Engineer

Аналітика даних

систематична обробка даних для пошуку закономірностей

Що це таке

Перетворювати записані факти на відповіді, за якими хтось може діяти, — ось завдання аналітики даних, яку здебільшого називають просто аналітикою. У рамковій моделі для великих даних, яку видав NIST, американський інститут стандартів, аналітика означає методичне опрацювання даних, а за потреби й їх перетворення, щоб виявити закономірності, зв'язки між даними й тенденції в часі та спробувати передбачити, що буде далі. Цей крок NIST вміщує в ширший *життєвий цикл аналітики даних*, який робить із сирих даних знання, придатне для організації: збирання, підготовка, аналіз, візуалізація й надання доступу.

Аналітика — різновид обробки даних, і від транзакційної обробки її відрізняє те, що саме вона читає. Додаток записує одне замовлення за раз, а аналітичний запит перебирає мільйони рядків, щоб повернути кілька. Інженерія даних — ще інша справа: вона будує конвеєри й сховища, з яких аналітика потім читає.

Під цим словом ховається дуже різна робота:

- **Звітність**: ті самі показники за розкладом, у PDF або листом.
- **Дашборди**: ті самі показники, що постійно оновлюються в BI-інструменті на кшталт Power BI.
- **Разовий аналіз**: SQL-запит під питання, якого досі ніхто не ставив, часто до сховища даних на зразок BigQuery.
- **Геопросторовий аналіз**: та сама робота з картографічними даними в GIS, де питають не лише «скільки», а й «де».
- **Аналітичні додатки**: модель або статистичний метод, прихований за вебформою, для людей, які не пишуть коду.

- **Веб- і продуктова аналітика**: підрахунок відвідувачів вебсайту чи користувачів продукту і того, що вони там роблять.

Останній різновид має ціну для приватності, і саме його мають на увазі сторінки цього сайту про конфіденційність, файли cookie, подяки й безпеку, коли пишуть «жодної аналітики» чи «жодного аналітичного тега». Ідеться про тег, піксель стеження, cookie й обліковий запис в аналітичній компанії: нічого з цього сайт не використовує. Проте себе сайт вимірює. Вебсервер веде журнал доступу, а програма перетворює його на щоденні лічильники й місячний рейтинг прочитаних сторінок, сайтів, звідки прийшли люди, і краулерів. Це лише числа, і маршрут окремого відвідувача сайтом із них ніколи не складають.

Застосуйте поняття до рішення

Застосуйте «Що таке аналітика даних?» до конкретного рішення. Ми можемо пов'язати поняття із системою, яку ви створюєте або купуєте. Дашборди і звіти, що показують ті кілька цифр, від яких залежить рішення, й оновлюються без нагадувань.

Переглянути наші послуги Зв'язатися з нами →

Перевірте припущення за терміном

Використайте пояснення та пов'язані роботи, щоб сформулювати запитання для проєкту. Ми можемо перевірити припущення разом до вибору інструмента. Комунікація бренду з урахуванням реакцій аудиторії — комунікація бренду з урахуванням реакцій аудиторії

Комунікація бренду з урахуванням реакцій аудиторії Зв'язатися з нами →

Чому це важливо

Більшу частину вартості аналітики визначають два рішення: де виконуються запити і як збирають дані про людей.

Аналітика безпосередньо на транзакційній базі даних — найдешевший початок: дані свіжі, копіювати нічого не треба, але кожен звіт змагається з додатком за пам’ять і диск. Репліка для читання знімає це навантаження з основного сервера ціною ще одного сервера. Сховище даних на зразок BigQuery тримає зберігання окремо від обчислень і за тарифом on-demand виставляє рахунок за кожен запит залежно від того, скільки байтів він прочитав, тож вузькі запити обходяться дешево. Натомість потрібен конвеєр, який переносить туди дані.

