セルフサービス型データプレパレーションの手引き

～機密情報を守るために必要なデータガバナンスとは～

データマイニング、抽出、クレンジング、結合、ブレンド、マスキング――これらはすべて、データプレパレーションの作業です。
データを操作し様々なソースから抽出したデータの結合方法を考えたり、静的なレポートやPDF のセッションキーを調整したり、正確に報告するためにデータ形式を整えたりといった作業に、データ分析のプロフェッショナルでさえ、1 週間に何時間も割いています。
Altair ® Monarch ® などのセルフサービス型のデータプレパレーションツールなら、データのインポート、変形、エクスポートを高速化および効率化できるだけでなく、自動処理機能によって作業時間のさらなる短縮も可能です。
この「セルフサービス型データプレパレーションの手引き」では、データプレパレーション作業を簡略化し、データの作成、利用、処理、変形を簡単に行う方法をご紹介します。

第 1 章

セルフサービス型データプレパレーションに欠かせない、見逃されがちな 5 つの要素

1 多構造化 / ストリーミングデータの活用

重要な情報は、往々にして多構造なドキュメントやデータソースに格納されており、キーを入力しなければ利用することはできません。
生産レポートや他社文書に眠るデータは分析することで大きな価値を持つにもかかわらず、データを取り出すための手段がないということは、少なくありません。

2 データマスキング

データ量が急増する一方、データ保護規制への適合と機密データ盗難防止を実現し、機密データを保護するための対策も拡大しなければなりません。
データマスキングは、あらゆる産業と分野で情報漏洩を防ぐための解決策として登場しました。

データマスキングとは、元のデータセットの形式は変えずにデータ値だけを変え、新たなデータを生成する技術です。
これを活用することにより、分析やトレーニング、検証といった用途においてはデータを隠し、権限を持つユーザーにのみ元データを表示することが可能になります。

3 プロセスの自動化

データプレパレーションは、自動化できる反復作業を常に探しながら行うべきです。
また、データに一貫性を持たせ生産性を高めるために、作業方法は組織内で共有しましょう。

4 ガバナンスによるリスク低減

セルフサービスを使用する最大の目的は、ビジネスユーザーが簡単に、かつ素早くデータを分析できるようにするためです。
しかし、それに伴い IT 部門の介入が減れば、自ずとリスクが増えてしまいます。
ガバナンスの強化は必要ですが、データプレパレーションで扱うデータの大半が CSV の抽出データや PDF 形式のレポート、社外からのデータであることを踏まえたうえで、ストレスフリーなガバナンスを実現しなければなりません。
企業向けソリューションには、処理済みのデータセット、再利用可能なモデル、分析結果の可視化、ダッシュボードのすべてを適切に保管および管理し、アクセス権を制御できることが求められます。

5 使いやすさ

スクリプト言語や複雑なフローダイアグラムを新たに習得することなく、自分でデータを直接操作し、定義されている 100 の関数、集計フィールド、データマスキングを実行できます。
VLOOKUP、マクロ、ピボットテーブルも不要になれば、さらに快適に作業ができるようになるでしょう。

第 2 章

フットワークは軽く、ガバナンスは固く

セルフサービス型データプレパレーションツールは、データディスカバリーや高度なデータ分析に欠かせないツールとして多くの人に利用されるようになりました。

セルフサービス型データプレパレーションとコーポレートガバナンスの両立

ほとんどの企業は、企業向けアプリケーションやデータウェアハウスなどの管理対象システムに保管されたデータを、厳密な体制のもと管理しています。
セルフサービス型分析ソリューションを導入する最大のメリットの1つは、様々なソースからのデータを迅速に結合および分析できる点にありますが、このアプローチは、ガバナンス的には非常に悩ましい問題をはらんでいます。というのも、分析に使うデータの約半分は、取引システムから抽出した CSV やテキストデータ、個人のスプレッドシート、他社からのレポート、半構造化データなど、IT 部門の管理が及ばないソースが起源なのです。
そのため、バージョン管理、データ漏洩、突合、監査などに関する問題を避けては通れません。企業向けデータプレパレーションプラットフォームには、ビジネスユーザーが求めるスピードと機敏さを損なうことなく、こうしたガバナンス上のリスクにスムーズに対処できることが求められます。

