

We need a Data Catalog

But which one?

ipt – Dr. Gerald Reif

Herausforderung bei der Produktauswahl

Dynamisch Produktkategorie

- Features der Produkte ändern sich laufend
- Laufende Veränderung am Markt
- Die Anforderungen ändern sich laufend

Hersteller mit unterschiedlichen Backgrounds

- Schwerpunktsetzung in den Produkten
- Entwickeln sich langsam zu all-in-one Produkten

Integration in die Datenarchitektur

- Integration in die Datenplattform
- Anbindung neuer Quellen
- Unterstützung für neue Services

Data Catalog

Stakeholder Anforderungen

Enabler für
Data-driven Enterprise

Data Analyst/Engineer/Scientist Stakeholder

- Suchen nach Daten für ihre Arbeit
- Informationen über Qualität und Aktualität der Daten
- Beziehung der Daten zu Personen, Applikationen und Organisationseinheiten
- Suche nach Synonymen (Thesaurus, Tags, Glossar)
- Self-Service Zugriff auf die Daten

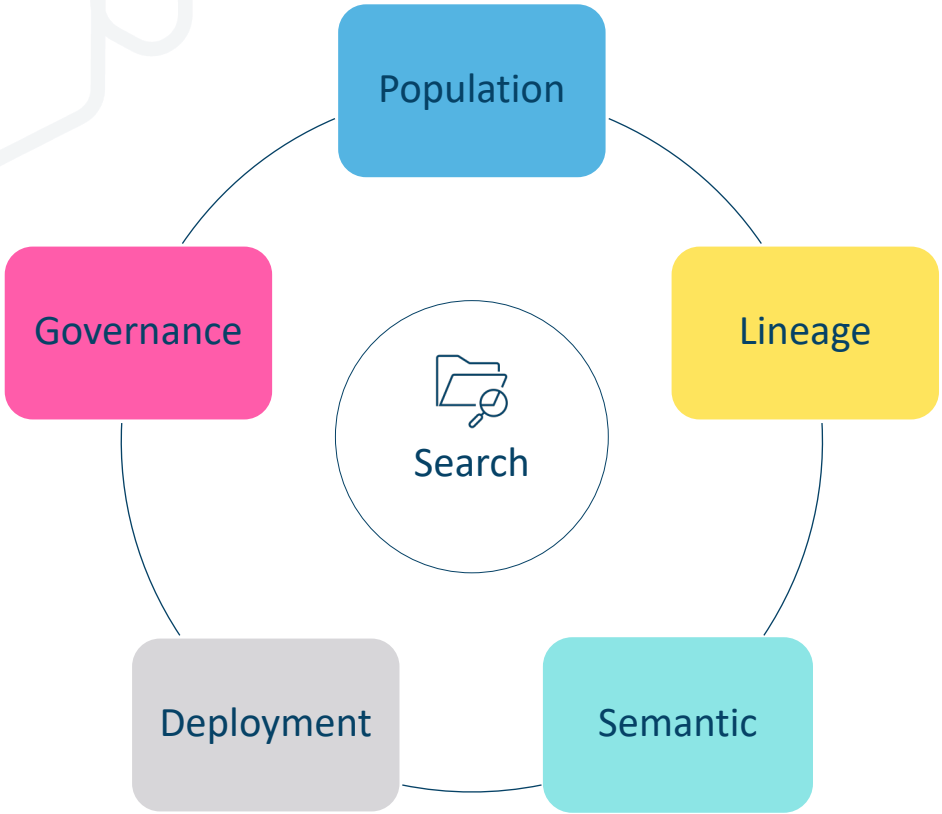
Data Governance Stakeholder

- Ownership (Verantwortung) für die Daten
- Geordnete Prozesse um Zugriff auf Daten zu bekommen
- Sicherstellen, dass gesetzliche und internen Regeln eingehalten werden
- Klassifikation von Daten zur Regelung des Zugriffs



Tech Talks: Data Governance
Dein Enabler für Business Value

5 Features of Data Catalogs





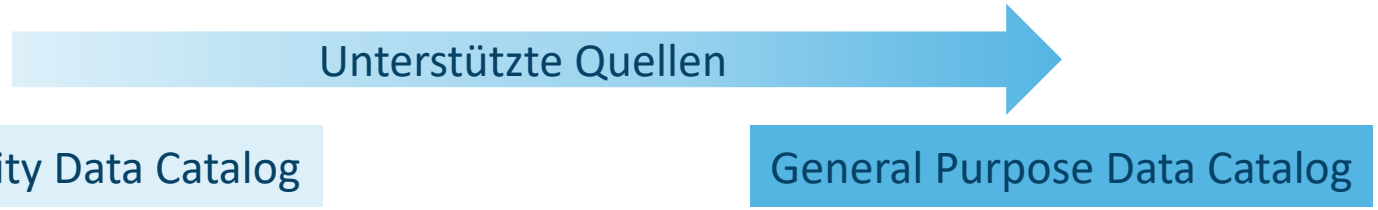
Population of a Data Catalog

Automatisiertes Erfassen der Metadaten (Tabellen, Spalten, Datentypen, Anzahl von Instanzen, letztes Update, etc.)

