

# Data Quo Vadis



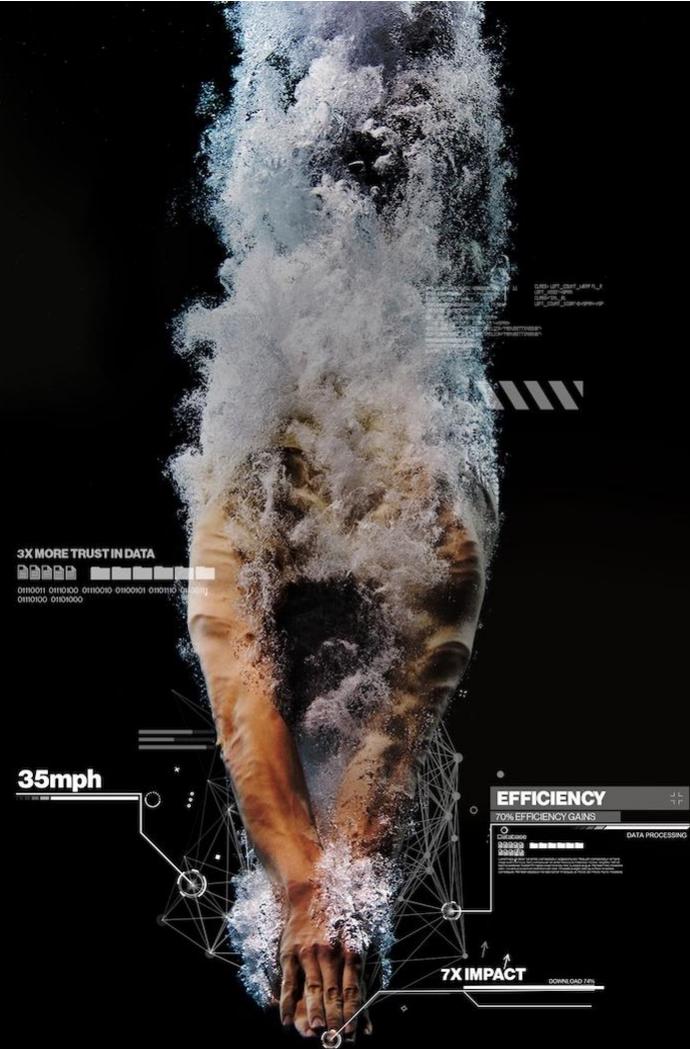
Data Lineage mit Pentaho

[Clemens.Rabe@Pentaho.com](mailto: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**

