Engineer

Datenherkunft

der Weg der Daten durch Quellen und Verarbeitungsschritte

Was es ist

Datenherkunft beschreibt, welchen Weg Daten von ihren Quellen über Verarbeitungsschritte bis zu einem Datensatz oder Bericht nehmen. Eine Umsatzangabe kann aus einem Bestellregister, einer Tabelle mit Erstattungen und einer Regel für die Währungsumrechnung stammen. Mit einer Herkunftsübersicht lässt sich der Wert zu seinen Eingaben zurückverfolgen. In der anderen Richtung zeigt die Übersicht, welche Berichte eine Änderung an einer Spalte oder an einem Liefertermin betreffen kann. Die Darstellung kann Verbindungen zwischen Tabellen zeigen oder für einen Lauf genau die gelesenen und geschriebenen Partitionen und Felder festhalten.

OpenLineage unterscheidet Datensätze, benannte Verarbeitungsaufgaben und einzelne Ausführungen. Eine Ausführung kann melden, welche Datensätze sie gelesen und erzeugt hat. Schemainformationen, Quellversion und Ausführungsumgebung können als zusätzliche Hinweise mitkommen. Die Ereignisse helfen, eine konkrete Lieferung auf den Lauf zurückzuführen, der sie tatsächlich erzeugte. Eine Auswertung des Quellcodes kann auch mögliche Verbindungen zeigen, für die kein Lauf beobachtet wurde. Der Leser sollte erkennen, auf welcher Art von Nachweis jede Beziehung beruht.

Wählen Sie den Detaillierungsgrad nach der Frage. Auf Tabellenebene sieht eine Analystin, dass ein Bericht zwei Quellen nutzt. Auf Spaltenebene kann sie einen Betrag durch eine Umrechnung oder ein Feld durch eine Umbenennung verfolgen. Auf Laufebene können Betriebsverantwortliche erkennen, welche Version eine Tageslieferung verarbeitet hat. Für die Untersuchung eines strittigen Werts können Zeitfenster, Partition und Regelversion ebenso wichtig wie die Tabellennamen sein.

Das Konzept für eine Entscheidung nutzen

Nutzen Sie „Was bedeutet Datenherkunft?“ für eine konkrete Entscheidung. Wir können helfen, das Konzept mit dem System in Verbindung zu bringen, das Sie entwickeln oder einkaufen. Zahlen kommen termingerecht dort an, wo sie gebraucht werden, ohne dass am Monatsende jemand eine Tabelle kopieren muss.

Unsere Leistungen entdecken Kontakt aufnehmen →

Die Annahmen hinter dem Begriff prüfen

Nutzen Sie die Erklärung und die verlinkten Arbeiten, um Fragen für Ihr Projekt zu formulieren. Wir können die Annahmen gemeinsam mit Ihnen prüfen, bevor ein Werkzeug ausgewählt wird. Ein Infrastrukturplan für 14 Abteilungen — 14 Forschungsabteilungen mit einem gemeinsamen Infrastrukturplan

Ein Infrastrukturplan für 14 Abteilungen Kontakt aufnehmen →

Warum es wichtig ist

Eine Herkunftsübersicht grenzt die Fehlersuche ein. Wenn der Tageswert plötzlich fällt, lässt sich prüfen, ob die Quelle weniger Datensätze geliefert hat, ein Join eine Gruppe ausließ oder der Bericht den falschen Zeitraum auswählte. Die Verbindungen geben eine Reihenfolge für die Untersuchung vor. Bestätigen Sie eine vermutete Unterbrechung mit dem Lauf und den tatsächlichen Daten. Ein lückenlos wirkendes Diagramm belegt für sich weder Aktualität noch Korrektheit.

Bekannte Empfänger erleichtern Änderungen. Eine Entwicklerin möchte eine Statusspalte umbenennen. Der Überblick zeigt einen Abgleich und mehrere Berichte, die sie verwenden. Die Beteiligten können vereinbaren, wer die Umbenennung prüft, wer vorübergehend beide Felder braucht und wann der frühere Name entfallen kann. Diese Absprache kommt rechtzeitig, solange sich die Änderung leicht anpassen lässt. Die Abhängigkeiten lassen sich sichtbar machen, während Berechtigungen und Dateninhalte geschützt bleiben.

Die Herkunft hilft auch, zwei unterschiedlich berechnete Werte mit demselben Namen zu erklären. Eine Auswertung rechnet abendliche Aufträge vielleicht noch zum laufenden Tag. Eine andere ordnet sie der Lieferung des nächsten Tages zu. Mit den Quellen und Verarbeitungsschritten lässt sich die Frage konkret stellen. Eine Prüfung der Daten zeigt anschließend, ob die zuständige Regel angewandt wurde.

Wie es in der Praxis aussieht

Die Arbeitsprobe zum Betrieb von Data Warehouses in PostgreSQL, MS SQL und BigQuery beschreibt Arbeit an mehreren Plattformen. Bei einer solchen Umgebung ist relevant, welche Verarbeitung in welchem System stattfindet. Das

folgende fiktive Beispiel zeigt eine Methode, den Weg einer einzelnen Auswertung festzuhalten.