管轄外のデータを管理するには

管轄外のデータを管理するには、その基盤となるコンテンツリポジトリを構築しなければなりません。その認識は企業の間で広く共有されていますが、では、どのようなガバナンス機能が必要なのでしょうか。

- **データリテンション**
  一貫性を保つには最低でもドキュメントのバージョン管理が必須ですが、規制や業務上の要件を満たすには、ソースデータやドキュメントを永続化し、保管することも必要になります。
- **データマスキング**
  データ漏洩の原因の大半は従業員によるものです。
  データディスカバリーツールは情報を集約して共有できる強力な手段ではありますが、元データが保護されていないことが多いだけでなく、社会保障番号などの個人を特定できる情報や、医療記録などのプライバシーに関わるデータ、あるいは営業秘密が含まれていることもしばしばです。
  データ漏洩に関しては業界や各国政府にいていくつもの規則が定められており、順守を怠った場合、漏洩インシデント1件につき多額の費用が生じ、個人や組織の法的責任が問われる事態にも発展しかねません。
  一方で、権限を持つユーザーがマスキングを自由に解除できる必要もあります。
- **データ系列**
  元のソースデータをリポジトリに保管するなら、完全なデータ系列を確保し、ソースドキュメントのセルにまで遡れるようにしておく必要があります。
  この機能は、データの監査や突合に不可欠です。
- **データキュレーション**
  使用頻度の高いデータソースや自動化した準備ルーチンは、ユーザーのロールに合わせて共有すべきです。
  これにより、重要な意思決定の基となるデータに信頼性と一貫性を持たせることができます。
- **ロールベースのアクセス制御**
  処理済みのデータセットについては、ユーザーの権限に合わせて適切な部分のみが提供されるよう、アクセス権を設定する必要があります。
- **監査**
  現場で求められるのは、監査ログと監査レポートの包括的な機能を備えたシステムです。

第 3 章

セルフサービス型データプレパレーションの価値を高める自動化機能

ご存知の通り、様々なソースからデータを抽出、結合し、整理整頓するのは容易ではありません。それどころか、適切な情報を適切なタイミングで、かつ信頼できる方法で収集しなければ、データディスカバリーや高度な分析ツールの真の価値を引き出すことはできません。
準備作業を迅速化できるセルフサービス型ツールがあれば、データの下準備に費やす時間を減らしその分の時間を分析に割けるのに…とは誰しもが考えたことがあるのではないでしょうか。
優れた分析結果を提供してこそ組織に貢献できるのであって、だらだらとデータ準備に時間をかけることに意味はありません。
このように、準備作業は、データアナリストの共通した悩みのひとつです。

データプレパレーション作業を自動化すれば、時間とコストを大幅に削減することができます。

Altair のセルフサービス型データプレパレーションツール

セルフサービスの自動化

第一の特徴は、IT 部門の手を借りることなく、データアナリストや一般ユーザーが自らデータプレパレーションを行える点にあります。
どのような作業も GUI ベースの再利用可能な“ワークスペース”に保存でき、次回以降の同じ作業をすべて自動化できます。
シンプルでわかりやすいプロセスデザイナーでは、プロセスの開始日時と準備完了後の送信先を指定できます。

信頼できる確かなデータ

わたしたちが準備するデータは、ビジネスにおける日々の重要な意思決定の根拠になります。
そのため、準備プロセスの透明性の確保が必須になりますが、それをワークスペースで共有し実現します。
また、認可済みのデータセットやプロセスを作成しておき、他ユーザーの作業を厳密に管理したいときはそれらの使用を義務付けることも可能です。
ダッシュボードを見たユーザーから、外れ値の原因を尋ねられた場合には、
データ系列の末端にある元データにまで遡り、たとえば顧客から届いた請求書のハイライトされた対象行を確認してもらうことができます。

