SAP AI Core

生成日付: 2025-05-12 05:23:38 GMT+0000

SAP AI Core | For External SAP Customers

公開

オリジナルコンテンツ: https://help.sap.com/docs/AI_CORE/2d6c5984063c40a59eda62f4a9135bee?locale=en-US&state=PRODUCTION&version=CLOUD

機械翻訳に関する免責事項

この PDF 文書は、利便性を考慮して機械翻訳されています。機械翻訳の正確さや完全性については、SAP は一切保証しません。元の英語バージョンが優先されます。

フィードバックをいただける場合は、この文書の HTML バージョンから SAP Help Portal のフィードバックオプションをご利用ください。

警告

この文書は SAP Help Portal から生成されたものであり、公式の SAP 製品文書の不完全バージョンです。カスタム文書に含まれる情報は、SAP Help Portal でのトピックの配置を反映していない可能性があり、重要な点や他のトピックとの相関関係が欠落している可能性があります。そのため、本稼動での使用には適していません。

詳細については、https://help.sap.com/docs/disclaimer を参照してください。

SAP AI Core システムの概要

SAP AI Core システムは、内部ツールと外部ツールを接続します。

ユーザは、SAP AI Core の使用時にさまざまなリポジトリ、システム、およびオブジェクトと対話します。これらのオブジェクトの一部は SAP によって提供されます。また、拡張制御 (権限) および継続的統合/継続的デプロイメント (CI/CD) を有効化するために、カスタマはこれらのコンポーネントを提供します。

