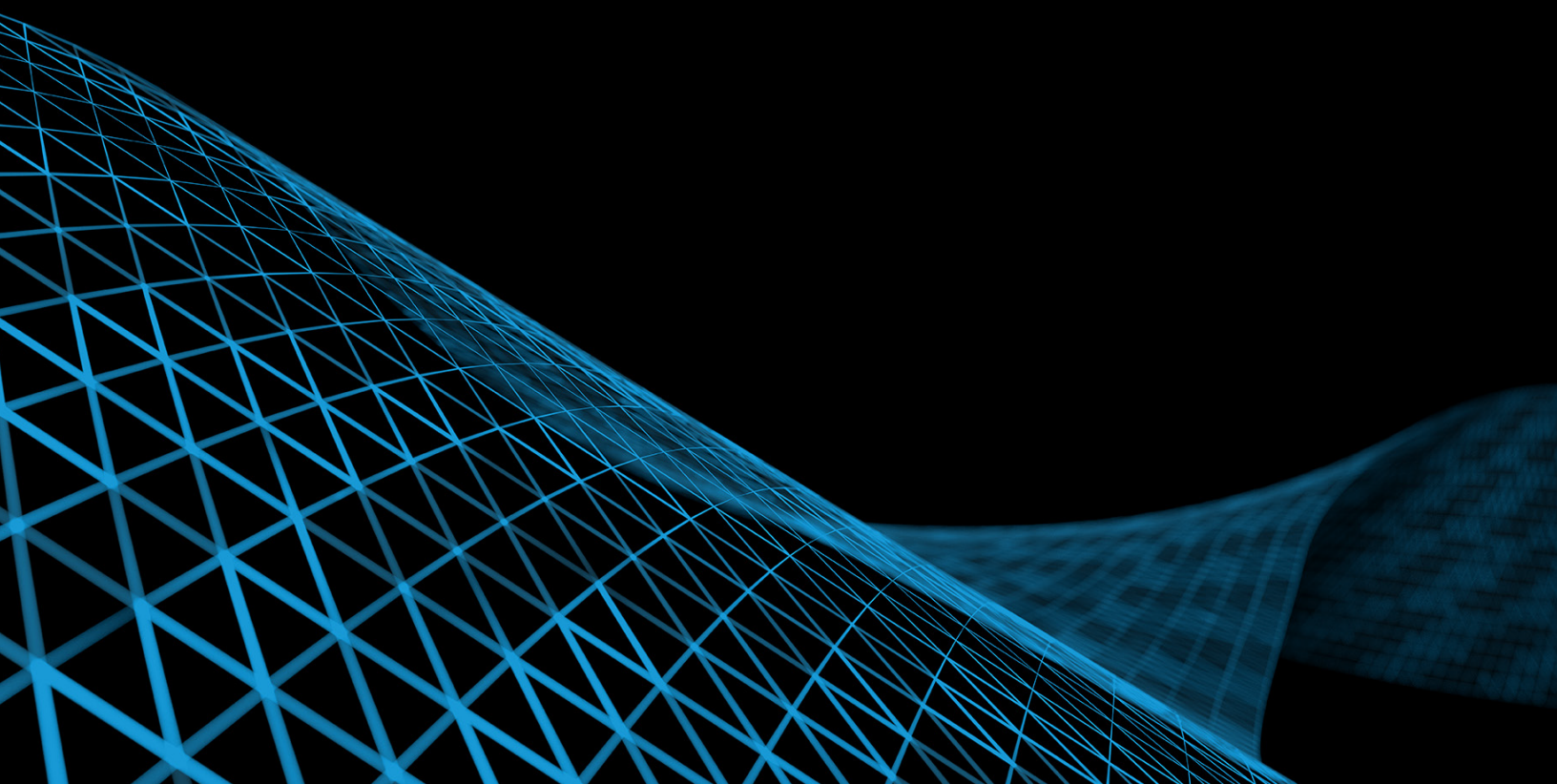
TechTarget | Custom Media
E-Book

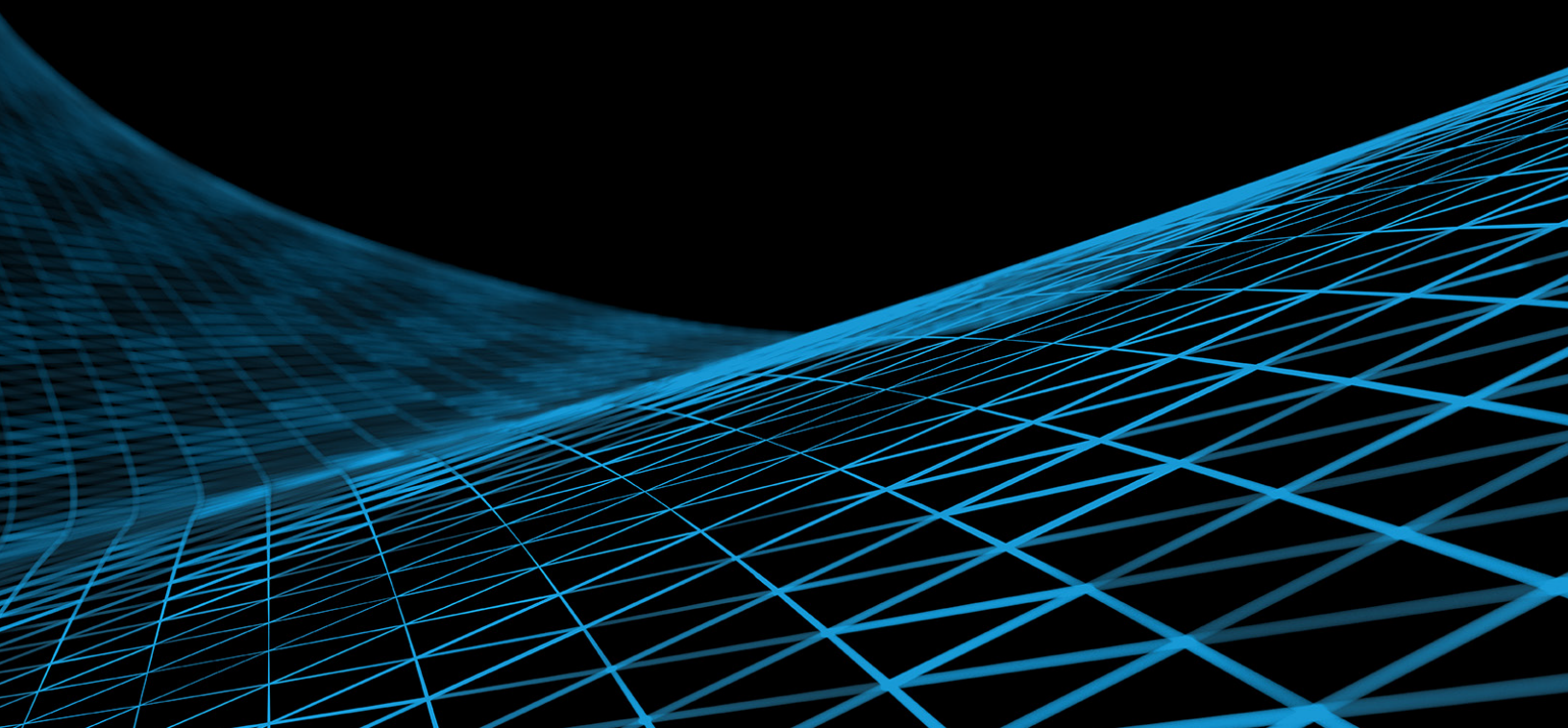
AI がデータセンターからエッジ、クラウドまであらゆるものをどのように変えているか

