Engineer

Dataanalyse

systematisk behandling av data for å finne mønstre

Hva det er

Å gjøre registrerte fakta om til svar som noen kan handle ut fra, er oppgaven til dataanalyse, i dagligtale ofte bare kalt analyse eller analytics. I rammeverket for stordata fra NIST, det amerikanske standardiseringsinstituttet, betyr analyse å gå metodisk gjennom data og bearbeide dem der det trengs, for å få fram mønstre, sammenhenger mellom data og utviklingstrekk over tid, og helst også forutsi hva som skjer videre. NIST plasserer dette trinnet i en større *livssyklus for dataanalyse*, som gjør rådata om til kunnskap en organisasjon kan bruke: innsamling, klargjøring, analyse, visualisering og tilgang.

Analyse er en form for databehandling, og det som skiller den fra den daglige transaksjonsbehandlingen, er hva den leser. En applikasjon skriver én ordre om gangen, mens en analytisk spørring leser millioner av rader for å returnere en håndfull. Data engineering er noe tredje: der bygger man pipelinene og lagringen som analysen henter fra.

Ordet dekker arbeid som ser svært ulikt ut:

- **Rapportering**: de samme tallene levert etter en fast plan, som PDF eller e-post.
- **Dashbord**: de samme tallene holdt oppdatert i et BI-verktøy som Power BI.
- **Ad hoc-analyse**: en SQL-spørring skrevet for et spørsmål ingen har stilt før, ofte mot et datavarehus som BigQuery.
- **Geografisk analyse**: det samme arbeidet på kartdata i et GIS, der spørsmålet er hvor og ikke bare hvor mye.
- **Analytiske applikasjoner**: en modell eller en statistisk metode bak et skjema, for folk som ikke skriver kode selv.
- **Web- og produktanalyse**: å telle besøkende på et nettsted eller brukere av et produkt, og se hva de gjør der.

Den siste typen har en pris i personvern, og det er den sidene om personvern, informasjonskapsler, kolofon og sikkerhet på dette nettstedet mener når de skriver «ingen analysekode», «ingen analyseverktøy» eller «ingen analyse». Det de sikter til, er sporingskode, sporingspiksler, informasjonskapsler og en konto hos et analyseselskap, og ingen av delene er i bruk her. Nettstedet måler seg selv likevel. Webserveren fører en tilgangslogg, og et program gjør den om til en daglig opptelling og en månedlig rangering av leste sider, henvisende nettsteder og søkeroboter. Det er tall som aldri settes sammen til én besøkendes vei gjennom nettstedet.

Bruk begrepet i en beslutning

Bruk «Hva er dataanalyse?» i en konkret beslutning. Vi kan knytte begrepet til systemet dere bygger eller kjøper. Dashbord og rapporter som viser de få tallene en beslutning faktisk hviler på, oppdatert uten at noen må be om det.

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. Merkevarekommunikasjon ut fra publikums reaksjoner — merkevarekommunikasjon ut fra publikums reaksjoner

Merkevarekommunikasjon ut fra publikums reaksjoner Ta kontakt →

Hvorfor det er viktig

To valg avgjør det meste av hva analysen kommer til å koste: hvor spørringene kjøres, og hvordan data om mennesker samles inn.

Analyse rett på transaksjonsdatabasen er den billigste starten. Dataene er ferske og ingenting må kopieres, men hver rapport konkurrerer med applikasjonen om minne og disk. En lesereplika flytter belastningen bort fra primærserveren, til prisen av en ekstra server å drifte. Et datavarehus som BigQuery holder lagring og beregning atskilt og fakturerer, med on demand-prisingen, hver spørring etter hvor mange byte den leser, slik at smale spørringer blir billige. Til gjengjeld må en pipeline kopiere dataene inn.

Leverandørbindingen gjemmer seg i definisjonene. Et BI-verktøy holder målene sine i sitt eget formelspråk, og et år senere står definisjonen av en aktiv kunde i en fil bare det verktøyet kan åpne. Definisjoner som ligger som SQL i repoet, overlever et bytte av verktøy.

Web- og produktanalyse reiser i tillegg et juridisk spørsmål. Det typiske driftede verktøyet virker gjennom sporingskode: et skript på hver side som setter en informasjonskapsel eller en lignende identifikator i nettleseren og melder hver

sidevisning tilbake til leverandøren. Artikkel 5 nr. 3 i EUs ePrivacy-direktiv gjør samtykke til vilkåret for å lagre informasjon på den besøkendes utstyr eller lese den tilbake derfra, med to unntak: lagring eller tilgang som bare har til formål å overføre en kommunikasjon over et nett, og lagring eller tilgang som er strengt nødvendig for å levere en tjeneste den besøkende uttrykkelig har bedt om. Kan hendelsene knyttes til en person, om så bare gjennom en cookie-ID eller en IP-adresse, gjelder også personvernforordningen (GDPR), med krav om rettslig grunnlag, et uttalt formål og en lagringstid. Og leverandøren er et annet selskap, som nå sitter med en oversikt over hva de besøkende dine gjorde.

Det reelle valget står dermed mellom tre ting: et driftet verktøy som er raskt å sette opp og har dashbord med på kjøpet, en førstepartsinnsamling man bygger og drifter selv, og tall fra loggene serveren allerede skriver, som svarer på færre spørsmål og rommer langt mindre.

