ホワイトペーパー

データ品質 / ガバナンス

# AI 対応データの 6 つの原則

## AI 向けの信頼できるデータ基盤の確立

# 目次

# 概要

- 本書では、人工知能（AI）を使用したデータ活用における 6 つの原則について解説します。
- 6 つの原則とは、データの「多様性」「タイムリー性」「正確性」「安全性」「発見可能性」「マシンにおける容易な活用性」の確保です。
- 本書では、AI Trust Score についても解説します。AI Trust Score で、自社のデータが 6 原則をどの程度遵守しているのかを評価することができます。

## はじめに

人工知能（AI）は、医療・製造・カスタマーサービスなどの業界を大幅に改善し、顧客と従業員の両方に、より質の高いエクスペリエンスをもたらすことが期待されています。データ利用者は、機械学習（ML）などの AI テクノロジーで数学的予測を実行し、インサイトを引き出して意思決定を改善しています。さらに、生成 AI などの新たな AI テクノロジーは、驚くほどリアルなコンテンツを生成し、ほぼすべてのビジネスの生産性を高める可能性を秘めています。

Gartner 社は、2025年までに 90% のグローバル企業が生成 AI を労働力として導入。2026年までに、80% 以上の企業が生成 AI 対応のアプリケーションを本番環境に導入すると予測しています。[1]

しかし、AI でビジネス成果を加速するには、質の高いデータが不可欠です。本書では、AI 対応データを確保するための 6 つの原則について解説します。

[1]「Analysts to Discuss Generative AI Trends and Technologies」Gartner 社 2023年 10月

# AI 対応データの 6 つの原則

データを放り込むだけで AI 戦略を成功に導くことはできませんが、多くのデータ利用者は、データを放り込むだけで済ませています。この方法は、最初の数件の AI プロジェクトでは成果を挙げているように思えるかもしれませんが、プロジェクトが進行していくと、データサイエンティストの多くの時間がデータの修正や準備に奪われることになります。

さらに、AI で利用されるデータは、優れたアプリケーションに対応できるよう、高品質なデータを適切に準備する必要があります。手作業でデータのクリーニングと強化を行ってデータの正確性と完全性を確保し、マシンがデータを容易に理解できるよう整理するには、多くの時間が必要です。また、多くの場合、自動学習向けに情報が持つ意味を補強し、AI がより効率的にタスクを実行できるよう定義やラベルなどの追加情報が必要です。

だからこそ、下流における AI プロセス向けのデータの準備が早い方が、より多くのメリットを得ることができるのです。食料品をそのまま調理担当者に渡すのではなく、前もって下準備をした食料品を渡せば、労力と時間を削減して迅速に料理を提供することができます。AI 対応データの準備と活用にも同じことが言えます。下図は、適切なデータの「準備」と AI 対応データを確保するために重要な 6 つの原則を定義しています。