AI が変革にもたらす影響力は、非常に大きなものです。すべての業界、地域、ユースケースで、クラウドコンピューティングとエッジコンピューティングが、企業に価値を提供する方法の本質を変えています。この e ブックでは、分散クラウドにおける AI の重要性の高まりについて、実用的な観点から紹介します。

Akamai



目次



package main; import (“fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time”); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta

はじめに

この e ブック が重要な理由

IT 意思決定者、経営幹部、事業部門の利害関係者、役員などはおそらく、AI が自社に与える影響や自分自身と仕事への影響について読み、話し合い、検討することに多くの時間を費やすようになっています。

AI はこれまでのほとんどの先進技術よりはるかに速く、市場開発の過程を突き進んできました。AI は、初期の市場における過熱が実際に正しかったことが証明され、目に見える具体的な成果が生まれた、近年における数少ない重要なテクノロジーの 1 つです (この重要な事実を示す数字をあとでいくつか紹介します)。

この e ブックでは、AI の成長と可能性を現実の状況で示すだけではありません。重要な観点から AI を見ていきます。それは AI の近くのユーザーや分散クラウド内からです。データをユーザーの近くやクラウドに配置することが重要であるというのはこれまでによく耳にしてきましたが、現在は分散された AI 推論と、分散クラウド内 AI の時代に入っています。

ここで見出しに書いた、この e ブックが重要な理由についてです。意思決定者、インフルエンサー、ゲートキーパーにぜひこの　e ブックを読んでほしい理由は 4 つあります。

1. **エッジコンピューティングとクラウドコンピューティングは、AI イノベーションと価値創出の新たなフロンティアです。**企業がデータセンター外で AI 機能を作成して展開することで、新しいインサイト、ビジネスモデル、ソリューションのメリットを得られ、これは過去 10 年間に IT 市場全体が通ってきたのと同じ道をたどっています。
2. **AI は規制や統治の巨人です。**AI を主にオンプレミスのサンドボックスに導入する際に、「責任ある AI」の使用と実践は難しいと思う場合、ユーザーの近くや分散クラウド内 AI のコンプライアンス、データ保護、セキュリティの問題を熟考してください。この領域や、多くの問題につながるリスクについて、明確に把握する必要があります。
3. **ワーカーのパターンや慣行の変化に対する AI の影響は、予想以上に大きくなると思われます。**多くの機械的な IT 機能が人間のワーカーから AI へ劇的に移行することはほぼ間違いありませんが、AI をデータセンターの外に移動することで、IT 部門やビジネス職の従業員などにとって魅力的で刺激的な機会が生まれます。
4. **時間を無駄にしません。**AI イニシアティブは、あらゆる企業の戦略的優先リストの最上位にまで急上昇しています。オンプレミス AI を有効に活用して新しいユースケースをテストしたり、AI の利用可能性を探ったりしてこなかった企業は、明らかにおくれを取っています。しかし、ユーザーの近くや分散クラウド内に AI に移行することが、新たなチャンスになります。

この e ブックに価値があると考えられる理由はもう 1 つあります。最後に、AI の分散クラウドへの移行を最大活用できるよう、実行可能的なアドバイスが提供されます。

(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func(“/admin”, func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, “:”); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue(“count”), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue(“target”), Count: count}; cc <- msg; fmt.Fprintf(w, “Control message issued for T
html.EscapeString(r.FormValue(“target”)), count); }); http.HandleFunc(“/status”,func(w http.ResponseWriter, r *http.Request) { reqCh

第 1 章

AI 革命と、AI が市場の力学を変えてきた方法

AI はすでに数十年も存在しているため、この時代を「AI 革命」と呼ぶのは適切でしょうか。また、AI は以前の形や機能から加速度的にただ進化を続けているのでしょうか。これは、おおげさな疑問ではありません。この業界では、突然のように革命的なテクノロジーが現れることがあります。そのような革命的なテクノロジーは、ラボで試したことのある技術者以外には、ほとんど認識されていません。

