Engineer

Lignage des données

le parcours des données, de leur source à leurs dérivés

Ce que c’est

Le lignage des données retrace le parcours des données depuis leurs sources au fil des transformations jusqu’à un jeu de données ou un rapport. Un chiffre d’affaires affiché peut provenir d’un registre des commandes, d’une table des remboursements et d’une règle de conversion monétaire. Le lignage donne au lecteur un chemin vers les données qui alimentent ce chiffre. Dans l’autre sens, il montre quels rapports dépendent d’une colonne ou de l’heure à laquelle une source fournit sa livraison. Selon le besoin, il va d’un diagramme de tables aux partitions et champs exacts lus et écrits lors d’une exécution.

OpenLineage distingue le jeu de données, le traitement nommé et une exécution de ce traitement. Celle-ci peut déclarer ce qu’elle a lu et produit. Elle peut y joindre le schéma, la version de la source ou l’environnement. Ces événements permettent de remonter à l’exécution qui a fourni un résultat particulier. L’examen du code peut aussi montrer des dépendances possibles, y compris pour des cas jamais observés. Il faut que le lecteur puisse savoir sur quoi repose chaque relation.

Le niveau de détail se choisit en fonction de la question. Au niveau des tables, une analyste voit que le rapport repose sur deux jeux de données. Au niveau des colonnes, elle suit un montant dans une formule ou un champ après un renommage. Au niveau des exécutions, l’exploitation retrouve la version qui a traité la livraison du mardi. Pour expliquer un chiffre contesté, le créneau, la partition et la version d’une règle peuvent compter autant que les noms des tables.

Utilisez le concept pour prendre une décision

Utilisez « Qu'est-ce que le lignage des données ? » pour éclairer une décision réelle. Nous pouvons vous aider à relier ce concept au système que vous construisez ou achetez. Des chiffres disponibles là où il faut, à la fréquence prévue, sans recopier un tableur à la fin du mois.

Découvrir nos prestations Contactez-nous →

Vérifiez les hypothèses derrière le terme

Appuyez-vous sur l'explication et les réalisations liées pour formuler les questions de votre projet. Nous pouvons examiner les hypothèses avec vous avant le choix d'un outil. Un plan d'infrastructure pour 14 départements — 14 départements de recherche sur un même plan d'infrastructure

Un plan d'infrastructure pour 14 départements Contactez-nous →

Pourquoi c’est important

Le lignage réduit le champ de recherche après une anomalie. Si un total quotidien chute, on peut examiner si la collecte a reçu moins de lignes, si une jointure a perdu un groupe ou si le rapport couvre la mauvaise période. Le parcours aide à choisir l’ordre des vérifications. Il reste à confirmer le point de rupture en regardant les données et le traitement. Un diagramme apparemment complet peut représenter une source périmée ou une transformation erronée.

Connaître les destinataires permet aussi de préparer une modification. Une développeuse compte renommer un champ de statut. La vue indique qu’un rapprochement et plusieurs rapports l’utilisent. Leurs responsables peuvent se répartir les vérifications, décider qui a temporairement besoin des deux noms et convenir du retrait de l’ancien. Cette discussion a plus de valeur quand le changement peut encore être adapté. La vue des dépendances peut être accessible à des personnes qui n’ont pas accès aux données elles-mêmes.

Une provenance soignée aide en outre à comparer deux résultats portant le même nom. Un rapport peut compter les commandes du soir dans la journée en cours, un autre dans les expéditions du lendemain. Une comparaison des sources, de l’heure de coupure et des transformations rend la différence explicite. Les responsables peuvent ensuite choisir la définition voulue et vérifier son application.

À quoi cela ressemble en pratique

Engineer ApS présente l’exploitation d’entrepôts dans PostgreSQL, MS SQL et BigQuery. Avec plusieurs moteurs, il devient utile de savoir où se déroule chaque transformation. L’exercice fictif suivant illustre le niveau de preuve possible pour un rapport unique.

Imaginons un rapport quotidien qui compte les commandes prêtes à partir. La boutique fournit les commandes, un autre flux fournit les événements d'expédition. Le traitement groupe par date d'envoi et retient le dernier statut applicable à chaque commande. Enregistrez les entrées et la sortie, et donnez un nom stable au traitement. Conservez un identifiant d'exécution, son début, sa fin, les partitions lues et la version de la règle de sélection. Si le traitement échoue puis recommence, gardez les deux tentatives visibles et indiquez laquelle a produit la sortie publiée.

Prenez ensuite une ligne du mardi. Remontez avec des requêtes autorisées de la date du rapport et du numéro de commande jusqu'à l'événement choisi. Vérifiez sa date par rapport à l'heure de coupure et la manière de trancher quand deux événements ont le même horodatage. Un flux retardé doit apparaître dans l'explication du rapport et dans la chronologie de l'incident. Si l'heure de coupure change, les responsables des rapports concernés peuvent tester une livraison normale et une livraison tardive.

Commencez par un rapport important et quelques relations examinées à la main. Demandez à une analyste de remonter une ligne et à une personne de l'exploitation de remonter un échec. Notez le premier point où l'une des deux doit deviner. Ajoutez l'identifiant, la version ou la relation qui manquait à ce moment. Vous pourrez ensuite juger la collecte automatique sur la base de la même recherche.

Conservez l'heure de l'événement, l'heure de disponibilité et la limite de publication. Ces trois repères expliquent le cas d'un événement qui a eu lieu à temps mais a été reçu en retard.

Où les problèmes surviennent

Un diagramme peut ignorer un export manuel, une procédure stockée ou un tableur modifié après son téléchargement. Indiquez les parties observées et l'endroit où le reste est documenté. Ajoutez une source et une date aux liens relevés manuellement. Relisez-les lorsque le traitement est déplacé.

Un nom court peut désigner une table d'archive et une autre table en production. Donnez à chaque jeu un identifiant stable avec son environnement et son emplacement. Gardez une correspondance avec les noms anciens lors d'un

déplacement. Protégez aussi les métadonnées. Les noms de tables, les requêtes et des exemples de lignes peuvent révéler des éléments à des lecteurs qui n'ont aucune raison d'y accéder.

Essayez enfin une petite modification réversible dont vous connaissez l'effet. Vérifiez la sortie et le parcours enregistré. Une valeur d'entrée peut changer le résultat alors que la dépendance reste identique. La relation doit changer lorsque le traitement utilise effectivement une autre source. Vous aurez ainsi un résultat observable sur le parcours réellement étudié.

Pour aller plus loin

- Le modèle OpenLineage pour les jeux de données et les exécutions
- L'ontologie de provenance PROV-O du W3C
- La documentation de Google Cloud sur le lignage des données

Glossaire

Choisissez selon vos contraintes réelles

Présentez un jeu de données représentatif, les questions auxquelles il doit répondre et le goulot d'étranglement actuel. Nous pouvons convenir de la manière de mesurer la qualité, le débit et la reprise.

Consulter les tarifs Contactez-nous →

Rendez la terminologie concrète

Demandez comment « Qu'est-ce que le lignage des données ? » s'applique à votre organisation. Un exemple concret tiré de votre système actuel suffit pour engager la discussion.

Découvrir nos prestations

welcome@engineer.company +45 50 10 61 84 Copenhague, Danemark

Emportez cette page avec vous

Fiche de contact

Créé avec soin et amour par D.

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

Droit d'auteur © 2025 – aujourd'hui · Engineer ApS · Tous droits réservés.

Imprimé depuis https://engineer.company/fr/glossary/data-lineage/