ORACLE

7 AIを成功に導く：データ基盤に関するチェックポイント

これらの重要なポイントに対処することで、リーダーはAIの導入を簡素化し、ROIを最大化し、ビジネス・イノベーションを推進することができます。

目次

AI戦略を支える「使えるデータ」になっていますか？

Jeffrey Erickson
シニア・ライター

量子物理学を説明し、フランス語の小説を要約できる大規模言語モデル（LLM）が、企業独自の運用データとナレッジ・ベースに関する深い専門知識を得られることを想像してみてください。これまでアナリストがキューに入れて順番に処理をしていたデータの検索、統合、文脈付けというタスクが、ビジネスユーザーとAIエージェントの自然言語による会話によって行えるようになり、そこからインサイトを得てアクションにまでつなげることが可能になりました。これによりITリソースが確保できるようになり、組織はよりデータドリブンになります。

魅力的な話ですよね。世界の最も野心的かつ人材が豊富で、資金調達に余裕のあるテクノロジー企業と、それらと競合するスタートアップ企業が、奇想いあうようにして生成AIの強力な機能を手に入れようとしていることも不思議ではありません。こうした生成AIモデルが大注目されている最中でも、組織全体の人々がデータにアクセスし、利用方法を根本的に変えることで、データの価値をどれだけ高めうるかという点は、いくら強調してもし過ぎることはありません。

これらすべてを支える基盤は、生成AIのニーズに対応したインフラストラクチャです。このインフラが備えるべき特長は、カスタマイズされたツール、基盤となるLLM、テクノロジー、コンピュート・システムを適切に組み合わせ、高速レスポンスを提供し、異常検知およびオブジェクト識別などのさまざまなユースケースをサポートします。成功は、複雑なAI運用を処理するために十分なパワーを備えているかにかかっていることがよくあります。まず、超高パフォーマンスとマイクロ秒のレイテンシを実現できるクラスタ・ネットワークで相互接続されたGPUについて考えてみましょう。また、ベクトル・データベースについても考えましょう。企業内のさまざまなナレッジ・ベースとLLMを組み合わせた検索拡張生成（RAG）を活用するためのAIエージェントの仕組みも重要です。そして、使いやすいAIエージェントのインターフェースも必要です。

> 生成AIモデルがデータの価値を何倍にも高めうる
> という点は、いくら強調してもしすぎることはありません。

エージェントAIとは

エージェントAIとは、情報を理解し、それに反応するだけでなく、
能動的に目的を追求する機能を持つ人工知能を指します。

エージェントAIの主な特徴

プロアクティブな振る舞い
AIは、外部からの指示に単純に反応するのではなく、自分から気づき自ら行動を起こすことができます。

適応力
AIは経験やフィードバックから学習し、目標達成に向けて適応することができます。

目的志向
AIは達成すべきと設定された目標に到達するためのステップを計画し、自律的に行動することができます。

自律性
AIはパラメータの範囲内で、独自に意思決定を行い、行動を起こすことができます。

生成AIは大きな可能性を秘めています。そして、その実現のためには多額のインフラストラクチャ投資が必要です。

AIサポートに向けたデータセンターの拡張に関する目を見張るような統計数字をご覧になったことがある方も多いのではないのでしょうか。そうした支出の多くは、優れたAIエクスペリエンスの提供を目指すハイパースケール・クラウド・プロバイダーによるものであり、先進的な企業は、来るべき事態に対応するためにデータ・インフラストラクチャの準備を進めています。

では、今すぐに準備するためにできることは何でしょうか。以下は、多くのITアーキテクトがAIへの期待を実現すべく取り組むにあたり、現在そして未来に向けて見直すべき7つのチェックポイントです。

1 これまで誰も考えつかなかったものも含め、生成AIのさまざまなユースケースから十分な価値を引き出す準備はできていますか？

高度にパーソナライズされたマーケティング、優れたカスタマーサービスのチャットボット、生産性を向上させるコーディング・ヘルパーなど、AIプロジェクトのいくつか例をもってインフラストラクチャのサポート・コストの正当性を証明することがよくあります。生成AIの活用可能なユースケースは日々拡大しています。率直に言えば、“こういうことができたらいいな”と思いついたことは、ほぼ実現可能になりつつあるのです。これは主に、高度なLLMがERP、HCM、SCMなどのエンタープライズ・アプリケーションに組み込まれ、ロボット工学、医療業界、ゲノム工学、航空宇宙工学などの研究を多く伴う分野で広く使用されていることによります。つまり、ほぼあらゆる場所で使われています。