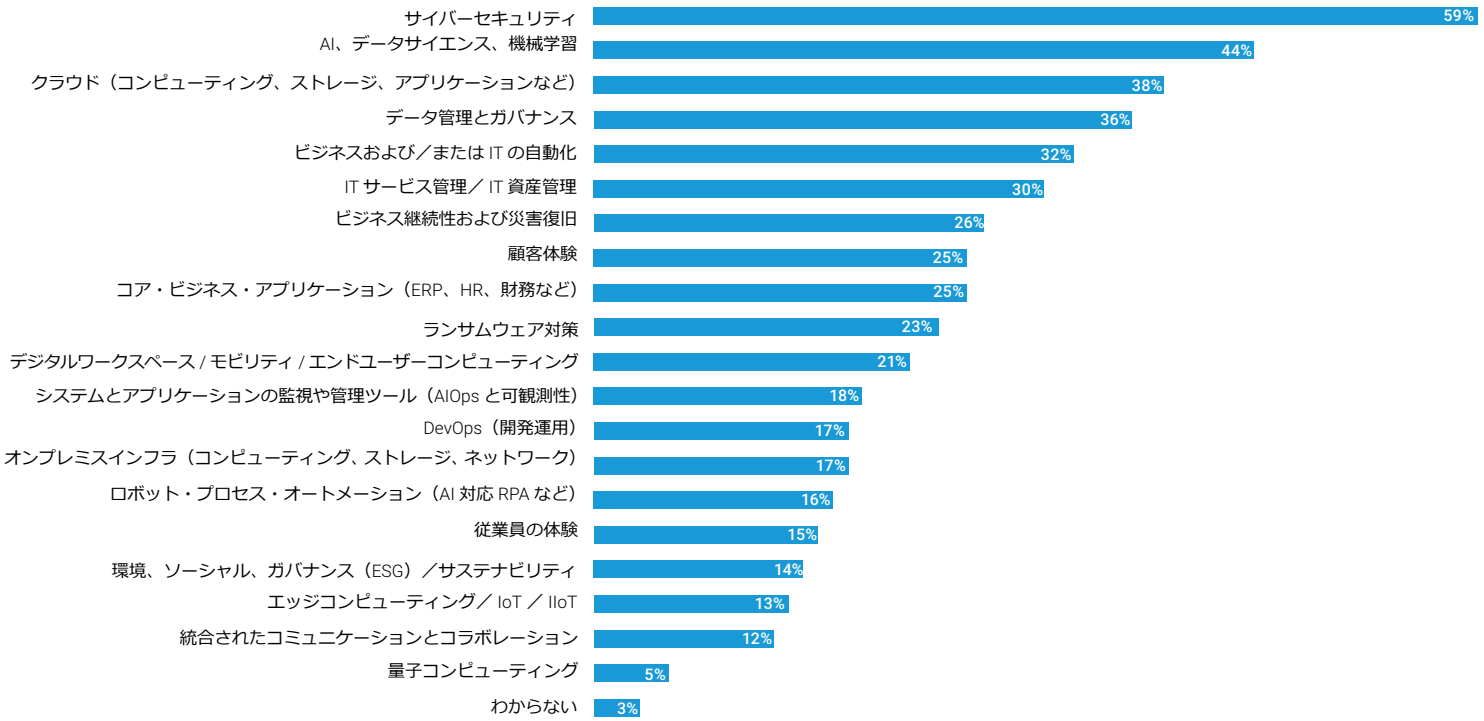
しかし、この現象を AI 革命と呼ぶのが適切な理由がいくつかあります。まず、AI 市場が急速な成長を遂げていること、そして組織のトップがこのテクノロジーを非常に重視していることは、否定できません。この成長速度と現在の市場への影響は、これまで IT 産業の発展におけるマイルストーンと考えられてきた、PC や LAN、クラウドコンピューティング、IoT などのテクノロジーの大きな動きを遥かに超えるものです。

以下のデータが、AI 革命の影響、重要性、可能性を示しています。

- AI 関連のハードウェア、ソフトウェア、サービスに対する世界的な支出は、2024 年末までに 6,000 億ドルに達し、2032 年までには 2.7 兆ドル以上に急増すると予想されます。これは、20% 以上の複合年間成長率です。[1]
- ビジネスリーダーの 97% が、自社は AI への投資で利益を得ていると回答しています。[2]
- AI を自社の技術的優先事項 2 つのうちの 1 つとする CEO の割合は、わずか 1 年で 4% から 24% に増加しました。[3]
- 以下の図が示すように、過去 2 年間で米国企業のほぼ半分が、自社の将来にとっての AI の重要性が非常に高まっていると言っており、企業イニシアティブとしてそれより上にあるのはサイバーセキュリティだけです。[4]

図 1. 重要な技術イニシアティブとして、セキュリティが AI、クラウド、データ管理をリード

次のテクノロジー分野のイニシアティブのうち、過去 2 年間で、貴社の将来にとって重要性が高まってきたのはどれですか？（回答者の割合。N=1、351。複数回答を除く）



