Engineer

Conjunto de datos

colección de datos publicada como una unidad

Qué es

Un conjunto de datos, en inglés dataset o data set, es una colección delimitada de datos que se trata como una unidad: tiene un nombre, una fuente, un periodo que cubre y una estructura común a sus registros, y esos cuatro datos establecen dónde empieza y dónde termina. Un mes de lecturas de contadores en un archivo es un conjunto de datos; también lo es una exportación nocturna de tres tablas relacionadas. La expresión no dice nada sobre el tamaño ni el formato, solo que los datos que contiene pertenecen al mismo conjunto.

Las páginas de este sitio utilizan la expresión en ese sentido práctico. Una habla de un conjunto histórico en formato JSONL y de conjuntos de 1 día y 30 días utilizados para pruebas de rendimiento; otra, de conjuntos depurados y tablas exportadas para cargarlos en otro lugar; una tercera, de enormes conjuntos generados por contadores inteligentes. En cada caso, la expresión delimita unos datos antes de hacer algo con ellos.

El vocabulario de catálogos de datos del W3C, DCAT, da a la expresión un marco formal. La versión 3 se convirtió en Recomendación del W3C el 22 de agosto de 2024 y define un conjunto de datos como una «colección de datos, publicada o mantenida por un único agente y disponible para su acceso o descarga en una o más serializaciones o formatos». La noción es amplia a propósito: incluye números, texto, imágenes y sonido. DCAT distingue después tres cosas que el habla cotidiana mezcla:

- **Conjunto de datos**: los datos como concepto, por ejemplo el consumo horario por zona de 2024.
- **Distribución**: una forma accesible de esos datos, como un archivo CSV descargable; un conjunto puede tener varias.

- **Serie de conjuntos de datos**: conjuntos publicados por separado con suficientes características comunes para agruparlos, como un archivo por mes.

Por tanto, dos archivos en formatos distintos pueden ser un solo conjunto de datos, y dos archivos mensuales bajo un mismo nombre pueden ser dos miembros de una serie.

La expresión tiene un segundo significado, más restringido, en Google BigQuery. Allí, un conjunto de datos es un contenedor de tablas, vistas y otros recursos dentro de un proyecto. Una tabla tiene que pertenecer a un conjunto, así que este se crea antes de cargar ningún dato, y se identifica como `project.dataset` en GoogleSQL y como `project:dataset` en la herramienta de línea de comandos bq. El contexto determina qué significado se está utilizando.

Utilice el concepto en una decisión

Aplique «¿Qué es un conjunto de datos?» a una decisión real. Podemos ayudarle a relacionar el concepto con el sistema que está construyendo o comprando. Paneles e informes con las pocas cifras de las que realmente depende una decisión, actualizados sin tener que pedirlo.

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. Comunicación de marca guiada por la respuesta del público — comunicación de marca guiada por la respuesta del público

Comunicación de marca guiada por la respuesta del público Póngase en contacto →

Por qué importa

Delimitar los datos permite compararlos. Dos conjuntos de lecturas solo pueden compararse cuando cubren el mismo periodo, los mismos lugares y el mismo intervalo entre registros, y dar nombre a un conjunto es la manera de fijar esas condiciones. DCAT asigna al intervalo una propiedad propia, la resolución temporal: la separación más pequeña entre los elementos del conjunto, escrita como una duración, por ejemplo `P1D` para un día. Un conjunto que declara su resolución indica al lector si sus filas representan horas, cuartos de hora o algo irregular antes de hacer la primera suma.

La descripción tiene que viajar con los datos. DCAT enumera lo que una distribución puede declarar sobre sí misma: un tipo de medio, un formato, un formato de compresión, un formato de empaquetado, un tamaño en bytes y una suma de comprobación; esta última permite a quien la recibe verificar que el

contenido no ha cambiado. El propio ejemplo de formato de compresión de la especificación es gzip. Un archivo que llega con esos datos puede comprobarse y abrirse; si llega sin ellos, hay que adivinarlos.

El formato de una distribución decide quién puede leerla. Tres son habituales:

- **CSV.** El RFC 4180, de 2005, documenta la estructura que parecen seguir la mayoría de las implementaciones: un registro por línea, una línea de cabecera opcional, comillas dobles alrededor de cualquier campo que contenga una coma, un salto de línea o una comilla doble, y las comillas dobles dentro de un campo escritas dos veces. Casi cualquier herramienta lo abre.
- **JSON Lines.** Un valor JSON válido por línea, en UTF-8, con `\n` como separador de líneas. Los archivos se guardan como `.jsonl` o, comprimidos con gzip, como `.jsonl.gz`. Se recomienda encarecidamente un salto de línea final, pero no es obligatorio. Los programas leen estos archivos registro a registro y las herramientas de texto de tipo Unix y las tuberías del shell los manejan bien.
- **Parquet.** Un formato de archivo de datos de código abierto que mantiene juntos los valores de cada columna. Está construido para almacenar y recuperar datos de forma eficiente, y sus esquemas de compresión y codificación permiten manejar datos complejos en grandes volúmenes.

