Engineer

Dataset

eine Datensammlung, die als eine Einheit veröffentlicht wird

Was es ist

Ein Dataset, im Englischen auch data set geschrieben, ist ein abgegrenzter Datenbestand, den Menschen als eine Einheit behandeln: Er hat einen Namen, eine Quelle, einen abgedeckten Zeitraum und eine gemeinsame Struktur seiner Datensätze. Diese vier Angaben bestimmen seine Grenzen. Eine Datei mit den Zählerständen eines Monats ist ein Dataset; ebenso ein nächtlicher Export von drei zusammengehörigen Tabellen. Der Begriff sagt nichts über Größe oder Format aus, sondern nur, dass die enthaltenen Daten zusammengehören.

Die Seiten dieser Website verwenden den Begriff in diesem praktischen Sinn. Eine spricht von einem historischen Dataset im JSONL-Format und von Datasets mit 1 Tag und 30 Tagen für Benchmarks, eine andere von kuratierten Datasets und Tabellen, die zum Laden in andere Systeme exportiert werden, eine dritte von riesigen Datasets aus intelligenten Stromzählern. Jedes Mal grenzt das Wort Daten ein, bevor etwas mit ihnen geschieht.

Das Data Catalog Vocabulary des W3C, DCAT, gibt dem Begriff einen formalen Rahmen. Version 3 wurde am 22. August 2024 zur W3C-Empfehlung. Sie definiert ein Dataset als eine „Sammlung von Daten, die von einem einzelnen Akteur veröffentlicht oder kuratiert wird und in einer oder mehreren Serialisierungen beziehungsweise einem oder mehreren Formaten zum Zugriff oder Herunterladen verfügbar ist“. Der Begriff ist bewusst weit gefasst: Zahlen, Text, Bilder und Ton zählen gleichermaßen dazu. DCAT unterscheidet dann drei Dinge, die im alltäglichen Sprachgebrauch ineinander übergehen:

- **Dataset**: die Daten als abstrakter Bestand, etwa der stündliche Verbrauch pro Zone für 2024.
- **Distribution**: eine zugängliche Form davon, etwa eine herunterladbare CSV-Datei; ein Dataset kann mehrere davon haben.

- **Dataset-Serie**: getrennt veröffentlichte Datasets, die genügend Gemeinsamkeiten für eine Gruppierung haben, etwa eine Datei pro Monat.

Zwei Dateien in unterschiedlichen Formaten können also ein Dataset sein, und zwei Monatsdateien unter einem Namen können zwei Mitglieder einer Serie sein.

In Google BigQuery hat das Wort eine zweite, engere Bedeutung. Dort ist ein Dataset ein Container für Tabellen, Views und andere Ressourcen innerhalb eines Projekts. Eine Tabelle muss zu einem Dataset gehören. Deshalb wird ein Dataset angelegt, bevor Daten geladen werden. In GoogleSQL wird es als `project.dataset` und im bq-Kommandozeilenwerkzeug als `project:dataset` angesprochen. Der Kontext entscheidet, welche Bedeutung gemeint ist.

Das Konzept für eine Entscheidung nutzen

Nutzen Sie „Was ist ein Dataset?“ für eine konkrete Entscheidung. Wir können helfen, das Konzept mit dem System in Verbindung zu bringen, das Sie entwickeln oder einkaufen. Dashboards und Berichte zeigen die wenigen Zahlen, von denen eine Entscheidung tatsächlich abhängt, und aktualisieren sich ohne Aufforderung.

Unsere Leistungen entdecken Kontakt aufnehmen →

Die Annahmen hinter dem Begriff prüfen

Nutzen Sie die Erklärung und die verlinkten Arbeiten, um Fragen für Ihr Projekt zu formulieren. Wir können die Annahmen gemeinsam mit Ihnen prüfen, bevor ein Werkzeug ausgewählt wird. Markenkommunikation anhand von Publikumsreaktionen — Markenkommunikation anhand von Publikumsreaktionen

Markenkommunikation anhand von Publikumsreaktionen Kontakt aufnehmen →

Warum es wichtig ist

Eine Abgrenzung macht Daten vergleichbar. Zwei Bestände von Messwerten lassen sich nur vergleichen, wenn sie denselben Zeitraum, dieselben Orte und dieselben zeitlichen Abstände zwischen den Datensätzen abdecken. Mit der Benennung eines Datasets werden diese Bedingungen festgelegt. DCAT gibt dem Abstand eine eigene Eigenschaft, die zeitliche Auflösung: den kleinsten Abstand zwischen den Elementen des Datasets, angegeben als Zeitdauer wie `P1D` für einen Tag. Ein Dataset mit angegebener Auflösung sagt dem Leser schon vor der ersten Summe, ob seine Zeilen Stunden, Viertelstunden oder unregelmäßige Intervalle darstellen.