次ページ以降では、6 つの原則について順番に詳しく解説します。

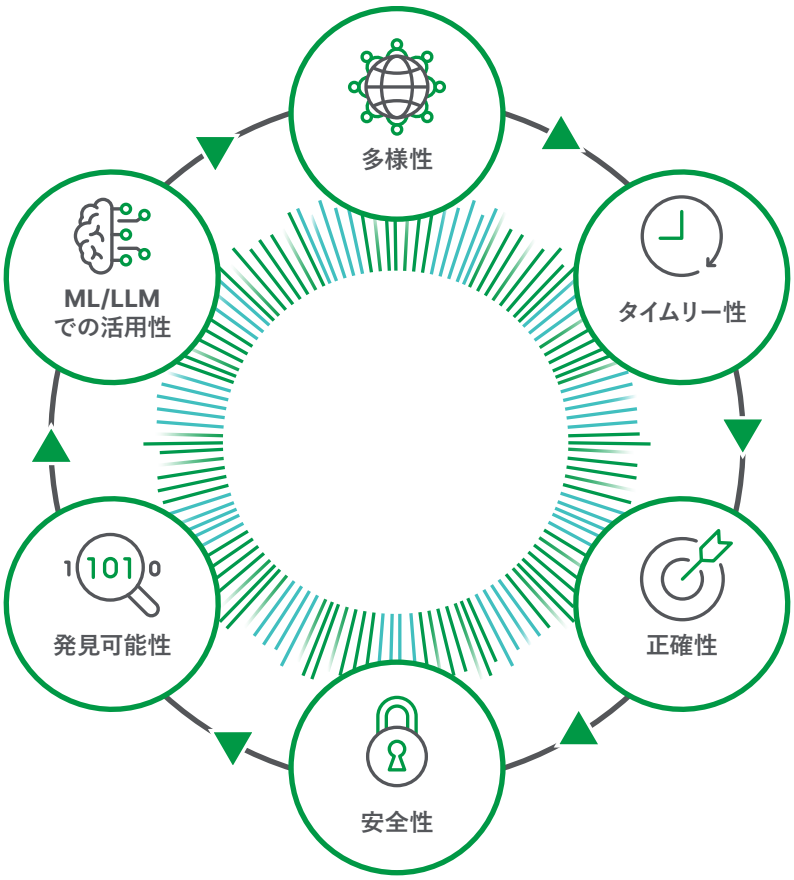


*図 1. AI 対応データの 6 つの原則*

# 1 データの多様性

AI システム（機械学習またはアルゴリズム）に対する偏見は、AI アプリケーションが社会的不平等などの人間の偏見を反映した結果を生成した場合に生じます。これは、アルゴリズムの開発プロセスに偏見的な仮定が含まれている場合、より一般的である学習用データに偏見がある場合に発生する可能性があります。たとえば、信用スコアのアルゴリズムが範囲の狭い財務属性を継続して使用している場合、ローン申請を却下する可能性があります。

こうした偏見を回避するために、第 1 原則では、AI モデルに多種多様なデータを提供することに重点を置いています。これにより、データの多様性を高めて偏見を軽減し、AI アプリケーションが不公正な判断を行う可能性を低くすることができます。

サイロ化された範囲の狭いデータセットに基づいて AI モデルを構築するのではなく、問題領域に関連するさまざまなパターン・見解・バリエーション・シナリオなど、広範なデータソースから多様なデータを取得してください。このようなデータは適切に構造化されており、クラウドまたはオンプレミスに保存されている可能性があります。メインフレーム / データベース / SAP システム / SaaS アプリケーション上にデータが保存されている場合もあります。一方、ソースデータが構造化されておらず、ファイルまたはドキュメントとして、社内のドライブ上に保存されている場合もあります。

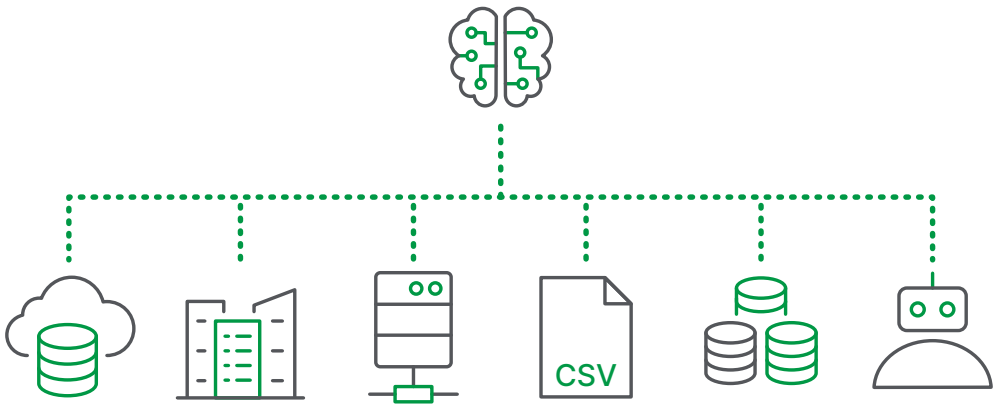


*図 2. データの多様性*

**機械学習や生成 AI アプリケーションにデータを統合するには、さまざまな形式の多様なデータを取得する必要があります。**

# 2 データのタイムリー性

多様なデータの活用は、機械学習や生成 AI アプリケーションの強化に不可欠なのは明白ですが、データの鮮度も非常に重要です。昨日の状況に基づいた天気予報が今日の旅行計画には使えないのと同様に、古い情報で学習された AI モデルは、不正確または的外れの結果を生成する可能性があります。最新のデータを活用することで、AI モデルは最新のトレンドを把握し、状況の変化に適応した最適な結果を提供できるようになります。そのため、重要な AI 対応データの第 2 の原則は、「データのタイムリー性」です。

