mdt.

Ex Datis, Claritas

Qué es el linaje de datos y por qué es esencial en el contexto actual

Qué es el linaje de datos (data lineage): cómo traza el viaje de los datos desde su origen hasta su destino y por qué garantiza su integridad y trazabilidad.

Manuel Delgado Tenorio · 22 de noviembre de 2023

Publicado originalmente en www.manueldelgado.com/blog/que-es-el-linaje-de-datos



Si tu estrategia de datos fuera una orquesta, el linaje de datos sería el primer violín.

El linaje de datos (*data lineage*) es un aspecto crítico de la disciplina del *data management* puesto que garantiza la integridad, trazabilidad y confiabilidad de la información dentro de una organización. Representa cada paso del viaje de los datos desde su origen hasta su destino, proporcionando una comprensión integral de cómo los datos se mueven, transforman, y se utilizan a lo largo de su paso por la plataforma de datos.

El linaje responde a preguntas fundamentales sobre los datos, como de dónde provienen, cómo se transforman y hacia dónde se dirigen. Esta visibilidad es indispensable para las empresas, ya que mejora la toma de decisiones, facilita el cumplimiento de regulaciones y refuerza el gobierno de los datos. En esencia, el linaje de datos es la base sobre la que las organizaciones mantienen la calidad de los datos, garantizan el cumplimiento normativo y ejercen el gobierno.

Gracias por leer mi blog. Suscríbase gratuitamente para recibir los nuevos artículos.

Las herramientas de linaje de datos y su evolución

La evolución de las prácticas de linaje de datos y de la tecnología en la que se apoyan ha estado estrechamente vinculada a la creciente complejidad de las plataformas de datos. Inicialmente, el linaje de datos era un concepto relativamente sencillo, a menudo documentado manualmente o mediante herramientas básicas. Sin embargo, a medida que las organizaciones comenzaron a lidiar con volúmenes masivos de datos de fuentes cada vez más diversas, se hizo evidente la necesidad de soluciones de linaje de datos automatizadas, escalables y más sofisticadas.

El estado actual de la tecnología de linaje de datos se caracteriza por capacidades avanzadas que brindan visibilidad en tiempo real todo a lo largo de arquitecturas de datos complejas. Las herramientas modernas de linaje de datos utilizan *Machine Learning* para descubrir, capturar y visualizar flujos de datos de manera automática. Estas herramientas no solo rastrean el movimiento de datos, sino que también documentan las transformaciones y cálculos aplicados en cada paso, ofreciendo una comprensión detallada del viaje de los datos.

Además, la evolución de las herramientas de linaje de datos ha experimentado un cambio transformador en los últimos años: desde la gestión de tablas completas al manejo de columnas individuales. En las etapas iniciales del desarrollo del linaje de datos, las herramientas proporcionaban principalmente una visión macro, describiendo el recorrido de *datasets* enteros a lo largo de cada etapa de procesamiento. A medida que las organizaciones buscaban mayor nivel de detalle en la gestión de sus datos, surgió la necesidad de herramientas capaces de analizar y rastrear el flujo de datos a nivel de columna.

Este cambio hacia la granularidad a nivel de columna ha sido impulsado por el reconocimiento de que no todos los datos dentro de un conjunto de datos se crean a la vez ni desde el mismo origen. Las columnas individuales a menudo representan elementos de datos distintos, cada uno con su propia importancia e impacto en los procesos posteriores. Las herramientas modernas de linaje de datos se han adaptado a esta demanda, permitiendo a las organizaciones enfocarse en columnas específicas y comprender así cómo se obtienen, transforman y consumen a lo largo del ciclo de vida de los datos.

Esta evolución no solo mejora la precisión de la información extraída del linaje de datos, sino que también permite a las organizaciones identificar problemas, dependencias y transformaciones a un nivel más refinado, contribuyendo en última instancia a estrategias de gestión de datos más efectivas.

El linaje de datos como catalizador de la calidad, el cumplimiento y el gobierno

Una de las razones clave por las que creo que el linaje de datos es la función más importante en la gestión de datos es su papel protagonista en garantizar la calidad de los datos. Al comprender la ruta que toman los datos, las organizaciones pueden identificar y solucionar problemas en la fuente, evitando la propagación de errores en todo el sistema. Este enfoque proactivo hacia la calidad de los datos no solo ahorra tiempo y recursos, sino que también mejora la confiabilidad de los datos organizativos, un activo crucial en el panorama *data-driven* actual.