重要なことは、LLMを補完して、AIを自社のビジネスの主要なエキスパートにすることです。履歴データ、運用データ、財務データ、ドキュメント・データ・ライブラリへのアクセスを提供することで、生成AIは企業全体の多くの人のために、様々な形で戦力として機能させることができます。

具体的にどういったことでしょうか。たとえばそれは、エグゼクティブがデータと「対話」し、その場でドリルダウンして新しいビジネス・インテリジェンスを導き出すようなケースです。あるいは、AIエージェントがソフトウェア開発や研究開発のブレインストーミングのパートナーになり、プロトタイプやモックアップを瞬時に作成するケースもありえます。また、見込み顧客の開拓、コミュニケーション、プロセスの細かい作業がAIによってこれまでにないレベルの自動化が可能になるため、営業担当者がより多くのリードを管理できるようにもなるでしょう。

これらはほんの一例にすぎません。製品、財務、人事、法務の各チームがアイデアを出し合うブレーンストーミング・セッションを行うことで、創造的な生成AIユースケースとデータ・インフラストラクチャ計画の確固たる出発点が得られる可能性があります。

同業他社によるアイデアをさらにお望みの場合は、遺伝子解析やスポーツ中継など、さまざまな現実のユースケースをご覧ください。

ファインチューニングとRAG：AIに自社データを「理解」させる2つの方法

ファインチューニングとRAGはどちらも、生成AIモデルがより状況に適し、組織に合わせてカスタマイズされた回答を提供できるよう支援します。その違いは次のとおりです。

ファインチューニングとは、CohereのCommandやMetaのLlama 3のような汎用モデルを、特定のドメインに特化した小規模のデータセットを使って追加トレーニングを行うことを指します。チューニングにより、モデルはそのドメイン特有の用語や文脈を学習し、コーディング、金融、医療のような特定タスクにとおいてより高い精度での応答が可能になります。マイナス面は、コストがかかることと、プロセスをサポートするためにGPUベスのインフラが必要なことです。

RAGは、汎用のAIモデルが特定の組織に役立つアウトプットを提供できるようにするアーキテクチャ上の仕組みです。クエリへの回答を作成する際に、あらかじめ厳選された関連性の高い内部データをLLMに提供することで、回答に業務の文脈や社内知識の要素を加えることができます。この結果、LLMの流暢な言語能力と社内データに関するインサイトを組み合わせたAIシステムが、より状況に適した実用的な応答が可能になります。ファインチューニングと異なり、RAGではLLM自体を変更する必要はありません。LLMはまた、回答後には提供したデータが保持されないため、データ漏洩のリスクを軽減できます。

RAGとファインチューニングのどちらを使用すべきでしょうか。

[要件に合わせた選び方を学ぶ]

2 非構造化および半構造化データをすべてAIで活用可能な状態にできていますか？

非構造化データおよび半構造化データをAIで使用するための処理は、データの収集と取り込み、ストレージ、処理（クレンジング、機能抽出、正規化、セグメンテーション、場合によってはその他のステップ）を含む複数のステップからなるプロセスです。こうして初めて、データレイクの画像、ドキュメント、音声および動画ファイルが、ファインチューニング、ベクトル化、RAGに向け準備されます。

ファインチューニング用のデータ: 特定のタスクのために生成AIモデルをファインチューニングするには、分野固有の新しいデータを表示することが必要な場合があります。たとえば、生成AIモデルを医療レポートを作成のためにチューニングする場合、モデルが適切な専門用語とコンテキストを学習できるように、高品質で関連性の高い医療レポートのデータセットが必要になります。

継続的なAIの出力のためのデータ: AI対応のデータ管理システムは、JSONドキュメントやリレーショナル・データ、半構造化データなど、幅広いデータ・ストアへのアクセスをLLMに提供する必要があります。ベクトル埋め込みをデータに効率的に適用し、埋め込みをデータベースにベクトルとして格納し、データベースをクエリして効率的なベクトル検索と自然言語処理を可能にする必要があります。そのためには、継続的なデータ収集と、AIモデルの出力の正確性と関連性の向上を支援するRAGアーキテクチャも備えた統合パイプラインが必要な場合があります。