AI 戦略向けのタイムリーなデータを確保するには、低遅延のリアルタイムのデータパイプラインを構築して展開することが重要です。**変更データキャプチャ（CDC）**は、リレーショナルデータベースシステムからタイムリーなデータを配信するのに多用されています。また、**ストリームキャプチャ**は、低遅延処理を必要とする IoT デバイスからのデータ取得に使用されています。データが取り込まれると、ターゲットリポジトリが更新され、**ほぼリアルタイムで継続的に変更が適用されます。**これにより、可能な限り最新のデータを取得することができます。

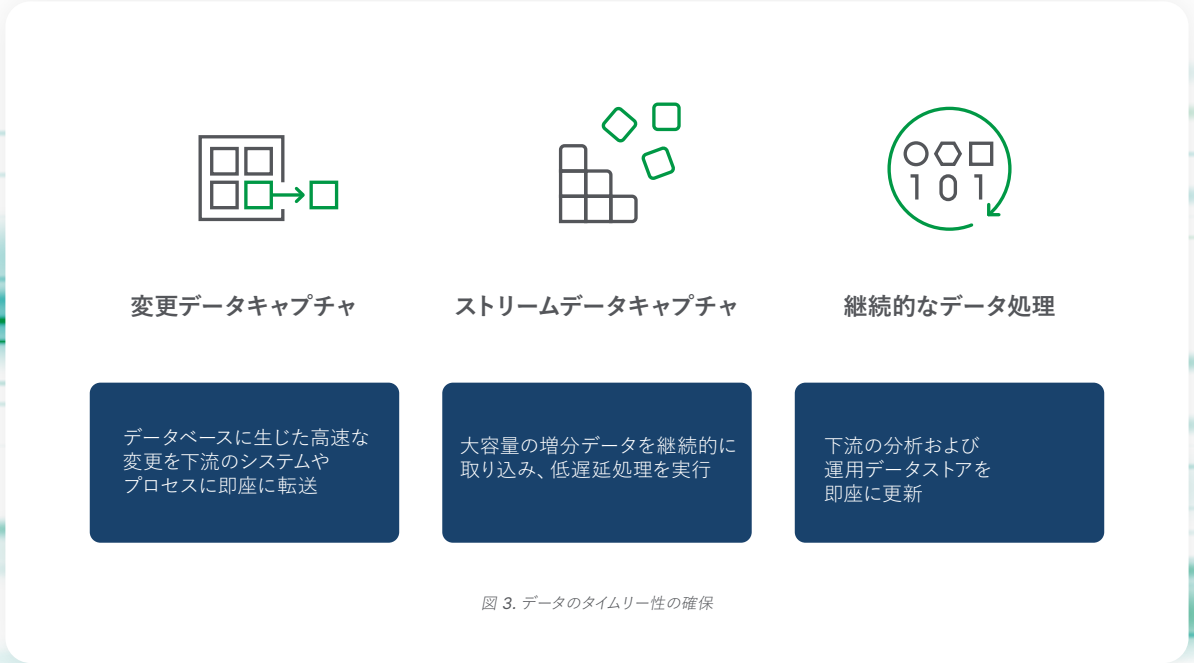


*図 3. データのタイムリー性の確保*

**データのタイムリー性を確保することで、より正確な情報に基づいた予測が可能になります。**

# 3 データの正確性

機械学習や生成 AI 戦略の成功は、データの正確性という重要な要素に左右されます。AI モデルは、スポンジのように膨大な情報を取得して学習し、タスクを実行するからです。スポンジが汚水を吸収するように、不正確な情報で学習した AI モデルは、偏った生成物や無意味なコンテンツを生成します。最終的には、AI システムの機能不全につながります。そのため、重要な AI 対応データの第 3 の原則は、「データの正確性」です。これは、信頼できる AI アプリケーションの構築に欠かせない基本原則です。

データの正確性には、3 つの側面があります。1 つ目は、ソースデータの特性・完全性・分布・冗長性・形式を理解する**ソースデータのプロファイリング**です。一般的に探索的データ分析（EDA）とも呼ばれています。

