Engineer

Dataanalys

systematisk bearbetning av data för att hitta mönster

Vad det är

Att göra registrerade fakta till svar som någon kan agera på är vad dataanalys handlar om, och i vardagligt tal heter det ofta bara analys eller analytics. I sitt ramverk för stordata beskriver NIST, det amerikanska standardiseringsinstitutet, analys som att metodiskt gå igenom data och vid behov bearbeta dem för att lyfta fram mönster, samband mellan data och trender över tid, och helst också förutsäga vad som händer härnäst. NIST lägger det steget i en större *livscykel för dataanalys*, som förvandlar rådata till kunskap som en organisation kan använda: insamling, förberedelse, analys, visualisering och åtkomst.

Analys är en form av databehandling som skiljer sig från den dagliga transaktionshanteringen genom vad den läser. En applikation skriver en order i taget, medan en analytisk SQL-fråga läser miljontals rader för att returnera en handfull. Datateknik, det som på engelska kallas data engineering, är något tredje: där byggs de pipelines och den lagring som analysen sedan läser från.

Ordet täcker arbete som ser mycket olika ut:

- **Rapportering**: samma siffror levererade med jämna mellanrum, som PDF eller e-post.
- **Dashboards**: samma siffror hållna aktuella i ett BI-verktyg som Power BI.
- **Ad hoc-analys**: en SQL-fråga skriven för något som ingen har undrat över förut, ofta mot ett analysdatalager som BigQuery.
- **Geografisk analys**: samma arbete på kartdata i ett GIS, där man frågar var och inte bara hur mycket.
- **Analytiska applikationer**: en modell eller en statistisk metod bakom ett formulär, för människor som inte skriver kod själva.
- **Webb- och produktanalys**: att räkna besökarna på en webbplats eller användarna av en produkt och se vad de gör där.

Den sista sorten har ett pris för integriteten, och det är den sorten som webbplatsens sidor om integritet, kakor, kolofon och säkerhet syftar på när de skriver ”ingen analys” eller ”ingen analystagg”. Det de menar är en tagg, en spårningspixel, en kaka och ett konto hos ett analysföretag, och inget av det används här. Webbplatsen mäter sig själv ändå. Webbservern för en åtkomstlogg, och ett program gör om den till dagliga summor och en rangordning för varje månad av lästa sidor, hänvisande webbplatser och sökrobotar. Det är siffror som aldrig sätts samman till en enskild besökares väg genom webbplatsen.

Använd begreppet i ett beslut

Använd ”Vad är dataanalys?” 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

Två val avgör det mesta av vad analysen kommer att kosta: var SQL-frågorna körs och hur data om människor samlas in.

Analys direkt på transaktionsdatabasen är den billigaste starten. Datan är färsk och ingenting behöver kopieras, men varje rapport konkurrerar med applikationen om minne och disk. En läsreplik flyttar belastningen bort från primärservern, till priset av ännu en server att sköta. Ett analysdatalager som BigQuery håller lagring och beräkning åtskilda och debiterar med sin on demand-prissättning varje fråga efter hur många byte den läser, så smala frågor blir billiga. I gengäld måste en pipeline kopiera in datan.

Inlåsningen gömmer sig i definitionerna. Ett BI-verktyg håller sina mått i sitt eget formelspråk, och ett år senare finns definitionen av en aktiv kund i en fil som bara det verktyget kan öppna. Definitioner som ligger som SQL i repot överlever ett byte av verktyg.

Webb- och produktanalys väcker dessutom en juridisk fråga. Det typiska hostade verktyget arbetar genom en tagg: ett skript på varje sida som sätter en kaka eller en liknande identifierare i webbläsaren och rapporterar varje sidvisning till leverantören. Artikel 5.3 i EU:s ePrivacydirektiv gör samtycke till villkoret för att

lagra information på besökarens utrustning eller läsa den därifrån, med två undantag: lagring eller åtkomst vars enda syfte är att överföra en kommunikation via ett nät, och lagring eller åtkomst som är strikt nödvändig för att leverera en tjänst som besökaren uttryckligen har begärt. Går händelserna att koppla till en person, om så bara genom identifieraren i en kaka eller en IP-adress, gäller även dataskyddsförordningen (GDPR), med krav på rättslig grund, ett angivet ändamål och en lagringstid. Och leverantören är ett annat företag, som nu sitter med en förteckning över vad dina besökare gjorde.

Det verkliga valet står alltså mellan tre saker: ett hostat verktyg som går fort att lägga in och har dashboards på köpet, en förstapartsinsamling som man bygger och driver själv, och siffror ur loggar som servern redan skriver, som besvarar färre frågor och rymmer betydligt mindre.

Så ser det ut i praktiken

För en professionell nätverksplattform byggde företaget produktanalys i PostgreSQL i stället för en tagg från tredje part. Schemat omfattar 7 tabeller och 47 funktioner. Händelserna ligger i tabeller uppdelade efter 22 partitionsdefinitioner som själva tar bort gammal data enligt en fast plan, och besökarens identitet har ersatts med en pseudonym, skapad med ett salt som byts ut regelbundet. Klienten skickar ingenting förrän besökaren har samtyckt, och trattar räknas ut i databasen, bredvid datan, det som NIST kallar datalokalitet. Sidan är ärlig om priset: när saltet roterar går kohortanalys över lång tid förlorad, och det följer inga gratis dashboards med.