これらのシステムが導入されると、どのような状況になるでしょうか。たとえばセールス・チームは、すべての通話音声を保存し、AIに顧客感情を分析させながら要約を作成することができます。営業担当者がメモを取りながら顧客対応するために、解読不能な走り書きをする必要もなくなります。顧客の次回の注文に対する見積もり依頼の詳細を忘れることもありません。細部へのこだわりは同じですが、簡単により正確かつ効率的になります。

成功のカギ:統合されたマルチモーダル・データベースにより、専門的なシステム全体にわたりデータを移動、再同期、再保護することなく、これらすべてのプロセスを実現します。このような統合データ・モデルは、IT部門が異なるサイロを統合する必要なく、AIベクトルを含むデータ型の横断的検索を可能にするため、業務を桁違いに高速化することができます。AIモデルは、一見無関係に見えるものの重要な情報間の微妙な関係を考慮するため、これにより、さらに豊かな結果を導き出すことができます。

3 機密データを保護しつつ、AIに活用できるガバナンスとアクセス制御の一元管理戦略は整っていますか？

生成AIツールに提供されるデータのセキュリティとプライバシーの維持は、モデルのファインチューニングとRAG、および日常的な使用の両方において必須のインフラストラクチャ設計の原則です。

生成AIプラットフォーム向けのデータは、ニーズに応じて匿名化、暗号化、マスキングが可能です。AIの出力生成に使用するためのデータも、依頼者の役割を考慮し、多要素識別プロセスを使用することで、厳重に管理する必要があります。組織によっては、専用インフラストラクチャ、またさらにはオンプレミスのデータセンターに格納されたAIモデルのコピーを使用して推論を実行し、プライベート・データをさらに保護することにする場合もあります。

また、テクノロジープロバイダーや標準化団体にサポートを求めることも可能です。
たとえば、オラクルは米国国立標準技術研究所（NIST）の米国AI安全研究所コンソーシアム(AISIC)[1]と連携する280以上の組織の1社として名を連ねています。AISICの目標は、「科学的根拠に基づき、実証に裏付けされたAIの測定と政策に関するガイドラインと基準を開発し、全世界にわたるAIの安全性の基礎を築く」ことです。メンバーは、コミュニティがAIをセキュアに開発し導入できるよう支援するためのガイドライン、ツール、方法、プロトコル、ベストプラクティスの開発に取り組みます。

AIを中心とした同様の取り組みは他にもあります。

[1] 米国AI安全研究所コンソーシアム

生成AIアーキテクチャ：コンバージド・データベースと複数の単一用途データベースの比較

マルチモーダル・データベースを使用すると、通常、生成AIを活用するための幅広いETLやデータ移動は必要ありません。むしろ、単一のデータ・プラットフォームを使用してさまざまなAIプロジェクトやワークロードをサポートすることで、データ移行に伴うリスクを最小限に抑えるとともに、複雑さと管理上のオーバーヘッドを大幅に削減することができます。マルチモーダル・データベースは、データが1つのフォーマットやモデルである必要があることではなく、単にアプリケーションの種類ごとにデータ・アーキテクチャが選択可能であることを示しています。

4 部門横断的に生成AI戦略を推進できるコラボレーション体制と標準化ルールはありますか？

適切な専門知識があれば、既存のモデルをあなたの組織独自の特注AIソリューションに変えることができます。人はAIの活用を望んでいます。そのために必要なツールを提供することで、生成AIのワークフローが単発のシャドーITプロジェクトではなく、企業の戦略の一部になる可能性が大幅に高まります。

取り組みを最大化する方法としては、部門を超えて透明性を提供し、一貫性を促進するセンター・オブ・エクセレンス（CoE）が挙げられます。CoEは、生成AIに関連するベストプラクティス、ツール、技術を統合し、それらを活用してビジネス上の問題を解決する取り組みをリードします。