Die Beschreibung muss mit den Daten weitergegeben werden. DCAT nennt die Angaben, die eine Distribution über sich machen kann: Medientyp, Format, Kompressionsformat, Verpackungsformat, Größe in Bytes und Prüfsumme. Mit Letzterer kann ein Empfänger prüfen, ob der Inhalt unverändert geblieben ist.

Das eigene Beispiel der Spezifikation für ein Kompressionsformat ist gzip. Eine Datei mit diesen Angaben lässt sich prüfen und öffnen; bei einer Datei ohne sie muss man raten.

Das Format einer Distribution entscheidet darüber, wer sie lesen kann. Drei sind verbreitet:

- **CSV.** RFC 4180 aus dem Jahr 2005 hält den Aufbau fest, dem die meisten Implementierungen offenbar folgen: ein Datensatz pro Zeile, eine optionale Kopfzeile, doppelte Anführungszeichen um jedes Feld mit Komma, Zeilenumbruch oder doppeltem Anführungszeichen und ein doppelt geschriebenes doppeltes Anführungszeichen innerhalb eines Feldes. Fast jedes Werkzeug kann es öffnen.
- **JSON Lines.** Ein gültiger JSON-Wert pro Zeile in UTF-8, mit `\n` als Zeilentrenner. Die Dateien werden als `.jsonl` oder gzip-komprimiert als `.jsonl.gz` gespeichert. Ein abschließender Zeilenumbruch wird dringend empfohlen, ist aber nicht vorgeschrieben. Programme lesen diese Dateien Datensatz für Datensatz, und Textwerkzeuge im Unix-Stil sowie Shell-Pipes kommen gut mit dem Format zurecht.
- **Parquet.** Ein quelloffenes Datendateiformat, das die Werte jeder Spalte zusammenhält. Es ist für effizientes Speichern und Abrufen ausgelegt. Seine Kompressions- und Kodierungsverfahren ermöglichen die Verarbeitung komplexer Daten in großen Mengen.

Hier steht ein Datensatz in den ersten beiden Formaten. Das Komma in der Notiz erfordert in CSV Anführungszeichen:

```
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"}
```

Wie es in der Praxis aussieht

Aus vielen Quellen ein Dataset zu formen war die Aufgabe, als Stromdaten im jeweiligen Meldetakt ihrer Quelle eintrafen: von stündlichen über 15-minütige Intervalle bis hin zu teils unregelmäßigen Zeitstempeln. Das erschwerte die Erstellung eines einheitlichen, vergleichbaren Datenbestands. Dabei wurden rohe Ereignisdaten aus einem historischen JSONL-Dataset in stündlichen Verbrauch umgewandelt: eine Zeile pro Stunde und Zone, indem der

Erzeugungsmix jedes Intervalls summiert, Importe addiert und Exporte abgezogen wurden. Drei Varianten entstanden, in Python, JQ und SQL. Die Python-Pipeline las die JSONL-Dateien in Data Frames ein, führte ein Resampling der Zeitreihen auf Stundenintervalle durch und exportierte die Ergebnisse als CSV oder lud sie zur Untersuchung in eine schlanke SQLite-Datenbank. Im Benchmark mit historischen Datasets von 1 Tag und 30 Tagen war sie die schnellste der drei: 34 ms für einen Tag und 180 ms für dreißig. Zwischen Ein- und Ausgabe ändert sich die Granularität: vom jeweiligen Abstand der Quelle zu einer festen Zeile pro Stunde und Zone.

Ein Dataset zu transportieren ist eine andere Aufgabe. Sechs ETL/ELT-Pipelines übertrugen Daten aus einem entfernten Warehouse hinter einem IPsec-VPN nach Google BigQuery. Die Daten kamen als komprimierte Archive mit Parquet-, CSV- und .bak-Dateien an. Bash- und Python-Skripte ordneten jede Datei nach Typ und Schema ein. Nachdem Prozeduren die Zeilen bereinigt, normalisiert und angereichert hatten, wurden kuratierte Datasets und Tabellen in Zwischenformate exportiert, mit gzip komprimiert und in einen Bucket in Google Cloud Storage übertragen. Von dort luden die bq-CLI und Python-Skripte sie in BigQuery. Die manuelle Übertragung dieser Datasets hatte 3+ Stunden gedauert; automatisiert dauerte sie unter 20 Minuten. Die Tabellen wurden nun täglich statt einmal pro Woche aktualisiert.

