Engineer

Jeu de données

collection de données publiée comme un tout

Ce que c’est

Un jeu de données, appelé dataset ou data set en anglais, est un ensemble délimité de données que l’on traite comme un tout : il possède un nom, une source, une période couverte et une structure commune à ses enregistrements. Ces quatre éléments définissent ses limites. Un mois de relevés de compteurs dans un fichier constitue un jeu de données ; un export nocturne de trois tables liées en constitue un autre. Le mot ne dit rien de la taille ni du format, seulement que les données qu’il contient forment un ensemble.

Les pages de ce site emploient le mot dans ce sens pratique. L’une parle d’un jeu historique au format JSONL et de jeux de 1 jour et de 30 jours utilisés pour des benchmarks ; une autre, de jeux préparés et de tables exportées pour être chargés ailleurs ; une troisième, de vastes jeux produits par des compteurs intelligents. Chaque fois, le mot délimite les données avant qu’on leur applique un traitement.

Le vocabulaire de catalogues de données du W3C, DCAT, donne au mot un cadre formel. Sa version 3 est devenue une recommandation du W3C le 22 août 2024. Elle définit un jeu de données comme une « collection de données, publiée ou organisée par un seul agent, et accessible ou téléchargeable dans une ou plusieurs sérialisations ou un ou plusieurs formats ». La notion est volontairement large : nombres, textes, images et sons en relèvent tous. DCAT distingue ensuite trois choses que l’usage courant confond :

- **Jeu de données** : les données en tant que concept, par exemple la consommation horaire par zone pour 2024.
- **Distribution** : une forme accessible de ce jeu, par exemple un fichier CSV téléchargeable ; un jeu peut en avoir plusieurs.

- **Série de jeux de données** : des jeux publiés séparément qui ont assez de points communs pour être regroupés, par exemple un fichier par mois.

Deux fichiers dans des formats différents peuvent donc constituer un seul jeu de données, et deux fichiers mensuels sous un même nom peuvent être deux membres d'une série.

Le mot possède un second sens, plus étroit, dans Google BigQuery. Un dataset y est un conteneur de tables, de vues et d'autres ressources, à l'intérieur d'un projet. Toute table doit appartenir à un dataset ; celui-ci est donc créé avant tout chargement de données. Il est désigné par `project.dataset` dans GoogleSQL et par `project:dataset` dans l'outil en ligne de commande bq. Le contexte détermine le sens employé.

Utilisez le concept pour prendre une décision

Utilisez « Qu'est-ce qu'un jeu de données ? » pour éclairer une décision réelle. Nous pouvons vous aider à relier ce concept au système que vous construisez ou achetez. Des tableaux de bord et des rapports présentant les quelques chiffres qui déterminent une décision, actualisés automatiquement.

Découvrir nos prestations Contactez-nous →

Vérifiez les hypothèses derrière le terme

Appuyez-vous sur l'explication et les réalisations liées pour formuler les questions de votre projet. Nous pouvons examiner les hypothèses avec vous avant le choix d'un outil. Une communication de marque éclairée par le public — une communication de marque éclairée par les réactions du public

Une communication de marque éclairée par le public Contactez-nous →

Pourquoi c'est important

Une frontière rend les données comparables. Deux ensembles de relevés ne se comparent que s'ils couvrent la même période, les mêmes lieux et le même espacement entre les enregistrements. Nommer un jeu de données permet de fixer ces conditions. DCAT donne à cet espacement sa propre propriété, la résolution temporelle : le plus petit intervalle entre les éléments du jeu, écrit comme une durée, par exemple `P1D` pour un jour. Un jeu qui indique sa résolution dit au lecteur, avant la première somme, si ses lignes correspondent à des heures, des quarts d'heure ou à des intervalles irréguliers.

