Azure OpenAI のデプロイの種類

[アーティクル] • 2024/09/03

Azure OpenAI では、お客様はビジネスと使用のパターンに合ったホスティング構造を選択できます。 このサービスで提供されるデプロイの 2 つの主要な種類は、**標準**と**プロビジョニング済み**です。 標準にはグローバル デプロイ オプションが用意されており、トラフィックをグローバルにルーティングしてスループットを向上させます。 実行される推論操作はどのデプロイもまったく同じですが、課金、スケール、パフォーマンスは大きく異なります。 ソリューション設計の一環として、2 つの重要な決定を行う必要があります。

- **データ所在地のニーズ**: グローバル リソースまたはリージョン リソース
- **呼び出しボリューム**: 標準またはプロビジョニング済み

グローバルとリージョンのデプロイの種類

標準デプロイの場合、リソース内で**グローバル**または**リージョン**の 2 種類の構成を選択できます。 グローバル標準は、開発と実験で初めて使用する場合に推奨されます。 グローバル デプロイでは、Azure のグローバル インフラストラクチャが利用され、お客様のトラフィックはお客様の推論要求に最適な可用性を持つデータ センターに動的にルーティングされます。 グローバル デプロイでは、初期スループットの制限は高くなりますが、待ち時間は高い使用レベルで異なる場合があります。 大規模なワークロードを使って変化の小さい待ち時間を必要とするお客様には、プロビジョニング済みスループットを購入することをお勧めします。

グローバル デプロイは、すべての新しいモデルと特徴の最初の場所になります。 非常に大きいスループットが必要なお客様は、プロビジョニングされたデプロイ オファリングを検討する必要があります。

デプロイのタイプ

Azure OpenAI には、3 種類のデプロイが用意されています。 これらで提供される異なるレベルの機能の間には、スループット、SLA、価格に関するトレードオフがあります。 オプションの概要と、それぞれの詳細な説明を次に示します。

テーブルを展開する

サービス	Global-Batch	グローバル標準	Standard	プロビジョニング済み
最適な用途	オフライン スコアリング 遅延に敏感ではなく数時間で完了できるワークロード。 データ処理の場所に関する要件がないユース ケース向け。	お客様に推奨される出発点。 Global-Standard では、Standard よりも高い既定クォータとより多くのモデルを利用できます。	データ所在地の要件があるお客様向け。中程度以下のボリューム用に最適化。	大きくて一貫したボリューム用のリアルタイム スコアリング。 最高のコミットメントと制限が含まれます。
動作のしくみ	ファイルを介したオフライン処理	世界中のどこにでもトラフィックをルーティングできます		
作業の開始	Global-Batch	モデル デプロイ	モデル デプロイ	プロビジョニング済みのオンボード
原価	最も安価なオプション Global Standard の価格と比べて 50% 低いコスト。 クォータ割り当てが大きい新しいモデルすべてにアクセス可能。	グローバル デプロイの価格	リージョンの価格	一貫した使用ではコストを節約できる可能性があります
取得内容	Global Standard と比較した場合の大幅な割引	最も高い既定の呼び出し単位の支払い制限で、すべての新しいモデルに簡単にアクセスできます。 使用量が多いお客様は、待ち時間の変動が大きくなる可能性があります	可用性に関するSLA で簡単にアクセスできます。 バースト性が高い中程度以下のボリューム用に最適化。 一貫して使用量が多いお客様は、待ち時間の変動が大きくなる可能性があります。	非常に高く予測可能なスループットでのリージョン アクセス。 提供されている容量計算ツールを使用して PTU あたりのスループットを決定します

サービス	Global-Batch	グローバル標準	Standard	プロビジョニング済み
得られないもの	❌リアルタイム呼び出しのパフォーマンス ❌データ処理の保証 保存されたデータは指定された Azure の地理的な場所に留まりますが、推論のためのデータ処理は任意の Azure OpenAI の場所で実行される可能性があります。 データ所在地の詳細を確認する	❌データ処理の保証 保存されたデータは指定された Azure の地理的な場所に留まりますが、推論のためのデータ処理は任意の Azure OpenAI の場所で実行される可能性があります。 データ所在地の詳細を確認する	❌一貫した低遅延での高いボリューム	❌呼び出し単位の支払いの柔軟性
呼び出しごとの待ち時間	該当なし (ファイルベースの非同期プロセス)	リアルタイムの呼び出しと、中程度以下の使用量に最適化。 使用量が多いお客様は、待ち時間の変動が大きくなる可能性があります。 モデルごとに設定されたしきい値	リアルタイムの呼び出しと、中程度以下の使用量に最適化。 使用量が多いお客様は、待ち時間の変動が大きくなる可能性があります。 モデルごとに設定されたしきい値	リアルタイム用に最適化。
コード内の SKU 名	`GlobalBatch`	`GlobalStandard`	`Standard`	`ProvisionedManaged`
課金モデル	トークン単位の支払い	トークン単位の支払い	トークン単位の支払い	月単位のコミットメント

