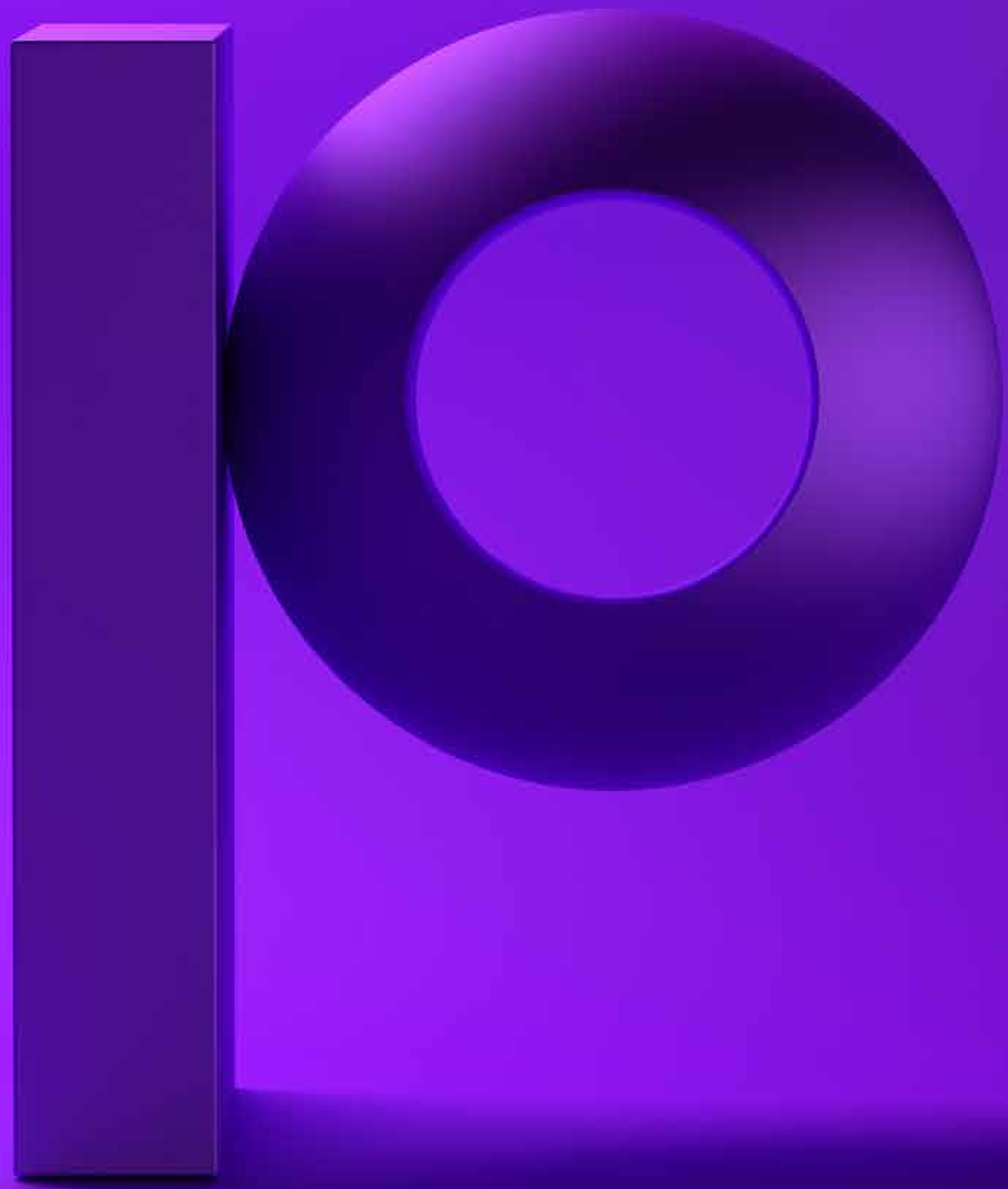
precisely

エンリッチメントデータ調達を成功させる6つのステップ

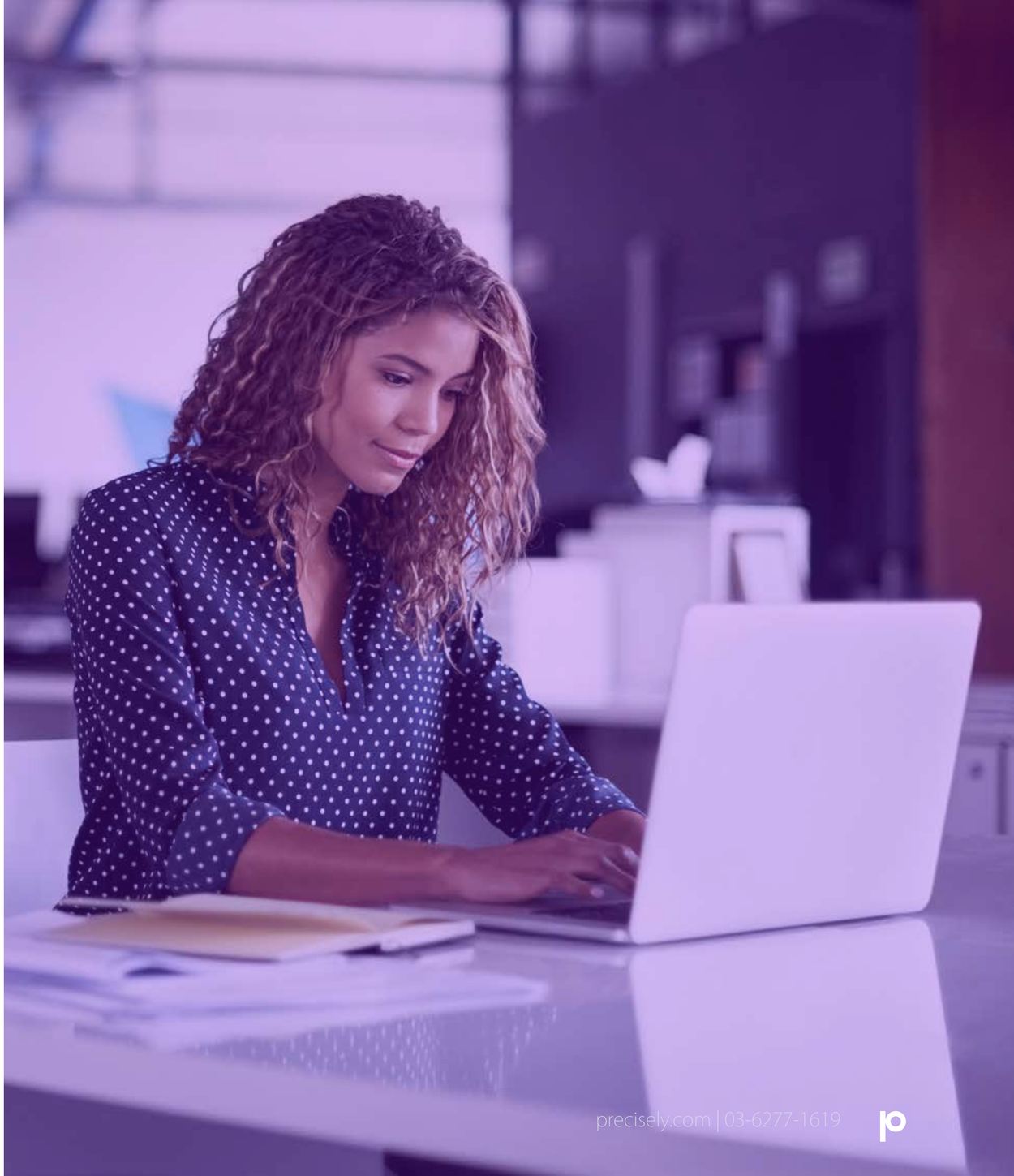


はじめに

今やデータは企業を支える原動力となっています。企業は、自社の顧客や業務に関する膨大な量のデータ (いわゆるファーストパーティデータ) を自力で生成しています。しかし、外部のサードパーティデータも欠かせません。企業は、こうした外部データによって顧客への理解を深め、オムニチャネルで充実したマーケティングキャンペーンを展開したり、営業担当者が「最善の次の一手」を打てるよう支援したりします。また、事業リスクの評価、コンプライアンスの遵守、確かな情報に基づく計画策定をはじめ、店舗やレストラン、インフラ施設などの最適な立地選定などにも、外部データが活用されています。

