Engineer

Dataflythistorikk

sporet data følger fra kilder gjennom bearbeiding

Hva det er

Dataflythistorikk, på engelsk data lineage, beskriver veien data har tatt fra kilder, gjennom behandling og fram til et datasett eller en rapport. Et omsetningstall i en oversikt kan stamme fra en ordretabell, en tabell for refusjoner og en regel for omregning av valuta. Historikken gir en vei fra tallet tilbake til disse inndataene. Den kan også leses motsatt vei for å finne rapportene som blir berørt når en eier endrer en kolonne eller flytter et leveringstidspunkt. En enkel oversikt kan vise hvilke tabeller som henger sammen. For en bestemt kjøring kan det være nødvendig å kjenne både versjon, felt og partisjoner.

OpenLineage skiller mellom et datasett, en navngitt jobb og en kjøring av den jobben. Jobben flytter eller behandler data. Kjøringen kan melde hvilke datasett som var inn- og utdata, og legge til detaljer som skjema, kildeversjon og kjøremiljø. Hendelsene gir et spor tilbake til den utførelsen som faktisk produserte en leveranse. Kildekoden kan i tillegg vise mulige avhengigheter også når ingen kjøring er registrert. Det bør gå fram hvilken slags grunnlag hver forbindelse bygger på.

Velg detaljnivå ut fra spørsmålet. På tabellnivå kan en analytiker se at en rapport har to kilder. På kolonnenivå kan man følge et beløp gjennom en beregning eller et felt gjennom et navnebytte. På kjøringsnivå kan driften se hvilken utgave av behandlingen som leste en bestemt dags leveranse. Skal et omstridt tall granskes, kan utvalg, tidsvindu og regelversjon ha like stor betydning som tabellnavnene.

Bruk begrepet i en beslutning

Bruk «Hva er dataflythistorikk?» i en konkret beslutning. Vi kan knytte begrepet til systemet dere bygger eller kjøper. Tall som kommer fram dit de trengs, til faste tider, uten at noen må kopiere et regneark ved månedsslutt.

Se tjenestene våre Ta kontakt →

Undersøk antakelsene bak ordet

Bruk forklaringen og det tilknyttede arbeidet til å formulere spørsmål for prosjektet. Vi kan undersøke antakelsene sammen før et verktøy velges. Én infrastrukturplan for 14 avdelinger — 14 forskningsavdelinger på én felles infrastrukturplan

Én infrastrukturplan for 14 avdelinger Ta kontakt →

Hvorfor det er viktig

Historikken avgrenser en feilsøking. Når dagens tall faller brått, kan man undersøke om kilden leverte færre rader, en kobling mistet en gruppe eller rapporten valgte feil tidsrom. Sporet peker ut de første stedene å se. Bekreft deretter hvert mulig brudd ved å observere dataene eller kjøringen. Et diagram kan stemme med oppsettet selv om en levering var forsinket eller beregningen var feil.

Det er også enklere å planlegge en endring når man kjenner mottakerne. En utvikler vil gi et statusfelt et nytt navn. Oversikten viser at en avstemming og flere rapporter bruker feltet. Eierne kan avtale hvem som skal teste det nye navnet, hvem som trenger gammel og ny variant en stund, og når den gamle kan fjernes. En slik samtale er mer verdifull mens endringen er enkel å justere. Oversikten kan vise avhengigheter samtidig som innhold og tilganger holdes avgrenset.

Kildehistorikk gjør det lettere å sammenligne to rapporter som bruker samme navn på ulike tall. Den ene kan knytte kveldens ordre til dagen de kom inn. Den andre kan telle dem som neste dags leveranser. Først må man finne kilden, reglene og tidssnittet for hver rapport. Så kan eierne diskutere hvilket spørsmål de vil besvare og kontrollere om reglene faktisk ble fulgt.