自社にAI CoEを構築する方法

この取り組みのコア・テクノロジーは、モデルのファインチューニングと導入の技術的ステップにおけるコラボレーションを推奨するデータ・サイエンス・プラットフォームです。ハイパースケール・クラウド・プロバイダーは、データ・サイエンティストがCohere、Meta、その他のサプライヤーから提供される強力な基盤モデルを活用および拡張できる生成AIプラットフォームを備えています。これらのプラットフォームは、データ・サイエンティストとITチームに、サービス全体でモデル、データセット、データ・ラベルを保存、交換、再利用できる可能性のある場所を提供します。

ハイパースケーラはまた、モデル・カタログを提供するほか、ファインチューニングされたLLMにそれらを実行するためのデータとコンピュート・インフラストラクチャを提供します。今では各部門が、他部門での生成AIの成功を踏まえ、それを活用・展開できる環境が整っています。

CoEと主要なデータ・サイエンス・プラットフォームの両方を使用することで、継続的な学習と改善の文化を促進することができます。オラクルを含むプロバイダーは、Oracle Cloud Infrastructure (OCI) Generative AIなどのフルマネージド・サービスを提供し、LLMをさまざまなユースケースにシームレスに統合しています。また、ベースモデルを自社のデータセットでファインチューニングしてカスタム・モデルを作成することもできます。

5 クラウド・プロバイダーの長所を組み合わせて、AIのためのデータ可用性を最大限に高められていますか？

組織ではGoogle GeminiやMicrosoftのCopilotを具体的に検討している可能性があります。あるいは、オープンソース・モデルをAWSでホストしたり、Cohereの基盤モデルをOCIでホストすることも選択肢にあるでしょう。現在、オラクルとその他のハイパースケーラーとの間で結ばれた新しい契約により、オンプレミスまたはクラウドでOracle Databaseテクノロジーを中心に企業データ・ガバナンス戦略を構築している企業にとって、可能性が大幅に広がりました。オラクルとAzure、Google Cloud、AWSの革新的なマルチクラウド関係により、OCIだけでなく、それぞれのプロバイダーのデータセンター内でOracle Databaseサービスを使用できるため、低いレイテンシと管理オーバーヘッドでこれらのモデルのいずれをも活用することが可能です。

データへの簡単でセキュアなアクセスを損なうことなく、必要な場所で必要なAIモデルを実行することができれば、まだ構想に至っていない、こうした生成AIのユースケースで成功を収める可能性が高まります。チームは、特定のタスクに最適なサービスを利用することができ、セキュリティと耐障害性の高さの維持とコスト削減を実現できます。

6 ファインチューニングや推論に必要なインフラシステムを調達、運用、コスト管理する計画はありますか？

生成AIのファインチューニングと導入は、コンピューティング集約型の取り組みです。それぞれのやりとりには、何十億ものパラメータを含むLLMが、情報の操作と分析を行うための専門的なコンピューティング・システムを利用した、膨大なデータセットに対する複雑な計算が含まれることがあります。モデルが複雑になるほど、データを処理して適切な出力を提供するために、より多くのコンピューティングが必要になります。モデルのファインチューニングと推論の両方で、高品質な結果を実現しつつコンピュート・コストを抑えるには計画が必要です。

多くの組織では、社内でシステムを構築するためにコストをかけたり専門知識を習得するよりも、主要なクラウド・プロバイダーが提供しているような一般的なデータ・サイエンス・プラットフォームを利用して、クリーンでタスクに関連性の高いデータセットを作成しています。

