Engineer

データ分析

パターンや傾向を見つけるためにデータを体系的に処理すること

どのようなものか

記録された事実を、行動に使える答えへ変えるのが、データ分析の仕事です。英語ではdata analytics、または単にanalyticsと呼びます。米国の標準機関NISTが公開するビッグデータの枠組みでは、データを順序立てて調べ、必要に応じて形を変え、パターン、データ間の関係、時間的な傾向を見いだし、その後を予測することを意味します。NISTはそれを、生データを組織が使える知識へ変える、より広い「データ分析のライフサイクル」の一段階としています。収集、準備、分析、可視化、アクセスの提供という流れです。

データ処理の一種ですが、トランザクション処理とは、読む対象が異なります。アプリケーションは注文を一件ずつ書きますが、分析クエリは数百万行を読んで少数の結果を返します。また、分析で使うデータを供給するパイプラインや保存先を構築する、データエンジニアリングとも異なります。

この語には、かなり異なる仕事が含まれます。

- **レポート作成：** 同じ数値を、PDFやメールで定期的に届けます。
- **ダッシュボード：** Power BIなどのビジネスインテリジェンスツールで、その数値を最新に保ちます。
- **個別の分析：** これまでなかった問いのためにSQLを書き、多くの場合はBigQueryなどのデータウェアハウスに問い合わせます。
- **地理空間分析：** GISの地図データに同じ作業を行い、数量だけでなく場所も問います。
- **分析アプリケーション：** コードを書かない人向けに、フォームの背後でモデルや統計手法を使います。
- **ウェブ・製品の利用分析：** サイトの訪問者や製品の利用者と、その行動を数えます。

最後の種類にはプライバシーの負担があり、このサイトのプライバシー、Cookie、クレジット、セキュリティのページで「アクセス解析なし」と述べるのは、その意味です。解析タグ、追跡ピクセル、Cookie、解析会社のアカウントはありません。ただし、サイト自身の測定は行います。ウェブサーバーがアクセスログを保持し、プログラムが日ごとの件数と、読まれたページ、参照元サイト、クローラーの月間順位に変えます。数値として保持し、個々の訪問者の経路には結び付けません。

概念を判断に生かす

「データ分析とは？」を実際の判断に生かしましょう。その概念を、開発または購入するシステムに結び付けて考えるお手伝いをします。 意思決定を左右する少数の指標を示すダッシュボードやレポートを、依頼しなくても更新される形で提供します。

サービスを見る お問い合わせ →

用語の背後にある前提を検証する

解説とリンク先の実績を参考に、プロジェクトについての質問を整理してください。ツールを選ぶ前に、前提を一緒に検討できます。 読み手の反応に基づくブランド発信 — 読み手の反応に基づくブランド発信

読み手の反応に基づくブランド発信 お問い合わせ →

なぜ重要なのか

分析の費用の大半は、クエリをどこで動かし、人に関するデータをどう集めるか、という二つの選択で決まります。

トランザクション用のデータベースで分析するのが、最も安く始める方法です。最新のデータを使え、コピーも不要ですが、各レポートが、メモリーとディスクをアプリケーションと取り合います。読み取りレプリカは、二台目のサーバーの費用と引き換えに、その負荷をプライマリーから外します。BigQueryなどのデータウェアハウスは、保存と計算を分離し、オンデマンド料金では、各クエリが読むバイト数で請求するため、狭く絞ったクエリが有利です。また、データをコピーするパイプラインが必要です。

乗り換えの難しさは、定義に隠れます。BIツールが独自の数式言語で指標を保持すると、一年後、「アクティブな顧客」の意味が、そのツールでしか開けない一つのファイルに置かれます。定義をSQLとしてリポジトリに保持すれば、ツールを変えても残ります。

ウェブや製品の利用分析には、法的な問いも加わります。一般的な外部ツールは、各ページのタグ、すなわちブラウザーにCookieなどの識別子を置き、各閲覧を提供事業者へ報告するスクリプトで動きます。EUのePrivacy指令第5条(3)は、訪問者の機器への情報の保存・読み取りを、二つの例外を除き、同意を条件としています。例外は、ネットワーク通信の伝送だけを目的とする保存・アクセス

と、利用者が明示的に求めたサービスの提供に厳密に必要な保存・アクセスです。Cookie識別子やIPアドレスを通じてであっても、事象を人に結び付けられるなら、GDPRも適用されます。適法な根拠、明示した目的、保存期間が必要です。そして、訪問者の行動記録を持つ会社が、もう一社増えます。

したがって実際の選択肢は、すぐに追加できダッシュボードも付く外部ツール、自社で構築・運用する収集の仕組み、またはサーバーがすでに書くログの集計です。最後は答えられる問いが少ない代わりに、保持する情報もずっと少なくなります。

実際の活用例

ビジネス上の人脈をつなぐプラットフォームでは、第三者のタグの代わりに、PostgreSQL内に自社管理の利用分析を構築しました。7テーブルと47関数のスキーマ、定期的に古い分を削除する22のパーティション定義を持つイベントテーブル、更新するソルトで仮名化した訪問者IDを備えます。訪問者が同意するまではクライアントから何も送らず、NISTがデータローカリティと呼ぶように、データがあるデータベースの中でファネルを計算します。そのページは代償も示しています。ソルトを更新するため長期のコホート分析はできず、無償のダッシュボードも付いてきません。