Crawling

Registrierte Datenquellen werden vom Data Catalog nach Metadaten durchsucht

- Unterstützt das Produkt die Datenquellen in meinem Unternehmen?



Push

- API des Data Catalog wird verwendet und Metadaten in den Catalog zu schreiben
- Streaming (Acryl Data unterstützt Streaming)

Data Governance

Prozesse, Richtlinien und Standards, zur Förderung der Verfügbarkeit, Qualität und Sicherheit von Daten in einer Organisation.

- Klassifikation von Daten um kritische Daten gezielt zu schützen
  - Automatisierung über Mustererkennung in Policies
- Unterstützt geordnete Prozesse
  - Self-Service Zugriff auf Daten nur nach Zustimmung des Data Owner
  - Definition von neuen Glossar Begriffen
- Technische Umsetzung der Prozesse
  - Integration mit IAM, Glossar, etc.
  - Integriertes Ticketing System

Data Lineage

Data Lineage ist der Prozess des Verstehens, Aufzeichnens und Visualisierens von Daten auf ihrem Weg von der Datenquelle zum Verbraucher.

Anwendungen

- Vertrauen in Daten
- Debugging
- Audit

Technische Umsetzung

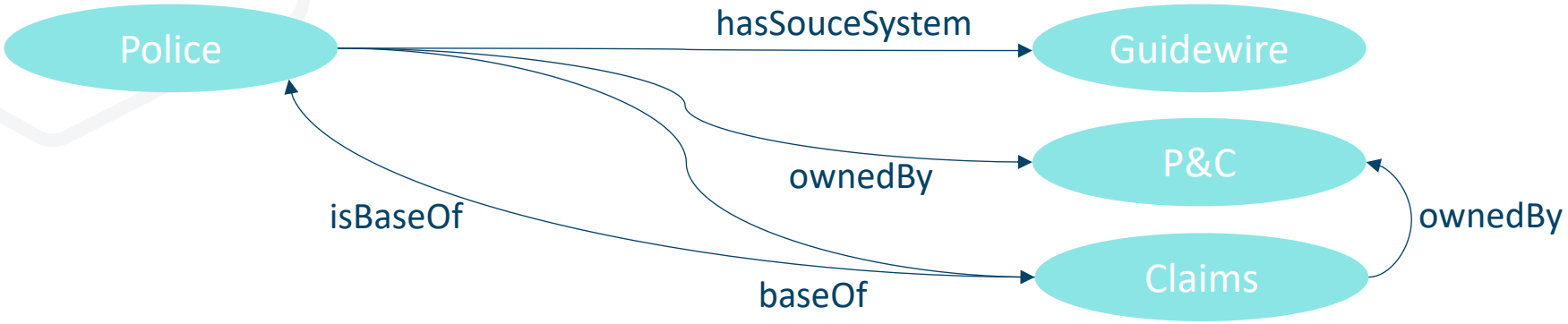
- Analyse von ETL / ELT Pipelines (Extraktion aus SQL und Mapping Definitionen)
- API zur Definition von Lineage (Apache Atlas REST API)
- Extraktion von Events aus dem Spark Execution Plan (Tools: OpenLineage, Spline)

Empfehlung für Produktentscheidung

- Detaillierte Analyse der Produktspezifikation (z.B. SQL unterstützt, Stored Procedures nicht)

Knowledge Graphs

Zwischen Daten, Applikationen, Verantwortlichen, Abteilungen, etc. gibt es viel Beziehungen. Diese lassen sich sehr flexibel in Knowledge Grafen abbilden.