エンリッチメントデータの用途は、ほぼ無限にあります。難しいのは、事業目標の達成に直結する最適なデータとデータソースをいかにして見つけ出すかということです。BI ダッシュボード、問題解決、アナリティクス、AI/ 機械学習など、サードパーティデータをどの目的で利用する場合でも、結果を左右するのは、使用するデータの品質にほかなりません。この eBook では、サードパーティのエンリッチメントデータの調達を成功させるための 6 つのステップを解説します。

1. データのユースケースを理解する
2. 必要なデータを特定する
3. データソースの候補を洗い出す
4. 有力候補を評価する
5. 利用規約を確認する
6. データの取得方法を検討する



precisely.com | 03-6277-1619

ステップ 01

データのユースケースを理解する

「データのユースケースを理解する」のは当然のことに思えるかもしれませんが、多くの企業では、データの調達担当者と、そのデータを実際に使う担当者やチームとの間に隔たりがあります。例えば、データ購入の決定権が購買部門や IT 部門にある場合です。意思決定者がプロジェクトオーナーから離れていればいるほど、この最初のステップで要件をしっかりと吸い上げておくことが重要になります。

例えば、想定される用途が概念実証プロジェクトであれば、データの長期的な可用性や鮮度よりも、すぐに使えてコストを抑えられることの方が重視されます。機械学習モデルのトレーニングに使うのであれば、決め手になるのはデータの量かもしれません。一方で、業務上きわめて重要な意思決定を支援するためのデータであれば、品質と更新頻度の高さが最優先の要件となるでしょう。

さらに、調達したデータを最終的にどのような形で利用するのかを理解しておくことも重要です。例えば、「社内分析に使うのか、それとも分析結果を社外に公表するのか」、「データ自体を直接的または間接的な方法で一般に公開するのか」、「データ収益化戦略として、自社のデータと組み合わせて再販するのか」などの点について検討します。

これらの要素がすべて、データ調達の意思決定に影響します。

ポイント：
用途に適したデータを確実に入手するためのベストプラクティスは、開発チーム、想定されるエンドユーザー、社内のデータ専門家から要件を聞き取ることです。



precisely.com | 03-6277-1619

ステップ 02
必要なデータを見極める

次のステップは、ユースケースに必要なデータポイントを見極めることです。例えば、住宅を購入する人の地区選びに役立つ消費者向け Web アプリケーションを構築するというユースケースを考えてみましょう。まず入手すべきなのは位置情報データ、特に地区の境界線データです。郵便番号のような大まかな区切りに比べ、実際の生活圏をより高い精度で特定できるためです。アクセスの良さや幹線道路への近さを把握するには、道路網データが必要です。また、データを地図上で表示するには、地理空間データが不可欠です。さらに、各地区の周辺にある学校、店舗、事業所、公園、娯楽スポット、ショッピングモール、官公庁などの、いわゆる POI も把握しておきたい要素です。

各地区の平均的な住宅価格や生活費を算出するには、過去の不動産取引データや経済統計データが必要です。また、各地区のプロファイルを作成するには、所得、持ち家率、年齢層、関心事などの人口統計データや消費者セグメンテーションデータも必要です。

このステップで押さえておきたいのは、最終的に達成したい目標は何か、そこに到達するために何を知る必要があるかを、この初期段階でブレインストーミングして掘り下げておくことが重要だという点です。また、必要なデータが構造化データなのか、非構造化データなのか、あるいはその両方の組み合わせなのかを理解しておくことも重要です。データ要件の洗い出しで漏れがあると、ソフトウェアやデータベースの設計が終わってから対応するのは格段に難しくなります。

ポイント：
最終的に達成したい目標は何か、そこに到達するために何を知る必要があるかを、ブレインストーミングして掘り下げましょう。



データの調達先は国内か国外か

このステップでは、地理的な要件も検討します。データの調達先を国外にまで広げると、それ以降のステップはさらに複雑になります。同一または類似のデータセットを入手できるかどうかは、国によって大きく異なる可能性が高いからです。仮に入手できるとしても、データの構造や利用規約が異なる可能性が高く、ユースケースに適合するとは限りません。

27

企業が現在、自社のデータアーキテクチャに組み込んでいるサードパーティ データソースの平均数 [1]

