Engineer

Dataafstamning

overblik over datas vej gennem kilder og bearbejdning

Hvad det er

Dataafstamning beskriver den vej, data tager fra kilder gennem bearbejdning til en rapport, et datasæt eller en anden modtager. Et omsætningstal kan komme fra ordrer, returneringer og en regel for omregning af valuta. Afstamningen gør det muligt at følge tallet tilbage til de indgange, det bygger på. Samme vej kan læses fremad, så en udvikler kan se, hvilke rapporter der bliver berørt, hvis et felt eller tidspunktet for en levering ændres. Detaljerne kan gå fra en enkel oversigt over tabeller til en log over præcis de felter, versioner og partitioner, som en kørsel læste og skrev.

I OpenLineage er et datasæt en indgang eller udgang. Et job er en navngiven opgave, der flytter eller behandler data. En kørsel er én udførelse af jobbet. Den kan indberette, hvilke datasæt den brugte, hvilke den skabte, og detaljer om for eksempel skema og kildeversion. Koden kan på samme tid afsløre, hvilke afhængigheder der kan forekomme. Den, der bruger afstamningen, skal kunne se, om en linje bygger på kode, en registreret kørsel eller en manuel undersøgelse.

Vælg detaljer efter spørgsmålet. På tabelniveau får en analytiker at vide, om en rapport henter kundeoplysninger fra to forskellige steder. På kolonneniveau kan man følge et beløb tilbage gennem omregningen. På kørselsniveau kan en driftsmedarbejder se, hvilken udgave af bearbejdningen der håndterede en bestemt leverance. Tidsinterval, partition og regelversion kan være lige så afgørende som tabellernes navne.

Brug begrebet i en beslutning

Brug »Hvad er dataafstamning?« i en konkret beslutning. Vi kan knytte begrebet til det system, I bygger eller køber. Tal, der lander, hvor der er brug for dem, efter en fast plan, uden at nogen kopierer et regneark ved månedens slutning.

Se vores ydelser Kontakt os →

Undersøg antagelserne bag ordet

Brug forklaringen og det tilknyttede arbejde til at formulere spørgsmål til jeres projekt. Vi kan undersøge antagelserne sammen, før et værktøj vælges. Én infrastrukturplan til 14 afdelinger — 14 forskningsafdelinger på én infrastrukturplan

Én infrastrukturplan til 14 afdelinger Kontakt os →

Hvorfor det betyder noget

Afstamning gør en fejlsøgning mindre. Hvis dagens resultat falder brat, kan man undersøge, om indsamlingen fik færre rækker, et sammenkædningsled tabte en gruppe, eller rapporten brugte det forkerte tidsinterval. Afstamningen peger på de første steder at undersøge. Derefter kan man bekræfte hvert fund med en måling. En sammenhængende tegning siger i sig selv ikke, at de underliggende data var rigtige eller friske.

At kende modtagerne hjælper også med at planlægge en ændring. En udvikler vil omdøbe et statusfelt, og oversigten viser en afstemning og flere rapporter, der læser feltet. Ejerne kan nu aftale, hvem der afprøver det nye navn, hvilke modtagere der i en periode behøver et kompatibelt felt, og hvornår det gamle kan fjernes. Den samtale er lettere, før ændringen er taget i brug. En oversigt til flere hold kan vise afhængigheder uden at afsløre personoplysninger eller give flere adgange.

God dokumentation af oprindelse hjælper, når to rapporter bruger samme navn om forskellige tal. Sammenlign kilderne og definitionerne bag dem. Det kan f.eks. vise, at den ene rapport tæller ordrer modtaget om aftenen samme dag, mens den anden placerer dem på næste dags forsendelsesliste. Vej og definition afklarer spørgsmålet. En egentlig test afgør, om reglen er fulgt.