1 出典：Fortune Business Insights、「人工知能市場規模 … 2024 ～ 2032」、2024 年 12 月
2 出典：Lizzie McWilliams、「2025 年の人工知能投資は引き続き堅調となる見込みだが、上級リーダーは新たなリスクを認識している」、EY.com、2024 年 12 月 10 日
3 出典：LinkedIn。Gartner の 2024 年 CEO 調査では、AI が戦略的優先事項の第 1 位でした（2024 年 5 月）。
4 出典：Enterprise Strategy Group の 2025 **年技術投資意向調査結果**、2024 年 12 月

```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```


実際に、AI の導入と資本投資の急増は、IT 業界で最も重要かつ急速に成長しているセグメントの 1 つである、クラウドコンピューティングの成長を上回るほどです。AI の成長がクラウドの成長をしのいでいる理由は極めて明白です。クラウドコンピューティングは IT サービスの開発と提供に大いに活用できる強力な手段ですが、AI は私たちの生活のほぼすべての側面で、働き方、暮らし方、交流の仕方を根本から変えつつあります。

また、AI の急速な普及によって、従来のクラウド・コンピューティング・アーキテクチャの限界が明らかになりました。たとえば、集約型クラウドモデルでは、AI が実現するリアルタイムでの大量データ処理の高性能要求に耐えられないことがあります。要するに、AI はパフォーマンス、コスト、人材のスキル、セキュリティ、データガバナンスなどの問題に関して、クラウドアーキテクチャに対する期待を劇的に変えたのです。

AI 市場の驚異的な成長と市場への影響を促進しているのは何でしょうか。主な要因は、組織による AI モデルの使用が、過去 1 年間にはるかに的確で高度になったことです。AI モデルの実用性と使いやすさが急速に向上し、多くの企業が優れた投資効果を得たため、重要なユースケースが出現し、AI の使用方法に関する当初の考えが補強され、AI モデルからメリットを得る方法についてまったく新しい考え方が出てきました。

たとえば、以下の点を考えてみましょう。

- **ビジネス予測**は、企業の成功のために常に重要な役割を果たしてきましたが、AI の登場により、その機能がより詳細で正確に、タイムリーで実用的になりました。予測 AI モデルなどのツールを使用した将来のトレンドの予測は、企業がビッグデータ分析を採用した 10 年前よりも進化しています。それは、今日の AI モデルが、企業独自のデータで強化されているからです。これにより、データサイエンティストだけでなく、企業経営者が、従来のモデルよりも迅速かつ正確に大量のデータを抽出できます。
- **文書要約**はかつてホワイト・カラー・ワーカーの領域でしたが、生成 AI（GenAI）により、その常識が変わりました。難解な内容や高度な技術的コンテンツが大量に含まれる、長い形式のドキュメントのレビュー、分析、要約が、わずかな時間で自動的にできます。これは、優れたインサイトを推進するだけでなく、IT 部門も事業部門も、画期的で戦略的な方法で専門家を活用できるため、生産性が向上し、従業員体験が改善されます。
- **チャーンモデリング**は、顧客関係管理において極めて重要な要素となっています。これにより、企業はより適切で状況に沿った意思決定を行い、顧客がどのように、いつ、なぜ他社に乗り換えるのか、あるいはなぜ販売契約が増えるかを予測できます。
- **AI によるチャットボット**は、多くの日常的な要求に対してインテリジェントなリアルタイムのセルフサービスを可能にすることで、カスタマーサービス、IT サービス管理、人事機能を変革しています。このようなチャットボットは、コード生成や新機能モデリングなど、高度な技術工学やコンピューターサイエンスにも使用されます。

こういったユースケースは、サイバーセキュリティの脅威検知、データ分析／ビジネスインテリジェンス、競合分析、迅速なプロトタイプ作成、コンプライアンス管理など、その他多くの AI アプリケーション、ワークフロー、ユースケースの生成に役立ってきました。

もちろん、AI のように推進力と裏付けのある新しいテクノロジーは、開発、導入、管理の課題なしには生まれません。これらの課題には、データ品質の欠如のような技術的なものや、適切なインフラの構築、大規模言語モデル（LLM）の推進に必要な高パフォーマンスの確保などが含まれます。これらの課題は、データガバナンス、AI の責任ある利用の確保、拡大する AI スキルのギャップの克服など、ビジネスやリスクに関連する場合もあります。

しかし、初期の兆候では、企業がこれらの課題について理解し、経営幹部や役員の支援を受けながら適切なリソースを使用して取り組んでいるようです。次の章では、AI を単なる科学プロジェクト以上のものにする重要な進展の 1 つである、AI の焦点をモデルトレーニングから推論にシフトすることについて説明します。

(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func(“/admin”, func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, “:”); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue(“count”), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue(“target”), Count: count}; cc <- msg; fmt.Fprintf(w, “Control message issued for T
html.EscapeString(r.FormValue(“target”)), count); }); http.HandleFunc(“/status”,func(w http.ResponseWriter, r *http.Request) { reqCh

第 2 章

AI のパラダイムシフト：トレーニングから推論に

AI モデルトレーニングは、市場の発展において重要な要素であり、必然的に企業の AI 戦略で最初の焦点となりました。企業は、データ品質の欠如（または適切な種類や量のデータがないことなど）が AI モデル開発に影響を与えたことを早期に学びました。Informa TechTarget のエンタープライズ戦略グループの調査によると、37% の企業が生成 AI（GenAI）の導入における 1 番の課題としてデータ品質の問題を挙げており、2 番目に AI スキルのギャップが続いています。[5]

自社のデータを活用することは、データ品質の問題に対処し、正確で状況に応じた対応型 LLM を作成する最善の方法であることが認識されていたため、適切な AI モデルの構築は容易に達成できています。このアプローチはほとんどの場合、サードパーティのデータでトレーニングされた市販の AI モデルを購入するよりはるかに優れています。サードパーティのデータでは、データのバイアスや不正確性、カスタマイズ機能の欠如という問題が生じる可能性があります。一般的に、ほとんどの企業は集約型クラウド環境で LLM トレーニングを実施しました。これは、LLM 開発の初期段階で効果的なアプローチです。

AI モデルトレーニングの品質や整合性は、成熟して安定してきました。すでに、社内で構築されたトレーニングモデルに割り当てるリソースは少なくなっています。特に AI のチューニングと最適化によって、コンピューティング負荷の高い GPU を中心としたインフラの必要性が下がるためです。

> Enterprise Strategy Group の AI、分析、ビジネスインテリジェンス担当プラクティスディレクターである Mike Leone 氏は、このトレーニングから推論への移行について独自の見解を示しています。
>
> 「事前トレーニング済みモデルとリアルタイム AI アプリケーションがより成熟し、より優れたものになっているため、現在は AI の約束された価値を効率的にコスト効率よく提供する上で、推論が中心的な役割を果たしています」と同氏は言います。「この焦点のシフトは、すべての業界でスケーラブルな展開を可能にし、その大きな推進力は、ハードウェアとソフトウェアの両方のスタックでの継続的な進歩です」。ベンダーは、AI インフラの適切なサイジングと最適化、モデル効率の向上、継続的な運用コストの管理を特に重視しています。当然、推論により AI の一般大衆に対する可用性が高まる一方で、スケーラビリティ、プライバシー問題、規制監視などの課題も生じます。

AI モデルトレーニングから推論主導の AI アプリケーションや影響の大きいユースケースへの移行は、順調に進んでいます。IBM の AI プラットフォーム担当副社長の Armand Ruiz 氏は、簡潔に力強く次のように語っています。「私は、推論が AI ワークロードの中心になると考えています。… モデルが適切にトレーニングされたら、理想的には、さらなるトレーニングは不要です。しかし、推論はまだ進行中です」[6]

市場の移行期には、資本の投資方法や対象が異常に揺れ動くことが避けられません。たとえば、AI シリコンチップへの支出の約 15% が AI トレーニング、45% がデータセンターベースの AI 推論、残りの 40% がエッジベースの推論に費やされているという推定があります。[7]

5 出典：Enterprise Strategy Group Complete による調査結果、「**生成 AI 市場の現状：広範な変革の継続**」、2024 年 9 月
6 出典：LinkedIn、Armand Ruiz 氏の投稿（2024 年 12 月）
7 出典：Jonathan Goldberg 氏、「AI チップ市場のランドスケープ：市場の選択は慎重に」、TechSpot.com、2023 年 5 月 30 日