2 つ目は、データ品質ルールの構築と展開、有効性を継続的に監視して、**戦略的改善を運用化**することです。データの重複を排除してデータ結合をサポートするデータスチュワードの関与が必要な場合があります。マシンが推奨するデータ品質の提案を取り入れたプロセスの自動化と高速化に、AI を利用することもできます。

3 つ目は、データエンジニアやデータサイエンティスト向けのツールを使用して、データ系統と影響分析を可能にすることです。データ変更による潜在的な影響を明らかにし、データの出所を追跡することで、AI モデルが使用する不測のデータ変更を防ぐことができます。


*図 4. データの正確性を確保する手順*

**高品質で正確なデータは、モデルが関連するパターンと関係を特定します。これにより、正確な意思決定・生成・予測が可能になります。**

# 4 データの安全性

AI システムでは、個人を特定できる情報・財務記録・独自のビジネス情報などの機密データが使用されることが多いため、こうしたデータの使用には責任が伴います。データを保護せずに AI アプリケーションを使用することは、金庫のドアを開けたままにしておくようなものです。悪意のある人が機密情報を盗んだり、学習用のデータを改ざんして偏見的な結果を生成したり、生成 AI システム全体を破壊する可能性があります。データの安全性は、プライバシーを保護してモデルの完全性を維持し、責任を持って強力な AI アプリケーションを開発するのに極めて重要です。そのため、重要な AI 対応データの第 4 の原則は、「データの安全性」です。

データの安全性を手作業で確保するのは、ほぼ不可能です。次の 3 つの方法で、大規模かつ自動化することができます。1 つ目は、**データの分類**です。次の段階に送るデータを検出・分類・ラベル付けします。2 つ目は、**データの保護**です。マスキング・トークン化・暗号化などのポリシーを定義し、データを難読化します。3 つ目は、**データセキュリティ**です。ユーザーのデータへのアクセスを管理して記述するポリシーを定義します。3 つの概念は、以下のように連携することで機能します。まず、プライバシー層を定義し、データにセキュリティタグ（機密・社外秘・制限付き）を付けます。次に、保護ポリシーを適用して制限されたデータをマスキングします。最後に、アクセス制御ポリシーでアクセス権を制限します。


*図 5. データの安全性の中核要因*

**3 つの方法は、データを保護して AI システム全体の信頼性を高め、評価価値を維持するのに不可欠です。**

# 5 データの発見可能性

ここまでの原則は、主に適切なデータを正しい形式で、適切な人やシステム、AI アプリケーションに迅速に提供することに重点を置いていますが、データを蓄積するだけでは不十分です。AI 対応データは、システム内で検出可能で、即座にアクセスできる必要があります。図書館のすべての書籍に鍵がかけられていたら、書籍から知識を獲得して活用することはできません。発見可能なデータは、機械学習と生成 AI の真の可能性を解放します。そして、機械学習と生成 AI のワークロードが学習・適応し、革新的な成果を生み出すのに必要な情報の発見を可能にします。そのため、重要な AI 対応データの第 5 の原則は、「データの発見可能性」です。

データの発見可能性の中核は、間違いなくメタデータに対する適切な対応です。AI データセットに関連付けられた技術的なメタデータだけでなく、ビジネスメタデータとセマンティックタイピングも定義する必要があります。**セマンティックタイピング**は、自動化されたシステムに追加の意味を提供します。追加されたビジネス用語は、人間の理解をサポートする追加の状況認識を提供します。データセット内の技術項目にビジネス用語をマッピングした**用語集**を作成し、概念の共通理解を確保することをお勧めします。AI サポートによる拡張も可能です。ドキュメントを自動生成し、用語集からビジネスセマンティクスを追加することができます。最終的に、すべてのメタデータがインデックス化され、**メタデータカタログ**からデータを発見できるようになります。


*図 6. データの発見可能性の中核要因*

**この方法は、現在の AI タスクに不可欠なデータを直接発見して適用することができ、実用性を確保します。**

# 6 マシンにおけるデータの容易な活用性

前述のとおり、機械学習や生成 AI アプリケーションは強力なツールですが、その潜在能力はデータ活用の容易な活用性に左右されます。人間は手書きのメモを解読したり、複雑なExcel ファイルを扱うことができますが、こうしたテクノロジーでは、情報を特定の形式で表示する必要があります。好き嫌いの多い子供が与えられた食事を摂らずに空腹になってしまうのと同様に、データが機械学習テストや大規模言語モデル（LLM）アプリケーションに適していなければ、AI 戦略は成功しません。データを容易にAI システムで利用できるようにすれば、その潜在能力を引き出すことができます。情報をスムーズに取り込み、クリエイティブな成果物を生み出す優れたアクションに変換することができます。そのため、重要な AI 対応データの 6 つ目の原則は、「マシンにおけるデータの容易な活用性」です。

