Engineer

Dataanalyse

systematisk behandling af data for at finde mønstre

Hvad det er

Dataanalyse er arbejdet med at gøre registrerede fakta til svar, som nogen kan handle på, og til hverdag kaldes det ofte bare analyse eller analytics. I sit rammeværk for big data beskriver det amerikanske standardiseringsinstitut NIST analyse som en metodisk gennemgang og bearbejdning af data, der skal bringe mønstre, sammenhænge mellem data og udviklinger over tid frem og gerne også forudsige, hvad der sker næste gang. Hos NIST er analysen ét trin i en større *livscyklus for dataanalyse*, som gør rådata til viden, en organisation kan bruge: indsamling, klargøring, analyse, visualisering og adgang.

Analyse er en form for databehandling, der skiller sig ud fra den daglige transaktionsbehandling ved det, den læser. En applikation skriver én ordre ad gangen, mens en analytisk forespørgsel læser millioner af rækker for at returnere en håndfuld. Data engineering er noget tredje: det bygger de pipelines og det lager, analysen læser fra.

Ordet dækker arbejde, der ser meget forskelligt ud:

- **Rapportering**: de samme tal leveret efter en fast plan, som PDF eller mail.
- **Dashboards**: de samme tal holdt opdateret i et BI-værktøj som Power BI.
- **Ad hoc-analyse**: en SQL-forespørgsel skrevet til et spørgsmål, ingen har stillet før, ofte mod et datawarehouse som BigQuery.
- **Geografisk analyse**: det samme arbejde på kortdata i et GIS, hvor spørgsmålet også er hvor, og ikke kun hvor meget.
- **Analytiske applikationer**: en model eller en statistisk metode bag en formular, til folk, der ikke selv skriver kode.
- **Web- og produktanalyse**: at tælle de besøgende på et website eller brugerne af et produkt og se, hvad de gør der.

Den sidste slags har en pris i privatliv, og det er den, dette sites sider om privatliv, cookies, kolofon og sikkerhed mener, når de skriver »ingen analytics«, »intet analyseværktøj« eller »ingen webanalyse«: intet analyse-tag, ingen sporingspixel, ingen cookie og ingen konto hos en analysevirksomhed. Sitet måler sig selv alligevel. Webserveren fører en adgangslog, og et program gør den til en daglig optælling og en månedlig rangliste over læste sider, henvisende sites og crawlere. Det er tal, som aldrig sættes sammen til én besøgendes vej gennem sitet.

Brug begrebet i en beslutning

Brug »Hvad er dataanalyse?« i en konkret beslutning. Vi kan knytte begrebet til det system, I bygger eller køber. Dashboards og rapporter, der viser de få tal, en beslutning faktisk afhænger af, opdateret uden at nogen beder om det.

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. Brandkommunikation med afsæt i publikums reaktioner — brandkommunikation med afsæt i publikums reaktioner

Brandkommunikation med afsæt i publikums reaktioner Kontakt os →

Hvorfor det betyder noget

To valg afgør det meste af, hvad analysen kommer til at koste: hvor forespørgslerne kører, og hvordan data om mennesker bliver indsamlet.

Analyse direkte på transaktionsdatabasen er den billigste begyndelse. Dataene er friske, og intet skal kopieres, men hver rapport konkurrerer med applikationen om hukommelse og disk. En læsereplika flytter belastningen væk fra den primære server, for prisen af en ekstra server at drive. Et datawarehouse som BigQuery holder lagring og beregning adskilt og afregner med sin on demand-pris hver forespørgsel efter, hvor mange bytes den læser, så smalle forespørgsler bliver billige. Til gengæld skal en pipeline kopiere dataene ind.

Bindingen til en leverandør gemmer sig i definitionerne. Et BI-værktøj holder sine målinger i sit eget formelsprog, og et år senere står definitionen af en aktiv kunde i en fil, som kun det værktøj kan åbne. Definitioner, der ligger som SQL i repositoryet, overlever et skifte af værktøj.

Web- og produktanalyse rejser desuden et juridisk spørgsmål. Det typiske hostede værktøj virker gennem et tag: et script på hver side, der sætter en cookie eller en lignende identifikator i browseren og melder hver sidevisning tilbage til leverandøren. Artikel 5, stk. 3, i EU's ePrivacy-direktiv gør samtykke til

betingelsen for at lægge oplysninger på den besøgendes udstyr eller hente dem derfra igen, med to undtagelser: lagring eller adgang, hvis eneste formål er at overføre en kommunikation over et netværk, og lagring eller adgang, der er strengt nødvendig for at levere en tjeneste, den besøgende udtrykkeligt har bedt om. Kan hændelserne knyttes til en person, om så kun gennem et cookie-id eller en IP-adresse, gælder databeskyttelsesforordningen, GDPR, også, med krav om et retsgrundlag, et oplyst formål og en opbevaringsperiode. Og leverandøren er en anden virksomhed, der nu ligger inde med en optegnelse over, hvad dine besøgende gjorde.

Det reelle valg står derfor mellem tre ting: et hostet værktøj, der er hurtigt at sætte op og har dashboards med, en førstepartsindsamling, man selv bygger og driver, og tal fra de logfiler, serveren allerede skriver, som besvarer færre spørgsmål og rummer langt mindre.

