Data Quo Vadis

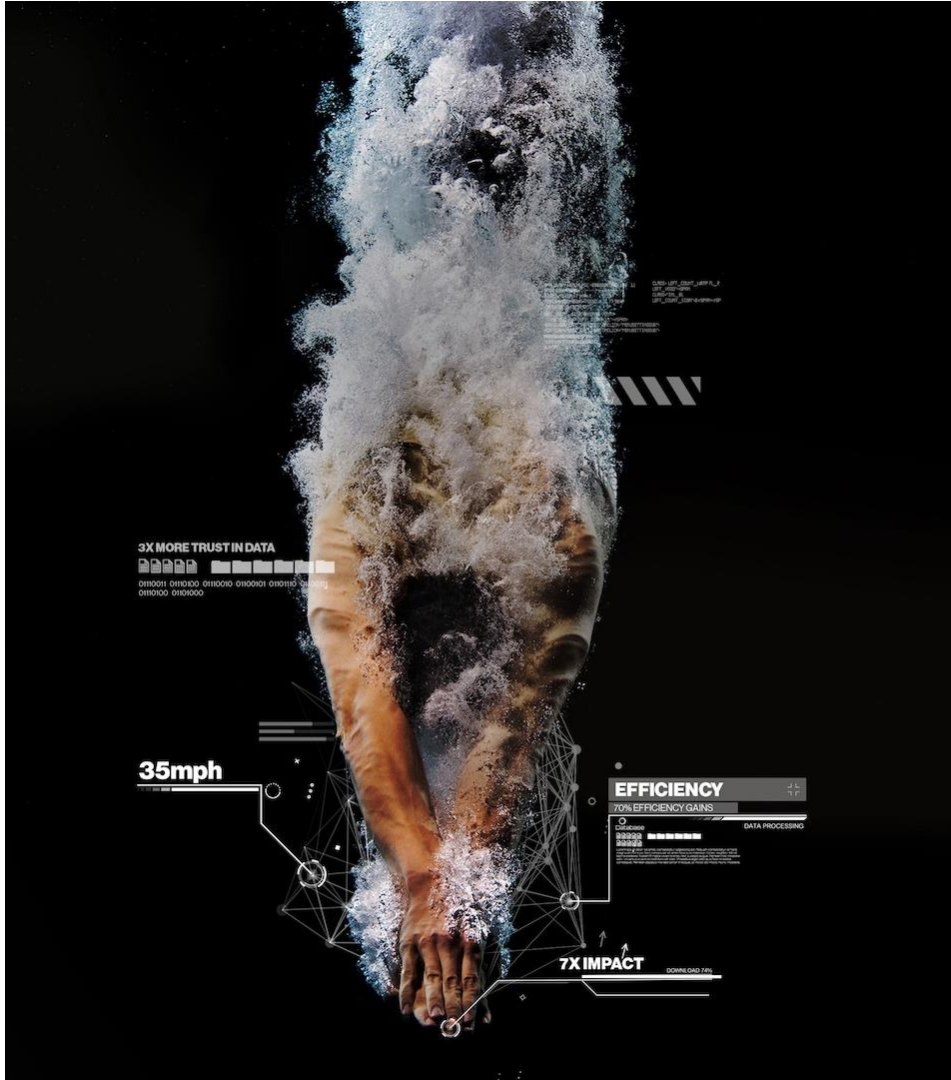
Data Lineage mit Pentaho

Clemens.Rabe@Pentaho.com – Solution Architect

Pentaho User Meeting Vol. 12

Herausforderungen in Zeiten von AI

- Akkurate Daten sind **schwer zu finden**
- Daten sind **nicht vertrauenswürdig**
- Datenbereitstellung für die Fachseite ist **aufwendig und komplex**
- Unmengen an Daten, allerdings viele **redundant, veraltet, obsolet**