[1]「Data Integrity Trends 2021 (データ完全性のトレンド 2021 年版)」、Corinium Research、2021 年



ステップ 03

データソースの候補を洗い出す

このステップがデータ調達における最大の難所かもしれません。サードパーティデータの提供元の数が非常に多いからです。手始めに、データプロバイダーのカテゴリにどのようなものがあるかを把握しておくと役立ちます。ただし、複数のカテゴリの特徴を併せ持つプロバイダーも多い点に注意してください。主なカテゴリは以下のとおりです。

- **一次データソース。**調査やデータ収集を自ら行い、それを外部に提供している組織です。学術機関や大学、政府機関、NGO、科学団体、業界団体、研究機関、シンクタンクなどがこれに該当します。また、ソフトウェアツールで Web ページからデータを抽出する「Web スクレイピング」と呼ばれるプロセスを使い、インターネットからデータを収集する組織もこれに当たります。一次ソースから得られるデータは「生データ」とも呼ばれます。

- **公共利用データおよびオープンソース データ。**通常、無償で入手できるデータで、多くの場合、一次データソースから提供されたものを元にしています。オープンソース データは、確かに商業目的でのデータ調達における有力な選択肢ですが、「無料」だからといって「コストがゼロ」だとは限りません。収集したデータの誤った使い方、データの欠落、不正確な統計的推論が問題を引き起こす可能性があります。また、オープンソース プロジェクトのデータを利用する場合、データに加えた改善内容を「還元」するように義務付けられることもあります。

- **データアグリゲーター。**複数のソースからデータを収集し、処理して新たな価値を付加してから新しいデータセットに結合し、利用しやすい形で再パッケージ化して提供する組織です。

- **データプロバイダーおよび単一ソース型のデータマーケットプレイス。**Precisely のようなデータプロバイダーは、データ集約を大規模に行い、市場で取り引きできるデータ製品を生み出しています。このようなデータ製品は商業利用を前提に作られており、簡単に統合できるよう標準化された形式、品質保証、データ ガバナンス、信頼性の高いメタデータ、製品ドキュメントなど、高い付加価値を備えています。また Precisely では、お客様がスムーズにデータを調達できるよう、データ製品を Precisely Data Experience というマーケットプレイスで提供しています。

- **データマーケットプレイス。**複数のデータプロバイダーが提供する市場性の高い製品やサービスを、検索可能な 1 つの環境に集約したプラットフォームです。Snowflake Data Marketplace や AWS Data Exchange がこれに該当します。また、セグメント化された消費者オーディエンスデータなど、業界や目的に特化したデータマーケットプレイスもあります。