Ein täglicher Bericht nennt versandbereite Bestellungen. Die Aufträge kommen aus dem Webshop, die Statusereignisse aus einem anderen Datenstrom. Die Auswertung gruppiert nach Versanddatum und wählt je Auftragsnummer den neuesten passenden Status. Erfassen Sie Ein- und Ausgaben und geben Sie der Verarbeitung einen dauerhaften Namen. Speichern Sie Kennung, Beginn und Ende des Laufs, die gelesenen Partitionen und die eingesetzte Version der Auswahlregel. Ein gescheiterter Versuch und eine spätere Wiederholung bleiben getrennte Ereignisse. Erkennbar bleibt, welcher Lauf das veröffentlichte Ergebnis lieferte.

Wählen Sie anschließend einen Auftrag aus dem Dienstagsbericht. Verfolgen Sie Datum und Referenz über freigegebene Abfragen bis zum Statusereignis, das der Lauf gewählt hat. Prüfen Sie seine Zeit im Verhältnis zum Annahmeschluss und die Regel für zeitgleiche Status. Ein verspäteter Datenstrom gehört in den Berichtshinweis und den Ablauf der Störung. Wenn ein Team die Grenze verschiebt, können die betroffenen Berichtseigner normale und verspätete Lieferungen prüfen.

Beginnen Sie mit einem wichtigen Bericht und überprüften Verbindungen. Lassen Sie eine Analystin eine einzelne Zeile und einen Betriebsmitarbeiter einen Fehlerlauf verfolgen. Halten Sie fest, an welcher Stelle eine der Personen raten muss. Ergänzen Sie zuerst die fehlende Kennung oder Verbindung. Das gleiche Beispiel hilft später, die Erfassung automatisch gemeldeter Beziehungen zu beurteilen.

Halten Sie die Ereigniszeit, die Verfügbarkeit der Eingabe und den Veröffentlichungsschluss fest. Damit lässt sich ein pünktliches Ereignis erklären, dessen Daten verspätet eintrafen.

Wo es schiefgeht

Eine Übersicht kann einen manuellen Export, eine gespeicherte Prozedur oder eine heruntergeladene und nachträglich bearbeitete Tabelle übersehen. Halten Sie fest, welcher Teil gemessen wird und wo der Rest dokumentiert ist. Eine manuell erfasste Beziehung braucht Quelle und Datum. Bei Änderungen muss jemand erneut prüfen, ob sie gilt.

Kurze Tabellennamen können in Test, Produktion und Archiv übereinstimmen. Eine dauerhafte Kennung sollte Umgebung und Speicherort berücksichtigen. Bei einem Umzug hilft die Zuordnung zu früheren Namen, alte Berichte zu erklären. Schützen Sie auch Metadaten. Tabellennamen, Abfragen und Beispielzeilen können Kenntnisse preisgeben, auf die Leser keinen Zugriff haben sollen.

Prüfen Sie wichtige Verbindungen schließlich durch eine kleine rücknehmbare Änderung mit erwarteter Wirkung. Prüfen Sie Ausgabe und aufgezeichneten Weg. Ein neuer Eingabewert kann das Ergebnis ändern, während die Abhängigkeit gleich bleibt. Eine andere Verbindung erwarten Sie, wenn die Aufgabe tatsächlich die Quelle wechselt. Dieser Nachweis gilt für den tatsächlich untersuchten Weg.

Bei zeitgleichen Ereignissen ist eine dokumentierte zweite Sortierung wichtig. Soll die jüngere Quellversion oder die höhere Ereigniskennung entscheiden? Legen Sie dafür einen kleinen Testdatensatz an. Die nächste Person kann damit denselben Status nachvollziehen, auch wenn sich die natürliche Reihenfolge der Zeilen nach einem Umzug geändert hat.

Weiterführende Lektüre

- Objektmodell für OpenLineage-Datensätze, Aufgaben und Läufe
- PROV-O-Ontologie des W3C für Herkunftsinformationen
- Google-Cloud-Dokumentation zur Datenherkunft

Glossar

Für Ihre tatsächlichen Rahmenbedingungen auswählen

Bringen Sie einen repräsentativen Datensatz, die Fragen, die er beantworten muss, und den aktuellen Engpass mit. Wir können vereinbaren, wie Qualität, Durchsatz und Wiederherstellung gemessen werden.

Preise ansehen Kontakt aufnehmen →

Die Begriffe praktisch nutzbar machen

Fragen Sie, wie „Was bedeutet Datenherkunft?“ auf Ihre Organisation zutrifft. Ein konkretes Beispiel aus Ihrem bestehenden System reicht aus, um das Gespräch zu beginnen.

Unsere Leistungen entdecken

welcome@engineer.company +45 50 10 61 84 Kopenhagen, Dänemark

Nehmen Sie diese Seite mit

Kontaktkarte

Mit Sorgfalt und Liebe gestaltet von D.

CVR 45969703 · Tagensvej 105C, 3. th, 2200 København N

Copyright © 2025 – heute · Engineer ApS · Alle Rechte vorbehalten.

Gedruckt von https://engineer.company/de/glossary/data-lineage/