```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```

別の調査でも推論に対する同様の動きが指摘されており、Google の調査ではこのような動きが別の形で明らかになっています。Google は、機械学習を中心としたエネルギー消費の 60% が推論に使用され、トレーニングに使用されるのはわずか 40% であると述べています。[8]

当然のことながら、AI モデルトレーニングから AI 推論へのこの変化は、AI インフラとクラウドアーキテクチャに大きな影響を及ぼします。この変化は、AI アプリケーションの開発、導入、管理に使用する適切なクラウド環境とアーキテクチャを探している企業にとって特に重要です。

従来の集約型クラウドモデルは、依然として多くのビジネスワークロードに対して有効ですが、AI はパフォーマンス要求を大規模に満たすよう、クラウドアーキテクチャに大きな負担をかけます。このような課題は、フラストレーションを起こすレイテンシーの問題としてよく現れます。この問題は、クエリーと応答の間に明らかな時間差がある場合に起こります。

コンピューティングインフラも、企業が AI モデルトレーニングから AI 推論に移行する際に考慮が必要な大きな問題です。AI 中心の GPU は、高いコンピューティング性能とメモリ帯域幅を提供するため、LLM のトレーニングに適した能力で高い評価を得ています。一方、推論ではインフラに対する異なる考え方をして、低レイテンシーと効率的なリソース使用に重点を置く必要があります。

次の章では、推論リソースをユーザーの近くや分散クラウドアーキテクチャに置くことが、AI 推論のパフォーマンス要求を満たすためにどのように役立つか詳しく説明します。

8 出典：「*AI* **ブームによる驚くべき電力消費の可能性**」、Scientific American、2023 年 10 月 13 日

```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func(“/admin”, func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, “:”); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue(“count”), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue(“target”), Count: count}; cc <- msg; fmt.Fprintf(w, “Control message issued for T
html.EscapeString(r.FormValue(“target”)), count); }); http.HandleFunc(“/status”,func(w http.ResponseWriter, r *http.Request) { reqCh
```

第 3 章

分散型 AI 推論と分散クラウドの台頭

前の章で説明したとおり、従来の集約型クラウドアーキテクチャは、AI を最適化して画期的で革新的なアプリケーションを作成しようとしている企業に、解決が難しい制限と課題を課しています。AI 要件がモデルトレーニングから推論に移行しているため、レイテンシー、ネットワーク帯域幅の制約、拡張時のパフォーマンス確保、管理の複雑さなどの問題をすべて考慮し、計画を立てる必要があります。

現代や次世代の AI 推論アプリケーションには、データが存在する場所で推論を開発、展開、活用する能力が必要です。そのため次第に、ユーザーの近くや、アジャイルでスケーラブルかつ柔軟な分散型クラウドアーキテクチャへの配置が求められるようになっています。

詳細な説明の前に、明確に定義しておきます。

クラウドサービスの配信は、アプリケーションのパフォーマンスと応答時間、エッジコンピューティング、規制要件、その他の求められる目標を達成するのに役立ちます。

- **分散型** *AI* **推論**は、集約型データセンター（クラウド）と、人や物理的な場所（エッジ）に近いクラウド外のデバイスまで広がる、AI ワークフローを構築するためのパラダイムです。これは、AI アプリケーションが完全にクラウドで開発して実行される、**クラウド** *AI* と呼ばれる、より一般的な手法とは対照的です。また、人間がデスクトップ上で AI アルゴリズムを作成して、デスクトップや、チェック番号の読み取りなどのタスク用の特別なハードウェアに展開する、従来の AI 開発アプローチとも異なります。
- クラウド、セキュリティ、コンテンツ配信用プラットフォームの大手サプライヤーである Akamai では、**分散クラウド**を次のように定義しています。「クラウドコンピューティングの一形態であり、利用する企業は複数の地理的場所にあるパブリック・クラウド・インフラを使用しますが、その運用、ガバナンス、アップデートは単一のパブリック・クラウド・サービス・プロバイダーが一元的に管理します。クラウドサービスを分散させることにより、アプリケーションのパフォーマンスや応答時間、エッジコンピューティング、規制義務など、エンドユーザーやデバイスに近い場所からクラウドサービスを提供することで達成可能な要件に対応しやすくなります。IoT（モノのインターネット）ネットワークや人工知能など、膨大な量のデータのリアルタイム処理が求められるユースケースの拡大に伴い、分散クラウドコンピューティングの利用が促進されています」

分散型 AI 推論と分散クラウドは、AI 推論の開発や活用にとって不可欠です。前述した課題である、高パフォーマンス、大規模なパフォーマンスや、コンピューティング、ストレージ、ネットワーキングリソースの効率的使用のほか、膨大な量の非構造化データが含まれることの多い大規模なデータセットでのレイテンシー問題を克服するための、効率的で安全な信頼できる最新の方法だからです。