precisely.com | 03-6277-1619

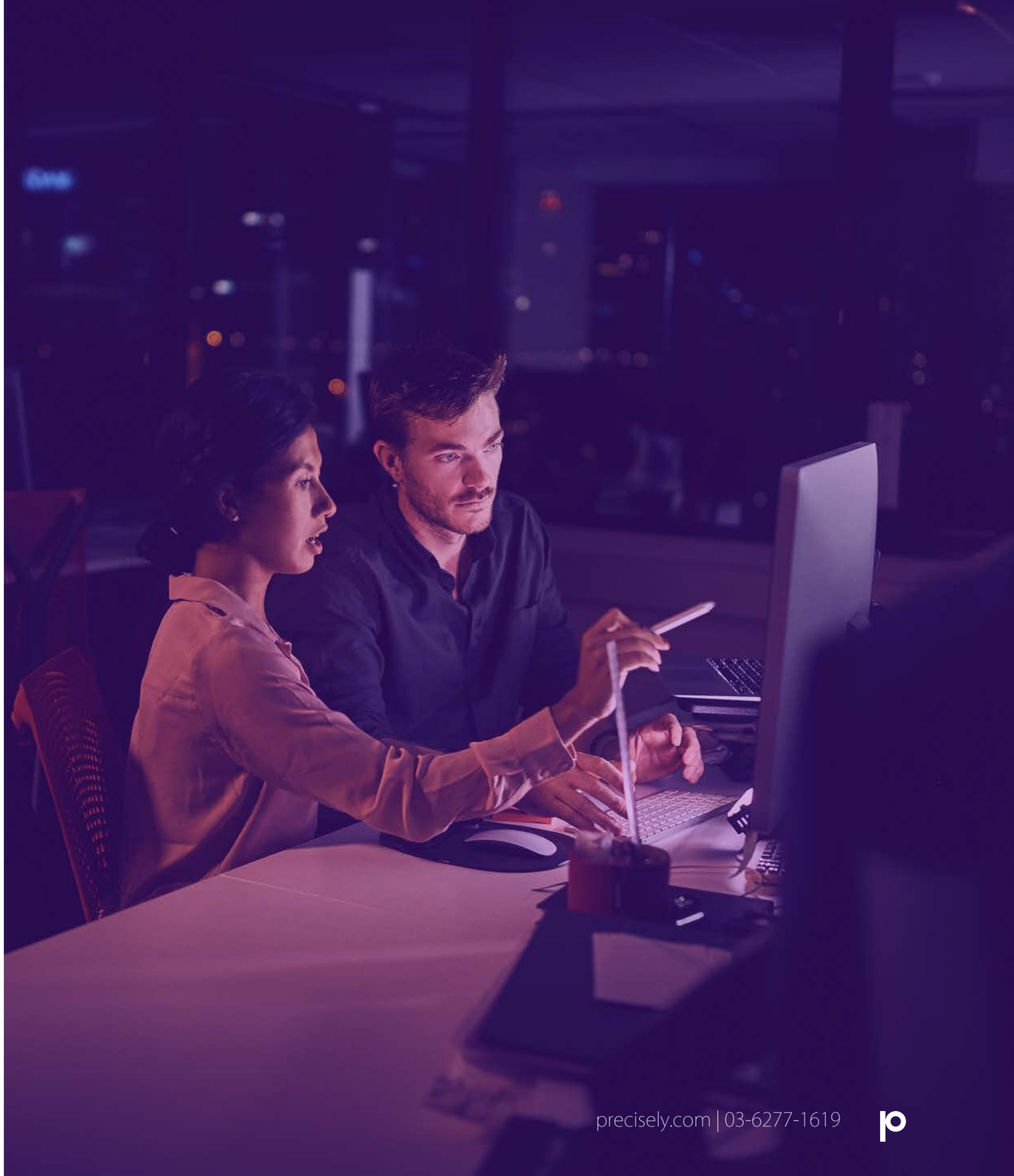
データソースの候補を見つける方法

まずはデータマーケットプレイスのサイトを見たり、Google で検索したりするのが、おすすめの方法です。プロバイダーのタイプ、自社の業界のプロバイダー、データのカテゴリ(「天気」、「農業」など)といった切り口で検索できます。さらに、データソースのさまざまな種類について理解を深めると、選定の判断に役立ちます。

ポイント:

以下の基準でデータプロバイダーやデータマーケットプレイスを評価しましょう。

- **データ製品の網羅性** — 必要なデータをすべて 1 か所で調達できるか。
- **サンプルデータセット** — データのサンプルを実際にダウンロードして評価できるか。
- **オンラインのソフトウェアツール** — 地図アプリなどを使い、オンラインでデータセットを確認できるか。
- **使いやすさ** — 検索機能の精度が高く、目的のデータを探しやすいサイト設計になっているか。



ステップ 04

候補を絞り込んで評価する

オープンソースデータを活用する場合でも、商用データプロバイダーから購入する場合でも、それぞれの候補を入念に評価する必要があります。このプロセスには非常に時間がかかります。ほとんどのユースケースでは、プロバイダーの異なる複数のデータセットが必要になるためです。

Precisely は、世界 130 社を超えるデータサプライヤーと提携しています。Precisely がサプライヤーに対して求めている、ユーザーにとって重要なデータの要件とは以下のとおりです。