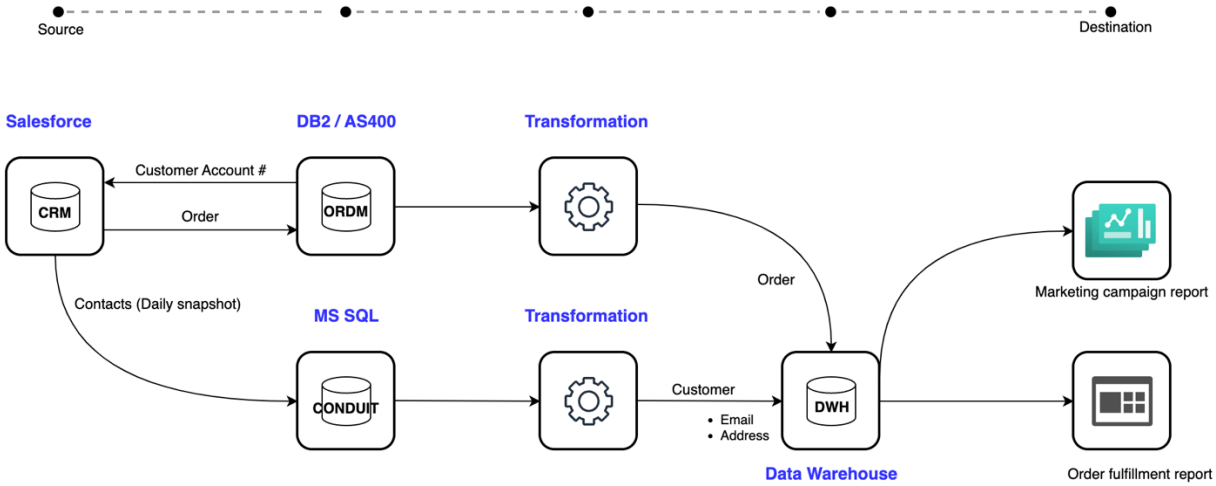


Was ist Data Lineage

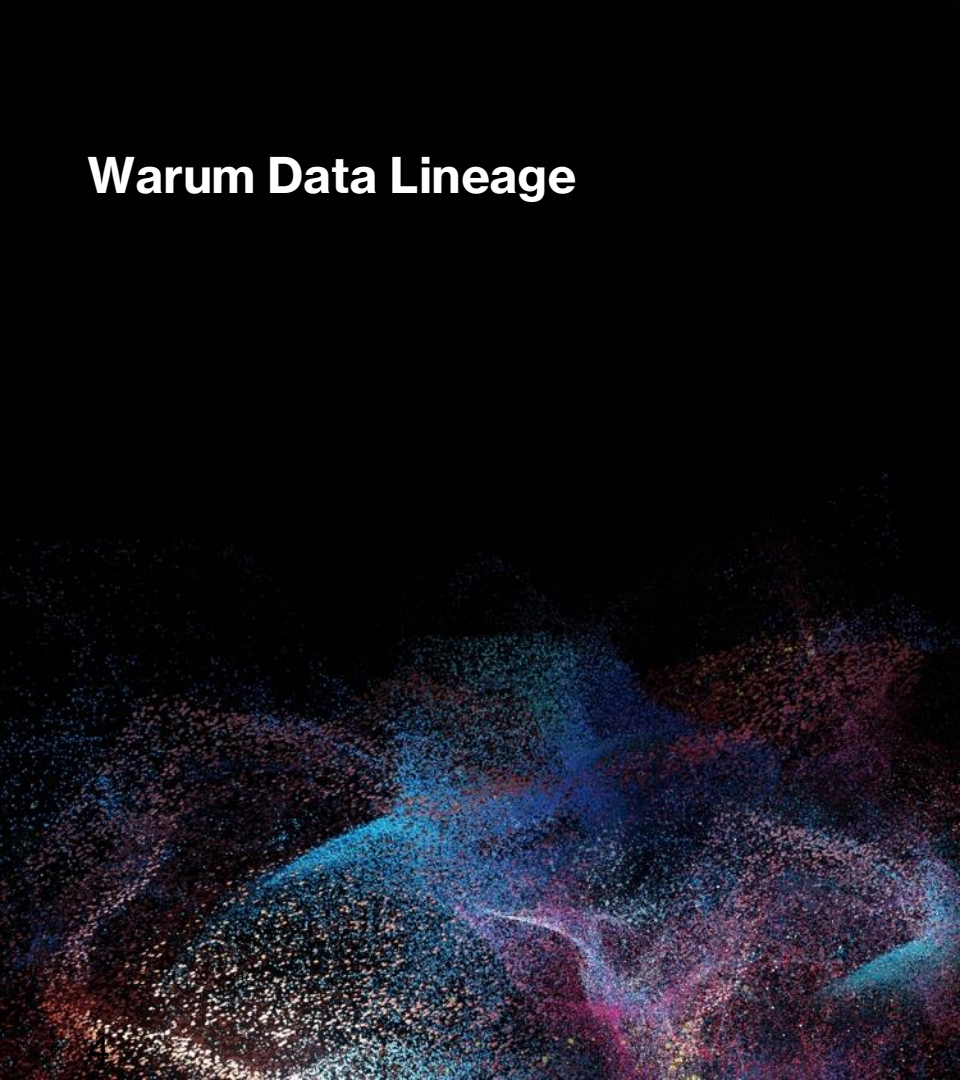
Ein Datenreise-Logbuch, von der Entstehung (Source) bis zum Zielpunkt (Destination) der Daten