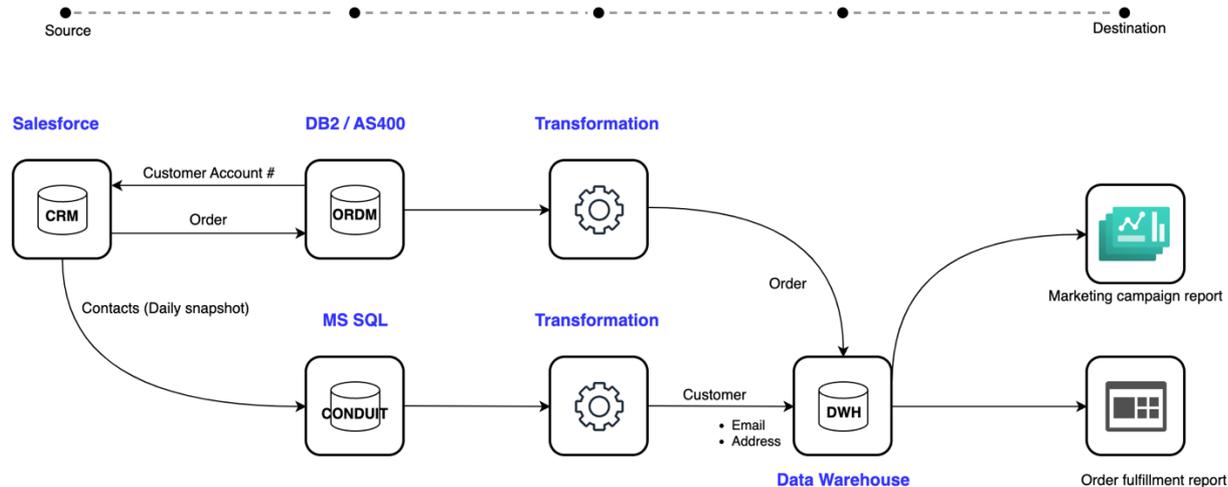


# Was ist Data Lineage

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

## 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"  
      } ]  
    }  
  }  
} ]
```

# Pentaho Data Integration OpenLineage Plugin

The screenshot displays the Pentaho Data Integration (PDI) Spoon interface. The main workspace shows a job design with the following components: **product\_catalog** (Input), **product\_id\_join** (Join), **total\_sales** (Statistics), **fieldnames** (Transform), and **Table output** (Output). A **sales\_raw** (Input) component is also present, connected to the **product\_id\_join** component. The interface includes a left-hand menu with categories like Input, Output, Metadata Discovery, Streaming, Transform, and Utility. Below the design workspace, the **Execution Results** panel shows a log of job steps, and the **Logging** panel at the bottom displays detailed execution logs for the job.

## 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

The screenshot displays the Pentaho Data Catalog interface for a table named 'product\_sales\_summary'. The interface is divided into several sections:

- Menu:** A dark sidebar on the left contains navigation options: Home, Discovery, Data Canvas (selected), Glossary, Reference Data, Data Mastering, Applications, Policy, Business Intelligence, Physical Assets, ML Models, Management, Data Operations, Workers, and Galaxy.
- Header:** Shows 'HITACHI Pentaho Data Catalog' and a search bar.
- Table Overview:** The main area shows 'Table: product\_sales\_summary' with tabs for Summary, Details, Properties, Glossary, and More. It includes a description field (currently empty), system information (Last Successful Scanned Date: May 21, 2025, 8:48 PM; Last Successful Profiled Date: May 21, 2025, 7:14 PM), and statistics (4 columns, 3 rows).
- Lineage:** A section titled 'Lineage' with a 'View Lineage' link. Below it, a diagram shows the data flow from source tables to the target table. The source tables are 'business\_apps\_db.salesanalytics', 'business\_apps\_db.salesanalytics', and 'business\_apps\_db.salesanalytics'. The target table is 'product\_sales\_summary' with columns: product\_id, product\_name, category, and total\_sales.
