Engineer

Походження даних

шлях даних від джерел через перетворення

Що це таке

Походження даних описує шлях від джерел через перетворення до набору даних, звіту або діаграми. Число виторгу на панелі може спиратися на реєстр замовлень, таблицю повернень і правило валютного перерахунку. Опис походження допомагає пройти від числа до кожного входу. У зворотному напрямку він указує, яких звітів торкнеться перейменування колонки або перенесення часу надходження. Рівень деталізації варіюється від огляду таблиць до відомостей про точні розділи й поля, які опрацював конкретний запуск.

OpenLineage виділяє набір даних, завдання з назвою й окремий запуск завдання. Завдання переносить або змінює дані. Запуск може повідомити, що саме прочитав і створив, а також додати відомості про схему, версію джерела й середовище. Такі події дають слід до запуску, який фактично підготував результат. Аналіз коду також може показувати можливі залежності, зокрема ще не спостережені. Читач має знати, на якому свідченні ґрунтується кожний зв’язок.

Вибирайте деталізацію за запитанням. На рівні таблиць видно, що звіт має два джерела. На рівні колонок можна простежити розрахунок суми або зміну назви поля. Для окремого запуску чергова людина побачить, яка версія опрацювала вівторкову партію. Щоб пояснити спірний результат, часто потрібні часовий інтервал, розділ і версія правила.

Застосуйте поняття до рішення

Застосуйте «Що таке походження даних?» до конкретного рішення. Ми можемо пов'язати поняття із системою, яку ви створюєте або купуєте. Цифри з'являються там, де потрібні, за розкладом, і ніхто не копіює таблицю наприкінці місяця.

Переглянути наші послуги Зв'язатися з нами →

Перевірте припущення за терміном

Використайте пояснення та пов'язані роботи, щоб сформулювати запитання для проєкту. Ми можемо перевірити припущення разом до вибору інструмента. Один інфраструктурний план для 14 підрозділів — 14 дослідницьких підрозділів, охоплених спільним інфраструктурним планом

Один інфраструктурний план для 14 підрозділів Зв'язатися з нами →

Чому це важливо

Історія походження звужує пошук причини. Якщо денний показник раптово впав, команда перевірить, чи джерело надіслало менше рядків, чи об’єднання відкинуло групу, чи звіт узяв хибний проміжок. Схема показує, де почати. Після цього треба перевірити дані та конкретне виконання. Повна на вигляд схема може описувати прострочений набір або некоректне перетворення.

Відомі споживачі полегшують зміни. Розробниця хоче перейменувати поле стану. Огляд показує звірку й кілька звітів, під’єднаних до цього поля. Власники можуть домовитися, хто перевірятиме нову назву, кому тимчасово потрібні обидва поля і коли можна прибрати старе. Ця розмова найбільш корисна, поки зміну легко скоригувати. Огляд залежностей можна надавати окремо від доступу до вмісту даних.

Відоме походження допомагає зрозуміти, чому два звіти з однаковим підписом дають різний результат. Один враховує вечірнє замовлення в день отримання, другий пов’язує його з відправленням наступної доби. Порівняння джерел і часового обмеження формулює точне питання. Потім власники можуть погодити потрібну відповідь і перевірити, чи дотримано правило.

Як це виглядає на практиці

У портфоліо Engineer ApS описано роботу зі сховищами в PostgreSQL, MS SQL та BigQuery. Коли обробка перетинає межі рушіїв, особливо важливо знати, де виконано кожний крок. Далі наведено вигадану вправу з документування одного звіту.

Щоденний звіт показує замовлення, готові до відправлення. Інтернет-магазин надсилає дані замовлень, окремий потік повідомляє про стан доставки. Обробка групує записи за датою відправлення й вибирає останню належну подію стану для кожного номера. Запишіть входи й вихід, дайте обробці стале ім’я. Збережіть ідентифікатор запуску, початок, кінець,

прочитані розділи та версію правила вибору. Невдалий запуск і успішна повторна спроба мають лишитися різними подіями. Позначте той запуск, що дав опублікований результат.

Виберіть запис вівторка. Простежте дату звіту й номер замовлення через дозволені запити до вибраної події зі складу. Перевірте її час щодо межі відправлення та подивіться, що відбувається з двома подіями однакового часу. Затримка потоку має бути помітною у відомостях про звіт і в описі ходу інциденту. Якщо часову межу змінять, відповідальні зможуть відтворити звичайне та пізнє надходження.

Почніть з одного важливого звіту та кількох зв’язків, перевірених вручну. Попросіть аналітикиню простежити один рядок, а чергового — невдалий запуск. Запишіть перше місце, де комусь довелося припустити відповідь. Додайте відсутній ідентифікатор, зв’язок або версію. Цей приклад згодом допоможе оцінити автоматичне збирання.

Зафіксуйте час події, появу даних у джерелі та межу публікації, щоб відтворити ситуацію, коли рання подія надійшла із запізненням.

Де виникають проблеми

Схема може пропустити ручний експорт, збережену процедуру або електронну таблицю, яку виправляють після завантаження. Зазначте кроки, які спостерігаєте автоматично, й місце для решти відомостей. Для ручного зв’язку потрібні джерело й дата. Після перенесення процесу його слід перевірити знову.

Назви таблиць можуть збігатися у випробувальному, виробничому й архівному середовищах. Ідентифікатор має включати місце та середовище. Після переміщення зберігайте зіставлення з попередніми назвами. Потурбуйтеся також про захист метаданих. Назви джерел, запити й приклади рядків можуть розкрити відомості людям, які не мають дозволу бачити самі дані.

Для перевірки важливого зв’язку виконайте малу оборотну зміну з відомим результатом. Перевірте результат і шлях, який записав запуск. Нова величина може змінити вихід за того самого зв’язку. Іншого зв’язку слід

очікувати, якщо завдання справді читає інше джерело. Це доказ для шляху, який ви справді простежили.

У подій часто є два різні часові поля. Одне описує, коли подія сталася, друге — коли її отримала система. Вони можуть опинитися по різні боки від межі добового звіту. Зазначте, за яким часом обчислюється підсумок і за яким вимірюється затримка доставки. Додайте контрольний запис, що надійшов після межі, хоча сама подія відбулася раніше. Так наступна перевірка відрізнить помилку часового правила від справді відсутньої доставки.

Що почитати далі

- Об’єктна модель OpenLineage для завдань, наборів і запусків
- Онтологія походження W3C PROV-O
- Документація Google Cloud про походження даних

Словник

Обирайте за фактичними умовами

Надайте репрезентативний набір даних, запитання до нього та поточне вузьке місце. Узгодимо вимірювання якості, пропускної здатності та відновлення.

Дивитися, скільки це коштує Зв'язатися з нами →

Зробіть термінологію практичною

Запитайте, як «Що таке походження даних?» стосується вашої організації. Конкретного прикладу з поточної системи достатньо для початку розмови.

Переглянути наші послуги

welcome@engineer.company +45 50 10 61 84 Копенгаген, Данія

Візьміть цю сторінку із собою

Візитівка

Дбайливо зроблено з любов'ю — D.

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

Авторське право © 2025 – дотепер · Engineer ApS · Усі права захищено.

Надруковано з https://engineer.company/uk/glossary/data-lineage/