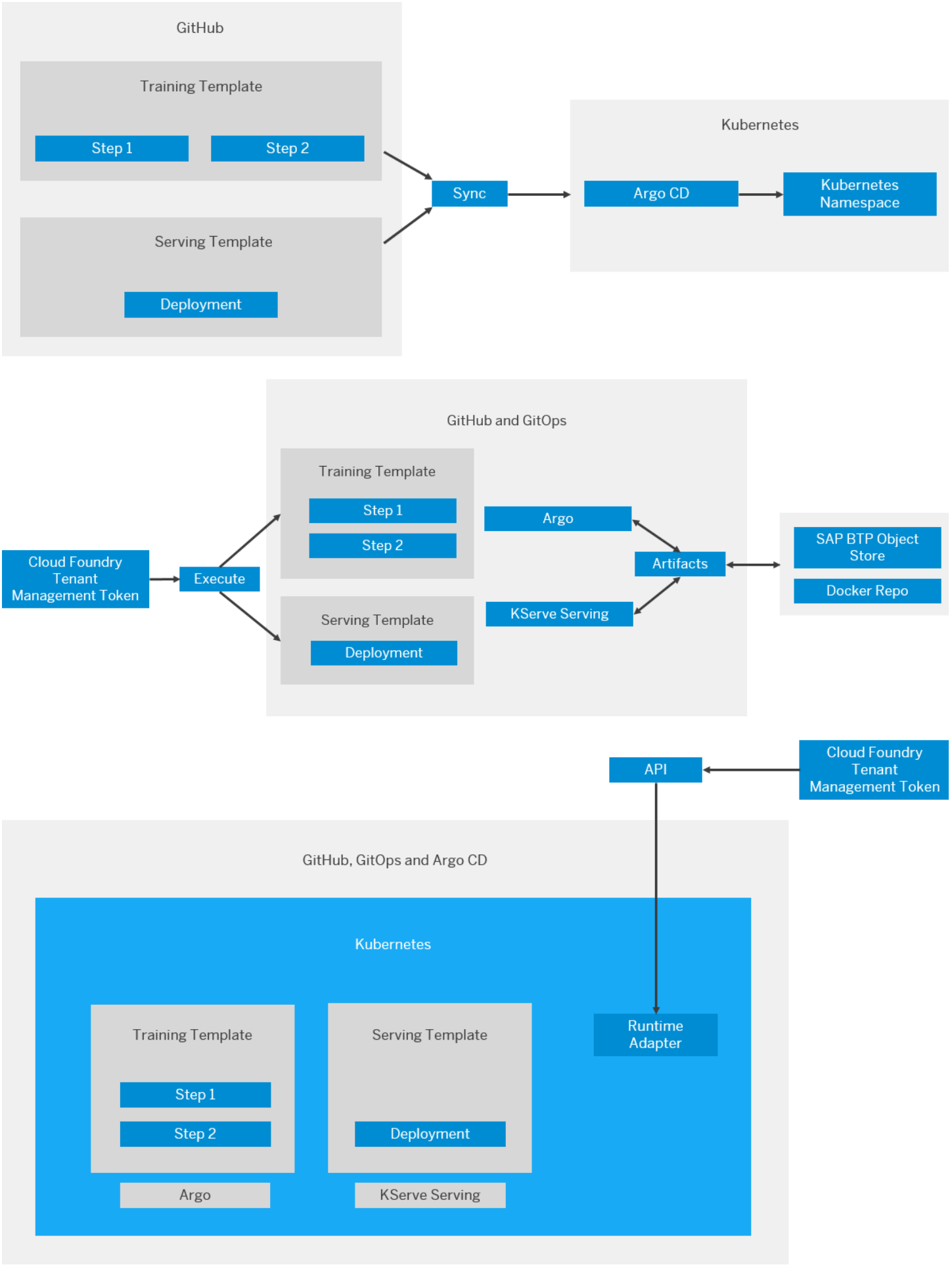
キーリポジトリ、システム、およびオブジェクト

要素の概要とその用途

内容	なぜ
Git リポジトリ	トレーニングの保存、ワークフローとテンプレートの提供
ハイパースケーラストレージ	トレーニングデータやモデルなどの入力アーティファクトおよび出力アーティファクト (SAP BTP オブジェクトストアサービスなど) の保存
Docker リポジトリ	テンプレートで参照されるカスタム Docker イメージの場合
Kubernetes (K8s)	K8s クラスタは、AI パイプラインで使用されるポッドを調整およびスケーリングします。リソースグループの分離は、K8s 名前空間に基づいています。
KServing (K)	機械学習モデルの最適なデプロイメント用。デプロイメントテンプレートでは、KServe 表記が使用されます。
AI API	複数のランタイムにわたるアーティファクトおよびワークフロー (トレーニングスクリプト、データ、モデル、モデルサーバなど) の管理 **ⅰ メモ** AI API は、他の機械学習プラットフォーム、エンジン、またはランタイムを AI エコシステムに統合するために使用することもできます。
Argo ワークフロー	Kubernetes のコンテナネイティブのワークフローエンジン。
SAP AI Launchpad	SAP AI Launchpad は、SAP Business Technology Platform (SAP BTP) 上のマルチテナントサービスとしてのソフトウェア (SaaS) アプリケーションです。カスタマおよびパートナは、SAP AI Launchpad を使用して、AI ランタイム (SAP AI Core など) の複数のインスタンスにわたる AI ユースケース (シナリオ) を管理することができます。SAP AI Launchpad では、生成 AI ハブを介して生成 AI 機能も提供されます。