## 機械学習におけるデータの容易な活用性

データの変換は、機械学習におけるデータの容易な活用性をサポートします。線形回帰などのアルゴリズムが注目されていますが、学習用のデータの品質と形式も同じく重要です。さらに、データをクリーニング・整理し、機械学習モデルで利用できるようになると、大きな成果を挙げることができます。適切なデータの準備は、効果的なモデル学習を可能にし、正確な予測や信頼性の高い生成物を実現します。最終的に、機械学習プロジェクト全体を成功へ導きます。

ただし、学習用データの形式は、基盤となる機械学習のインフラに大きく依存します。従来の機械学習システムはディスクベースです。データサイエンティストの業務のほとんどが、膨大なファイルを処理する最善策と手動のコーディング手順の確立に重点を置いています。最近は、データレイクハウスベースの機械学習システムに、データベースのような機能ストアを使用するようになっています。
データサイエンティストの業務は、第一級言語として SQL に移行しています。そのため、整形された質の高い表形式のデータ構造は、機械学習システムにとって最も使いやすく便利なデータ形式です。

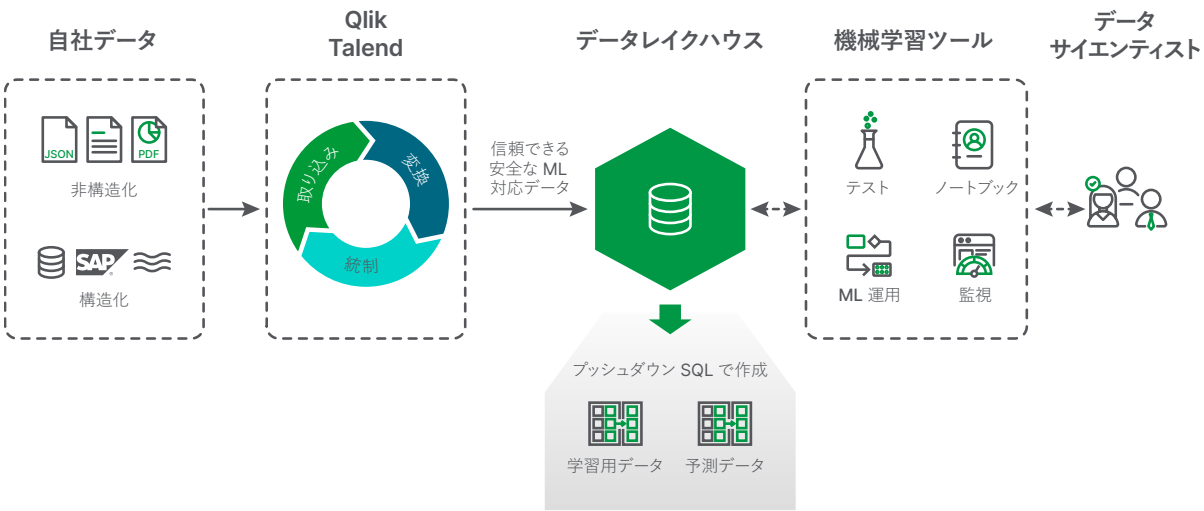


*図 7. 機械学習におけるデータの容易な活用性の確保*

## 生成 AI におけるデータの容易な活用性の確保

膨大なテキストデータで事前学習されている OpenAI 社の GPT-4、Anthropic 社の Claude、Google AI の LaMDA や Gemini などの大規模言語モデルは、生成 AI の中核の存在です。OpenAI 社の GPT-3 モデルは、3,000 億トークンを超える約 45 TB のデータで学習されたモデルだと推定されています。このような膨大な情報があっても、自社のデータにアクセスできない大規模言語モデルでは、自社のビジネスに関する具体的な質問に答えることができません。解決策は、こうしたモデルを自社の情報で強化し、より正確で関連性が高く、信頼性の高い AI アプリケーションを開発することです。