- Key Metrics:** A section on the right showing 'Data Quality Not Computed', 'Data Lineage Unverified', 'Sensitivity UNKNOWN', and 'Trust Score 0 Untrusted'.
- Business Terms:** A section with an 'Add Terms' button.
- Properties:** A section showing 'Type TABLE' and 'Columns 4'. Below it, 'Rows 3' is displayed.
- Labels:** A section at the bottom right.

# Data Pipes in Pentaho Data Catalog

The screenshot displays the Pentaho Data Catalog interface for configuring a Data Pipe. The interface is divided into a left-hand navigation menu and a main configuration area.

**Navigation Menu (Left):**

- Home
- Discovery
- Data Canvas
- Glossary
- Reference Data
- Data Mastering
- Applications
- Policy
- Business Intelligence
- Physical Assets
- ML Models
- Management
- Data Operations
- Workers
- Galaxy

**Main Configuration Area:**

- Header:** HITACHI | Pentaho Data Catalog | Search | CR
- Title:** Data Pipes | Set up a template for your data pipes
- Name:** AI training data | [Save & Run](#) | [Schedule Run](#)
- Engine:** Data Integration | [Clear Data Pipe](#)
- Configuration Steps:**
  - 1 Scope:** Album (PostgresSQL/sample/Album) | [Add Subset](#) | [+ Add Scope](#)
  - 2 Main Actions:**
    - Duplicate Data:** A copy of the selected data will be created in the chosen destination. (Selected)
    - Move Data:** Copy the data to the destination and delete it from the original source.
    - Purge Data:** Permanently delete the data during the migration process.
  - 3 Optional Actions:**
    - Tag Source:** Add a tag so people know why the data was moved. (Off)
    - Tag Destination:** Add a tag so people know why the data was moved. (On) | AI-training-data X
    - Send Notification:** Notify a person or group of people. (On) | 1
    - Privacy:** Configure the privacy rules to secure sensitive information. (On) | [Configure Privacy](#)
  - 4 Destination:** public (Postgres Synthea/public) | [Clear](#)

**Bottom Right:** [Save](#) | [Cancel](#)

# 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

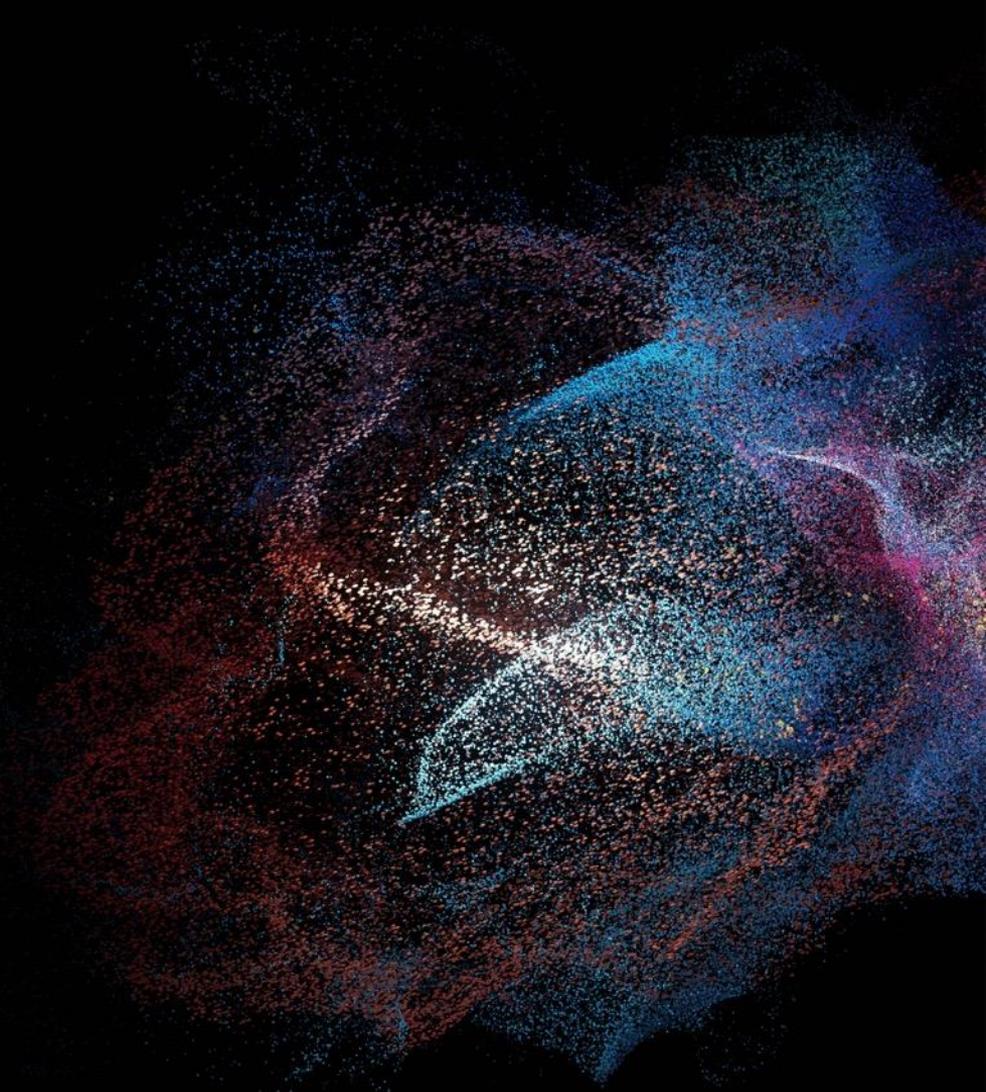
⌘ Self-Service Data Marketplace via **Pentaho Data Pipes**

⌘ 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