På rapporteringssidan hämtade ett rapporteringssystem i Python datan, körde analysen och levererade enligt ett förutsägbart schema samma PDF-rapport till dem som fattar de affärsmässiga besluten. Enligt sidan var det rapporteringen som lyfte mjukvaruintäkterna per kvartal med 400 %.

Det tredje exemplet är den här webbplatsen. Integritetssidan beskriver mätningen: en åtkomstlogg som byter ut den sista delen av IP-adressen mot nollor innan raden skrivs, hänvisande webbplatser som bara sparas som namn, och rader som tas bort efter omkring trettio dagar. Ett program gör om loggen till summorna och rangordningen, som behålls på obestämd tid utan något som kan foga ihop siffrorna till ett besök igen. Utan tagg, kaka eller tredje part visar sidan heller ingen samtyckesruta, och den anger själv skälet: det finns inget att samtycka till. Gränsen står där också: så länge en rad finns kvar kan en sällsynt

webbläsare i ett litet nät peka ut en besökare tydligare än webbplatsen önskar. Man får färre svar och inga trattar, men i gengäld lagras ingenting på läsarens enhet och ingenting skickas vidare.

Var det går fel

Två verktyg ger två besökarsiffror

Det hostade verktyget visar en siffra för förra månaden och serverloggen det dubbla, och båda har rätt om var sin sak. En tagg räknar bara webbläsare som körde dess skript, och bakom en samtyckesruta bara dem som sa ja. En logg räknar varje förfrågan, sökrobotar inräknade, tills någon filtrerar bort dem. Skriv ner definitionen av varje mått en gång, i SQL, med vad det tar med och vad det utelämnar.

Rapporterna gör applikationen långsam

Varje måndag går applikationen trögt medan veckorapporten skannar hela tabeller i transaktionsdatabasen. Flytta rapporten till en replik eller ett analysdatalager, eller sammanställ raderna i en liten sammanfattningstabell varje natt och låt fönsterfunktioner räkna på den:

```
SELECT day,
       signups,
       sum(signups) OVER (ORDER BY day
                          ROWS BETWEEN 6 PRECEDING AND CURRENT ROW)
         AS last_7_days
FROM daily_counts
ORDER BY day;
```

Samtyckesrutan försvinner, spårningen blir kvar

Ett team byter ut kakor mot en ”kakfri” produkt, plockar bort samtyckesrutan och fortsätter att följa besökarna. Artikel 5.3 gäller åtkomst till information på utrustningen, oavsett teknik. Europeiska dataskyddsstyrelsen (EDPB) skriver i sina riktlinjer om artikelns tekniska tillämpningsområde att fingeravtryck utifrån HTTP-huvuden, eller spårning av navigering enbart via IP-adressen, också kan omfattas, och påminner samtidigt om att regelns tillämplighet inte i sig betyder att samtycke krävs. Bedöm ett verktyg efter om det kopplar ihop en förfrågan med nästa, inte efter om det sätter en kaka.

Samtycke som en tillsynsmyndighet skulle underkänna

En samtyckesruta spärrar sidan tills besökaren godkänner, räknar rullning som ett ja eller gör det svårare att återkalla samtycket än det var att ge det. EDPB:s riktlinjer om samtycke säger att en cookie wall, där innehållet låses upp först efter ett ja, gör samtycket ofrivilligt, och att rullning aldrig är en entydig bekräftande handling. Artikel 7.3 i GDPR kräver att det ska vara lika lätt att återkalla samtycket som att ge det. Ge ett nej samma vikt som ett ja, och skicka ingenting innan svaret har kommit.

Pseudonyma händelser sparade för alltid

Händelsetabellen växer med miljontals rader i månaden, och de hashade användar-id:na i den behandlas som anonyma. Enligt skäl 26 i GDPR är pseudonymiserade uppgifter som kan kopplas till en person med hjälp av ytterligare information fortfarande personuppgifter, och skäl 30 nämner IP-adresser och kakor bland de identifierare på nätet vars spår, kombinerade med annan information, kan användas för att profilera och identifiera någon. Bestäm lagringstiden innan den första händelsen kommer in, partitionera efter tid så att en månad som har löpt ut kan raderas i ett enda steg, och spara långsiktiga siffror som aggregat utan identifierare.

Läs mer

- NIST SP 1500-1r2, Big Data Interoperability Framework: Definitions — utgångspunkten för artikeln: definitionerna av analys, livscykeln för dataanalys och datalokalitet.
- Directive 2002/58/EC, consolidated — artikel 5.3, samtyckesregeln för allt som lagras på eller läses från en besökares utrustning, med dess två undantag.
- Regulation (EU) 2016/679, the GDPR — definitionen av samtycke, artikel 7 om att återkalla det och skäl 26 och 30 om pseudonymiserade uppgifter och identifierare på nätet.
- EDPB Guidelines 05/2020 on consent — cookie walls, rullning och återkallande, vart och ett med konkreta exempel.
- EDPB Guidelines 2/2023 on Article 5(3) — varför spårning via IP-adress, fingeravtryck och spårningspixlar kan falla under samma regel som kakor.
- PostgreSQL tutorial: Window Functions — rangordningar och löpande summor över relaterade rader, beräknade inne i databasen.



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 dataanalys?” 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/analytics/