Slik ser det ut i praksis

I porteføljen omtaler Engineer ApS drift av datavarehus i PostgreSQL, MS SQL og BigQuery. Når data krysser plattformgrenser, er det relevant å vite hvilken jobb som behandler hver leveranse. Eksemplet nedenfor er en fiktiv øvelse, uavhengig av det omtalte arbeidet.

Tenk deg en dagsrapport over ordre som er klare for utsending. Ordrene kommer fra nettbutikken, mens en annen strøm leverer status fra lageret. Rapportjobben sorterer etter utsendingsdato og velger siste status per ordrenummer. Registrer inn- og utdata hver for seg, og gi koblingen et jobbnavn. Behold kjøringens identifikator, start og slutt, partisjonene den leste og

versjonen av regelen som valgte siste status. Den mislykkede kjøringen og et senere vellykket forsøk er forskjellige hendelser. Vis hvilken av dem som fikk levere resultatet.

Velg så en ordre fra tirsdagens rapport. Følg rapportdato og ordrenummer med godkjente spørringer tilbake til lagerhendelsen som ble valgt. Sjekk når den kom i forhold til grensen for dagens utsending, og hvordan to hendelser med samme tid blir sortert. Hvis statusene ankom sent, må det komme fram i forklaringen av rapporten og i hendelsens tidslinje. Ved en ny tidsgrense kan eierne teste både en normal levering og en forsinkelse.

Start med én viktig rapport og noen kontrollerte koblinger. La en analytiker følge én rad og la en driftsmedarbeider følge en mislykket kjøring. Finn første sted der en av dem må gjette. Legg til den manglende identifikatoren eller forbindelsen først. Samme fremgangsmåte gir en målestokk når registreringen senere automatiseres.

Lagre både tidspunktet da statusen oppsto og når dataene ble tilgjengelige for jobben. Noter også publiseringsgrensen, slik at en tidlig hendelse som kom sent, kan forklares.

Hvor det går galt

Et diagram kan se komplett ut selv om en manuell eksport, lagret prosedyre eller nedlastet regneark mangler. Tell hvilke deler som blir målt, og si hvor den øvrige veien er dokumentert. En manuelt ført kobling bør ha kilde og dato. Den må gjennomgås på nytt hvis arbeidet flytter.

Tabeller med samme navn kan ligge i test, produksjon eller arkiv. Bruk en stabil identifikator som omfatter miljø og plassering. Behold relasjonen til tidligere navn når noe flyttes. Beskytt også selve metadataene. Spørringer, eksempelrader og tabellnavn kan avsløre mer enn leseren trenger eller har tillatelse til.

Avslutt med en enkel, reversibel endring der utfallet er kjent. Kontroller resultatet og sporet for kjøringen. En ny verdi kan endre resultatet mens avhengigheten fortsatt er den samme. Forvent en ny kobling når jobben faktisk leser en annen kilde. Grunnlaget gjelder den veien dere faktisk observerte. Det kan utvides med den neste øvelsen.

Les mer

- OpenLineages modell for datasett, jobber og kjøringer
- W3Cs PROV-O-modell for proveniens
- Google Clouds dokumentasjon om dataflythistorikk

Ordliste

Velg ut fra deres faktiske rammer

Ta med et representativt datasett, spørsmålene det skal besvare og dagens flaskehals. Vi kan avtale målinger av kvalitet, gjennomstrømning og gjenoppretting.

Se hva det koster Ta kontakt →

Gjør begrepene praktiske

Spør hvordan «Hva er dataflythistorikk?» gjelder i organisasjonen deres. Et konkret eksempel fra dagens system er nok til å starte samtalen.

Se tjenestene våre

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

Ta med deg denne siden

Visittkort

Laget med kjærlighet av D.

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

Opphavsrett © 2025 – nå · Engineer ApS · Alle rettigheter forbeholdt.

Skrevet ut fra https://engineer.company/nb/glossary/data-lineage/