Jeder Zwischenstopp liefert Informationen zu den Daten

- Woher sie kommen, wohin sie weitergeleitet wurden
- Welche Anwendung (z.B.. ERP, AI-Modell, BI-Bericht)
- Was verändert (selektiert, vermengt, etc.) wurde
- Wann die Daten transformiert wurden
- Ob Fehler aufgetreten sind



Warum Data Lineage



- **Impact Analyse:** Auswirkungen von Änderungen am Datenfluß und –Schema besser einschätzen
- **Datenqualität:** Analyse wann und wodurch ein Datenqualitätsproblem verursacht wurde
- **Bessere Daten-gestützte Entscheidungen:** Kenntnis des Ursprungs der Daten ist essenziell für die Entscheidungsfindung
- **Effizientes Daten-Management:** Data Lineage gibt den Überblick über sämtliche Datenbewegungen und ermöglicht Redundanzen zu beseitigen
- **Data Governance & Compliance:** Lineage Dokumentation für die Anforderungen von GDPR, BCBS 239, CCPA, und weiteren Regulatorien

Pentaho Data Integration OpenLineage Plugin

Op

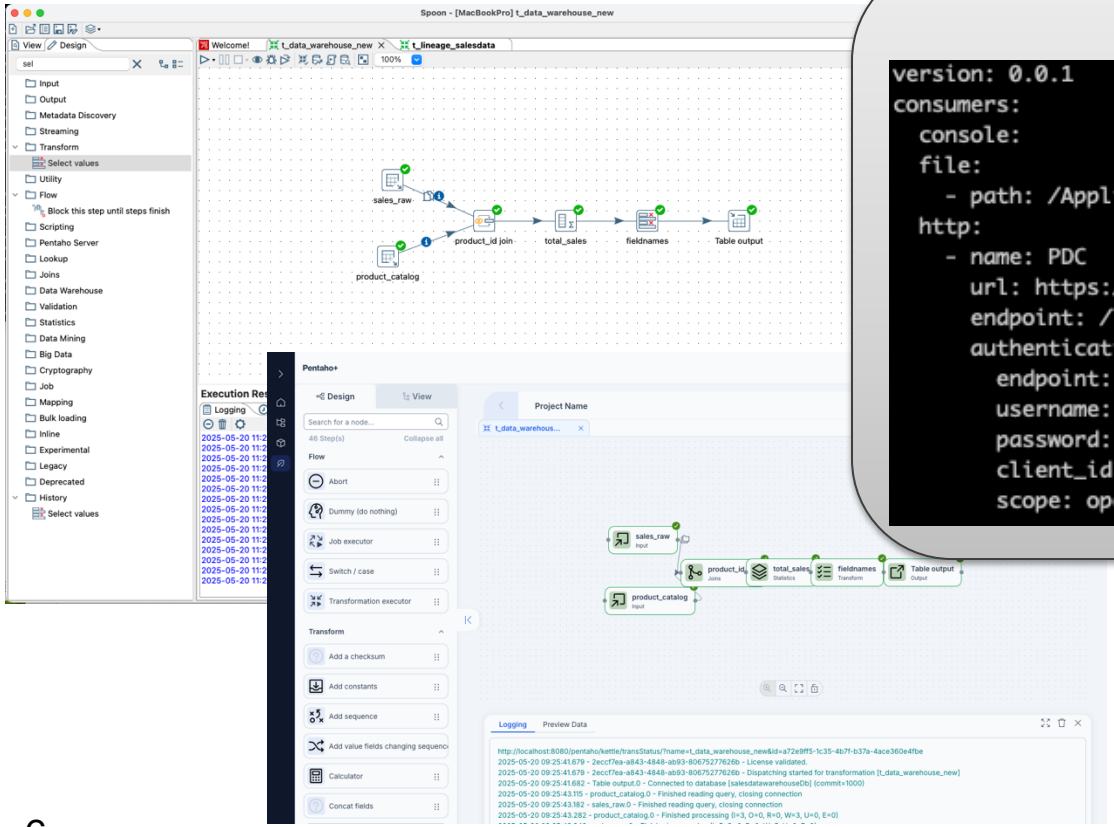
J

Input/Output Dataset


```
"outputs" : [ {
  "namespace" : "postgres://pdcdemo.demolab:5432",
  "name" : "business_apps_db.salesanalytics.product_sales_summary",
  "facets" : {
    "schema" : {
      "_producer" : "https://github.com/pentaho/pdi-plugins-ee/tree/pdi-openlineage-plugin-ee/pdi-openlineage-plugin",
      "_schemaURL" : "https://openlineage.io/spec/facets/1-1-1/SchemaDatasetFacet.json#/$defs/SchemaDatasetFacet",
      "fields" : [ {
        "name" : "product_id",
        "type" : "integer"
      }, {
        "name" : "product_name",
        "type" : "character varying"
      }, {
        "name" : "category",
        "type" : "character varying"
      }, {
        "name" : "total_sales",
        "type" : "double precision"
      } ]
    }
  }
} ]
```