レポート側では、Pythonによる分析レポートシステムが、データを取得し、分析し、同じPDFレポートを、事業上の判断を行う人に、あらかじめ決まった日程で届けました。そのページは、四半期のソフトウェア売上が400%増えた要因として、このレポートを挙げています。

このサイトが三つ目の例です。プライバシーのページで、測定内容を説明しています。IPアドレスの末尾を行の書き込み前にゼロにするアクセスログ、名前だけを保持する参照元サイト、約三十日の行の保存期間です。プログラムがログから日次集計と月間順位を作り、どちらも、単一の訪問に結び直せる情報のない件数として無期限に保持します。タグもCookieも第三者もないため、同意バナーを表示せず、「同意する対象がない」と説明しています。限界も明かしています。ログの行が残っている間は、小規模なネットワークと珍しいブラウザーの組み合わせから、サイト側の望む以上に訪問者を絞り込めてしまいます。読者の機器に何も保存せず、他者へ何も送らない代わりに、得られる答えは少なく、ファネルもありません。

問題が起きるところ

二つのツールで訪問者数が違う

外部ツールの先月の数値に対し、サーバーログは倍の数を示します。どちらも、異なる対象について正しい場合があります。タグはスクリプトを動かしたブラウザーだけを数え、バナーがあれば同意した訪問者だけを数えます。ログは、除外するまでクローラーも含め、すべてのリクエストを数えます。含むものと除くものを含め、各指標の定義をSQLで一度、書き留めます。

レポートがアプリケーションを遅くする

毎週月曜、週次レポートがトランザクション用データベースでテーブル全体を走査し、アプリケーションが遅くなります。レポートをレプリカやデータウェアハウスへ移すか、毎晩、生の行を小さな集計テーブルへまとめ、その上でウィンドウ関数を使います。

```
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;
```

バナーだけを除いて、追跡を残す

Cookieを「Cookieなし」の製品に置き換え、同意バナーを外して、訪問者の追跡は続けます。第5条(3)が扱うのは、仕組みにかかわらず、機器上の情報へのアクセスです。技術的な範囲に関するEDPBのガイドラインは、HTTPヘッダーによるフィンガープリンティングや、IPアドレスだけによる移動の追跡も対象になり得るとしています。同時に、規則の適用だけで、直ちに同意が必要になるわけではないことも確認しています。Cookieを設定するかではなく、一つの要求を次へ結び付けるかで、ツールを判断します。

規制当局が認めない同意

同意するまでバナーがページを塞ぎ、スクロールを同意と扱い、または撤回を同意より難しくします。EDPBの同意ガイドラインは、Cookieウォールでは自由に与えられた同意にならず、スクロールは明確な積極的行為には決してならないとします。GDPR第7条(3)は、撤回を同意と同じ容易さにするよう求めます。拒否に同意と同じ重みを与え、答えが来るまで何も送りません。

仮名化したイベントを永久に保持する

毎月数百万行増えるイベントテーブルで、ハッシュ化した利用者IDを匿名と扱います。GDPRの前文26は、追加情報で本人に結び付けられる仮名化データを個人データとし、前文30は、IPアドレスやCookie識別子をオンライン識別子の例に挙げ、その痕跡を他の情報と組み合わせると、人物のプロファイリングや特定に使える場合があるとしています。最初のイベントが届く前に保存期間を定め、期限切れの月を丸ごと削除できるよう時間で分割し、長期の数値は識別子のない集計として保持します。

参考資料

- NIST SP 1500-1r2『Big Data Interoperability Framework: Definitions』：この記事の出発点となる、分析、分析のライフサイクル、データローカリティの定義です。
- 指令2002/58/ECの統合版：訪問者の機器に保存・読み取りするものの同意要件を定める第5条(3)と、二つの例外です。
- 規則(EU) 2016/679、GDPR：同意の定義、撤回に関する第7条、仮名化データとオンライン識別子に関する前文26・30です。
- 同意に関するEDPBガイドライン05/2020：Cookieウォール、スクロール、撤回を実例とともに説明します。
- 第5条(3)に関するEDPBガイドライン2/2023：IP追跡、フィンガープリンティング、追跡ピクセルにも、Cookieと同じルールが適用され得る理由を説明します。
- PostgreSQLチュートリアル：ウィンドウ関数：関連する行の順位や累計を、データベース内で計算する方法です。

用語集

実際の制約に合わせて選ぶ

代表的なデータセット、そこから答えを得たい問い、現在のボトルネックをお知らせください。品質、処理量、復旧の測り方を合意できます。

料金を見る お問い合わせ →

用語を実務に結び付ける

「データ分析とは？」がお客様の組織にどう当てはまるかをご相談ください。現在のシステムの具体例が一つあれば、話を始められます。

サービスを見る

welcome@engineer.company +45 50 10 61 84 デンマーク、コペンハーゲン

このページを持ち歩く

連絡先カード

心を込めて丁寧に制作： D.

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

著作権 © 2025 – 現在 · Engineer ApS · すべての権利を留保します。

印刷元： https://engineer.company/ja/glossary/analytics/