同期

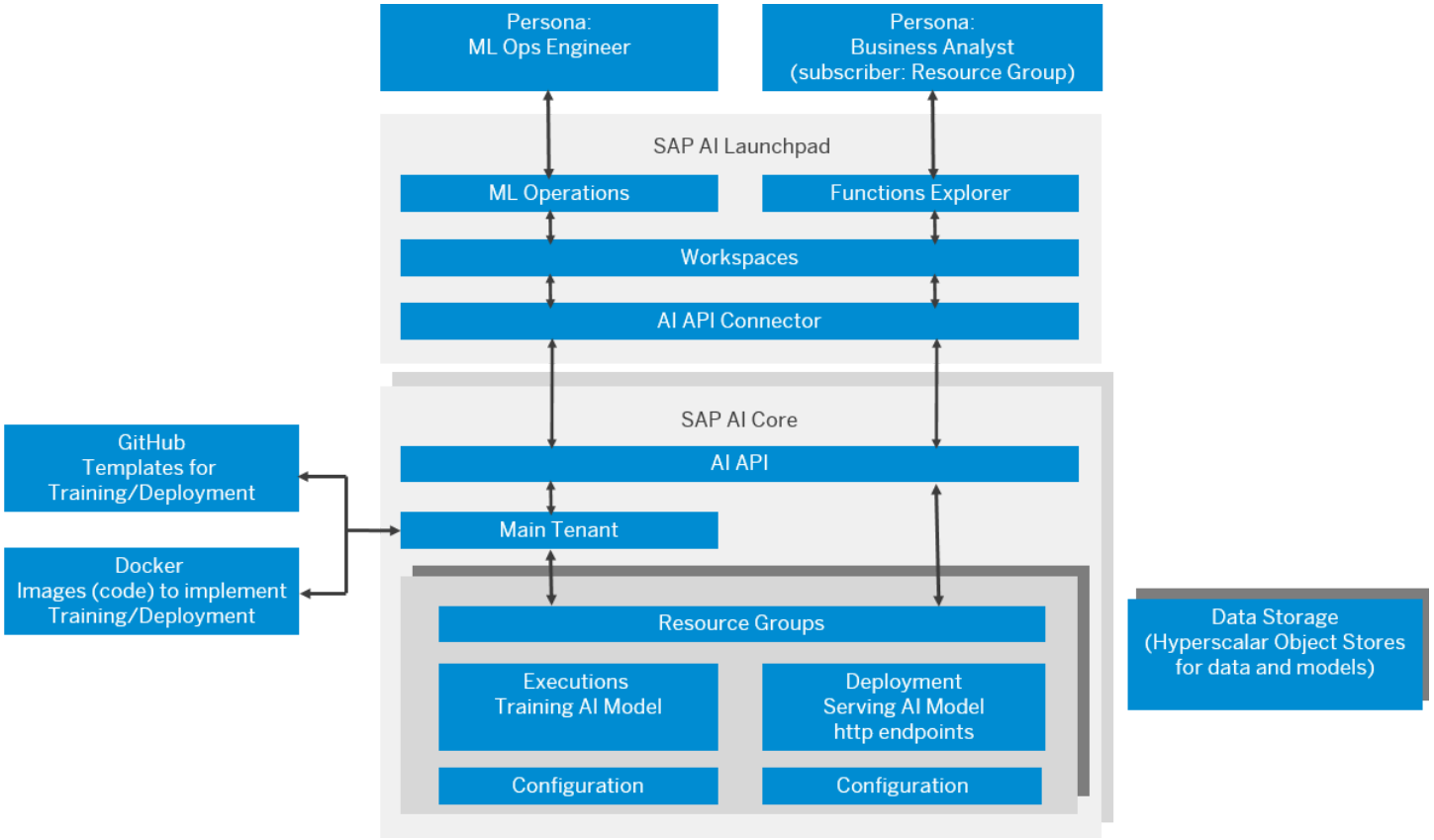
- GitOps 実装は、CI/CD を有効化するために SAP AI Core と統合されています。
- テンプレートは、Git リポジトリから Kubernetes クラスタに同期されます。
- トレーニングと提供テンプレートは、定期的に (数分ごとに) 同期されます。
- SAP AI Core によってテンプレート構文がチェックされ、同期に失敗すると、エラーメッセージが表示されます。



プロセスの概要

- SAP BTP トークンは、`AI API` で API コールの認証に使用されます。

- ユーザーは、AI API を介してテンプレートを実行します。
- トレーニングテンプレートは、Argo ワークフローを使用して実行されます。トレーニングパイプラインは、データ入力アーティファクトを利用し、モデルアーティファクトを出力します。
- トレーニングテンプレートおよび AI API 設定を使用して、実行が登録されます。その結果、トレーニングジョブになります。
- サービステンプレートは、KServing を使用してデプロイされます。
- デプロイメントは、サービステンプレートおよび AI API 設定を使用して作成されます。結果は推論サーバです。
- アーティファクト (データセット、モデルなど) は、ハイパースケーラストレージとの間でコピーされます。
- 必要なイメージが登録された Docker リポジトリからプルされます。
- シナリオおよび実行可能オブジェクトの詳細は、`AI API` によって Kubernetes クラスタから取得されます。



AI API の概要

AI API により、複数のランタイムにわたる AI アセット (トレーニングスクリプト、データ、モデル、モデルサーバなど) を管理できます。

Argo ワークフローと提供テンプレート、およびそれらの実行とデプロイメントは、AI API の SAP AI Core 実装を使用して管理されます。SAP AI Core では、Argo ワークフローとサービステンプレートは `Executable` のコンセプトの下でマッピングされます。マッピングメカニズムが機能するには、Argo ワークフローおよび提供テンプレートで YAML ファイルの metadata セクションに特定の属性が必要です。これらの属性は、両方のテンプレートタイプで共有されます。

SAP AI Core では、ランタイム固有の追加 API が提供されます。これらは、AI API 仕様の拡張である AI Core API 仕様で提供されています。

