Engineer

Dataset

en samling data som publiceras som en enhet

Vad det är

Ett dataset, ibland skrivet data set och i svensk förvaltning kallat datamängd, är en avgränsad mängd data som hanteras som en sak. Det har ett namn, ett ursprung, en period som det täcker och en form som alla dess poster delar, och de fyra uppgifterna säger var det börjar och slutar. En månads mätaravläsningar i en enda fil är ett dataset, och det är en nattlig export av tre tabeller som hör ihop också. Ordet säger ingenting om storlek eller format, bara att datan inuti hör samman.

Sidorna på den här webbplatsen använder ordet i den vardagliga betydelsen. En talar om ett historiskt dataset i JSONL-format och om dataset över 1 och 30 dagar som används för att mäta prestanda, en annan om kurerade dataset och tabeller som exporteras för att läsas in någon annanstans, en tredje om enorma dataset från smarta elmätare. Varje gång drar ordet en gräns runt data innan något görs med dem.

W3C ger ordet en formell ram i Data Catalog Vocabulary, DCAT. Version 3 blev en W3C-rekommendation den 22 augusti 2024 och definierar ett dataset som en samling data som en enda aktör har publicerat eller kurerat, och som går att komma åt eller ladda ned i en eller flera serialiseringar eller format. Begreppet är brett med avsikt: siffror, text, bilder och ljud räknas alla. DCAT skiljer sedan på tre saker som vardagsspråket gärna blandar ihop:

- **Dataset**: datan som idé, t.ex. timförbrukning per zon för 2024.
- **Distribution**: en form som datan går att hämta i, t.ex. en CSV-fil för nedladdning. Ett och samma dataset kan ha flera.
- **Serie av dataset**: dataset som publiceras var för sig men har tillräckligt gemensamt för att grupperas, t.ex. en fil per månad.

Två filer i olika format kan alltså vara ett enda dataset, och två månadsfiler under samma namn kan vara två medlemmar i en serie. Myndigheten för digital förvaltning har gjort en svensk profil av den europeiska tillämpningen DCAT-AP, kallad DCAT-AP-SE, som i sin tur bygger på rekommendationen från W3C. Där heter de tre klasserna datamängd, distribution och datamängdsserie, och det är de orden en svensk myndighet använder när den beskriver sina data.

I Google BigQuery har ordet en andra, snävare betydelse. Där är ett dataset en behållare för tabeller, vyer och andra resurser, och det ligger inuti ett projekt. En tabell måste höra till ett dataset, så datasetet skapas innan någon data läses in. I GoogleSQL adresseras det som `project.dataset` och i kommandoradsverktyget bq som `project:dataset`. Sammanhanget avgör vilken betydelse som gäller.

Använd begreppet i ett beslut

Använd ”Vad är ett dataset?” i ett konkret beslut. Vi kan knyta begreppet till systemet ni bygger eller köper. Dashboards och rapporter som visar de få siffror ett beslut faktiskt hänger på, uppdaterade utan att någon behöver be om det.

Se våra tjänster Hör av er →

Granska antagandena bakom ordet

Använd förklaringen och det länkade arbetet för att formulera frågor till projektet. Vi kan undersöka antagandena tillsammans innan ett verktyg väljs. Varumärkeskommunikation utifrån publikens respons — varumärkeskommunikation utifrån publikens respons

Varumärkeskommunikation utifrån publikens respons Hör av er →

Varför det spelar roll

En gräns gör data jämförbara. Två serier av avläsningar går att jämföra bara när båda täcker samma period, samma platser och samma avstånd mellan posterna, och att namnge ett dataset är sättet att låsa de villkoren. DCAT ger avståndet en egen egenskap, temporal upplösning, som anger det minsta avståndet mellan posterna i datasetet och skrivs som en varaktighet, t.ex. `P1D` för en dag. Ett dataset som anger sin upplösning talar om för läsaren om raderna är timmar, kvartar eller något oregelbundet, redan innan den första summan räknas ut.

Beskrivningen måste följa med datan. DCAT räknar upp vad en distribution kan säga om sig själv: medietyp, format, komprimeringsformat, paketeringsformat, storlek i byte och en kontrollsumma. Den sista låter mottagaren verifiera att innehållet är oförändrat. Specifikationens eget exempel på ett komprimeringsformat är gzip. En fil som kommer med de uppgifterna går att granska och öppna; en fil som kommer naken måste gissas fram.

Formatet för en distribution avgör vem som kan läsa den. Tre är vanliga:

- **CSV.** RFC 4180 från 2005 beskriver den uppställning som de flesta implementationer verkar följa: en post per rad, en valfri rubrikrad, dubbla citattecken runt varje fält som innehåller ett kommatecken, en radbrytning eller ett dubbelt citattecken, och ett dubbelt citattecken inuti ett fält skrivs två gånger. Nästan alla verktyg öppnar det.
- **JSON Lines.** Ett giltigt JSON-värde per rad, kodat i UTF-8, med `\n` som radavskiljare. Filerna sparas som `.jsonl` eller, komprimerade med gzip, som `.jsonl.gz`. En avslutande radbrytning rekommenderas starkt men är inget krav. Program läser en post i taget, och textverktyg och skalpipelines i Unix-stil hanterar formatet väl.
- **Parquet.** Ett kolumnorienterat filformat med öppen källkod, byggt för effektiv lagring och hämtning, med metoder för komprimering och kodning som klarar komplexa data i stora mängder.

Här är en och samma post i de två första formaten. Kommatecknet i fältet note tvingar fram citattecken i CSV-raden:

```
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"}
```

Så ser det ut i praktiken

Tre projekt på webbplatsen visar ett dataset som formas, flyttas och läses.