https://github.com/OpenLineage/OpenLineage/blob/main/spec/OpenLineage.md

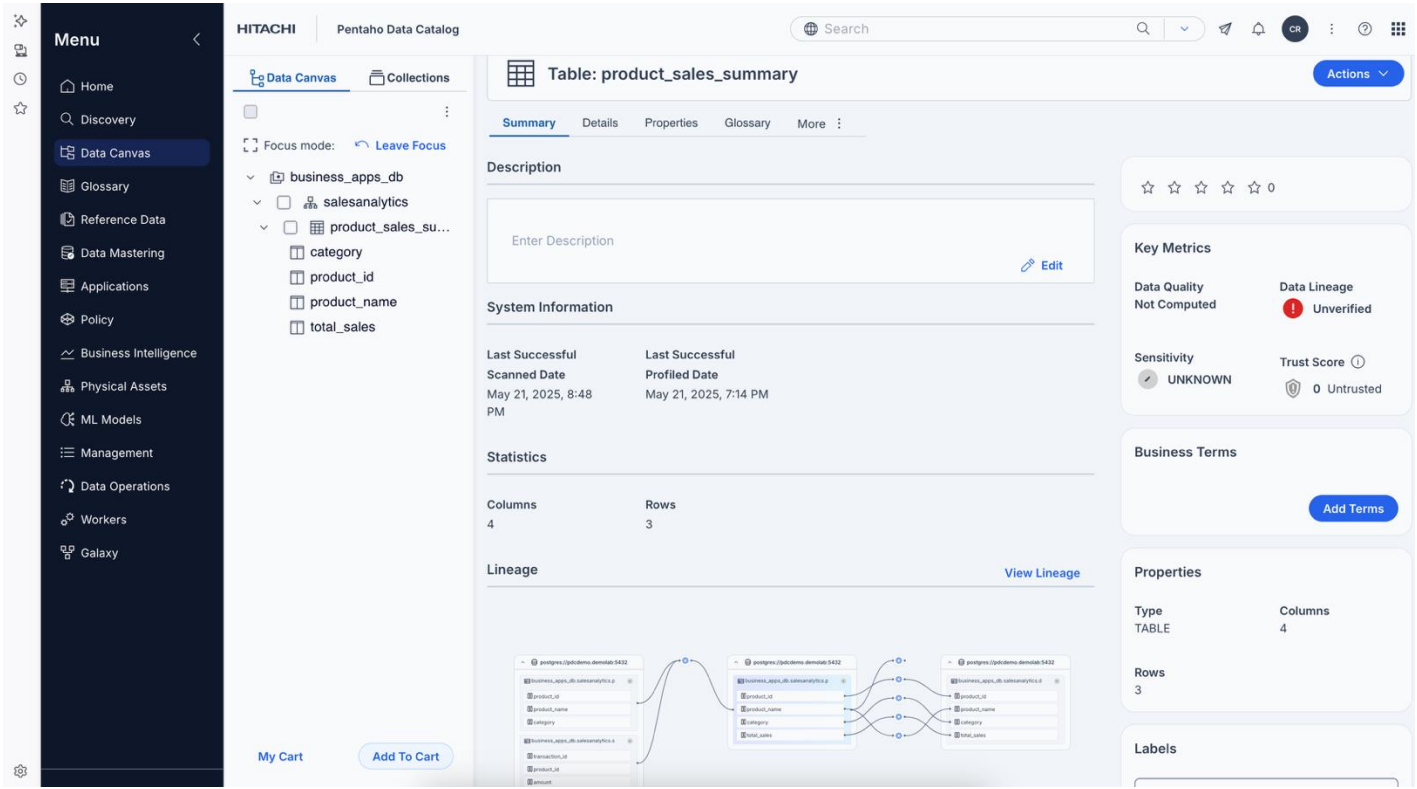
Pentaho Data Integration OpenLineage Plugin