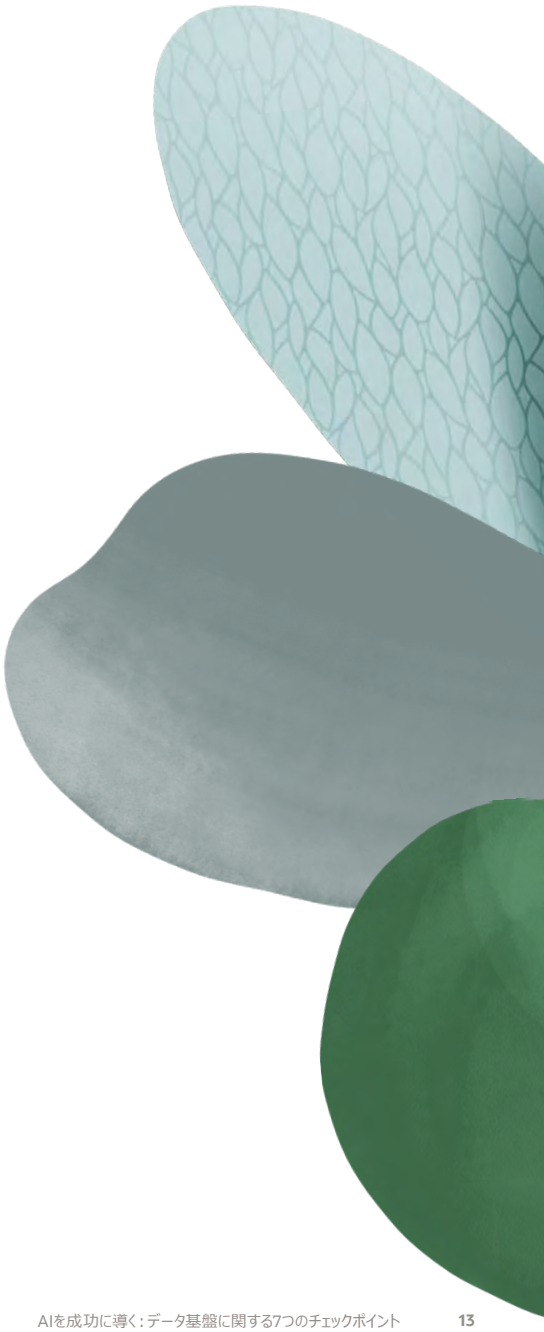
次に、クラウド・プロバイダーのAIインフラストラクチャ・サービスについて理解しましょう。多くの場合、一般的なオープンソース・ツールおよび生成AI基盤モデルや他の必要なテクノロジーにすぐにアクセスしたり、統合することができます。

オープンソース・コミュニティとハイパースケーラは、生成AIのコスト削減を支援することに熱心に取り組んでいることにも留意が必要です。クラウドでは、AIワークロードの高速化に必要な強力なGPUやクラスタ・ネットワーキングに加え、自動スケーリング、スポット・インスタンス、リザーブド・インスタンス、最適化されたジョブ・スケジューリング、効率的なワークロード分散、マルチテナント、モデル最適化などの機能が利用できます。その時々のワークロードの正確な需要にコンピュートパワーを適合させる技術により、リソースの利用率を最大化し、企業は使用した分だけ料金を支払うことができます。

ファインチューニング自体も「微調整」が進んでいます

T-Fewとして知られる方法は、従来のファインチューニングの方法と比較して、高い精度を維持しながら、トレーニング期間を短縮し、必要な計算リソースを削減することができます。

最終的には、モデルとインフラストラクチャが最高の効率で稼働していることを証明できるようにする必要があります。クラウド・プロバイダーの生成AIサービス、またはTensorFlow ServingやTorchServeなどの効率的なモデル・サービング・フレームワークを通じて、推論プロセスをモニターおよび最適化できます。



7 AI計画に対するエグゼクティブの支援は得られていますか？実行に必要なチームやリーダーシップとの連携、リソースの確保はできていますか？

AI戦略は、多くの場合、現在のエンタープライズ・システムから開始します。CRM、HCM、ERP、その他のコア・アプリケーションのプロバイダーは、生成AI機能とAIエージェントを一般的なワークフローに組み込んでいますサプライヤーからの請求書の処理速度向上と手作業の大幅削減を支援するインテリジェント文書認識、レポートと分析をより深く理解するための生成ナラティブ、リスクやパフォーマンス分析などの領域におけるさまざまなインテリジェントな自動化をお考えください。これらにより道は開けますが、自社独自のワークフローにAIを導入するには、より広範な戦略が必要になります。これには、先に説明したセンター・オブ・エクセレンス(CoE)が含まれる可能性があります。

生成AIサービスとAIエージェントを業務に導入するには、エグゼクティブ、部門長、IT、法務、コンプライアンス・チーム、そしてもちろん新しいテクノロジーを使用する従業員など、組織全体にわたる賛同とリソースが必要です。それぞれの事業部門のステークホルダーと協力して、生産性の向上と競争上の優位性に焦点を当てたビジネス・ケースを策定することで、協調的なアプローチを策定することができます。