- **データ品質。** データが正確 (値が正しく、ジオメトリがクリーンで、不正な重複データがないこと) で、完全 (網羅性が十分で、属性情報が完備されており、欠損値がないこと) で、最新 (データが新鮮で、適切に更新されること) であること。
- **データ構造。** 製品の構造 (ダウンロード場所、フォルダー構成、ファイルとファイル名、フィールド名など) が過去にダウンロードしたものと常に一貫していること。
- **構造変更の管理。** 何らかの変更を行う場合は、少なくとも 90 日前に通知すること。
- **ドキュメントとメタデータ。** フィールドのレイアウト、定義、製品メトリクス (ファイルサイズ、レコード件数など)、リリースノート、データの生成時期、ソース (一次、二次、三次データの区別) が明記された製品ドキュメントが整備されていること。
- **トラブル発生時の解決プロセス。** 大規模なデータセットを扱う際に想定されるトラブルの発生に備えて、問題を報告するための窓口や体制が確立されていること。トラブル発生時には、迅速に解決策を提示して実装すること。
- **更新スケジュールとロードマップ。** 更新スケジュールを一貫して公開しており、コンテンツの改善に向けた明確な製品ロードマップを示していること。

一般的に、生データに近ければ近いほど、この評価プロセスの重要性は高くなります。そのため、生データやオープンソースデータを利用する場合は、そのデータセットが安定して供給されるか、適切にメンテナンスされているか、品質基準を満たしているかを見極める必要があります。また、マーケットプレイスで調達したものなどの商用データ製品を購入する場合であっても、データプロバイダーがサプライヤーに対して上記のような基準を課していることを、自ら確認しておくべきです。

ポイント :

利用するデータソースが生データやオープンソースデータに近ければ近いほど、データセットが安定して供給されるか、適切にメンテナンスされているか、品質基準を満たしているかを見極めることが重要になります。

「多くの企業と同様に、当社もデータ信頼性の向上を優先課題として掲げており、データ品質などの重要な戦略の柱に注力しています」

– Fisher Funds、データ担当役員、Gladwin Mendez 氏 [2]

[2]「Data Integrity Trends 2021 (データ完全性のトレンド 2021 年版)」、Corinium Research、2021 年

ステップ 05

利用規約とコンプライアンスリスクを理解する

このステップは、評価プロセス中でもきわめて重要な部分であり、その重要性ゆえに一つの独立したステップとして解説します。サードパーティデータがユースケースに適合するかどうかは、そのデータに適用される利用規約によって決まります。これは、オープンソースデータの場合も同様です。利用規約は法的な拘束力を持ち、そのデータの利用について、何がどのような条件下で認められるかを定めたものです。

ポイント：

利用規約は、そのサードパーティデータの利用について、どのような条件下で、何が許可され、何が禁止されているのかを定めたものです。データに投資する前に、自社のユースケースが許可されていることを必ず確認しましょう。

「企業は、サードパーティソースのデータを活用して自社データを拡充しようとしています。しかし大半の企業は、それを大規模に一貫して行うのは難しいと述べています」[3]

[3]「Data Integrity Trends 2021 (データ完全性のトレンド 2021 年版)」、Corinium Research、2021 年

リスク要因

このステップでの確認を怠ると、取得したデータを意図した用途に使えないことが後から発覚するという事態になりかねません。さらに深刻なケースでは、データの利用後に問題が発覚し、法的な責任を問われることもあります。また、一次データを、そのデータソースの知的所有権を侵害する形で不正に利用してしまうというリスクもあります。このようなリスクを防ぐためにも、特に大企業においては、必ず法務部門と共に検討を進めることが望ましいと言えます。

コンプライアンス上の懸念事項

サードパーティデータを利用する場合は、コンプライアンスへの配慮が不可欠です。特に、顧客データや消費者データが関係する場合はなおさらです。データセットに個人を特定できる情報 (PII) が含まれるかどうか、また、含まれる場合は誰がその匿名化に責任を負うのかを確認しておく必要があります。さらに、そのデータが、Open Financial Exchange (OFX) のような業界標準、欧州の一般データ保護規則 (GDPR) や米国のカリフォルニア州消費者プライバシー法 (CCPA) などの政府規制に準拠した方法で収集されたものかどうかも確認しましょう。本人とコンテンツプロバイダーのいずれの同意も得ずにスクリーンスクレイピングで収集されたデータは、コンプライアンス面で特に大きなリスク要因となります。

precisely.com | 03-6277-1619

ステップ 06

データの取得方法を検討する

データ調達の最終ステップでは、データの取得方法には多くの選択肢があることを踏まえ、自社の運用やユースケースに最適な方法を見極めましょう。

プロバイダーがデータをダウンロード可能な形で提供する手段には、FTP サイト、クラウドストレージサイト、Web ページなどがあります。Snowflake Data Marketplace や AWS Data Exchange などのデータマーケットプレイスはデータの取得フローを効率化するように設計されており、Precisely をはじめとする複数のプロバイダーに 1 つの画面からアクセスして、データを取得できます。