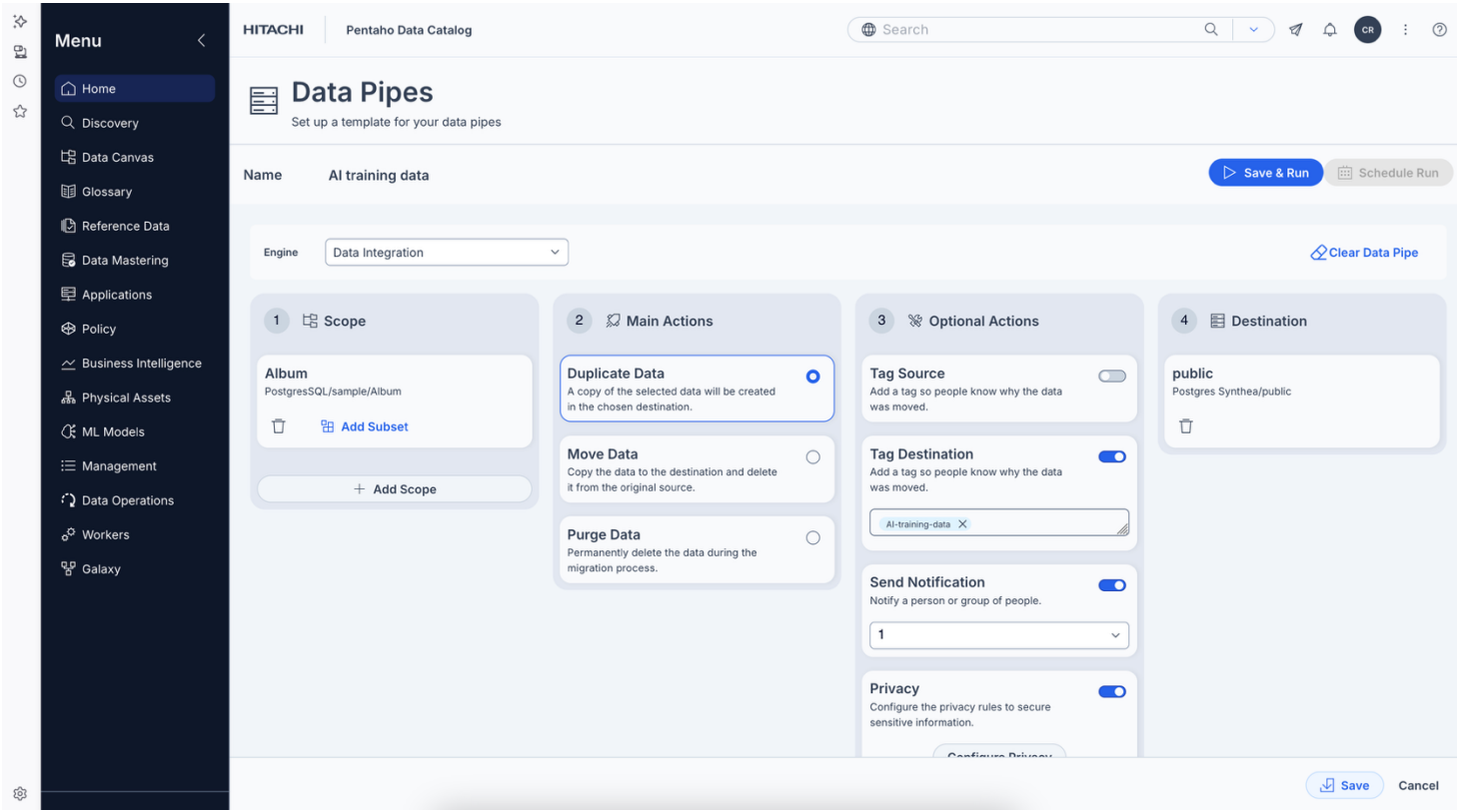
config.yml

```
version: 0.0.1
consumers:
  console:
  file:
    - path: /Applications/Pentaho/openlineage/openlineage.out
  http:
    - name: PDC
      url: https://pdcdemo.demolab
      endpoint: /lineage/api/events
      authenticationParameters:
        endpoint: /keycloak/realms/pdc/protocol/openid-connect/token
        username: admin@hv.com
        password: Encrypted 2be98afc86af09788a816a3758fc0fc9b
        client_id: pdc-client
        scope: openid
```