Sådan ser det ud i praksis

For en professionel netværksplatform byggede virksomheden produktanalyse i PostgreSQL i stedet for et tag fra en tredjepart. Skemaet har 7 tabeller og 47 funktioner, hændelsestabellerne er partitioneret med 22 partitionsdefinitioner, der rydder op i sig selv efter en fast plan, og identiteten er pseudonymiseret bag et salt, der roterer. Klienten sender intet, før den besøgende har givet samtykke, og funnels beregnes i databasen ved siden af dataene, det, NIST kalder datalokalitet. Siden er ærlig om prisen: det roterende salt koster kohorteanalyser over lang tid, og der følger ingen gratis dashboards med.

På rapporteringssiden hentede et rapporteringssystem i Python dataene, kørte analysen og leverede den samme PDF-rapport efter en forudsigelig tidsplan til dem, der træffer de kommercielle beslutninger. Ifølge siden var det rapporteringen, der øgede den kvartalsvise softwareomsætning med 400 %.

Det tredje eksempel er dette site. Privatlivssiden beskriver målingen: en adgangslog, der erstatter den sidste del af IP-adressen med nuller, før linjen skrives, henvisende sites gemt som navn alene, og linjer, der slettes efter omkring tredive dage. Et program gør loggen til optællingen og ranglisten, som gemmes på ubestemt tid uden noget, der kan samle tallene til ét besøg igen. Uden tag, cookie eller tredjepart viser siden intet samtykkebanner og giver selv grunden: der er ikke noget at samtykke til. Grænsen står der også: mens en linje

findes, kan en usædvanlig browser på et lille netværk indsnævre en besøgende mere, end sitet bryder sig om. Man får færre svar og ingen funnels, men til gengæld gemmes intet på læserens enhed, og intet sendes videre.

Hvor det går galt

To værktøjer giver to besøgstal

Det hostede værktøj viser ét tal for sidste måned og serverloggen dobbelt så mange, og begge har ret om hver sin ting. Et tag tæller kun browsere, der kørte dets script, og bag et banner kun dem, der sagde ja. En log tæller hver forespørgsel, crawlere med, indtil nogen filtrerer dem fra. Skriv hvert måltals definition ned én gang, i SQL, med hvad det tager med og udelader.

Rapporterne gør applikationen langsom

Hver mandag bliver applikationen sløv, mens ugerapporten scanner hele tabeller i transaktionsdatabasen. Flyt rapporten til en replika eller et datawarehouse, eller saml rækkerne i en lille opsummeringstabel hver nat, og lad vinduesfunktioner 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 forsvinder, og sporingen bliver

Et team skifter cookies ud med et »cookiefrit« produkt, tager samtykkebanneret ned og fortsætter med at følge de besøgende. Artikel 5, stk. 3, handler om adgang til oplysninger på udstyret, uanset teknikken. Det Europæiske Databeskyttelsesråd, EDPB, skriver i sine retningslinjer om artiklens tekniske anvendelsesområde, at fingerprinting ud fra HTTP-headere eller sporing af navigation alene via IP-adressen også kan være omfattet, og minder samtidig om, at samtykke ikke er påkrævet, blot fordi reglen gælder. Bedøm et værktøj efter, om det forbinder én forespørgsel med den næste, og ikke efter, om det sætter en cookie.

Samtykke, som tilsynsmyndighederne ville afvise

Et banner spærrer siden, indtil den besøgende accepterer, regner scrolling for et ja eller gør det sværere at trække samtykket tilbage, end det var at give det. EDPB's retningslinjer om samtykke siger, at en cookievæg betyder, at samtykket ikke er frivilligt, og at scrolling aldrig er en klar bekræftende handling. Artikel 7, stk. 3, i GDPR kræver, at det er lige så let at trække samtykket tilbage som at give det. Giv et nej samme vægt som et ja, og send intet, før svaret foreligger.

Pseudonyme hændelser gemt for evigt

Hændelsestabellen vokser med millioner af rækker om måneden, og de hashede bruger-id'er i den bliver behandlet som anonyme. Ifølge betragtning 26 i GDPR er pseudonymiserede data, som yderligere oplysninger kan henføre til en person, stadig personoplysninger, og betragtning 30 nævner IP-adresser og cookie-id'er blandt de onlineidentifikatorer, hvis spor, kombineret med andre oplysninger, kan bruges til at profilere og identificere nogen. Fastlæg opbevaringsperioden, før den første hændelse ankommer, partitionér efter tid, så en udløbet måned kan fjernes i ét hug, og gem langsigtede tal som aggregater uden identifikatorer.

Læs mere

- NIST SP 1500-1r2, Big Data Interoperability Framework: Definitions — artiklens udgangspunkt: definitionerne af analyse, livscyklussen for dataanalyse og datalokalitet.
- Directive 2002/58/EC, consolidated — artikel 5, stk. 3, reglen om samtykke til alt, der gemmes på eller læses fra en besøgendes udstyr, med dens to undtagelser.
- Regulation (EU) 2016/679, the GDPR — definitionen af samtykke, artikel 7 om at trække det tilbage og betragtning 26 og 30 om pseudonymiserede data og onlineidentifikatorer.
- EDPB Guidelines 05/2020 on consent — cookievægge, scrolling og tilbagetrækning, hver med konkrete eksempler.
- EDPB Guidelines 2/2023 on Article 5(3) — hvorfor sporing via IP-adresse, fingerprinting og sporingspixels kan falde ind under samme regel som cookies.
- PostgreSQL tutorial: Window Functions — ranglister og løbende summer over beslægtede rækker, beregnet inde i databasen.

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 dataanalyse?« 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/analytics/