Sådan ser det ud i praksis

Engineer ApS beskriver arbejde med datavarehuse i PostgreSQL, MS SQL og BigQuery. Når behandlinger og modtagere spænder over flere motorer, er det relevant at kende deres grænser og leverancer. Det følgende er en fiktiv øvelse i at dokumentere én rapport.

Forestil dig en dagsrapport over ordrer, der er klar til afsendelse. Ordrene kommer fra webshoppen, og et andet feed giver besked om afsendelser. Rapportens behandling grupperer på afsendelsesdato og finder den seneste tilladte statushændelse for hver ordre. Registrér indgange og udgang, og navngiv jobbet, der forbinder dem. Gem en identifikator for kørslen, dens start

og slutning, de læste partitioner og den udgave af reglen, der vælger status. Et mislykket forsøg og en senere omkørsel skal kunne ses hver for sig, men det skal fremgå, hvilket resultat der blev godkendt.

Vælg nu en ordre fra tirsdagens rapport. Følg rapportdato og ordrenummer gennem de godkendte forespørgsler til den hændelse, som kørslen valgte. Kontrollér, at hændelsen faldt på den tilsigtede side af dagens skæringspunkt, og find reglen for to hændelser på samme tidspunkt. Hvis feedet var forsinket, skal det fremgå af rapportens forklaring og fejlhistorikken. Når skæringspunktet senere ændres, kan ejerne af de berørte rapporter afprøve både en normal levering og en, der ankommer for sent.

Begynd gerne med én vigtig rapport. Bed en analytiker følge en enkelt række og en driftsmedarbejder følge et mislykket forsøg. Angiv, hvor de første gang må gætte. Den manglende identifikator, version eller forbindelse bliver næste konkrete opgave. Automatisk registrering kan derefter følge samme spørgsmål og gøre det klart, hvor den har huller.

Gem både statushændelsens tidspunkt og tidspunktet, hvor den blev tilgængelig for rapporten. Aftal også tidspunktet for publicering, så en tidlig hændelse med sen levering kan spores.

Hvor det går galt

En tegning kan se komplet ud, selv om den mangler et manuelt udtræk, en lagret procedure eller et regneark, som en person retter efter download. Opgør, hvilke dele der måles automatisk, og angiv, hvor resten er dokumenteret. En manuel forbindelse bør have en kilde og en dato. Gennemgå den igen, når arbejdet flyttes.

Datasæt med samme korte navn kan ligge i test, produktion eller arkiv. Brug en stabil identifikator, der omfatter miljø og placering. Gem tidligere identifikatorer, når en tabel flyttes, så en gammel rapport stadig kan forstås. Undlad også at overfylde hændelsesloggen med forespørgslers tekst eller dataeksempler, hvis metadata læses af flere end dem, der må læse kilden.

Prøv til sidst en enkel, reversibel ændring med et kendt resultat. Kontrollér både udgangen og den vej, kørslen registrerede. En ny værdi kan ændre resultatet, mens afhængigheden består. Vent først en anden forbindelse, når jobbet faktisk

skifter kilde. Så kan man tage stilling til den faktiske vej gennem systemet.

Læs mere

- OpenLineages model for datasæt, job og kørsler
- W3C's PROV-O-model for proveniens
- Google Clouds dokumentation om dataafstamning

Ordliste

Vælg ud fra jeres faktiske vilkår

Medbring et repræsentativt datasæt, de spørgsmål det skal besvare, og den aktuelle flaskehals. Vi kan aftale målinger af kvalitet, gennemløb og gendannelse.

Se, hvad det koster Kontakt os →

Gør begreberne praktiske

Spørg, hvordan »Hvad er dataafstamning?« gælder i jeres organisation. Et konkret eksempel fra jeres nuværende system er nok til at begynde samtalen.

Se vores ydelser

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

Tag denne side med dig

Visitkort

Lavet med kærlighed af D.

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

Copyright © 2025 – nu · Engineer ApS · Alle rettigheder forbeholdes.

Udskrevet fra https://engineer.company/da/glossary/data-lineage/