プロビジョニング済み

プロビジョニング済みデプロイを使うと、デプロイで必要なスループットの量を指定できます。 その後、サービスは必要なモデル処理容量を割り当て、その準備が整っていることを確認します。 スループットは、デプロイのスループットを表す正規化された方法であるプロビジョニング スループット ユニット (PTU) という観点で定義されます。 各モデルバージョン ペアでは、デプロイして PTU ごとにさまざまな量のスループットを提供するために、さまざまな量の PTU が必要となります。 詳しくは、プロビジョニング済みスループットの概念に関する記事をご覧ください。

Standard

標準デプロイでは、選択されたモデルで呼び出し単位の支払いの課金モデルが提供されます。 消費した分だけ支払うので、最も早く使い始めることができます。 各リージョンで使用できるモデルとスループットは、制限される場合があります。

標準デプロイは、バースト性が高い中程度以下のボリューム用に最適化されています。 一貫して使用量が多いお客様は、待ち時間の変動が大きくなる可能性があります。

グローバル標準

> **ⓘ 重要**
>
> 保存されたデータは指定された Azure の地理的な場所に留まりますが、推論のためのデータ処理は任意の Azure OpenAI の場所で実行される可能性があります。 **データ所在地の詳細を確認する** 。

Global デプロイは、非グローバル デプロイ タイプと同じ Azure OpenAI リソースで利用できます。ただし、Azure のグローバル インフラストラクチャを利用して、トラフィックを要求ごとに最適な可用性のデータ センターに動的にルーティングできます。 グローバル標準では、最大の既定クォータが提供され、複数のリソース間での負荷分散の必要がなくなります。

一貫して使用量が多いお客様は、待ち時間の変動が大きくなる可能性があります。 しきい値はモデルごとに設定されます。 詳しくはクォータに関するページを参照してください。 大規模なワークロードの使用時に、変動の少ない待ち時間を必要とするアプリケーションには、プロビジョニング済みスループットを購入することをお勧めします。

Global Batch

> **ⓘ 重要**
>
> 保存されたデータは指定された Azure の地理的な場所に留まりますが、推論のためのデータ処理は任意の Azure OpenAI の場所で実行される可能性があります。**データ所在地の詳細を確認する**。

Global Batch は、大規模で大量の処理タスクを効率的に処理するように設計されています。 個別のクォータ、24 時間のターゲット ターンアラウンド、Global Standard と比較した場合の 50% 低いコストで要求の非同期グループを処理します。 バッチ処理では、一度に 1 つの要求を送信するのではなく、1 つのファイル内で多数の要求を送信します。グローバル バッチ要求には、オンライン ワークロードの中断を回避する個別のエンキュー トークン クォータがあります。

主なユース ケースは次のとおりです。

- **大規模なデータ処理**: 広範なデータセットを並列ですばやく分析します。
- **コンテンツ生成**: 製品の説明や記事など、大量のテキストを作成します。
- **ドキュメントの校閲と要約**: 長いドキュメントの校閲と要約を自動化します。
- **カスタマー サポートの自動化**: 多数の問い合わせを同時に処理して迅速な対応を実現します。
- **データの抽出と分析**: 膨大な量の非構造化データから情報を抽出して分析します。
- **自然言語処理 (NLP) タスク**: 大規模なデータセットに対して感情分析や翻訳などのタスクを実行します。
- **マーケティングとパーソナル化**: パーソナル化されたコンテンツとレコメンデーションを大規模に生成します。