開始するAIワークフローが決まったら、より広範なITチームとデータサイエンス・チームにエンゲージメントして、どのようなアーキテクチャが必要になるかを理解します。次のような点を確認すべきです。LLMはどこに配置しますか。その費用はどこの負担になりますか。データ・フローはどのようになりますか。新しいベクトル・データベースやRAGアーキテクチャは必要でしょうか。特定の部門やタスクに合わせたLLMのファインチューニングは必要ですか。データ・ストレージと処理能力はどこから確保しますか。ネットワーク設計とその他のアーキテクチャ上の意思決定により、継続的なAI推論に必要な低レイテンシと高スループットは実現できるでしょうか。

スケジュール、マイルストーン、リソース要件、およびデータ・プライバシー、セキュリティ、コンプライアンス、コストに関する考慮事項などのリスクの概要を記載した詳細な提案書を作成しましょう。信頼できるクラウド・サービス・プロバイダーがいかに役立つかも検討します。

> いかなるAIモデルも、クリーンで整然としたデータの安定したフローなしには効果的なものにならないことをご留意ください。

AIの導入は自社データ戦略全体の延長線上にあると考えることが重要です。経営層・エグゼクティブの賛同を得るためには、AI計画が組織全体のビジネス戦略およびデータガバナンス計画と密接に連携していることを確認し、データ移行や管理負荷の増大といった不要な複雑さが生じないようにする必要があります。

オラクルのソリューション

データ・インフラストラクチャで、世界最先端の生成AIモデルの高速かつコスト効果の高いファインチューニングと推論を実行する必要がある場合、オラクルがお役に立ちます。

Oracle Cloud Infrastructure

Oracle AI インフラストラクチャ

OCIは、包括的なAIサービスと最先端の生成AIイノベーションを、クラス最高のAIインフラストラクチャ上で提供しますデフォルトのモデルを選択することも、オープンソースまたは自社開発のLLMからモデルを選択できる生成AIサービスを活用して、モデルをファインチューニングし、自社のエンタープライズ・データで補強することもできます。また、業界をリードするデータベース・システムにアクセスすることも可能です。

詳細

Oracle Database 23ai

Oracle Database 23aiは、トランザクションと分析のためのマルチモーダル・データベースであり、JSON、リレーショナル、グラフ、空間データに加え、データベース内の機械学習とともに、組み込みのベクトル・データベースを提供します。また、OCIとマルチクラウド・パートナーのデータセンターでOCIサービスとしてのみ利用可能な、独自に最適化されたプラットフォーム上でベクトル処理とOracle Databaseを実行することもできます。

詳細

OCI Data Science および生成AIサービス

OCI Data Scienceは、お気に入りのオープンソース・フレームワークを使用して、データ・サイエンティストが機械学習（ML）モデルを共同で構築、トレーニング、導入、管理したり、データベース内のMLを活用したりできるクラウド・サービスです。OCIの生成AIサービスは、選択した言語モデルをさまざまなユースケースに統合するためのフルマネージド・サービスです。

詳細

OCI上で、より高速でコスト効果の高いファインチューニング、推論、バッチ処理を実現できます。OCIは、大規模なAIワークロードを、過剰なコンピュート・コストをかけずに高速に実行するためのスケールとパフォーマンスを提供します。

オラクルが支援できること

独自のマルチクラウド関係により、データ共有が容易になるよう支援します。

全世界の組織がOCI上で稼働するOracle Databaseサービスを使用して、Oracle Exadata Database Service上でアプリケーションを迅速に構築および実行し、Oracle Autonomous Database機能を活用しています。また、マルチクラウド関係により、他のハイパースケール・クラウドでアプリケーションを実行している組織も、単一の運用環境のシンプルさ、セキュリティ、低レイテンシにより、Oracle Databaseと同じメリットを得ることができます。

AWSではExadata Database ServiceとAutonomous DatabaseをAmazon Bedrockなどのサービスと組み合わせて、AzureではExadata Database ServiceとAutonomous DatabaseをAzure OpenAIなどのサービスと組み合わせて、Google CloudではExadata Database ServiceとAutonomous DatabaseをGoogle CloudのVertex AIやGemini基盤モデルなどと組み合わせて使用して、アプリケーションを迅速に構築および実行します。

詳細

オラクルへのお問い合わせ

050-3615-0035にお電話いただくか、oracle.com/jp にアクセスしてください

北米以外の地域では、oracle.com/jp/contact で最寄りのオフィスをお探しください