Прив’язка до постачальника ховається у визначеннях. BI-інструмент зберігає свої міри власною мовою формул, і за рік визначення активного клієнта опиняється у файлі, який відкриває тільки цей інструмент. Визначення, що лежать як SQL у репозиторії, переживають зміну інструмента.

Веб- і продуктова аналітика додає ще й юридичне питання. Типовий хмарний сервіс працює через тег: скрипт на кожній сторінці ставить у браузері cookie або подібний ідентифікатор і звітує постачальникові про кожен перегляд сторінки. Стаття 5(3) європейської Директиви ePrivacy робить згоду умовою для того, щоб покласти інформацію на пристрій відвідувача або прочитати її звідти, з двома винятками: зберігання чи доступ, єдина мета яких — передати повідомлення мережею, і зберігання чи доступ, суворо необхідні для надання послуги, яку відвідувач явно замовив. Щойно події можна пов'язати з людиною, хай навіть через ідентифікатор cookie чи IP-адресу, діє ще й GDPR: потрібні правова підстава, заявлена мета і строк зберігання. А постачальник — інша компанія, у якої тепер є запис того, що робили ваші відвідувачі.

Тож вибір насправді з трьох: хмарний сервіс, який швидко підключити і який має готові дашборди; власне збирання першою стороною, яке будують і обслуговують самі; і підрахунки з журналів, які сервер і так пише, — вони відповідають на менше питань і містять значно менше.

Як це виглядає на практиці

Для професійної соцмережі компанія побудувала продуктову аналітику в PostgreSQL замість стороннього тега. Схема налічує 7 таблиць і 47 функцій; події лежать у таблицях, розбитих на партиції за 22 визначеннями, що за розкладом самі прибирають старе, а особу відвідувача псевдонімізовано сіллю, яку регулярно змінюють. Клієнт нічого не надсилає, доки відвідувач не дав згоди, а лійки конверсії рахуються там само, де лежать дані, — NIST називає це локальністю даних. Сторінка чесно називає ціну: через ротацію солі втрачається когортний аналіз на довгих горизонтах, а безкоштовних дашбордів у комплекті немає.

На боці звітності система звітності на Python забирала дані, проводила аналіз і за передбачуваним розкладом надсилала той самий PDF-звіт людям, які ухвалюють комерційні рішення. Сторінка пов'язує з цією звітністю зростання квартального доходу від програмного забезпечення на 400%.

Третій приклад — цей сайт. Його сторінка про конфіденційність описує вимірювання: журнал доступу, де остання частина IP-адреси стає нулями ще до запису рядка, сайт, звідки прийшов відвідувач, лишається тільки назвою, а рядки зберігаються приблизно тридцять днів. Програма зводить журнал у щоденний підсумок і місячний рейтинг, які зберігаються безстроково як

числа, без того, що дало б знову скласти з них один візит. Тега, cookie і третьої сторони тут немає, тож банера згоди сторінка не показує й пояснює це тим, що погоджуватися нема на що. Межу вона теж називає: поки рядок існує, браузер, яким мало хто користується, у малій мережі може вказати на відвідувача точніше, ніж сайтові хотілося б. Компроміс такий: менше відповідей і жодних лійок, зате на пристрої читача нічого не зберігається і нікому нічого не передається.

Де виникають проблеми

Два інструменти дають два числа відвідувачів

Хмарний сервіс показує одне число за минулий місяць, а журнал сервера — удвічі більше, і кожне правильне щодо свого. Тег рахує лише браузери, які виконали його скрипт, а за банером — лише тих, хто погодився. Журнал рахує кожен запит, краулерів теж, доки їх хтось не відфільтрує. Запишіть визначення кожної метрики один раз, на SQL, разом із тим, що вона враховує і що відкидає.

