Engineer

データリネージ

データの出所・処理・行き先の記録

どのようなものか

データリネージは、データがどの入力元からどの処理を通り、どのデータセットやレポートに届いたかを示します。売上の数字が、注文表、返金表、通貨換算の規則から算出されるとします。リネージを使うと、数字から入力を逆にたどれます。順方向にたどれば、列名や配信時刻を変えたときに影響を受ける集計が分かります。必要な粒度は、表どうしの関係を表す図から、処理が読み書きした列とパーティションまでさまざまです。

OpenLineage は、データセット、名前のあるジョブ、そのジョブを実行した一回分の記録を区別します。ジョブはデータを変換したり移動したりします。個別の実行は、読み込んだデータセットと出力したデータセットを報告し、スキーマ、入力のバージョン、実行環境などの情報を添えられます。記録されたイベントを使えば、特定の成果物に至った実行を確認できます。ソースコードを調べて、発生し得る依存関係を把握する方法もあります。関係を表示する際には、観察、コードの解析、手動確認のどれが根拠かを示します。

粒度は知りたいことに合わせます。表の単位なら、レポートが二つの個別データセットを使うことが分かります。列の単位なら、金額がどんな計算を通過したか、項目名がどこで変わったかを確認できます。実行の単位なら、火曜日のデータをどの処理のバージョンが読み込んだかが分かります。観測された数値を検証するためには、表の名前だけでなく、時間範囲、パーティション、適用された規則も必要な場合があります。

概念を判断に生かす

「データリネージとは？」を実際の判断に生かしましょう。その概念を、開発または購入するシステムに結び付けて考えるお手伝いをします。 月末に誰かが表計算ファイルをコピーしなくても、決まった時刻に必要な場所へ数値が届く仕組みをつくります。

サービスを見る お問い合わせ →

用語の背後にある前提を検証する

解説とリンク先の実績を参考に、プロジェクトについての質問を整理してください。ツールを選ぶ前に、前提を一緒に検討できます。 14部門を支える共通インフラ計画 — 14 共通インフラ計画の対象となった研究部門

14部門を支える共通インフラ計画 お問い合わせ →

なぜ重要なのか

リネージがあると、障害を調査する範囲を絞れます。日計が急に下がったとき、入力が少なかったのか、結合処理であるグループが除かれたのか、集計に誤った期間が指定されたのかを確認できます。経路が最初の調査箇所を教えてくれます。見つけた疑いは、対象データと実行結果を見て検証します。依存関係をすべて描いたように見える図でも、入力が遅れたり計算が間違ったりする場合があります。

下流の利用者が分かると、変更を準備しやすくなります。開発者がステータス欄の名前を変えるとき、照合処理と複数のレポートがその欄を読んでいることに気付くかもしれません。新しい名前を誰が試すか、どの利用者に旧名も一定期間必要か、いつ旧名を廃止するかを話し合えます。仕様をまだ直せる段階で役立つ会話です。データ本体を閲覧する権限とは別に、依存関係だけを公開することもできます。

出どころが明らかであれば、同じ名前の数字が二つのレポートで異なる理由も説明できます。一方は夕方の注文を当日分に含め、もう一方は翌日の発送分に入れることがあります。入力元、締切時刻、変換を比較すると、何が違うのかを言葉にできます。その後、使うべき定義を決め、実際の計算を確認します。

実際の活用例

Engineer ApS のポートフォリオには、PostgreSQL、MS SQL、BigQuery のデータウェアハウスの運用が掲載されています。複数の処理基盤で運用するときは、各処理がどこで行われるかを知ることに意味があります。次は、個別のレポートを検証するための架空の例です。

発送可能な注文の日次レポートを考えます。注文は通販サイトから、発送ステータスは倉庫から別の経路で届きます。処理は発送日で集計し、注文番号ごとに条件を満たす最新のステータスを選びます。入力と出力を別々に記録し、ジョブに固定の名前を付けます。実行の識別子、開始と終了、読み込んだパーティション、選択規則の版を残します。再実行があった場合、失敗した試行と成功した試行を区別し、どちらが公開結果を作ったかを示します。

火曜日のレポートから一件を選びます。アクセスを許された照会を使い、レポートの日付と注文番号から選ばれたイベントへ進みます。締切時刻に対してイベントがどちら側にあるか、同じ時刻の複数イベントがどう扱われるかを確かめます。入力が遅れた場合は、レポートの説明と障害の時系列の両方に残します。後で締切を変える際には、関係する担当者が正常時と遅延時の両方を試せます。

初めは重要なレポート一つと、手で確かめた少数の関係で十分です。分析担当者に一行を、運用担当者に失敗した一回をたどってもらいます。初めて推測に頼った箇所を記録し、足りなかった識別子、版、関係を加えます。収集を自動化する場合にも、同じ問いで記録の価値を確かめられます。

イベントが発生した時刻と、入力として使えるようになった時刻を両方残します。公開の締切も記録すると、締切前に起きたイベントが遅れて届いた例を説明できます。

問題が起きるところ

図が完成しているように見えても、手動の書き出し、ストアドプロシージャ、ダウンロード後に書き換えられた表計算ファイルが抜ける場合があります。自動で観察している範囲と、残りをどこに記録しているかを示します。人の手で加えた関係には、情報源と確認日を付けます。処理が移動したら見直します。

同じ短い名前の表が試験、本番、保管先に存在することがあります。安定した識別子に環境と保存場所を含めます。移動後も以前の名前との対応があれば、過去の結果を説明できます。メタデータも保護します。保存先の名前、照会内容、行の例には、閲覧権限を持たない人に知らせるべきでない情報が含まれることがあります。

重要な経路は、結果を予想できる小さな変更で確かめます。安全に戻せる変更を加え、実際の出力と記録された経路を照合します。入力値だけを変える場合、出力は変わっても依存関係は同じです。入力元を変えた場合は、経路にも新しい関係が現れることを確認します。試せた経路に対して、再確認できる証拠を残せます。

参考資料

- OpenLineage のデータセット・ジョブ・実行のモデル
- W3C PROV-O による来歴情報のモデル
- Google Cloud のデータリネージの説明

用語集

実際の制約に合わせて選ぶ

代表的なデータセット、そこから答えを得たい問い、現在のボトルネックをお知らせください。品質、処理量、復旧の測り方を合意できます。

料金を見る お問い合わせ →

用語を実務に結び付ける

「データリネージとは？」がお客様の組織にどう当てはまるかをご相談ください。現在のシステムの具体例が一つあれば、話を始められます。

サービスを見る

welcome@engineer.company +45 50 10 61 84 デンマーク、コペンハーゲン

このページを持ち歩く

連絡先カード

心を込めて丁寧に制作： D.

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

著作権 © 2025 – 現在 · Engineer ApS · すべての権利を留保します。

印刷元： https://engineer.company/ja/glossary/data-lineage/