:** Veredelung der Daten in SAP Datasphere zur Erstellung eines Analytic Models und zur Visualisierung in einem SAP Analytics Cloud Dashboard.

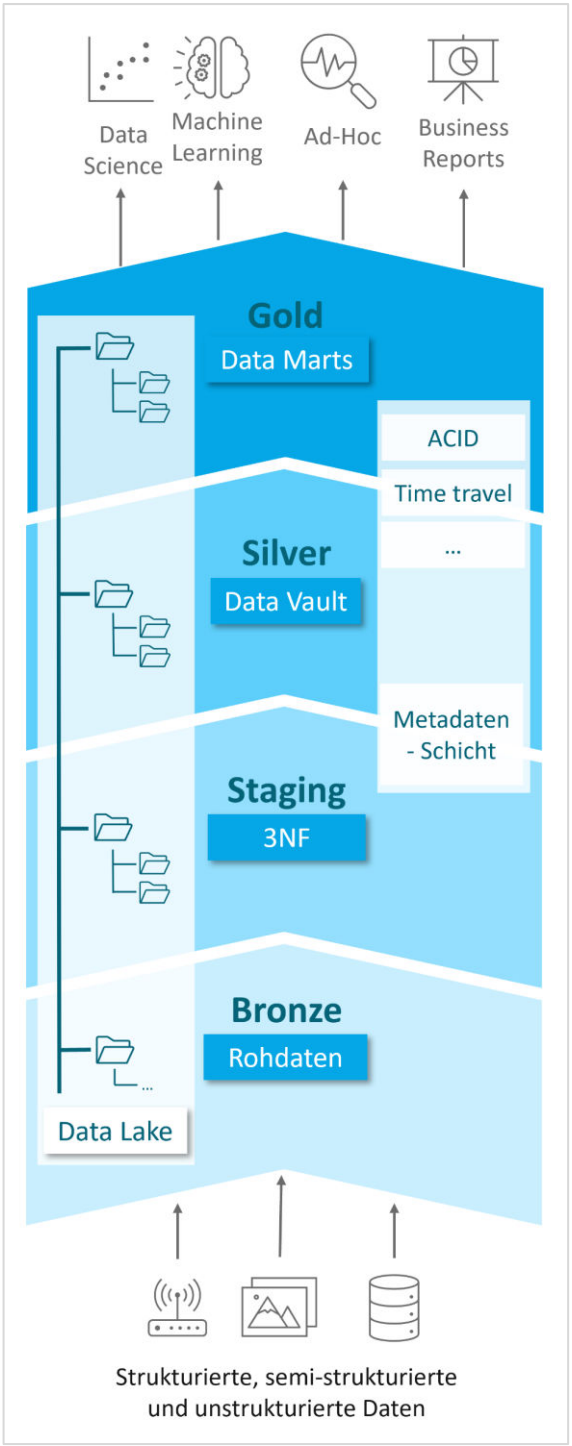

Abbildung 1: Die Medaillenstruktur des Data Lakehouse | ISR