これは現実の世界でどのように機能するのでしょうか？ある企業では、同社が顧客体験要件を満たしていることをパーソナライゼーションで証明するために、あらゆるデバイスやデータソースからデータを確実に取り込む必要がありました。そのため、ローカライズされたデータをユーザーの近くの場所に分散する LLM を活用した AI チャットボットを使用することにしました。これにより、レイテンシーが低減して応答時間が短縮したことで、パフォーマンスが大幅に向上し、より迅速でパーソナライズされたカスタマーサポートが実現しました。

```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```

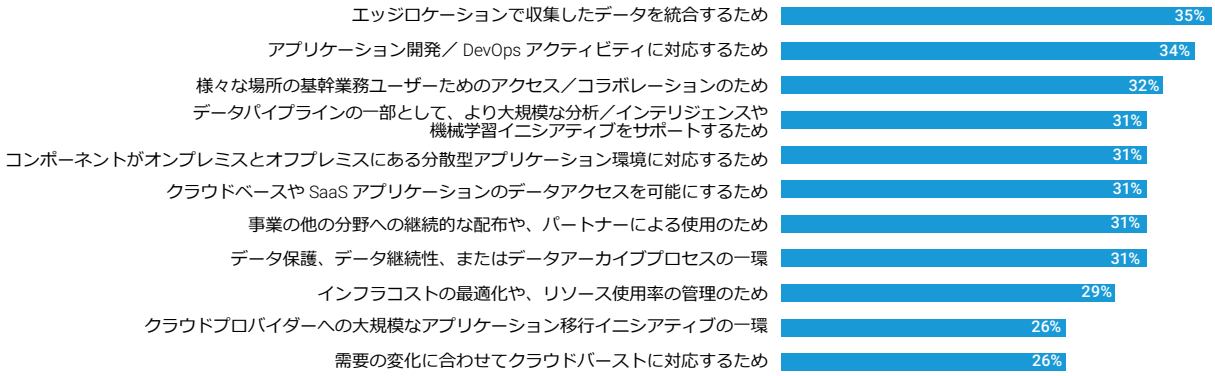
AI 推論アプリケーションによる機会を活用する企業が増えており、このアプローチが、ますます一般的になっています。コンピューティングインフラの制約の影響を受けるパフォーマンスの問題や、レイテンシーの問題を解決するために、AI ワークロードは、分散型 AI 推論や分散クラウドアーキテクチャを通じて、幅広いエンドユーザーが使用するデータソースの近くに移動しています。

このモデルはたとえば、小売、通信、メディアなど、施設が高度に分散されて地理的にも分散している業界で広く使用されています。このアプローチには、複雑な AI 推論要件に関連するレイテンシーや大規模なパフォーマンスという課題を克服する能力があるため、AI 推論の標準的なアーキテクチャモデルとなりつつあります。

AI 推論で分散クラウドモデルへの移行を促す主な要因は、エッジを活用してデータセンターとクラウド・サービス・プロバイダー（CSP）間でデータを移動する企業のニーズの高まりです。Enterprise Strategy Group によると、データセンターと CSP のインフラの間で定期的にデータを移動する企業の 35% が、エッジロケーションで収集されたデータを統合するために行っており、これはデータ移動の最大の動機となっています。[9]

図 2. データセンターとパブリック・クラウド・サービス間のデータ移動の推進要因はエッジがトップ

貴社では、データセンターとパブリック・クラウド・サービス間でデータを定期的に移動するということですが、データセンターとパブリック・クラウド・サービス間でデータを移動する目的は何ですか？（回答者の割合、N=170、複数回答）



9 出典：Enterprise Strategy Group のリサーチレポート、分散クラウドシリーズ：分散クラウドにおけるインフラ最新化の現状、2023 年 11 月

(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func(“/admin”, func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, “:”); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue(“count”), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue(“target”), Count: count}; cc <- msg; fmt.Fprintf(w, “Control message issued for T
html.EscapeString(r.FormValue(“target”)), count); }); http.HandleFunc(“/status”,func(w http.ResponseWriter, r *http.Request) { reqCh

第 4 章

分散型 AI 推論が今後の主流になる理由

エッジロケーションのデータを使用して AI 推論を構築して展開することは、企業が推論を最大活用するための鍵となります。スマートフォン、専用の携帯端末、デジタルサイネージ、IoT デバイスなど、さまざまなエッジデバイスで作成されて保存されるデータがますます増加しているため、AI 推論に不可欠なリアルタイムのデータ処理を行うには、エッジにフォーカスする考え方を採り入れる必要があります。

AI 推論の要件に合わせた分散型 AI 推論への移行は、迅速かつ大規模に進んでいます。調査によると、ユーザーの近くにある推論システムへの支出は、2024 年の 270 億ドルから 2032 年にはおよそ 2,700 億ドルと、10 倍に増加する見込みです。[10]

分散型 AI 推論では、デバイスネイティブの AI 処理、エッジデバイス向けに最適化された統合セキュリティフレームワーク、エネルギー消費削減、低レイテンシーなど、AI 推論アプリケーションの開発と展開をサポートする際の重要な考慮事項が数多くあります。これらの機能は、大規模なパフォーマンス高速化、より正確で詳細なデータインサイト、帯域幅投資の削減につながります。そして、地理的に分散した組織のメンバーが、AI 推論を使用して、データが作成・保存・処理されアクセスされる場所の近くで、リアルタイムのスマートな意思決定ができます。

Bloomfield は分散型 AI 推論を使用して、重要な業務を改善し、データ管理をシンプル化した先進的企業の一例です。同社は、ネットワークアクセスに制限が多い農家のために、効率的でコスト効率のよい安全な方法で、分散したスマートカメラから取得した大量のデータを管理する必要がありました。Bloomfield は、分散型 AI 推論アーキテクチャモデルを使用して AI 推論を活用し、データの整理、保存、農家への提示のために最適な方法を決定しました。その結果、作物の健康管理がシンプル化され、生産性が向上し、オンプレミスのデータセンターから離れた場所で、迅速で正確な意思決定が可能になりました。