Звіти гальмують додаток

Щопонеділка додаток пригальмовує, поки тижневий звіт сканує цілі таблиці в транзакційній базі даних. Перенесіть звіт на репліку чи в сховище даних або щоночі зводьте сирі рядки в невелику підсумкову таблицю й рахуйте по ній віконними функціями:

```
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;
```

Банер зникає, а стеження лишається

Команда міняє cookie на продукт «без cookie», знімає банер згоди і далі стежить за відвідувачами. Стаття 5(3) стосується доступу до інформації на пристрої, хоч би яким був механізм. Європейська рада із захисту даних (EDPB) у настановах щодо технічної сфери застосування цієї статті пише, що під неї можуть підпадати й фінгерпринтинг за HTTP-заголовками, і відстеження навігації лише за IP-адресою, та водночас нагадує: те, що стаття

застосовна, саме по собі ще не означає, що потрібна згода. Оцінюйте інструмент за тим, чи пов’язує він один запит із наступним, а не за тим, чи ставить він cookie.

Згода, яку регулятор відхилив би

Банер блокує сторінку, доки відвідувач не погодиться, вважає прокручування згодою або робить відкликання згоди складнішим, ніж було її надання. Настанови EDPB щодо згоди кажуть, що cookie wall, тобто доступ до вмісту лише в обмін на згоду, робить її невільною, а прокручування ніколи не буває чіткою ствердною дією. Стаття 7(3) GDPR вимагає, щоб відкликати згоду було так само легко, як її надати. Дайте відмові ту саму вагу, що й згоді, і нічого не надсилайте, поки не отримаєте відповіді.

Псевдонімні події, що зберігаються вічно

Таблиця подій росте на мільйони рядків щомісяця, а хешовані ідентифікатори користувачів у ній вважають анонімними. Пункт 26 преамбули GDPR відносить до персональних даних і псевдонімізовані дані, які додаткова інформація дає змогу пов’язати з людиною, а пункт 30 називає IP-адреси та ідентифікатори cookie серед онлайн-ідентифікаторів, сліди яких у поєднанні з іншою інформацією можна використати, щоб скласти профіль людини й ідентифікувати її. Встановіть строк зберігання ще до першої події, розбийте таблицю на партиції за часом, щоб місяць, строк зберігання якого сплив, видалявся цілим, а довгострокові показники зберігайте як агрегати без ідентифікаторів.

Що почитати далі

- NIST SP 1500-1r2, Big Data Interoperability Framework: Definitions — визначення аналітики, життєвого циклу аналітики даних і локальності даних, від яких відштовхується ця стаття.
- Directive 2002/58/EC, consolidated — стаття 5(3), правило згоди для всього, що зберігають на пристрої відвідувача або читають із нього, з двома винятками.
- Regulation (EU) 2016/679, the GDPR — визначення згоди, стаття 7 про її відкликання та пункти 26 і 30 преамбули про псевдонімізовані дані й онлайн-ідентифікатори.
- EDPB Guidelines 05/2020 on consent — cookie walls, прокручування й відкликання згоди, кожне з розібраними прикладами.
- EDPB Guidelines 2/2023 on Article 5(3) — чому відстеження за IP-адресою, фінгерпринтинг і пікселі стеження можуть підпадати під те саме правило, що й cookie.

- PostgreSQL tutorial: Window Functions — рейтинги й накопичувальні суми за пов’язаними рядками, обчислені в самій базі даних.

Словник

Обирайте за фактичними умовами

Надайте репрезентативний набір даних, запитання до нього та поточне вузьке місце. Узгодимо вимірювання якості, пропускної здатності та відновлення.

Дивитися, скільки це коштує Зв'язатися з нами →

Зробіть термінологію практичною

Запитайте, як «Що таке аналітика даних?» стосується вашої організації. Конкретного прикладу з поточної системи достатньо для початку розмови.

Переглянути наші послуги

welcome@engineer.company +45 50 10 61 84 Копенгаген, Данія

Візьміть цю сторінку із собою

Візитівка

Дбайливо зроблено з любов'ю — D.

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

Авторське право © 2025 – дотепер · Engineer ApS · Усі права захищено.

Надруковано з https://engineer.company/uk/glossary/analytics/