Engineer

Datalinjage

datas väg från källor genom bearbetningar

Vad det är

Datalinjage beskriver den väg data tar från källorna, via bearbetningar och fram till ett dataset eller en rapport. Ett intäktstal kan komma från order, återbetalningar och en regel för valutaväxling. Linjaget ger en väg från talet bakåt till underlaget. Vägen går också att följa framåt för att se vilka rapporter som berörs när en ägare ändrar en kolumn eller flyttar en leverans. En enkel översikt kan visa förhållandet mellan tabeller. En dokumenterad körning kan dessutom visa exakt vilka versioner, fält och partitioner den läste och skrev.

I OpenLineage skiljer man mellan ett dataset, ett namngivet jobb som flyttar eller bearbetar data och en viss körning av jobbet. Körningen kan ange vilka dataset den läste och vilka den skapade. Den kan också bifoga detaljer om exempelvis schema, källversion och miljö. Registrerade händelser berättar vilket försök som faktiskt producerade en leverans. Källkod kan dessutom visa vilka relationer som är möjliga. Den som tittar på en länk behöver veta om den kommer från kod, en observerad körning eller en manuell granskning.

Välj nivå efter frågan. På tabellnivå ser en analytiker att rapporten har två källor. På kolumnnivå går det att följa ett belopp genom en beräkning eller en personuppgift genom ett namnbyte. På körningsnivå kan driften förstå vilken version som bearbetade tisdagens indata. Om ett tal behöver förklaras kan tidsspann, partition och regelversion vara lika viktiga som tabellnamnen.

Använd begreppet i ett beslut

Använd ”Vad är datalinjage?” i ett konkret beslut. Vi kan knyta begreppet till systemet ni bygger eller köper. Siffror som kommer fram dit de behövs, på fasta tider, utan att någon kopierar ett kalkylblad i slutet av månaden.

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. En infrastrukturplan för 14 avdelningar — 14 forskningsavdelningar under en och samma infrastrukturplan

En infrastrukturplan för 14 avdelningar Hör av er →

Varför det spelar roll

Linjaget gör en felsökning mindre. Om dagsvärdet plötsligt faller kan teamet undersöka om insamlingen tog emot färre rader, om en koppling tappade en grupp eller om rapporten avgränsade fel period. Förbindelserna ger en ordning att söka i. Bekräfta en misstänkt brytpunkt genom att titta på den aktuella körningen och dess data. Ett tydligt diagram kan nämligen finnas även när leveransen är gammal eller beräkningen visar fel.

De som känner mottagarna kan också planera en ändring bättre. En utvecklare vill döpa om ett statusfält. Källhistoriken visar att en avstämning och flera rapporter använder fältet idag. Ägarna kan bestämma vem som provar det nya namnet, vem som behöver båda namnen samtidigt en period och när det gamla kan tas bort. En sådan diskussion går lättare när ändringen fortfarande kan revideras. Tillgång till översikten kan hållas skild från tillgång till personuppgifter och rapportinnehåll.

Bra proveniens ger en gemensam grund när två rapporter använder samma ord om olika tal. Den första räknar kanske in kvällens order på samma dag. Den andra räknar dem till nästa leveransdag. Frågan blir tydligare när man har jämfört källor, tidsgränser och bearbetningar. Därefter kan ägarna välja vad rapporten ska visa och kontrollera om regeln tillämpades.

Så ser det ut i praktiken

I Engineer ApS:s portfölj beskrivs arbete med datalager i PostgreSQL, MS SQL och BigQuery. Över flera plattformar behöver den som granskar en rapport kunna peka ut vilken behandling som körs var. Följande fiktiva exempel visar en sådan granskning för en rapport.

Tänk dig en daglig lista över order som kan skickas. Orderinformationen kommer från webbutiken, och utsändningsstatus kommer från ett annat flöde. Jobbet grupperar på utsändningsdatum och väljer den senaste status som uppfyller villkoren för varje order. Registrera indata och utdata och ge behandlingen ett

jobbnamn. Spara körningens identitet, start, slut, källpartitioner och vilken version av urvalsregeln som användes. Ett misslyckat försök och en senare omkörning ska gå att skilja åt. Markera vilket resultat som faktiskt publicerades.

Välj sedan en rad i tisdagens rapport. Följ datumet och ordernumret genom godkända frågor tillbaka till den valda statushändelsen. Kontrollera om händelsen låg på rätt sida om dagens tidsgräns och hur två händelser med samma tid hanteras. Om flödet kom sent ska det gå att se i rapportens beskrivning och i händelsens tidslinje. Om tidsgränsen flyttas kan ägarna använda samma test med en normal leverans och en försenad.

Börja med en viktig rapport och granska ett litet antal förbindelser. Be en analytiker följa en rad och en operatör följa en misslyckad körning. Notera var de börjar behöva gissa. Lägg till den identifikator eller den förbindelse de saknar. Samma fråga ger sedan stöd för att verifiera automatiska insamlare.

Spara tidpunkten för statushändelsen, tidpunkten då den blev tillgänglig och gränsen för publicering. Då går det att förklara en tidig händelse som levererades sent.

Var det går fel

En översikt kan se fullständig ut även om en manuell export, en lagrad procedur eller ett kalkylblad som rättas efter nedladdning saknas. Beskriv vad som samlas in automatiskt och var resten dokumenteras. En manuell förbindelse behöver källa och datum. Granska den på nytt när arbetet flyttas.

Korta tabellnamn kan vara identiska i test, drift och arkiv. Använd en stabil beteckning som omfattar både miljön och platsen. Bevara den tidigare beteckningen när data flyttas. Skydda även metadata. Tabellnamn, innehållet i frågor och exempelrader kan avslöja känsliga förhållanden för den som saknar tillgång till själva uppgifterna.

Pröva till sist en enkel ändring med känt och reversibelt resultat. Kontrollera både utdata och den väg som körningen registrerade. Ett ändrat ingångsvärde kan ge ett nytt resultat med samma beroende. En annan relation är förväntad om jobbet faktiskt byter källa. Det ger stöd för den väg ni faktiskt observerade och ett konkret underlag för nästa kontroll.

Läs mer

- OpenLineages modell för dataset, jobb och körningar
- W3C:s proveniensmodell PROV-O
- Google Clouds dokumentation om datalinjage

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 datalinjage?” 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/data-lineage/