サブスクリプションでグローバル デプロイへのアクセスを無効にする方法

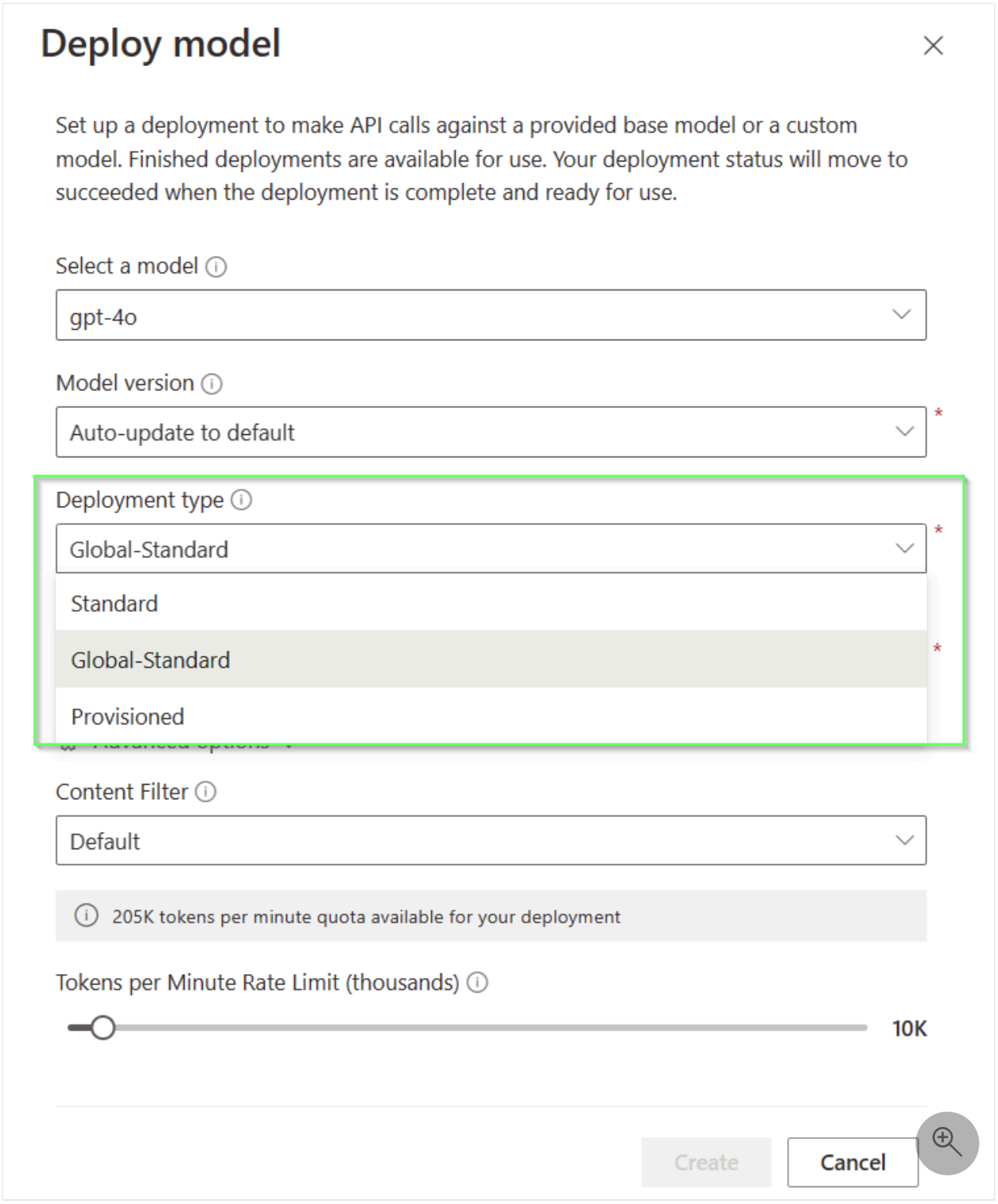
Azure Policy は、組織の標準を適用し、コンプライアンスを大規模に評価するのに役立ちます。 コンプライアンス ダッシュボードを通じて、環境の全体的な状態を評価するための集計ビューを提供します。これには、リソースごと、およびポリシーごとの粒度でドリルダウンできる機能が備わっています。 既存のリソースの一括修復と新しいリソースの自動修復を使用して、お客様のリソースでコンプライアンスを実現するのにも便利です。 AI サービスに関する Azure Policy と具体的な組み込みコントロールの詳細を参照してください。

次のポリシーを使用して、Azure OpenAI のグローバル標準デプロイへのアクセスを無効にできます。

JSON

```
{
    "mode": "All",
    "policyRule": {
        "if": {
            "allOf": [
                {
                    "field": "type",
                    "equals":
"Microsoft.CognitiveServices/accounts/deployments"
                },
                {
                    "field":
"Microsoft.CognitiveServices/accounts/deployments/sku.name",
                    "equals": "GlobalStandard"
                }
            ]
        }
    }
}
```

モデルをデプロイする



リソースの作成とモデルのデプロイについては、リソース作成ガイドに関する記事をご覧ください。

関連項目

- クォータと制限
- プロビジョニング スループット ユニット (PTU) のオンボード
- プロビジョニング スループット ユニット (PTU) の概要