Este es un registro en los dos primeros formatos, donde la coma dentro de la nota obliga a usar comillas en CSV:

```
id,read_at,kwh,note
M-17,2024-01-08T10:00:00Z,3.21,"late, manual"

{"id":"M-17","read_at":"2024-01-08T10:00:00Z","kwh":3.21,"note":"late, manual"}
```

Cómo se aplica en la práctica

Construir un conjunto de datos a partir de muchas fuentes fue la tarea cuando los datos eléctricos llegaban con la frecuencia de envío propia de cada fuente, desde intervalos horarios hasta intervalos de 15 minutos y, en algunos casos, marcas de tiempo irregulares, lo que dificultaba producir un conjunto coherente y comparable. El trabajo tomó datos de eventos en bruto de un conjunto histórico JSONL y los convirtió en consumo horario, con una fila por hora y zona, sumando la combinación de fuentes de producción en cada intervalo, añadiendo importaciones y restando exportaciones. Se escribieron tres variantes, en Python,

JQ y SQL. El flujo de procesamiento Python analizaba los archivos JSONL para convertirlos en data frames, remuestreaba las series temporales a intervalos horarios y exportaba los resultados como CSV o los cargaba en una base de datos SQLite ligera para inspeccionarlos. En las pruebas con conjuntos históricos de 1 día y 30 días, fue el más rápido de los tres, con 34 ms para un día y 180 ms para treinta. El trabajo entre entrada y salida cambia el grano, es decir, el nivel de detalle: pasa del intervalo propio de cada fuente a una fila fija por hora y zona.

Trasladar un conjunto de datos es otra tarea. Seis flujos ETL/ELT llevaron datos de un almacén remoto detrás de una VPN IPsec a Google BigQuery. Los datos llegaban en archivos comprimidos que contenían archivos Parquet, CSV y .bak, y los scripts Bash y Python clasificaban cada archivo por tipo y esquema. Una vez que los procedimientos habían limpiado, normalizado y enriquecido las filas, los conjuntos depurados y las tablas se exportaban a formatos intermedios, se comprimían con gzip y se transferían a un bucket de Google Cloud Storage, desde el que la CLI bq y los scripts Python los cargaban en BigQuery. Llevar esos conjuntos allí a mano había requerido 3 horas o más; de forma automatizada, el traslado tardaba menos de 20 minutos, y las tablas se actualizaban a diario en lugar de una vez por semana.

Leer un conjunto es la tercera tarea. Los contadores inteligentes de varias regiones geográficas produjeron enormes conjuntos de datos que sustentaron más de 500 informes de análisis escritos en SQL puro sobre PostgreSQL y BigQuery. Los datos contenían coordenadas de geolocalización, identificadores de contador, consumo energético con marca de tiempo y metadatos ambientales. Cada informe incluía SQL anotado para que los compañeros pudieran reproducirlo y auditarlo, y las notas de resolución de problemas registraban incidencias habituales de calidad de los datos, como coordenadas GPS ausentes o valores de contador corruptos, con recomendaciones para tratarlas. Los conjuntos que los compañeros pueden reproducir y auditar son conjuntos en los que confiarán los planificadores, y el trabajo contribuyó a una mejora del 10–15 % en la eficiencia de la planificación operativa.

Dónde surgen los problemas

Detrás de la mayoría de los problemas de un conjunto de datos hay información sobre los propios datos que se quedó en la cabeza de una persona.

No se declara el grano. Sumar filas horarias y filas de un cuarto de hora cuenta algunos intervalos cuatro veces. Escriba el grano en una frase, por ejemplo una fila por hora y zona en UTC, y ajuste todas las fuentes a él antes de sumar.

Una ausencia se interpreta como cero. Un campo vacío, un valor null y un 0 significan tres cosas distintas, y un total trata silenciosamente una lectura ausente como consumo cero. Las coordenadas GPS ausentes y los valores de contador corruptos eran lo bastante habituales en el proyecto de contadores inteligentes como para documentar su tratamiento. Decida por columna si una ausencia se descarta, se rellena o se marca, e indique la decisión en las notas del conjunto.

El dialecto CSV se desvía. El RFC 4180 documenta la estructura común y los archivos reales se apartan de ella: otro separador, una cabecera ausente, un campo con una coma al que le faltan las comillas. El RFC deja la cabecera como opcional, señalada solo por un parámetro MIME que se pierde al copiar el archivo a disco. Compruebe que la primera línea de cada archivo nuevo coincide con la cabecera esperada antes de cargarlo.

Una copia olvida de dónde procede. Un extracto depurado cargado en un segundo sistema parece el original un mes después. Registre la fuente, el periodo, el momento de extracción y una suma de comprobación junto a cada copia para que un informe pueda indicar qué extracto leyó.

Un conjunto de BigQuery acaba en el lugar equivocado. La ubicación de un conjunto se fija al crearlo; moverlo exige copiarlo o volver a crearlo. Todas las tablas de una consulta deben estar en conjuntos de la misma ubicación, así que elíjala antes de la primera carga.

Lecturas adicionales

- Vocabulario de catálogos de datos (DCAT), versión 3 — las definiciones del W3C de conjunto de datos, distribución y serie de conjuntos, con las propiedades que describen cada uno.
- BigQuery: introducción a los conjuntos de datos — el significado de contenedor: proyectos, ubicaciones, nombres y límites.
- JSON Lines — las tres reglas del formato y sus convenciones de nombres de archivo.

- RFC 4180: formato común y tipo MIME para archivos CSV — la estructura CSV que siguen la mayoría de las herramientas, sus reglas de comillas y el tipo text/csv.
- Apache Parquet: descripción general — qué es el formato de archivo orientado a columnas y dónde está su especificación.

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 un conjunto 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/dataset/