



AUS DER FORSCHUNG

Was ist Data Lineage? Die Transformationen hinter einem Ergebnis

Data Lineage verfolgt ein Ergebnis durch Eingaben, Joins, Filter und Transformationen. Benennen Sie jede Version, bevor eine Zahl reproduzierbar heißt.

Sinan Isoglu, MBA (Quantic)

Führungskraft für Umsatzwachstum, Dozent und Doktorand

VERÖFFENTLICHT

6. September 2026

LESEZEIT

5 Min.

WÖRTER

1.022

<https://isoglu.com/de/insights/journal/was-ist-data-lineage/>

Management Summary	2
Was bedeutet Data Lineage?	3
Welche Felder machen eine Lineage-Spur reproduzierbar?	3
Wie sieht ein Lineage-Graph aus?	4
Wo kann Data Lineage scheitern?	5
Wie prüft ein Team ein gemeldetes Ergebnis?	6
Referenzen	7

MANAGEMENT SUMMARY

Data Lineage ist die Spur von einer erklärten Ausgabe oder Entscheidung zurück durch Eingabedatensätze, Snapshots, Transformationen, Joins, Filter, Aggregationen und die nachgelagerte Nutzung. Sie hängt mit Provenienz zusammen, ist aber nicht dasselbe Objekt: Provenienz hält Entitäten, Aktivitäten und Verantwortliche fest, während eine angewandte Lineage den Entstehungspfad eines Datenergebnisses prüfbar macht. Dieser Beitrag nutzt das W3C-Modell PROV-DM als begrenzten Standardanker und baut einen synthetischen Lineage-Graphen mit Versionen, Schlüsseln, verspäteten Daten, manuellen Schritten und Ungültigkeitszuständen. Graph und Prüfregeln sind eigene Synthese. Sie ermöglichen Rekonstruktion, beweisen aber weder korrekte Quelldaten noch valide Messung oder Kausalität.

Schlagwörter: Data Lineage · Datenherkunft · Datenprovenienz · Transformation · Reproduzierbarkeit · Prüfpfad

Ein Umsatzdashboard kann eine Zahl zeigen, obwohl sich der Pfad dahinter bereits verändert hat. Die Eingabe wurde aktualisiert. Ein Join-Schlüssel wurde neu zugeordnet. Ein Filter wanderte aus der Abfrage in eine manuell gepflegte Tabelle. Die Zahl trägt noch dasselbe Etikett, aber eine andere Person kann nicht erkennen, ob sie noch dasselbe Objekt ist wie im Vormonat.

Data Lineage ist die Spur von einer erklärten Ausgabe oder Entscheidung zurück durch die Versionen der Eingaben, Transformationen, Joins, Filter, Aggregationen und nachgelagerten Nutzungen, die sie erzeugt haben. Sie markiert eine Rekonstruktionsgrenze, kein Wahrheitsprädikat.

Der Beitrag zum Event-Schema für Umsatzanalysen besitzt die Gestaltung unveränderlicher kommerzieller Ereignisse. Der Beitrag zur Evidenzprüfung besitzt die Frage, welche Evidenz eine Entscheidung überstehen soll. Diese Seite besitzt den Pfad zwischen Datenobjekt und Ergebnis.

Was bedeutet Data Lineage?

Lineage kann einen Katalog von Tabellen, eine Liste vorgelagerter Systeme, eine Warehouse-Funktion oder ein Diagramm an der Wand meinen. Das kann nützlich sein. Eine reproduzierbare Spur entsteht erst, wenn Objekte, Versionen, Operationen und Nutzung identifiziert sind.

Das Modell PROV-DM des World Wide Web Consortium liefert eine allgemeine Sprache dafür. Es beschreibt Provenienz über Entitäten, Aktivitäten und Verantwortliche und umfasst Beziehungen wie Nutzung, Erzeugung, Ableitung, Zuordnung und Zuschreibung (World Wide Web Consortium, 2013). Der Standard beschreibt, wer und was an der Entstehung beteiligt war. Er bescheinigt nicht, dass eine Quelle oder Transformation korrekt ist.

Die praktische Unterscheidung lautet:

Tabelle 1 Was bedeutet Data Lineage?