Lineage in Pentaho Data Catalog



Data Pipes in Pentaho Data Catalog



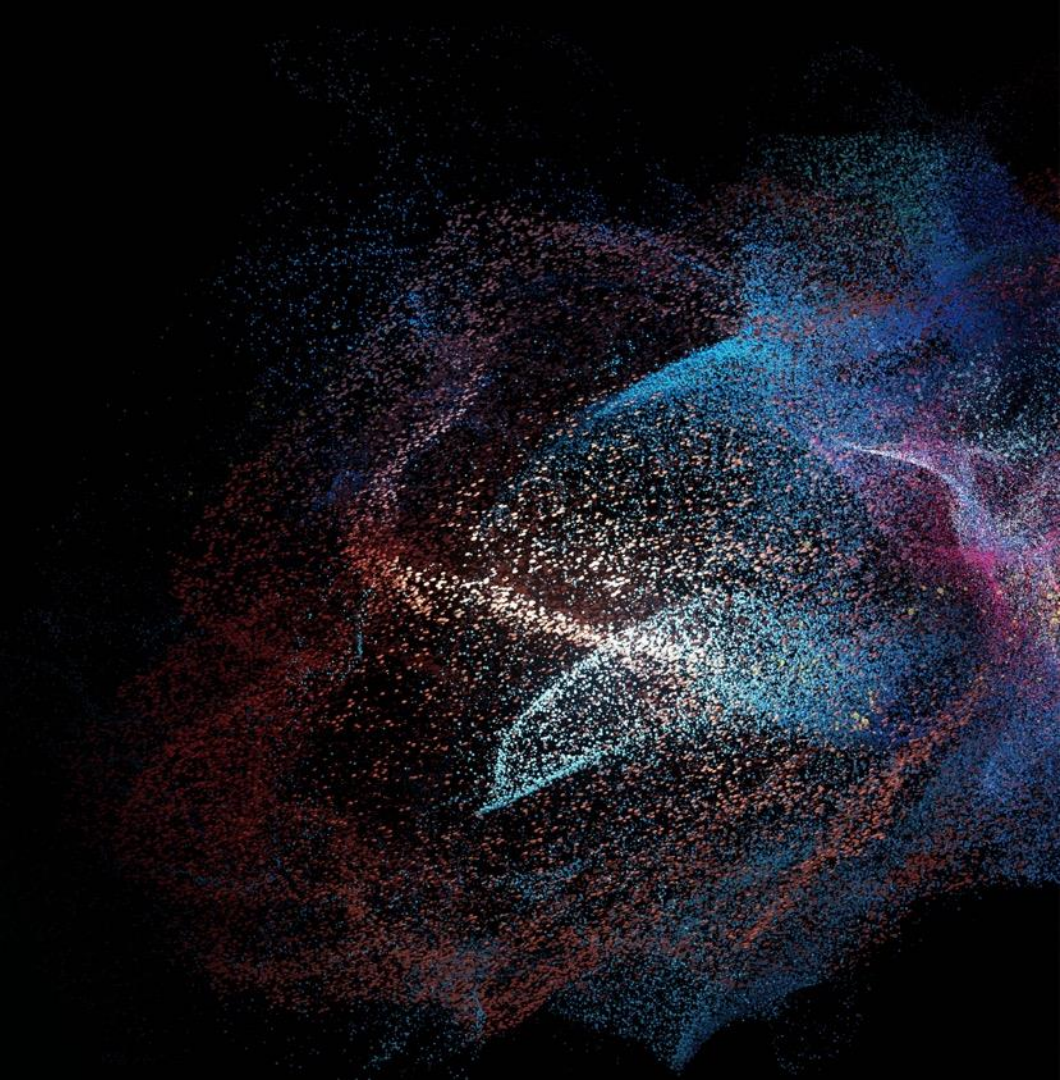
Zusammenfassung

- Akkurate Daten sind **schwer zu finden**
- Daten sind **nicht vertrauenswürdig**
- Datenbereitstellung für die Fachseite ist **aufwendig und komplex**
- Unmengen an Daten, allerdings viele **redundant, veraltet, obsolet**

- **Data Lineage** als wichtigen Datenqualitäts-Aspekt
- **OpenLineage Standard**
  - Quelle: Pentaho Data Integration/Kettle
  - Ziel: Pentaho Data Catalog
- Self-Service Data Marketplace via **Pentaho Data Pipes**
- Fingerprinting und Lineage im Pentaho Data Catalog zur **automatisierten Entdeckung von Redundanzen**

Get data-fit. Pentaho

Thank You



Data quality isn't just an IT concern – it's a business imperative.

Your customers are making big bets on:

- AI and GenAI initiatives
- Customer 360° and personalization
- Self-service analytics and automation
- Risk-based pricing and predictive models
- Regulatory compliance and audit readiness

Modern Data Quality Is:

Continuous – Always evolving with new inputs

Contextual – Aligned with specific business outcomes

Collaborative – Involves data scientists, engineers, business users

AI-Critical – Feeds trustworthy, explainable outputs