別の検討事項は、どのファイル形式でデータを取得できるかということです。例えば、CSV ファイル、ASCII ファイル、拡張タブ形式、各種データベース (Microsoft Access、MySQL など) ファイル 、区切りテキストファイルなどがあります。多くの場合、ファイル形式は、データプロバイダーではなく、データのタイプによって決まります。

データ活用が進んでいる企業であれば、データ取得を効率化・自動化できるアプリケーションプログラミングインターフェース (API) やソフトウェア開発キット (SDK) を提供しているプロバイダーを探すことも検討すべきです。

ポイント :

購入予定のデータが、自社にとって利用しやすい方法と形式で取得できることを確認しましょう。



まとめ

世界はデータで溢れており、そのデータを見つけ、評価し、取得するための要となる経路がインターネットです。選択肢は多岐にわたり、無償のオープンソースデータや公共利用データのファイルを (多くの場合「利用者の自己責任」で) ダウンロードして利用する方法から、1 社または複数の実績ある商用プロバイダーからデータ購入する方法までさまざまです。また、複数の商用データプロバイダーとその製品を 1 つのデジタルプラットフォームに集約し、検索プロセスをシンプルにするデータマーケットプレイスも登場してきています。

Precisely を活用すれば、データ活用にかかる時間とコストを削減できます。一般に、企業のデータ専門家がデータの探索、前処理、管理に費やしている時間は、業務時間の約 80% に上ります。[4] データを業務で実際に活用すること、つまり、リスクや機会の分析モデルの実行、効率化、顧客体験やリテンション率の向上や、データ専門家を雇用した本来の目的である他のすべてのユースケースには、わずか 20% の時間しか使われていないのが現状です。Precisely のデータ製品は、Precisely Data Experience で簡単に見つけることができます。さらに、Precisely は、データ製品に組み込むデータのサプライヤーに対して事前に審査を実施しています。サプライヤーには、データ品質、鮮度、ドキュメントに加え、ガバナンス、プロビナンス、利用規約、コンプライアンスに至るまで、高い基準を求めています。貴社のデータ専門家が最小限の前処理を行うだけですぐに使えるデータをご用意していますので、データを活用して事業目標を達成することに専念していただけます。

市場には数多くのデータプロバイダーが存在します。信用できるデータに基づき、より確信を持って意思決定を行うために、ぜひ Precisely をお役立てください。

「得られた知見を信頼できなければ、それに基づいて行動を起こすことはできません。その知見が自分の直感と一致しない場合は、なおさらです」

– State Street、グローバルマーケット部門、データガバナンス担当マネージングディレクター、Dan Power 氏 [5]

[4]「The 80/20 Data Science Dilemma (データサイエンスにおける 80 対 20 のジレンマ)」、IDG/InfoWorld、2021 年。

[5]「Data Integrity Trends 2021 (データ完全性のトレンド 2021 年版)」、Corinium Research、2021 年

precisely

Precisely について

データ整合性のグローバルリーダーである Precisely は、お客様のデータの正確性、一貫性、そして文脈性を保証します。Precisely Data Integrity Suite をはじめとする当社のポートフォリオは、ソフトウェアだけでなくエンリッチ データやデータ戦略コンサルティングサービスまで網羅し、お客様が Agentic-Ready Data (AI エージェント対応のデータ - 適切に統合・ガバナンス・エンリッチされ、企業規模の AI や自動化、分析を可能とする最高品質のデータ) を実現するために必要な、すべてのものを提供します。Fortune 100 企業のうち 95 社を含む、100 ヵ国、12,000 社以上のお客様が Precisely の各種サービスを利用して、厳しい規制やミッションクリティカルな要件にも対応したデータ環境を実現しています。詳細は www.precisely.com をご覧ください。

www.precisely.com

©2022 Precisely. All rights reserved worldwide. 本文書に記載されるその他すべての企業名および製品名は、それぞれの企業の商標である可能性があります。