Objekt	Beantwortete Frage	Mindestdatensatz
Provenienz	Welche Entitäten, Aktivitäten und Verantwortlichen waren beteiligt?	Entität, Aktivität, Verantwortlicher, Beziehung, Zeitpunkt
Data Lineage	Welcher Pfad und welche Operation erzeugten dieses Feld oder Ergebnis?	Eingabe- und Ausgabeversion, Operation, Schlüssel, Zeit, Verantwortung
Datenqualität	Passt das Ergebnis zur erklärten Entscheidung?	Regel, beobachteter Zustand, Nachweis, Ausnahme, Entscheidung
Ereignishistorie	Welche Geschäftseignisse traten wann auf?	Unveränderliches Ereignis, Subjekt, Typ, Wirksamkeitszeit, Erfassungszeit

Tabelle aus diesem Essay. Quellen und Einordnung stehen im Beitrag.

Provenienz ist der breitere Kontext. Lineage ist der angewandte Pfad. Datenqualität ist die Eignungsbeurteilung. Ereignishistorie ist der Beobachtungsdatensatz. Keines ersetzt die anderen.

Welche Felder machen eine Lineage-Spur reproduzierbar?

Beginnen Sie mit dem Ergebnis, nicht mit dem Diagramm. Schreiben Sie zuerst Einheit und Entscheidung auf und halten Sie dann die verbindenden Objekte fest:

- 1 Ausgabe: das Feld, die Tabelle, die Kennzahl oder das Entscheidungsobjekt.
- 2 Eingabe: jedes Quellobjekt mit Verantwortung sowie exakter Version oder Snapshot-Zeit.
- 3 Operation: Filter, Standardisierung, Join, Aggregation, Berechnung oder manuelle Änderung.
- 4 Schlüssel: Feld oder zusammengesetzter Schlüssel samt Eindeutigkeitsprüfung.
- 5 Zeit: Ereignis-, Erfassungs-, Verarbeitungszeit und Reporting-Stichtag.
- 6 Prüfung: Test für erwartete Zeilenanzahl, Datensatzebene oder Grenze.
- 7 Nutzung: Dashboard, Bericht, Modell oder Entscheidung, die die Ausgabe verwendet.
- 8 Ausnahme: verspätete Daten, fehlgeschlagener Test, geänderte Regel, manuelles Override oder offene Abweichung.

Auch die Ausgabe braucht eine Version. Eine aktuelle Tabelle ist nicht stabil, nur weil ihr Name gleich bleibt. Ein zeitgestempelter Snapshot ist mit einem anderen Snapshot vergleichbar. Ein Ereignisstrom beantwortet möglicherweise eine andere Frage über das, was dazwischen geschah.

Wie sieht ein Lineage-Graph aus?

Die folgenden sechs Zeilen sind synthetisch. Sie zeigen die Detailtiefe für die Rekonstruktion eines kleinen Ergebnisses. Sie sind keine Warehouse-, Kunden- oder Unternehmensdaten.

Abbildung 1 Die synthetische Tabelle zur Lineage-Rekonstruktion

ID	Eingabeobjekt und Version	Transformation	Schlüssel oder Ebene	Ausgabeobjekt	Prüfung und Nutzung
L-01	Ereignis-Snapshot E-2026-08-31	Filter event_type = qualified	Eine Ereignis-ID; Eindeutigkeit bestanden	Tabelle zulässiger Ereignisse V1	Zeilenanzahl abgeglichen; Routingbericht
L-02	Konto-Snapshot A-2026-08-31	Land und Segment normalisieren	Konto-ID; Dublettenprüfung offen	Konto-Segment-Tabelle V2	Bis zur Prüfung gehalten; Gebietsansicht
L-03	Rechnungszeilen I-2026-08-31	Join zur Produktzuordnung P-17	Produktschlüssel; viele-zu-eins erwartet	Umsatz je Position V3	Join-Kardinalität bestanden; PVM-Brücke
L-04	Nutzungsergebnisse U-2026-08-31	Nach Konto und 14-Tage-Fenster aggregieren	Konto-ID und Fensterende	Aktivierungskohorte V1	Stichtag dokumentiert; Aktivierungsanalyse
L-05	Manuelles Ausnahmeblatt X-2026-09-01	Genehmigten Ausschluss ergänzen	Konto-ID; Verantwortung benannt	Zugelassene Kohorte V2	Regeländerung protokolliert; Dashboard
L-06	Nachgelieferte Datei E-late-09-02	Vorherigen Snapshot nachladen	Ereignis-ID; V1 wird ersetzt	Neu ausgegebenes Ergebnis V2	Altes Ergebnis ungültig; Prüfung erneut öffnen

Synthetische Lineage-Tabelle des Autors auf Grundlage von PROV-DM des World Wide Web Consortium (2013). Das Provenienzvokabular ist quellengebunden; alle Objekte, Versionen, Operationen, Schlüssel, Werte und Entscheidungen sind illustrativ.