El linaje de datos también es clave en el cumplimiento normativo. Con un número creciente de leyes de protección de datos y regulaciones industriales, las organizaciones deben demostrar control y transparencia sobre sus procesos de datos. Las prácticas de linaje de datos proporcionan la documentación necesaria para demostrar el cumplimiento, mostrando cómo se manejan, transforman y aseguran los datos a lo largo de su ciclo de vida. Esto es particularmente crucial en sectores como el financiero o el farmacéutico, donde regulaciones estrictas rigen el uso y protección de información sensible.

Además, el linaje de datos respalda una gobernanza de datos efectiva. Al visualizar las relaciones y dependencias entre diferentes elementos de datos, las organizaciones pueden establecer y hacer cumplir políticas de gobierno de datos de manera más eficiente. Esto garantiza que los datos se utilicen adecuadamente, con pautas claras sobre acceso, uso y modificación. Un marco de gobierno de datos bien implementado, sustentado por un linaje de datos robusto, fomenta la adopción de una cultura basada en datos y alinea las prácticas de datos de la organización con sus objetivos estratégicos.

El linaje de datos en entornos multi-*cloud*

En el contexto de un entorno multi-*cloud* o híbrido (on-prem + nube), la importancia del linaje de datos se vuelve aún más evidente. A medida que las organizaciones adoptan cada vez más estrategias multi-*cloud* (aunque no siempre lo hagan por las razones correctas), la complejidad del movimiento e integración de datos aumenta. Los datos pueden atravesar diversas plataformas en la nube, infraestructuras locales y configuraciones híbridas. En tales escenarios, comprender el recorrido de los datos se vuelve fundamental para mantener el control, la seguridad y el cumplimiento.

El linaje de datos ayuda a las organizaciones a navegar la intrincada red de interacciones entre diferentes entornos *cloud*, garantizando que las políticas de gobierno de datos se apliquen de manera consistente y que se cumplan los requisitos regulatorios en todo su universo de datos. Por supuesto, el panorama multi-*cloud* introduce nuevos desafíos para el linaje de datos, como problemas de interoperabilidad, esquemas diversos de almacenamiento de datos y posibles latencias en el movimiento de datos. Superar estos desafíos requiere soluciones de

linaje de datos sofisticadas que puedan mapear y rastrear flujos de datos de manera fluida a través de diversas plataformas *cloud*, proporcionando una vista unificada que capacita a las organizaciones para operar en arquitecturas multi-*cloud* sin comprometer la visibilidad y el control de los datos.

El linaje de datos es cada vez más relevante

En conclusión, el linaje de datos destaca como la función más importante en la gestión de datos debido a su papel crucial en garantizar la calidad de los datos, facilitar el cumplimiento normativo y respaldar un gobierno de datos efectivo. La evolución del linaje de datos desde la documentación manual hasta herramientas automatizadas y avanzadas refleja la creciente complejidad de los entornos modernos de datos. A medida que las organizaciones continúan navegando los desafíos de *big data*, la computación en la nube y las fuentes de datos cada vez más diversas y numerosas, la importancia de soluciones robustas de linaje de datos solo se intensificará, consolidando su posición como piedra angular de estrategias exitosas de gestión de datos.

PARA SEGUIR LEYENDO

Qué pasa con los datos que generan las aplicaciones internas mal diseñadas, y todo lo que he escrito sobre estrategia y gestión de datos.



Sobre el autor

Manuel Delgado Tenorio

Manuel Delgado Tenorio es asesor independiente de estrategia de datos e inteligencia artificial, profesor e investigador. Cofundó y vendió una *startup* de inteligencia artificial, fue socio de la práctica de Data & Analytics para servicios financieros en EY y responsable de Data Analytics para el sur de Europa en AWS. Es profesor asociado en IE University, donde imparte la primera asignatura de grado sobre Customer Lifetime Value en España, investiga sobre el CLV en su doctorado en la Universidad Complutense de Madrid y es autor del libro "Customer Lifetime Value (CLV): No todos tus clientes valen lo mismo".

manueldelgado.com · LinkedIn · X

Recibe cada artículo en tu correo con la newsletter Ex Datis, Claritas: manueldelgado.com/newsletter/suscribete