



アーキテクチャ図

マルチモデル推論ワークフローオーケストレーション



マルチモデル推論ワークフローオーケストレーション: アーキテクチャ図

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon の商標およびトレードドレスは Amazon 以外の製品およびサービスに使用することはできません。また、お客様に誤解を与える可能性がある形式で、または Amazon の信用を損なう形式で使用することもできません。Amazon が所有していないその他のすべての商標は Amazon との提携、関連、支援関係の有無にかかわらず、それら該当する所有者の資産です。

Table of Contents

マルチモデル推論 i

マルチモデル推論ワークフローオーケストレーション 1

詳細情報 2

図の履歴 2

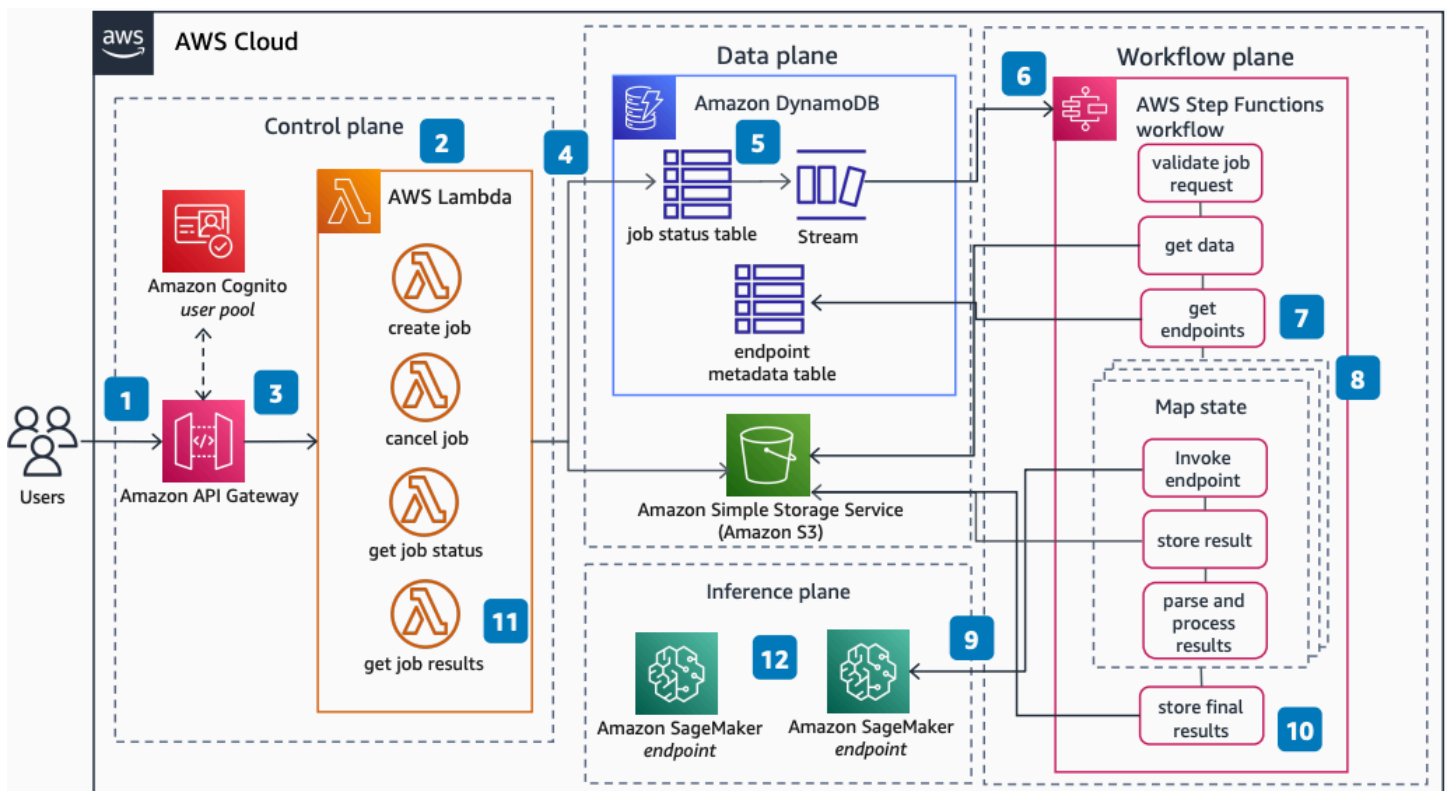
..... iv

マルチモデル推論ワークフローオーケストレーション

公開日: 2022 年 3 月 28 日 ([図の履歴](#))

このアーキテクチャは、複数の ML モデルの実行をオーケストレーションして、複雑な ML 主導のインサイトを得る方法を示しています。[AWS Step Functions](#) と [Amazon SageMaker AI](#) を使用すると、並列推論ジョブを実行して結果を集計できます。

マルチモデル推論ワークフローオーケストレーション



次の手順では、アーキテクチャについて説明します。

1. [Amazon API Gateway](#) は、ユーザーと管理者に RESTful インターフェイスを提供します。[Amazon Cognito](#) ユーザープールは認証を提供します。
2. 新しいジョブは、インサイトを実行するデータを含む API /create-job の へのPOSTリクエストによって作成されます。これにより、ジョブ作成[AWS Lambda](#)関数が呼び出されます。
3. この関数は、[Amazon Simple Storage Service](#) バケットにデータをアップロードし、[Amazon DynamoDB](#) テーブルにジョブ情報を追加して追跡します。

4. Lambda 関数は、API Gateway を介して公開される API メソッドのロジックを提供します。これにより、ジョブの作成、モニタリングが可能になり、結果の取得が容易になります。