Die Tabelle behauptet nicht, dass die Ausgabe richtig ist. Sie macht die Aussage prüfbar. L-02 wird gehalten, weil die erwartete Schlüsselbeziehung noch nicht bestanden ist. L-05 zeichnet eine manuelle Operation auf, statt sie in einem Notizbuch zu verstecken. L-06 zeigt, dass eine verspätete Eingabe ein bereits ausgegebenes Ergebnis ungültig machen und die Entscheidung erneut öffnen kann.

Wo kann Data Lineage scheitern?

Die häufigen Brüche sind strukturell:

- Snapshot-Austausch: Eine aktuelle Tabelle wird ohne gemeinsamen Stichtag mit einer früheren Ereignismenge verglichen.
- Join-Vervielfachung: Ein Viele-zu-viele-Join erzeugt zusätzliche Zeilen, während der Nenner unverändert bleibt.
- Schlüssel-Drift: Umbenannte Konten, Produkte oder Gebiete brechen den Pfad zwischen Versionen.
- Verspätete Daten: Nachlieferung ändert ein historisches Ergebnis, ohne Ungültigkeit oder Neuausgabe zu vermerken.

- Verborgene manuelle Arbeit: Eine Tabellenkorrektur ändert die Ausgabe ohne Verantwortung, Regel oder Version.
- Unbezeichneter Filter: Abfrage oder Dashboard filtert, ohne dass die Kennzahldefinition es nennt.

Ein Lineage-Graph kann jeden Bruch sichtbar machen. Er kann keine schlechte Quelle reparieren, entscheiden, ob ein Proxy das Zielkonstrukt misst, oder ein beschreibendes Ergebnis in einen Kausaleffekt verwandeln.

Wie prüft ein Team ein gemeldetes Ergebnis?

Bitten Sie die verantwortliche Person, ein Ergebnis rückwärts abzuspielen:

- 1 Ausgabe, Einheit, Reporting-Stichtag und Entscheidung benennen.
- 2 Jede Eingabe auf eine gehaltene Version oder einen zeitgestempelten Snapshot zurückführen.
- 3 Jede Operation in Reihenfolge aufschreiben, einschließlich manueller Änderungen und Ausschlüsse.
- 4 Jeden Join-Schlüssel auf erwartete Ebene und Kardinalität prüfen.
- 5 Eingabe- und Ausgabeanzahlen, Summen und Grenzänderungen abstimmen.
- 6 Verspätete Daten, fehlgeschlagene Tests, ersetzte Versionen und offene Kanten markieren.
- 7 Ergebnis als reproduzierbar, gehalten, neu ausgegeben oder nicht rekonstruierbar einstufen.

Wenn der Pfad nicht abspielbar ist, begrenzen Sie das Ergebnis auf den rekonstruierbaren Teil. Wenn er abspielbar ist, halten Sie Quellenqualität, Messung, Kausalität und Transfer getrennt. Lineage beantwortet, wie ein Ergebnis erzeugt wurde. Sie beantwortet nicht, ob es wahr ist oder was es verursacht hat.

World Wide Web Consortium. (2013). PROV-DM: The PROV Data Model. W3C Recommendation. Quellenseite

ENDE DES BEITRAGS

ZITIERWEISE

Isoglu, S. (2026, 6. September). Was ist Data Lineage? Die Transformationen hinter einem Ergebnis. Sinan Isoglu. <https://isoglu.com/de/insights/journal/was-ist-data-lineage/>

WEITERLESEN

Aus der Forschung

Was ist Datenprovenienz? Der Kontext hinter einer Zahl

<https://isoglu.com/de/insights/journal/was-ist-datenprovenienz/>

Aus der Forschung

Was ist Reproduzierbarkeit? Kann eine andere Person das Ergebnis rekonstruieren?

<https://isoglu.com/de/insights/journal/was-ist-reproduzierbarkeit/>

Aus der Forschung

Was ist Quellenprüfung? Von der Aussage zur Version of Record

<https://isoglu.com/de/insights/journal/was-ist-quellenpruefung/>



Sinan Isoglu, MBA (Quantic)

Führungskraft für Umsatzwachstum, Dozent und Doktorand