自社のデータを大規模言語モデルベースのアプリケーションに統合する、検索拡張生成（RAG）と呼ばれるテクノロジーがあります。通常、プレゼンテーション・メールアーカイブ・テキスト文書・PDF・トランスクリプトなど、構造化されていないファイルベースのソースから取得したテキスト情報を使用します。テキスト情報を扱いやすい単位に分割し、埋め込みと呼ばれるプロセスで、大規模言語モデルが使用する数値表現に変換します。これらの埋め込みは、Chroma / Pinecone / Weviate などのベクトルデータベースに保存されます。PostgreSQL / Redis / SingleStoreDB など、多くの従来型データベースの事業者も、ベクトルをサポートしています。最近、Databricks / Snowflake / Google BigQuery などのクラウドプラットフォームにも、ベクトルのサポートが追加されました。

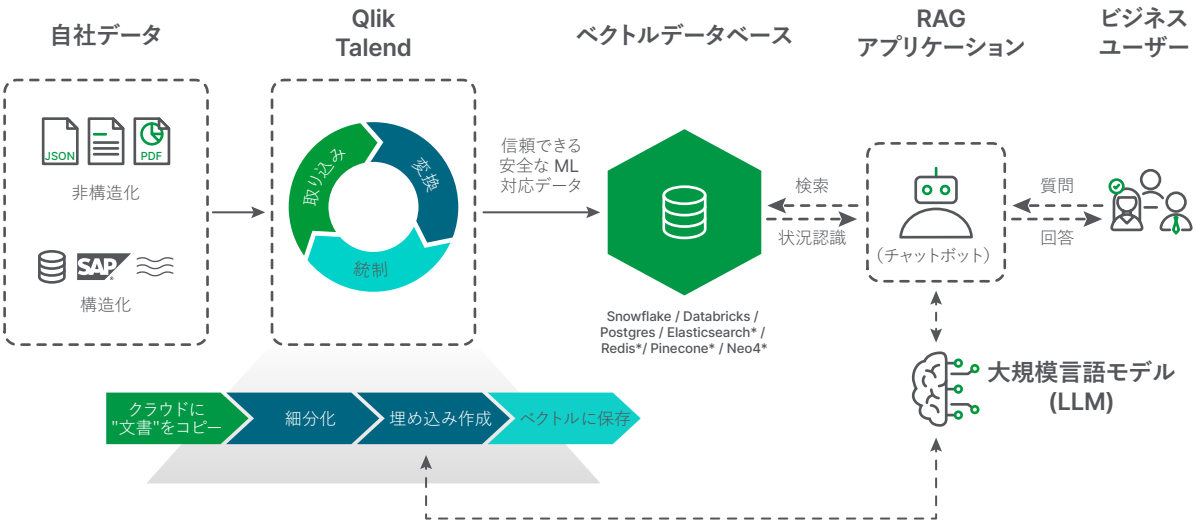


*図 8. 生成 AI におけるデータの容易な活用性の確保*

**Qlik は、構造化・非構造化データを問わず、生成 AI / RAG / 大規模言語モデルベースのアプリケーションで、質の高いデータを即座に利用できるようにします。**

# AI Trust Score

データの準備と AI 対応に関する 6 つの基本原則を定義してきましたが、いくつかの疑問が残ります。「この原則を体系化して、日々の業務に容易に応用できるのか?」「AI 対応に値するデータなのかを迅速に判断するにはどうすればよいか?」グローバルレベルでわかりやすい指標として、Qlik の AI Trust Score で測定することをお勧めします。

各原則にレベルを割り当て、各値を集計して総合点を算出します。これは、AI 対応に値するデータなのかを迅速かつ容易に評価できる信頼性の高い方法です。絶え間ない自社データの変化に対応するため、信頼スコアには定期的な見直しと頻繁な再調整が行われます。これにより、データの準備に関する傾向を追跡することができます。