5. テーブルの新しい項目によって DynamoDB ストリームがトリガーされ、推論のために Step Functions ワークフローがトリガーされます。
6. Step Functions ワークフローは、指定されたデータオブジェクトに対して複数の ML 推論ジョブを実行するために必要なすべてのステップを調整します。
7. ML エンドポイントに関するメタデータ情報は、ワークフローでできるように DynamoDB テーブルに保存されます。
8. ステップ関数のマップ状態は、各 ML エンドポイントを呼び出し、結果を並行して保存します。これにより、ワークフローは任意の数の ML エンドポイント呼び出しにスケールできます。
9. SageMaker AI モデルエンドポイントは ML モデルをホストします。ワークフローは、カテゴリごとにこれらのエンドポイントを呼び出します。
10. データ用に収集されたすべての ML インサイトの最終結果は、Amazon S3 に保存されます。
11. ワークフローが完了したら、への GET リクエストを通じて最終結果を取得できます /get-job-results。これにより、Amazon S3 バケットを読み取るジョブ結果の取得 Lambda 関数が呼び出されます。
12. マルチモデルエンドポイントは、カテゴリごとの言語固有の推論など、追加の最適化で各モデルタイプを拡張できます。

詳細情報

詳細については、次のリソースを参照してください。

- [AWS アーキテクチャアイコン](#)
- [AWS アーキテクチャセンター](#)
- [AWS Well-Architected](#)
- [Amazon SageMaker AI 製品ページ](#)
- [AWS Step Functions 製品ページ](#)

図の履歴

このリファレンスアーキテクチャ図の更新について通知を受け取るには、RSS フィードにサブスクライブします。

変更	説明	日付
初版発行	リファレンスアーキテクチャ図が最初に公開されました。	2022 年 3 月 28 日

 Note

RSS 更新をサブスクライブするには、使用しているブラウザで RSS プラグインが有効になっている必要があります。

翻訳は機械翻訳により提供されています。提供された翻訳内容と英語版の間で齟齬、不一致または矛盾がある場合、英語版が優先します。