- Durch das Verfolgen der Pfade in einem Graph lassen sich weitere Informationen zu einem Asset gewinnen.
- Knowledge Graphs sind nicht Lineage, da sie nicht darstellen, wie sich Daten bewegen, sondern wie Daten konzeptionell zusammenhängen und wie sie organisiert sind.
- Klassischen Graph-Datenbanken drängen in den Data Catalog Markt poolparty CAMBRIDGE SEMANTICS
- Bestehende Data Catalogs integrieren Knowledge Graphs data.world Collibra

Catalog Deployment

Cloud Data Platform Native Catalogs

→ Für die eigene Plattform bestens integriert

Produkte der grossen Hyperscaler

- Azure: Purview
- GCP: DataPlex
- AWS: Glue Data Catalog

- Snowflake: Alation, Atlan, Collibra, Data.world, DvSum, OvalEdge, Select Star (Certified Partners)
- Databricks: Unity Catalog (nur für Databricks Ressourcen)

On-premises Deployment

Nur wenige Produkte unterstützen on-premises Deployments.

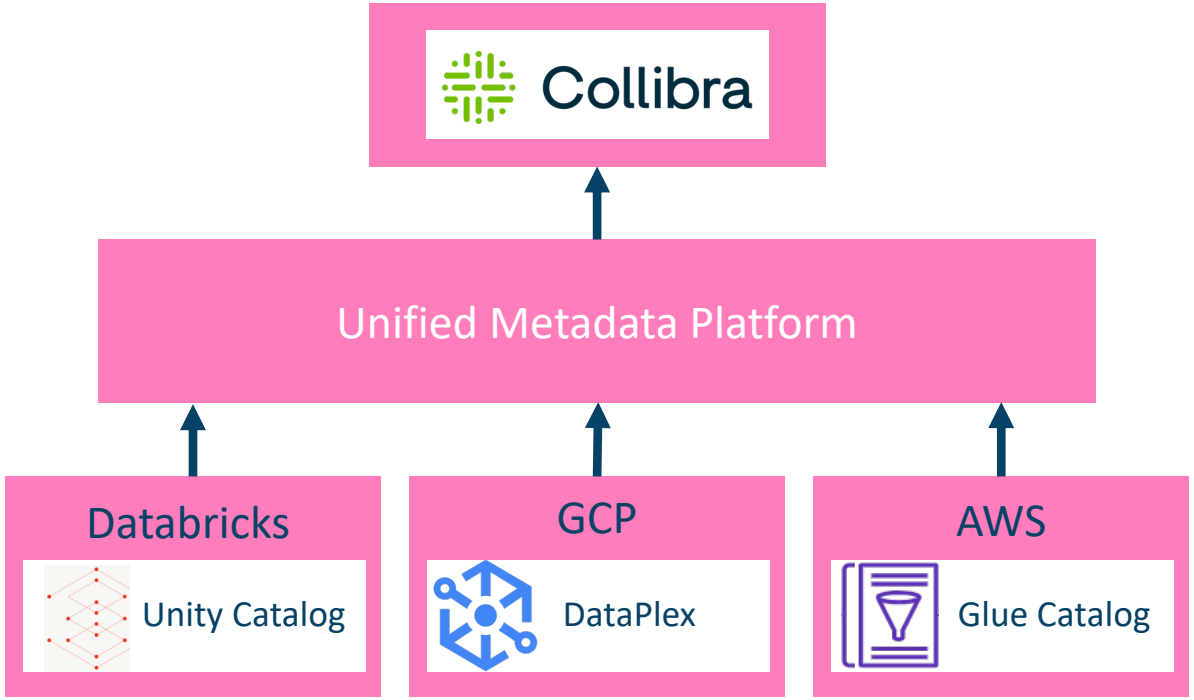
Trend geht in Richtung Cloud (z.B. Collibra hat on-premise Support Ende 2022 eingestellt)

Produkte: Informatica, Alation

Catalog of Catalogs

Data Catalog in heterogener Datenlandschaft

- Einsatz von spezialisierten Data Catalog
  - pro Plattform Vendor
  - pro Deployment Environment (Cloud, on-premises)
- Genereller Catalog für einheilticher Sicht auf die Metadtaten
  → Catalog of Catalogs



Wie ich meinen passenden Data Catalog finde

Fazit

- Vor der Entscheidung müssen die eigenen Anforderugen mit den Fähigkeiten der Produkte abgeglichen werden → **Suche**, Population, Governance, Lineage, Deployment, Semantic

Vorgehen

1. Die Anforderungen an den Data Catalog dokumentieren
2. Eine Shortlist an Produkten erstellen (Beginnen mit Cloud Native Produkt)
3. In einem PoC überprüfen, ob die Anforderungen erfüllt werden.

Mit einem Data Catalog beginnen

In heterogenen Umgebungen evtl. auf mehrere Produkte setzen (Catalog of Catalogs)