Formas: eldata kom från flera källor, var och en med sin egen rapporteringsfrekvens, från timvisa intervall till kvartsintervall och i vissa fall oregelbundna tidsstämplar, vilket gjorde det svårt att få fram ett sammanhängande och jämförbart dataset. Arbetet gjorde om rå händelsedata ur ett historiskt JSONL-dataset till förbrukning per timme, en rad per timme och zon, där produktionsmixen i varje intervall summerades, importen lades till och exporten drogs ifrån. Tre varianter skrevs, i Python, JQ och SQL. Den datapipeline som skrevs i Python tolkade JSONL-filerna till dataframes, omsamplade tidsserien till timintervall och exporterade resultatet som CSV eller läste in det i en lättviktig SQLite-databas för granskning. Mätt på historiska dataset över 1 och 30 dagar var den snabbast av de tre, med 34 ms för en dag och 180 ms för trettio. Arbetet mellan indata och utdata är ett byte av granularitet, från varje källas egna intervall till en fast rad per timme och zon.

Flyttas: sex datapipelines för ETL/ELT förde data till Google BigQuery från ett analysdatalager på distans, bakom ett IPsec-VPN. Datan kom som komprimerade arkiv med Parquet-, CSV- och .bak-filer, och skript i Bash och Python sorterade varje fil efter typ och schema. När procedurer hade rensat, normaliserat och berikat raderna exporterades kurerade dataset och tabeller till mellanformat, komprimerades med gzip och fördes över till en bucket i Google Cloud Storage. Därifrån läste kommandoradsverktyget bq och Python-skript in dem i BigQuery. Manuellt arbete på över 3 timmar blev under 20 minuter från början till slut, och datan gick från veckovis till daglig synkronisering.

Läses: smarta elmätare i flera geografiska regioner gav upphov till enorma dataset bakom fler än 500 analysrapporter skrivna i ren SQL på PostgreSQL och BigQuery. Datan innehöll geografiska koordinater, mätar-id, tidsstämplad energiförbrukning och metadata om miljön. Varje rapport hade kommenterad SQL-kod så att kollegor kunde återskapa och granska den, och felsökningsanteckningar beskrev vanliga problem med datakvaliteten, som saknade GPS-koordinater eller korrupta mätvärden, tillsammans med rekommenderad hantering. Sidan tillskriver arbetet en 10–15 % effektivare operativ planering.

Var det går fel

De flesta problem med ett dataset är en uppgift om datan som stannade i en enda persons huvud.

Granulariteten står ingenstans. Timrader som läggs ihop med kvartsrader räknar vissa intervall fyra gånger. Skriv ned granulariteten i en mening, t.ex. en rad per timme och zon i UTC, och anpassa varje källa till den innan något summeras.

En lucka läses som noll. Ett tomt fält, ett null och en nolla betyder tre olika saker, och en summa behandlar i tysthet en saknad avläsning som ingen förbrukning alls. Saknade GPS-koordinater och korrupta mätvärden var så vanliga i projektet med de smarta elmätarna att de fick en dokumenterad hantering. Bestäm för varje kolumn om en lucka ska tas bort, fyllas i eller flaggas, och skriv i datasetets anteckningar vilket som gäller.

CSV-dialekten glider. RFC 4180 beskriver den vanliga uppställningen, och verkliga filer avviker från den: en annan avgränsare, en rubrikrad som saknas, ett fält med kommatecken där citattecknen har fallit bort. Enligt RFC:n är

rubrikraden valfri, och den signaleras bara av en MIME-parameter som försvinner när filen kopieras till disk. Jämför den första raden i varje ny fil med den förväntade rubrikraden innan något läses in.

En kopia glömmer var den kom ifrån. Ett kurerat utdrag som läses in i ett annat system ser en månad senare ut precis som originalet. Skriv ned källan, perioden, tidpunkten för uttaget och en kontrollsumma bredvid varje kopia, så att en rapport kan säga vilket utdrag den läste.

Ett dataset i BigQuery hamnar på fel plats. Platsen för ett dataset bestäms när det skapas, och att flytta det innebär att kopiera eller skapa om det. Alla tabeller i en och samma SQL-fråga måste ligga i dataset på samma plats, så välj platsen före den första inläsningen.

Läs mer

- Data Catalog Vocabulary (DCAT) - Version 3 — definitionerna från W3C av dataset, distribution och serie av dataset, med de egenskaper som beskriver var och en.
- DCAT-AP-SE — den svenska profilen från Myndigheten för digital förvaltning, med de svenska namnen datamängd, distribution och datamängdsserie.
- BigQuery: Introduction to datasets — behållarbetydelsen av ordet: projekt, platser, namngivning och gränser.
- JSON Lines — formatets tre regler och dess konventioner för filnamn.
- RFC 4180: Common Format and MIME Type for CSV Files — den uppställning för CSV som de flesta verktyg följer, reglerna för citattecken och typen text/csv.
- Apache Parquet: Overview — vad det kolumnorienterade filformatet är och var dess specifikation finns.

Ordlista

Välj utifrån era faktiska förutsättningar

Ta med ett representativt dataset, frågorna det ska besvara och dagens flaskhals. Vi kan avtala mått på kvalitet, genomströmning och återställning.

Se vad det kostar Hör av er →

Gör begreppen praktiska

Fråga hur ”Vad är ett dataset?” gäller i er organisation. Ett konkret exempel från ert nuvarande system räcker för att börja samtalet.

Se våra tjänster

welcome@engineer.company +45 50 10 61 84 Köpenhamn, Danmark

Ta med dig den här sidan

Visitkort

Omsorgsfullt skapad med kärlek av D.

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

Upphovsrätt © 2025 – nu · Engineer ApS · Alla rättigheter förbehållna.

Utskrivet från https://engineer.company/sv/glossary/dataset/