La description doit accompagner les données. DCAT énumère ce qu'une distribution peut indiquer sur elle-même : un type de média, un format, un format de compression, un format d'empaquetage, une taille en octets et une somme de contrôle. Cette dernière permet au destinataire de vérifier que le

contenu n'a pas changé. L'exemple de format de compression donné par la spécification est gzip. Un fichier accompagné de ces informations peut être vérifié et ouvert ; un fichier livré sans elles impose de les deviner.

Le format d'une distribution détermine qui peut la lire. Trois formats sont courants :

- **CSV.** La RFC 4180, de 2005, consigne la structure que semblent suivre la plupart des implémentations : un enregistrement par ligne, une ligne d'en-tête facultative, des guillemets doubles autour de tout champ contenant une virgule, un saut de ligne ou un guillemet double, et tout guillemet double interne écrit deux fois. Presque tous les outils l'ouvrent.
- **JSON Lines.** Une valeur JSON valide par ligne, en UTF-8, avec `\n` comme séparateur de lignes. Les fichiers sont enregistrés sous `.jsonl` ou, compressés avec gzip, `.jsonl.gz` . Un saut de ligne final est fortement recommandé, mais pas obligatoire. Les programmes lisent ces fichiers un enregistrement à la fois ; les outils de texte de type Unix et les tubes du shell traitent bien ce format.
- **Parquet.** Un format de fichiers de données open source qui conserve ensemble les valeurs de chaque colonne. Il est conçu pour stocker et récupérer efficacement les données ; ses mécanismes de compression et d'encodage permettent de traiter des données complexes en masse.