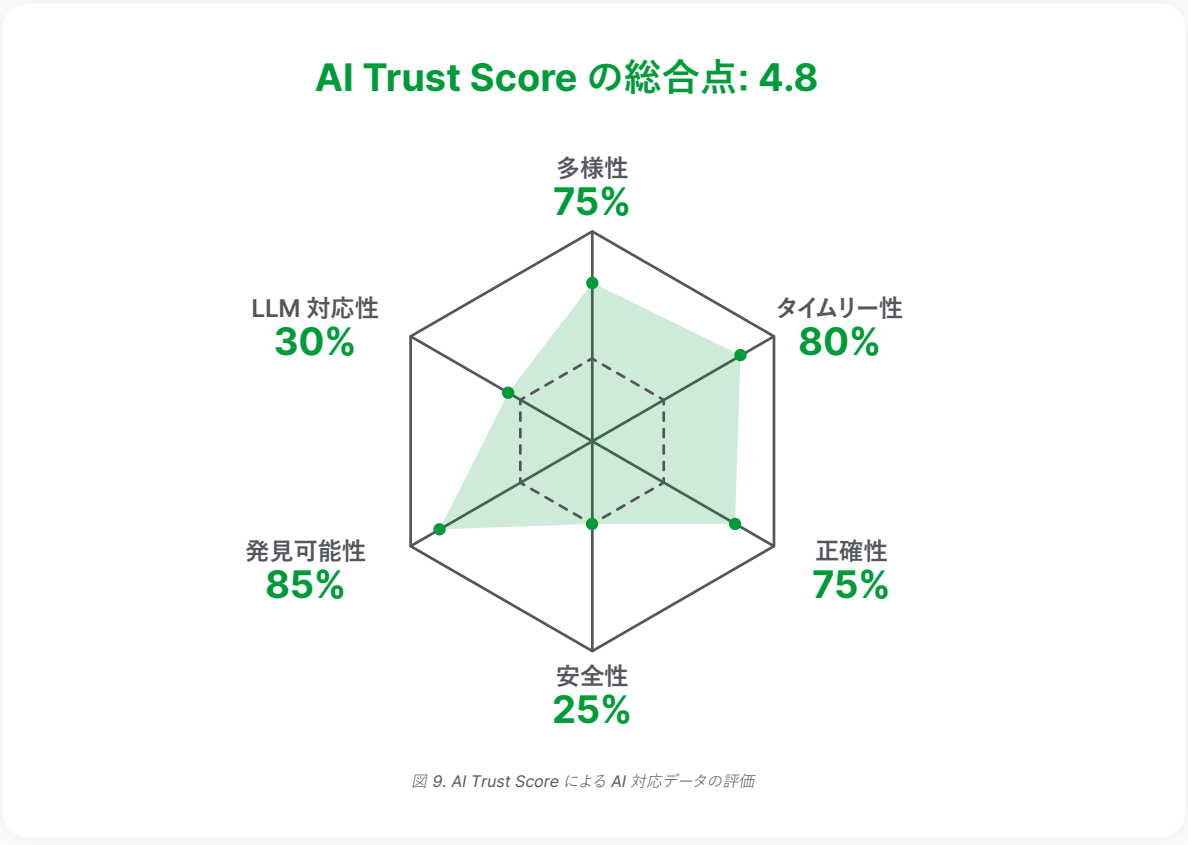


*図 9. AI Trust Score による AI 対応データの評価*

**AI Trust Score は、複数の指標を集計して判断しやすい単一の点数を算出します。**

# Qlik Talend の AI 向けデータ基盤

より適切な意思決定、業務の効率化、ビジネス革新を推進する質の高いリアルタイムのデータの必要性は、かつてないほど高まっています。成功している企業は、データウェアハウス・データレイク・その他のエンタープライズデータプラットフォームに信頼できるデータを効率よく提供するために、市場をリードする Qlik Talend のデータ統合・品質ソリューションを求めています。Qlik の包括的かつ業界最高水準のサービスは、自動化されたパイプライン、優れたデータ変換、信頼できる高品質のデータセットを提供します。これにより、データ専門家が切望する俊敏性と、企業が期待するガバナンスやコンプライアンスが可能になります。

Qlik Talend は、豊富なインサイトによる分析を実現するデータウェアハウスやデータレイクの構築、ビジネスを効率化する運用データインフラストラクチャの刷新、マルチクラウドデータを活用する AI 戦略をサポートします。