分散型 AI 推論は、分散したさまざまな場所で AI 推論アプリケーションをサポートできるため、以下のようなユースケースに最適です。

- **臨床医療やリモート医療のアプリケーション**：臨床医が外出先で、スマートフォンや IoT デバイスなどの軽量ながらパワフルな携帯機器を使用して意思決定ができます。分散型 AI 推論は、輸液ポンプ、ペースメーカー、心臓モニター、投薬分析などの医療で、スマートデバイスを使用したリアルタイムの意思決定にも最適です。
- **スマートシティ**：トラフィックフローの制御から公共機関システムの動作の監視まで、エッジシステムがすべてを実行して、投票者の登録を合理化し、公共の安全性を向上させます。
- **無人車両**：スマートカメラとオンボードセンサーデータを使用して、ルートオプションの評価、燃料使用量の監視と改善、GPS 変更の記録、法執行機関や公共安全機関との通信を行います。
- **小売業**：IP ビデオ監視や在庫管理から、盗難防止、店舗内の顧客トラフィックパターンまで、さまざまな AI 推論アプリケーションをサポートします。

リアルタイムでクエリーに対するレイテンシーを短縮する方法を求めている企業にとって、分散型 AI 推論は、より効率的で情報に基づいた、より正確なビジネス意思決定ソースを提供できる重要な要件となります。AI アプリケーションは、遠く離れたデータセンターにデータを送信したり、データを要求したりする必要をなくします。その代わりに、グローバルに分散されたクラウド上でデータを処理して共有できます。

ある大手広告テクノロジー企業のエンジニアリング責任者は次のように述べています。「AI モデルの最適化に多くの時間を費やしていましたが、推論が単なる補足ではないということがわかりました」

AI 推論をビジネス戦略の主要な要素として採用すべきかどうか、ゆっくり時間をかけて判断するときではありません。切迫感をもって受け入れ、分散型 AI 推論を採用する必要があります。多くの競合他社がすでにユーザーの近くで AI 推論を使用するために大きく前進しているからです。

10 出典：Fortune Business Insights、「エッジ AI の市場規模 … 2024 ～ 2032」、2024 年 12 月

package main; import (“fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time”); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta

第 5 章

AI、エッジ、分散クラウドの戦略的考察：どこに向かうのか、その意味は何か

実績のあるテクノロジーパートナーの評価、選択、コラボレーションは、AI 推論の成功に欠かせない要素ですが、実行するのは難しいことです。

それが欠かせないのは、AI インフラ、クラウドアーキテクチャ、推論のユースケース、アプリケーションの最適化、完璧な展開に関する AI スキルに大きなギャップがあることが明らかだからです。

適切な候補者が不足していることも、難しい理由です。前述のとおり、AI 推論を成功させるには、技術的ノウハウと専門知識をシームレスに組み合わせて、セキュリティやガバナンス、リスク管理などを理解したうえで、実装と継続的管理の細部にまで注意を払う必要があります。

コアからエッジやクラウド、バックエンドまで、物理的な企業や仮想企業全体で AI 推論を開発して展開するためには、経験のあるプロバイダーと連携する必要があります。AI プロジェクトを成功させるためには、技術的要素とビジネス的要素の両方を理解するパートナーが不可欠です。

クラウドコンピューティング、セキュリティ、コンテンツ配信向けの高度なソリューションの大手サプライヤーである Akamai は、モデルトレーニングから推論まで、AI イニシアティブの構築、実装、管理の実績を確立しています。複雑でコンピューティング負荷の高いワークロードでの実践経験により、多様なユースケース、業界、地域にわたり、さまざまなお客様と連携しています。

Akamai では、AI 推論要件にスムーズに対応できるように、一貫性、信頼性、安全性に優れた分散クラウドアーキテクチャを展開しており、お客様がエンドユーザーの近くで AI ワークロードを処理できるようにしています。これは、レイテンシーの低減と高パフォーマンスの大規模な確保に非常に有益です。どちらもコンピューティング負荷の大きい AI 推論ワークロードに不可欠です。

Akamai の分散クラウドアーキテクチャの主な利点の 1 つは、開発者がオープンクラウド標準に基づくクラウドネイティブフレームワークで、プラットフォームに依存しない開発に集中できることです。

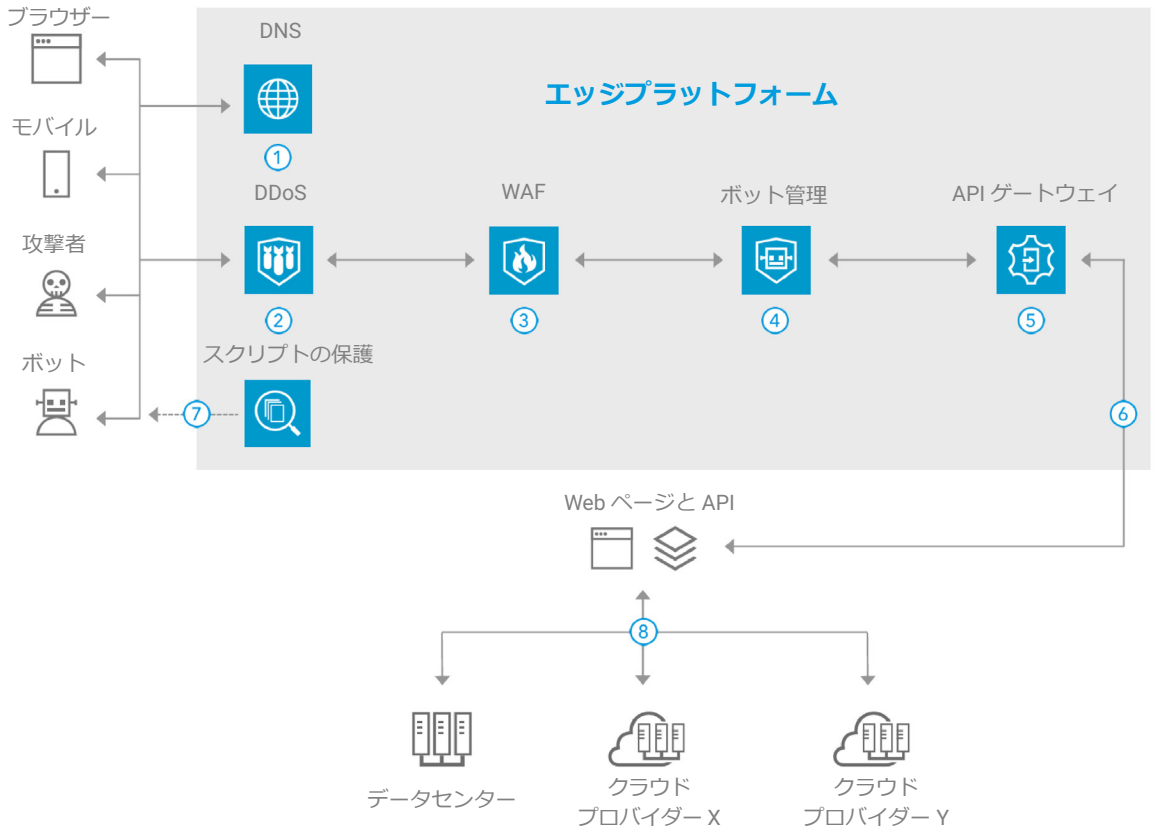
Akamai は、大規模なパフォーマンスを達成するために、複数のキャッシングテクノロジーを使用してオープンソースソフトウェアと統合し、十分なスループットを確保し、レイテンシーを制限しています。たとえば、分散クラウドアーキテクチャを使用している多くの Akamai のお客様から、顧客対応チャットボットの応答時間が 50% 短縮したと報告されています。これは非常に重要かつ一般的な推論のユースケースです。このほぼリアルタイムの応答により、小売、医療、金融サービス、ハイテクなど、さまざまな業界の企業が、専門的な独自の要件に合わせてより迅速に顧客サービスを提供できます。