Voici un enregistrement dans les deux premiers formats. La virgule dans la note impose les guillemets 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"}
```

À quoi cela ressemble en pratique

Former un seul jeu à partir de nombreuses sources était l'objectif lorsque les données d'électricité arrivaient à la fréquence propre à chaque source, d'intervalles horaires à des intervalles de 15 minutes, voire avec des horodatages irréguliers. Il était donc difficile de produire un jeu cohérent et comparable. Le travail a transformé les événements bruts d'un jeu historique JSONL en consommation horaire, à raison d'une ligne par heure et par zone, en additionnant les composantes du mix de production de chaque intervalle, en

ajoutant les importations et en retranchant les exportations. Trois variantes ont été écrites, en Python, JQ et SQL. Le pipeline Python convertissait les fichiers JSONL en structures de données tabulaires (data frames), rééchantillonnait les séries temporelles par intervalles horaires, puis exportait les résultats en CSV ou les chargeait dans une base SQLite légère pour inspection. Mesuré sur des jeux historiques de 1 jour et de 30 jours, il était le plus rapide des trois : 34 ms pour un jour et 180 ms pour trente. Le travail entre l'entrée et la sortie est un changement de granularité : de l'espacement propre à chaque source à une granularité fixe d'une ligne par heure et par zone.

Déplacer un jeu de données est un autre travail. Un ensemble de six pipelines ETL/ELT a transporté des données d'un entrepôt distant derrière un VPN IPsec vers Google BigQuery. Les données arrivaient dans des archives compressées contenant des fichiers Parquet, CSV et .bak ; des scripts Bash et Python classaient chaque fichier par type et par schéma. Après nettoyage, normalisation et enrichissement des lignes par des procédures, les jeux préparés et les tables étaient exportés dans des formats intermédiaires, compressés avec gzip et transférés dans un bucket Google Cloud Storage. L'outil bq et des scripts Python les chargeaient ensuite dans BigQuery depuis ce bucket. Le transfert manuel de ces jeux avait pris au moins 3 heures ; automatisé, il prenait moins de 20 minutes, et les tables étaient actualisées chaque jour plutôt qu'une fois par semaine.

Lire un jeu de données est le troisième travail. Des compteurs intelligents répartis dans plusieurs régions géographiques ont produit les vastes jeux à la base de plus de 500 rapports d'analyse écrits en SQL pur sous PostgreSQL et BigQuery. Les données contenaient des coordonnées géographiques, des identifiants de compteurs, des consommations d'énergie horodatées et des métadonnées environnementales. Chaque rapport comportait du SQL annoté pour que les collègues puissent le reproduire et l'auditer. Des notes de dépannage consignaient les problèmes courants de qualité, comme les coordonnées GPS manquantes ou les valeurs de compteurs corrompues, avec les traitements recommandés. Des jeux que les collègues peuvent reproduire et auditer sont des jeux auxquels les responsables de la planification peuvent se fier, et le travail a contribué à un gain de 10–15 % d'efficacité dans la planification opérationnelle.

Où les problèmes surviennent

La plupart des difficultés avec les jeux de données viennent d'une information sur les données restée dans la tête d'une seule personne.

La granularité n'est pas indiquée. Additionner des lignes horaires à des lignes par quart d'heure compte certains intervalles quatre fois. Écrivez la granularité en une phrase, par exemple une ligne par heure et par zone en UTC, et alignez toutes les sources sur elle avant de calculer les totaux.

Une absence se lit comme un zéro. Un champ vide, une valeur nulle et un 0 signifient trois choses différentes. Une somme traite pourtant silencieusement une mesure manquante comme une consommation nulle. Sur le projet de compteurs intelligents, les coordonnées GPS manquantes et les valeurs corrompues étaient assez fréquentes pour justifier un traitement documenté. Décidez pour chaque colonne si une absence est écartée, comblée ou signalée, et précisez ce choix dans les notes du jeu.

Le dialecte CSV dérive. La RFC 4180 documente la structure courante, mais les fichiers réels s'en écartent : autre séparateur, en-tête absent, champ contenant une virgule sans les guillemets nécessaires. La RFC laisse l'en-tête facultatif ; sa présence n'est signalée que par un paramètre MIME, qui ne suit pas le fichier lorsqu'on le copie sur disque. Vérifiez la première ligne de chaque nouveau fichier par rapport à l'en-tête attendu avant de le charger.

Une copie oublie sa provenance. Un extrait préparé chargé dans un second système semble identique à l'original un mois plus tard. Consignez la source, la période, l'horodatage de l'extraction et une somme de contrôle à côté de chaque copie, pour qu'un rapport puisse indiquer l'extrait qu'il a lu.

Un dataset BigQuery arrive au mauvais endroit. L'emplacement d'un dataset est fixé à sa création ; le déplacer impose de le copier ou de le recréer. Toutes les tables d'une même requête doivent appartenir à des datasets situés au même emplacement. Choisissez donc celui-ci avant le premier chargement.

Pour aller plus loin

- Vocabulaire de catalogues de données (DCAT), version 3 — les définitions W3C du jeu de données, de la distribution et de la série de jeux, avec les propriétés qui décrivent chacun.
- BigQuery : introduction aux datasets — le sens de conteneur : projets, emplacements, nommage et limites.
- JSON Lines — les trois règles du format et ses conventions de noms de fichiers.

- RFC 4180 : format commun et type MIME des fichiers CSV — la structure CSV suivie par la plupart des outils, ses règles de guillemets et le type text/csv.
- Apache Parquet : présentation — ce qu'est le format de fichiers en colonnes et où se trouve sa spécification.

Glossaire

Choisissez selon vos contraintes réelles

Présentez un jeu de données représentatif, les questions auxquelles il doit répondre et le goulot d'étranglement actuel. Nous pouvons convenir de la manière de mesurer la qualité, le débit et la reprise.

Consulter les tarifs Contactez-nous →

Rendez la terminologie concrète

Demandez comment « Qu'est-ce qu'un jeu de données ? » s'applique à votre organisation. Un exemple concret tiré de votre système actuel suffit pour engager la discussion.

Découvrir nos prestations

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

Emportez cette page avec vous

Fiche de contact

Créé avec soin et amour par D.

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

Droit d'auteur © 2025 – aujourd'hui · Engineer ApS · Tous droits réservés.

Imprimé depuis https://engineer.company/fr/glossary/dataset/