Slik ser det ut i praksis

For en profesjonell nettverksplattform bygde selskapet produktanalyse i PostgreSQL i stedet for sporingskode fra en tredjepart. Skjemaet teller 7 tabeller og 47 funksjoner. Hendelsene ligger i tabeller delt opp etter 22 partisjonsdefinisjoner som selv fjerner gamle data etter en fast plan, og den besøkendes identitet er byttet ut med et pseudonym laget med et salt som skiftes jevnlig. Klienten sender ingenting før den besøkende har samtykket, og trakter regnes ut i databasen, ved siden av dataene, det NIST kaller datalokalitet. Siden er ærlig om prisen: når saltet roterer, mister man kohortanalyse over lang tid, og det følger ingen gratis dashbord med.

På rapporteringssiden hentet et rapporteringssystem i Python dataene, kjørte analysen og leverte den samme PDF-rapporten etter en forutsigbar tidsplan til dem som tar de forretningsmessige beslutningene. Ifølge siden var det rapporteringen som løftet programvareinntektene per kvartal med 400 %.

Det tredje eksempelet er dette nettstedet. Personvernsiden beskriver målingen: en tilgangslogg som bytter ut siste del av IP-adressen med nuller før linjen skrives, henvisende nettsteder lagret bare som navn, og linjer som slettes etter omtrent tretti dager. Et program gjør loggen om til opptellingen og rangeringen, som oppbevares på ubestemt tid uten noe som kan sette tallene sammen til ett besøk igjen. Uten sporingskode, informasjonskapsel eller tredjepart viser siden heller ikke noe samtykkebanner, og den gir selv grunnen: det er ikke noe å samtykke til. Grensen står der også: så lenge en linje finnes, kan en uvanlig

nettleser på et lite nettverk peke ut en besøkende tydeligere enn nettstedet ønsker. Man får færre svar og ingen trakter, men til gjengjeld lagres ingenting på leserens enhet, og ingenting sendes videre.

Hvor det går galt

To verktøy gir to besøkstall

Det driftede verktøyet viser ett tall for forrige måned og serverloggen dobbelt så mange, og begge har rett om hver sin ting. Sporingskoden teller bare nettlesere som kjørte skriptet, og bak et banner bare dem som sa ja. En logg teller hver forespørsel, søkeroboter inkludert, til noen filtrerer dem bort. Skriv ned definisjonen av hvert måltall én gang, i SQL, med hva det tar med og hva det utelater.

Rapportene gjør applikasjonen treg

Hver mandag går applikasjonen tregt mens ukerapporten skanner hele tabeller i transaksjonsdatabasen. Flytt rapporten til en replika eller et datavarehus, eller slå sammen radene til en liten oppsummeringstabell hver natt og la vindusfunksjoner regne 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;
```

Banneret forsvinner, og sporingen blir

Et team bytter ut informasjonskapsler med et «cookiefritt» produkt, tar ned samtykkebanneret og fortsetter å følge de besøkende. Artikkel 5 nr. 3 handler om tilgang til informasjon på utstyret, uansett teknikk. Det europeiske personvernrådet (EDPB) skriver i retningslinjene om artikkelens tekniske virkeområde at fingeravtrykk ut fra HTTP-hodefelt, eller sporing av navigasjon bare via IP-adressen, også kan omfattes, og minner samtidig om at samtykke ikke er påkrevd bare fordi regelen gjelder. Vurder et verktøy etter om det knytter én forespørsel til den neste, ikke etter om det setter en informasjonskapsel.

Samtykke som tilsynsmyndighetene ville avvise

Et banner sperrer siden til den besøkende godtar, regner rulling som et ja eller gjør det vanskeligere å trekke samtykket tilbake enn det var å gi det. EDPBs retningslinjer om samtykke sier at en cookievegg gjør samtykket ufrivillig, og at rulling aldri er en klar bekreftende handling. Artikkel 7 nr. 3 i GDPR krever at det skal være like lett å trekke samtykket tilbake som å gi det. Gi et nei samme vekt som et ja, og send ingenting før svaret foreligger.

Pseudonyme hendelser lagret for alltid

Hendelsestabellen vokser med millioner av rader i måneden, og de hashede bruker-ID-ene i den behandles som anonyme. Ifølge fortalepunkt 26 i GDPR er pseudonymiserte data som kan knyttes til en person ved hjelp av tilleggsopplysninger, fortsatt personopplysninger, og fortalepunkt 30 nevner IP-adresser og cookie-ID-er blant nettidentifikatorene der sporene, kombinert med andre opplysninger, kan brukes til å profilere og identifisere noen. Fastsett lagringstiden før den første hendelsen kommer inn, partisjoner etter tid så en utløpt måned kan fjernes i én operasjon, og ta vare på langsiktige tall som aggregater uten identifikatorer.

Les mer

- NIST SP 1500-1r2, Big Data Interoperability Framework: Definitions — utgangspunktet for artikkelen: definisjonene av analyse, livssyklusen for dataanalyse og datalokalitet.
- Directive 2002/58/EC, consolidated — artikkel 5 nr. 3, samtykkeregelen for alt som lagres på eller leses fra en besøkendes utstyr, med de to unntakene.
- Regulation (EU) 2016/679, the GDPR — definisjonen av samtykke, artikkel 7 om å trekke det tilbake og fortalepunkt 26 og 30 om pseudonymiserte data og nettidentifikatorer.
- EDPB Guidelines 05/2020 on consent — cookievegger, rulling og tilbaketrekking, hver med konkrete eksempler.
- EDPB Guidelines 2/2023 on Article 5(3) — hvorfor sporing via IP-adresse, fingeravtrykk og sporingspiksler kan falle inn under samme regel som informasjonskapsler.
- PostgreSQL tutorial: Window Functions — rangeringer og løpende summer over beslektede rader, regnet ut inne i databasen.

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 dataanalyse?» 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/analytics/