Ein Dataset zu lesen ist die dritte Aufgabe. Intelligente Stromzähler in mehreren geografischen Regionen erzeugten riesige Datasets als Grundlage für mehr als 500 Analyseberichte in reinem SQL auf PostgreSQL und BigQuery. Die Daten enthielten geografische Koordinaten, Zähler-IDs, Energieverbrauch mit Zeitstempeln und umgebungsbezogene Metadaten. Jeder Bericht enthielt kommentiertes SQL, damit Kollegen ihn reproduzieren und prüfen konnten. Hinweise zur Fehlerbehebung dokumentierten häufige Datenqualitätsprobleme wie fehlende GPS-Koordinaten oder beschädigte Zählerwerte samt empfohlenem Umgang damit. Datenbeständen, die Kollegen reproduzieren und prüfen können, vertrauen auch Planer. Die Arbeit trug zu einer um 10–15 % höheren Effizienz der betrieblichen Planung bei.

Wo es schiefgeht

Die meisten Probleme mit Datasets gehen auf eine Dateneigenschaft zurück, die nur im Kopf einer Person geblieben ist.

Die Granularität bleibt unausgesprochen. Wer stündliche Zeilen zu Viertelstundenzeilen addiert, zählt manche Intervalle vierfach. Halten Sie die Granularität in einem Satz fest, etwa eine Zeile pro Stunde und Zone in UTC, und gleichen Sie jede Quelle daran an, bevor Sie summieren.

Eine Lücke wird als null gelesen. Ein leeres Feld, ein Nullwert und eine 0 bedeuten drei verschiedene Dinge. Eine Summe behandelt einen fehlenden Messwert stillschweigend als keinen Verbrauch. Fehlende GPS-Koordinaten und beschädigte Zählerwerte kamen im Smart-Meter-Projekt häufig genug vor, um den Umgang damit zu dokumentieren. Entscheiden Sie pro Spalte, ob eine Lücke weggelassen, gefüllt oder markiert wird, und halten Sie diese Entscheidung in den Dataset-Hinweisen fest.

Der CSV-Dialekt verändert sich. RFC 4180 dokumentiert den üblichen Aufbau, doch echte Dateien weichen davon ab: ein anderes Trennzeichen, eine fehlende Kopfzeile, ein Feld mit Komma, dessen Anführungszeichen fehlen. Der RFC lässt die Kopfzeile optional und kennzeichnet sie nur über einen MIME-Parameter, der beim Kopieren einer Datei auf die Festplatte verloren geht. Vergleichen Sie vor dem Laden bei jeder neuen Datei die erste Zeile mit der erwarteten Kopfzeile.

Eine Kopie verliert ihre Herkunft. Ein kuratierter Extrakt, der in ein zweites System geladen wurde, sieht einen Monat später genauso aus wie das Original. Halten Sie neben jeder Kopie die Quelle, den Zeitraum, den Extraktionszeitpunkt und eine Prüfsumme fest, damit ein Bericht angeben kann, welchen Extrakt er gelesen hat.

Ein BigQuery-Dataset landet am falschen Ort. Der Speicherort eines Datasets wird beim Anlegen festgelegt; ein Umzug erfordert das Kopieren oder Neuerstellen des Datasets. Jede Tabelle einer Abfrage muss in Datasets am selben Speicherort liegen. Wählen Sie den Ort deshalb vor dem ersten Laden.

Weiterführende Lektüre

- Data Catalog Vocabulary (DCAT) – Version 3 — die W3C-Definitionen von Dataset, Distribution und Dataset-Serie samt den Eigenschaften, die sie beschreiben.
- BigQuery: Einführung in Datasets — die Containerbedeutung des Wortes: Projekte, Speicherorte, Benennung und Grenzen.
- JSON Lines — die drei Regeln des Formats und seine Dateinamenskonventionen.

- RFC 4180: Gemeinsames Format und MIME-Typ für CSV-Dateien — der CSV-Aufbau, dem die meisten Werkzeuge folgen, seine Regeln für Anführungszeichen und der Typ text/csv.
- Apache Parquet: Überblick — das spaltenorientierte Dateiformat und der Ort seiner Spezifikation.

Glossar

Für Ihre tatsächlichen Rahmenbedingungen auswählen

Bringen Sie einen repräsentativen Datensatz, die Fragen, die er beantworten muss, und den aktuellen Engpass mit. Wir können vereinbaren, wie Qualität, Durchsatz und Wiederherstellung gemessen werden.

Preise ansehen Kontakt aufnehmen →

Die Begriffe praktisch nutzbar machen

Fragen Sie, wie „Was ist ein Dataset?“ auf Ihre Organisation zutrifft. Ein konkretes Beispiel aus Ihrem bestehenden System reicht aus, um das Gespräch zu beginnen.

Unsere Leistungen entdecken

welcome@engineer.company +45 50 10 61 84 Kopenhagen, Dänemark

Nehmen Sie diese Seite mit

Kontaktkarte

Mit Sorgfalt und Liebe gestaltet von D.

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

Copyright © 2025 – heute · Engineer ApS · Alle Rechte vorbehalten.

Gedruckt von https://engineer.company/de/glossary/dataset/