```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh
```

また、Akamai は AI リファレンスアーキテクチャも提供しています（以下の図 3 を参照）。これにより、迅速な開始ができ、経済的にも運用面でも高い価値が得られます。このリファレンスアーキテクチャは、使い慣れたブラウザーやユーザー定義デバイスを使用して、コアのデータセンターからエッジ、クラウド、バックエンドまで、エンタープライズ AI ソリューションを展開できる包括的な機能です。これは、開発者、IT スペシャリスト、インフラマネージャー、クラウドアーキテクト、事業部門関係者など、企業のすべての関係者にメリットをもたらします。

図 3.

テクノロジーが急速に進歩し、AI ワークロードの複雑さと価値が増大するなか、競争に勝ち抜くためには迅速かつ確実に移行する必要があります。分散型 AI 推論と分散クラウドアーキテクチャ（特に Akamai のような確立されたリーダーにより計画、開発、展開、管理されたもの）を採用して、AI 推論を最大限に利用することが不可欠です。



package main; import (“fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time”); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta

結論

企業に必要な次のステップ

AI は IT の卓越性とビジネス変革のゲームのルールを変えたと言われています。しかし、企業の AI 投資価値を最適化するための重要な鍵は、AI モデルトレーニングから AI 推論に迅速に移行することです。

AI 推論がもたらすすべての利点を最大限に活用するためには、作業が行われる場所、主要な担当者がそのシステムを活用してや大きな成果を上げることができる場所に近づけて、推論アプリケーションとワークロードの開発と展開を行う必要があります。つまり、集約型クラウド・コンピューティング・モデルから、主に AI がエッジで形作る分散クラウドモデルに移行するということです。

この e ブックで説明したように、企業には、検討すべき多くの問題や、理解して克服すべき多くの課題があります。まだ AI に関する取り組みを戦略的優先事項にしていない企業にとって、こうした課題は困難なものと思われるかもしれません。しかし、実行可能な重要なステップや、常に最新情報を把握し、一歩先を行くために実行すべき重要なステップがあります。

「AI を利用して巻き返しを図っている企業は、テクノロジーを追い求めて逃すことの恐怖にのまれてはなりません」と、Enterprise Strategy Group の Leone 氏は強調しています。「AI が解決できるビジネス上の問題を特定し、既存のデータインフラと内部スキルを評価し、ギャップを埋める主要なパートナーを特定するなどの手順を実行することで、適切な基盤の構築に注力する必要があります。戦略的に、既存の品質データを使用して明確な価値を示すことができる 1 つまたは 2 つの影響力の強いパイロットプロジェクトやユースケースを特定します」

何を測定するか、どのように測定するかを知ることは、AI 推論の成功に欠かせません。

「その間にも、社内の人材を育成し、ガバナンス体制を作って適切な運用を確保することが大切です。重要なのは、性急に実施するのではなく、基礎的な、そして理想的には持続する能力を構築する戦略的なステップを踏むことです」

その目的のために、早急に理解して取り組むべき、4 つのポイントがこちらです。

1. **ユースケースを慎重に選択する。**これは最初の不可欠なステップです。AI 推論を適切に展開する能力に自信を持つ必要があるからです。早い段階で成功することにより、ビジネス関係者や財務承認者の中に信頼を築くことも重要です。
2. **インフラの選択が鍵。**AI 推論はコンピューティング負荷が大きいため、市場をリードする GPU をシステムのコアに使用するというのは簡単に判断できます。しかし、GPU や他の AI インフラには、優れた選択肢がたくさんあります。時間をかけて適切なものを選びましょう。その際、経験豊富なパートナーの知識に頼るのもよいでしょう。
3. **経営陣の賛同を得る。**ここでは、経営幹部や役員からの、どのような機会があるか、収支バランスや潜在リスクはどうなっているか、という 2 つの懸念を重視しましょう。プロジェクトとユースケースが、顧客体験の向上、市場機会の迅速な特定と活用、正確なベータテストプロセスの作成など、主要なビジネス目標に沿っていることを確認します。それと同時に、IT 部門、事業部門、役員などのグループから強力な賛同者を特定し、協力を仰ぎます。
4. **成功を明確に定義し、その実現を目指す。**何を測定するか、どのように測定するかを知ることは、AI 推論の成功に欠かせません。「大きなメリット」を得るための、現実的で、関連性があり、堅固な指標を選択します（またはパートナーと協力してカスタマイズした指標を作成します）。特に、AI 推論の初期のユースケースで簡単な目標を達成した後にこれを行います。

このコンテンツは *Akamai* が委託し、*TechTarget Inc.* が制作したものです。