ウェアハウスもアップデート可能

処理済みのデータセットは、データウェアハウスやデータマートなどのシステムに送信することも可能です。
新しいデータは、様々なシステムに対応した業界標準のデータベースドライバーを使って既存のシステムに簡単に追加できます。

処理プロセスの共有

使用頻度の高いデータソースや自動化した準備作業は、簡単に共有および検索できます。
作成済みのものを共有し活用できるため、各人が作業手順を一から組み立てる必要はありません。
信頼できるプロセスを改良したり、処理済みのデータセットに自分の作業用データを追加したりすることも可能です。

ユーザーごとに適切なデータのみを表示

処理済みのデータセットは、ユーザーまたは部署単位でアクセスを制限できます。
自動化されたプロセスを 1 つ実行するだけで、毎回適切なユーザーに適切なデータのみを届けます。
指定したユーザーに対して特定のデータフィールドをマスキングしたり、伏せ字にしたりすることも可能です。
規制の厳しい産業、たとえば医療機関では、個人を特定できるような患者情報は隠しつつ、患者名に偽名をあてて患者単位でデータをロールアップしたい場合があります。
あるいは、管理者権限が割り当てられた病院関係者以外には社会保障番号を完全に伏せるといったことも、必要になるかもしれません。
Altair のツールなら、機密情報の細かな管理を完全に自動化することができ情報の漏洩を防ぎます。

リアルタイム更新

新しいソースデータが入手可能になると、データプレパレーションプロセスが自動的に呼び出されます。
たとえば、顧客の請求書内のデータを使用している場合、いくつかの場所を指定し(CMS、共有ディレクトリ、自分のメールボックスなど)、そこに新たな請求書が格納されたら自動的に検知するように設定することが可能です。
こうすることで、情報を常に最新の状態に保つことができます。

関連のデータ分析ソリューション

Altair は、個々のスキルレベルにかかわらず簡単に分析アプリケーションを構築し、既存のアプリケーションに分析機能を持たせて意思決定の質を高めるソリューションを提供しています。

データ分析と機械学習

Altair の機械学習・予測分析ソリューションは、データを迅速に可視化し、獲得した知見を組織内で素早く共有できるのが特長です。
コーディングは一切不要なため、データサイエンティストもビジネスアナリストも、モデル構築に時間を浪費することなく、データ分析に時間と労力を注ぐことができます。

ビッグデータ分析

Apache Spark フレームワーク上に構築された Altair の高度な分析ソフトウェアは、比類ない分析機能とデータ処理能力により、ビッグデータの活用やインサイト発掘の妨げになる課題克服をサポートします。
直感的でインタラクティブな操作性を備えており、数十億個のデータポイントの変形と分析も数分で完了するため、インサイトを素早く引き出し、確かな情報に基づいて決断を下すことができます。

データ可視化とストリーム処理

Altair のソリューションはビジネスユーザー、つまりデータ分析業務の直接的な担当者のために設計されたツールです。コードを書くことなく、データストリーム分析とデータ可視化のためのアプリケーションを構築、変更、導入することができます。
Panopticon はリアルタイムのストリームフィードや時系列データベースに対応しており、接続できないデータソースは皆無と言っても過言ではありません。複雑なデータストリーム処理プログラムを作成し、視覚的なユーザーインターフェースをデザインすることも可能です。時々刻々と変化する大量のデータを様々な観点で分析し、確かな根拠に基づいて懸命な判断を下せるようになります。

Altair について（Nasdaq : ALTR）

Altair は、製品開発、ハイパフォーマンスコンピューティング（HPC）、およびデータアナリティクスの分野において、ソフトウェアやクラウドソリューションを提供するグローバル企業です。多種多様な業界におけるお客様が、持続可能な未来を創造しコネクティッドな世界において力を発揮するためのテクノロジーを提供します。詳細については、www.altairjp.co.jp をご覧ください。

詳細はこちら

Web サイト：
www.altairjp.co.jp/data-prep

お問い合わせ：
www.altairjp.co.jp/data-analytics-contact-us

試用版の申請：
Altair のセルフサービス型データプレパレーションツール Altair Monarch を、30 日間無償でお試しいただけます。
www.altairjp.co.jp/monarch-free-trial