Engineer

Linaje de datos

el recorrido de los datos por orígenes y transformaciones

Qué es

El linaje de datos describe el recorrido de los datos desde sus orígenes, a través de transformaciones, hasta un conjunto de datos o un informe. Una cifra de ingresos puede proceder de los pedidos, de una tabla de reembolsos y de una regla de conversión de monedas. El linaje permite retroceder desde esa cifra hasta cada entrada. También se puede seguir hacia delante para saber qué informes se verán afectados cuando cambien una columna o la hora de una entrega. El detalle puede ir de un diagrama de tablas a una ejecución registrada con las particiones y campos exactos que leyó o escribió.

OpenLineage distingue un conjunto de datos, una tarea con nombre y una ejecución concreta de esa tarea. La tarea transforma o mueve los datos. La ejecución puede informar de las entradas que utilizó, las salidas que produjo y datos adicionales como el esquema, la versión de origen o el entorno. Esos eventos dejan un rastro hasta el trabajo que hizo una entrega. El análisis estático del código puede mostrar también dependencias posibles. El lector debe poder reconocer el método utilizado para establecer cada conexión.

El nivel de detalle depende de la pregunta. A nivel de tabla, una persona puede ver que el informe depende de dos fuentes. A nivel de columna, puede rastrear un importe a través del cálculo, o un campo después de un cambio de nombre. A nivel de ejecución puede comprobar qué versión de una tarea procesó la remesa de un día. Cuando se discute una cifra, el intervalo, la partición y la regla aplicada pueden importar tanto como los nombres de las tablas.

Utilice el concepto en una decisión

Aplique «¿Qué es el linaje de datos?» a una decisión real. Podemos ayudarle a relacionar el concepto con el sistema que está construyendo o comprando. Las cifras llegan donde hacen falta según un calendario, sin que nadie tenga que copiar una hoja de cálculo al terminar el mes.

Explore nuestros servicios Póngase en contacto →

Compruebe los supuestos detrás del término

Utilice la explicación y los trabajos enlazados para formular preguntas sobre su proyecto. Podemos examinar los supuestos con usted antes de elegir una herramienta. Un plan de infraestructura para 14 departamentos — 14 departamentos de investigación con un único plan de infraestructura

Un plan de infraestructura para 14 departamentos Póngase en contacto →

Por qué importa

El linaje ayuda a acotar una investigación. Si un indicador diario cae de golpe, se puede examinar si la carga recibió menos filas, si una unión descartó un grupo o si el informe usó un intervalo equivocado. El recorrido indica por dónde empezar. Después hay que confirmar los posibles fallos con datos de la carga observada. Un diagrama completo en apariencia puede coexistir con datos atrasados o una regla incorrecta.

Saber quién utiliza una entrada ayuda también a preparar cambios. Una desarrolladora quiere cambiar de nombre un campo de estado. El linaje indica que lo usan una conciliación y varios informes. Los responsables pueden acordar quién probará el nombre nuevo, qué consumidores necesitan los dos nombres durante la transición y cuándo se retirará el anterior. Esa conversación es más útil mientras se pueden ajustar los cambios. Al compartir la vista de dependencias, conviene mantener las restricciones sobre el contenido de los datos.

Una procedencia bien documentada ayuda a explicar por qué dos informes con el mismo rótulo dan cifras distintas. Uno puede incluir los pedidos de la noche en el total del día. El otro los asigna a la expedición del día siguiente. Identificar las fuentes, el corte horario y la transformación hace posible hablar de la diferencia con precisión. La definición acordada permite después verificar la cifra que se produjo.

Cómo se aplica en la práctica

El portafolio de Engineer ApS describe la operación de almacenes de datos en PostgreSQL, MS SQL y BigQuery. Trabajar en varios motores hace pertinente comprender sus límites y procesos. El ejercicio siguiente es ficticio y muestra qué conviene registrar para una consulta concreta.

Imaginemos un informe diario de los pedidos listos para salir. Los pedidos llegan de la tienda y otro flujo comunica los estados desde el almacén. El trabajo agrupa por fecha de expedición y elige el último estado pertinente de cada referencia de pedido. Registre por separado las entradas y la salida. Dé un nombre estable al

trabajo y anote su identificador de ejecución, principio, fin, particiones de origen y revisión de la regla. Si hay un fallo y se repite la operación, conserve ambos intentos y deje indicado cuál generó el informe publicado.

Después elija una fila del martes. Con consultas autorizadas, siga la fecha y el número de pedido hasta el estado elegido. Compruebe si ocurrió antes o después de la hora de corte y qué regla desempató dos eventos con la misma hora. Si el flujo llegó tarde, deje constancia de ello en la explicación del informe y en el registro de la incidencia. Cuando cambie la hora de corte, pida a los responsables de los informes afectados que prueben un día ordinario y una entrega tardía.

Puede empezar con un único informe importante y unas cuantas relaciones revisadas a mano. Pida a una analista que siga una fila y a una persona de operaciones que siga una ejecución fallida. Anote dónde empezaron a depender de una suposición. Añada primero el identificador, la versión o la relación que les faltó. Una vez clara la evidencia necesaria, tendrá una manera de evaluar la recogida automática.

Registre cuándo ocurrió el estado, cuándo quedó disponible para el trabajo y la hora de publicación. Así podrá explicar un evento anterior al corte que llegó tarde.

Dónde surgen los problemas

Un diagrama puede omitir una exportación manual, un procedimiento almacenado o una hoja de cálculo que alguien modifica después de descargarla. Identifique los pasos con observaciones y explique dónde se documenta el resto. Una relación manual necesita fuente y fecha. Cuando cambie el flujo, la relación debe comprobarse otra vez.

Una tabla de archivo y otra de producción pueden compartir el mismo nombre corto. Use un identificador estable que incluya el entorno y la ubicación. Mantenga la correspondencia con nombres anteriores si los datos se trasladan. Proteja los propios metadatos, ya que nombres de conjuntos, consultas de ejemplo y notas de operación pueden revelar información a personas que carecen de permiso para verla.

Termine con un cambio reversible cuyo resultado se conozca. Revise la salida y el recorrido registrado. Un valor nuevo puede cambiar el resultado y conservar la dependencia. La relación cambia cuando el trabajo pasa a utilizar otra fuente. La prueba aporta evidencia de la ruta realmente observada y un buen punto de partida para la siguiente comprobación.

Lecturas adicionales

- Modelo de objetos de OpenLineage para conjuntos, tareas y ejecuciones
- Ontología W3C PROV-O para la procedencia
- Documentación de Google Cloud sobre linaje de datos

Glosario

Elija según sus restricciones reales

Traiga un conjunto de datos representativo, las preguntas que debe responder y el cuello de botella actual. Podemos acordar cómo medir la calidad, la capacidad de procesamiento y la recuperación.

Vea cuánto cuesta Póngase en contacto →

Haga práctica la terminología

Pregunte cómo se aplica «¿Qué es el linaje de datos?» a su organización. Basta un ejemplo concreto de su sistema actual para iniciar la conversación.

Explore nuestros servicios

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

Llévese esta página

Tarjeta de contacto

Creado con esmero y cariño por **D.**

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

Derechos de autor © 2025 – actualidad · Engineer ApS · Todos los derechos reservados.

Impreso desde https://engineer.company/es/glossary/data-lineage/