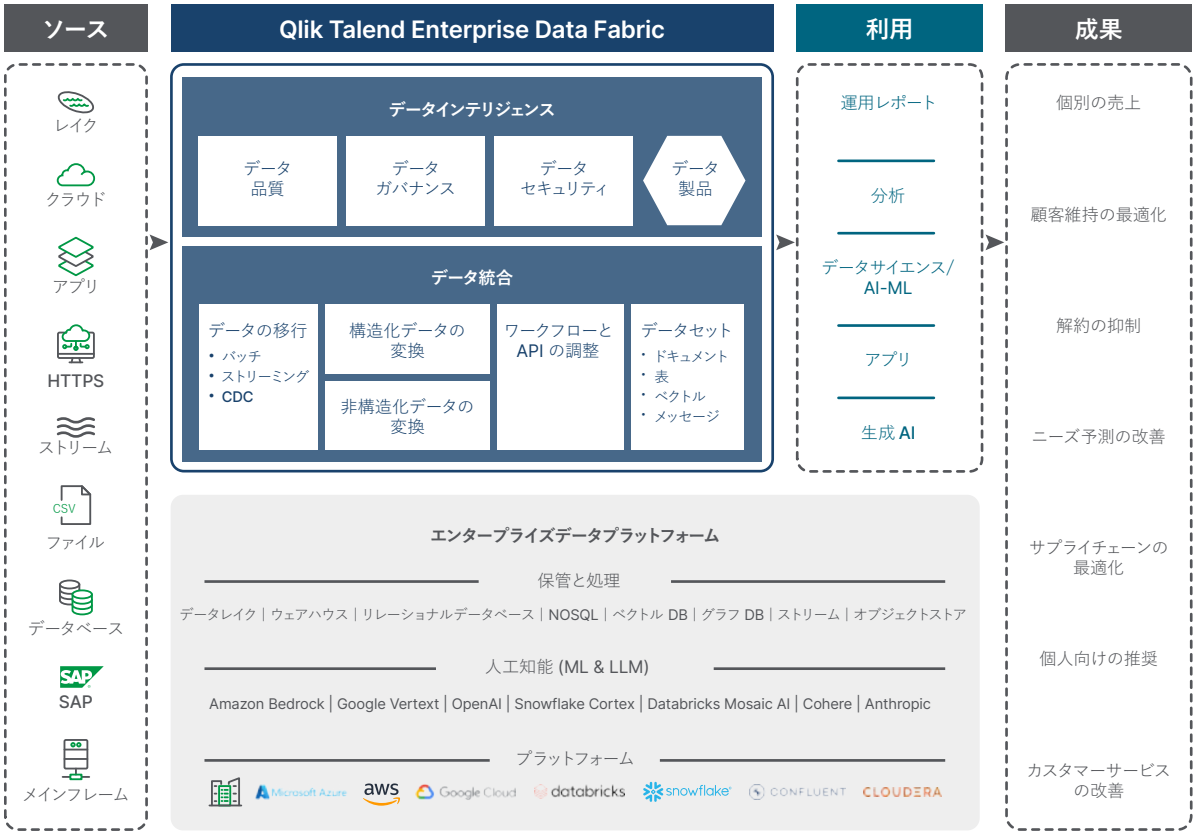


*図 10. Qlik Talend の AI / 分析向けエンタープライズデータファブリック*

## まとめ

機械学習が変革をもたらし、生成 AI が劇的に成長していく中、AI 対応データの準備も AI 活用の成功に不可欠な要素です。本書では、信頼できる強固なデータ基盤を確立し、企業が AI の真の可能性を引き出すのに重要な 6 つの原則について解説しました。

AI の潜在能力を
引き出すには?

詳細を見る

## Qlik について

Qlik は、複雑なデータ状況を実用的なインサイトに変換し、戦略的なビジネス成果を促進します。世界 40,000 社以上の顧客にサービスを提供している Qlik の製品ポートフォリオは、最先端かつエンタープライズ水準の AI / 機械学習と広範なデータ品質を基盤としています。また、優れたデータ統合およびデータ統制、多様なデータソースに対応する包括的なソリューションを提供します。Qlik の直感的でリアルタイムの分析は、隠れたパターンの発見や複雑なビジネス課題の解決、新たなビジネスチャンスの獲得を支援します。さらに、実用的で高度な拡張性を備えた Qlik の AI / 機械学習ツールで、適切で迅速な意思決定を可能にします。Qlik は顧客の戦略的パートナーとして、プラットフォームに依存しないテクノロジーと専門知識で、顧客の競争力を高めます。

**qlik.com**



24-DQ-0001-01-WHM