関連情報

AI Core API
AI API

AI API ランタイム実装

AI API 仕様は、機械学習アーティファクトのライフサイクル管理の一般的な仕様です。SAP AI Core は、AI API 仕様の特定の実行時実装の 1 つです。SAP AI Core に関係なく、AI API 仕様の他のランタイム実装を提供することもできます。このセクションでは、必要な境界条件および実装要件について説明します。

AI API を使用する利点は、クライアントがすべての AI API 対応ランタイム実装と統合できることです。たとえば、SAP AI Launchpad は、同じ API が提供されている限り、カスタムランタイム実装を操作できます。インテリジェントシナリオライフサイクル管理は、AI API 対応のランタイムと統合することもできます。AI API クライアント SDK (Python) を使用することもできます (詳細については、SAP AI Core SDK を参照してください)。

AI API 仕様

AI API の仕様は、以下の部分で構成されます。

- 主要スペック
- 拡張:
    - 分析拡張
    - リソースグループ拡張
    - データセット管理拡張
        - メトリック拡張

→ 推奨事項

少なくとも主要仕様を実装してから、ユースケースに基づいて拡張仕様を実装します。

AI API ランタイム機能エンドポイント

`Meta API` は AI API 仕様の一部です (エンドポイント `/lm/meta`)。この実装は、AI API ランタイム実装の機能を指定する設定応答を返す必要があります。

`Meta API` により、AI API クライアントは、利用可能なコマンドまたはユーザインタフェースを選択できるように、AI API 実装の機能をクエリすることができます。たとえば、一部の AI API ランタイムでは実行が提供されますが、デプロイメントは提供されません。また、デプロイメントではなく、実行のログが提供されることもあります。たとえば、SAP AI Launchpad などの SAP AI Core のクライアントが SAP AI Core の `Meta API` エンドポイントをクエリする場合、応答は以下のようになります。

```
json

{
        "aiApi": {
                "capabilities": {
                        "logs": {
                                "deployments": true,
                                "executions": true
                        },
                        "multitenant": true,
                        "shareable": true,
```

```
                        "staticDeployments": true,
                        "timeToLiveDeployments": true,
                        "userDeployments": true,
                        "userExecutions": true
                        "executionSchedules": true
                },
                "limits": {
                        "deployments": {
                                "maxRunningCount": -1
                        },
                        "executions": {
                                "maxRunningCount": -1
                        },
                        "minimumFrequencyHour": 1,
                        "timeToLiveDeployments": {
                                        "minimum": "10m",
                                        "maximum": -1
                        }
                },
                "version": "2.18.0"
        },
        "extensions": {
                "analytics": {
                        "version": "1.0.0"
                },
                "metrics": {
                        "capabilities": {
                                "extendedResults": true
                        },
                        "version": "1.0.0"
                },
                "resourceGroups": {
                        "version": "1.2.0"
                }
        },
        "runtimeApiVersion": "2.21.0",
        "runtimeIdentifier": "aicore"
}
```


その後、SAP AI Launchpad およびその他のクライアントが適宜対応し、この AI API の実行時実装のユーザインタフェースでデプロイメントを非表示にすることができます。

以下のような機能があります。

機能	**true の場合、ユーザは以下の処理を行うことができます。**
`logs.executions`	実行のログの表示
`logs.deployments`	デプロイメントのログの表示
`multitenant`	SAP AI Launchpad をメインテナントユーザとして使用する (リソースグループをサポート)
`shareable`	クライアントは 1 つのインスタンスを共有可能

機能	true の場合、ユーザは以下の処理を行うことができます。
staticDeployments	推論のために常に実行される静的エンドポイントは、ユーザがデプロイメントを開始しなくても利用できます。
userDeployments	デプロイメントの停止、更新、または削除
userExecutions	実行の停止または削除
timeToLiveDeployments	ランタイムエンジンでは、デプロイメントが自動的に削除されるまでの時間を定義することができます。
analytics	すべてのテナントの概要情報のレビュー
bulkUpdates	最大 100 の実行またはデプロイメントを同時に停止または削除します。
executionSchedules	スケジュールの作成

制限には、以下が含まれます。

制限	詳細
deployments.maxRunningCount	リソースグループ内で同時に実行されるデプロイメントの数を制限する (存在する場合)
executions.maxRunningCount	リソースグループで実行中の同時実行の数を制限します (存在する場合)。
timeToLiveDeployments.minimum	デプロイメントの ttl パラメータの最小可能値 (サポートされている場合)
timeToLiveDeployments.maximum	デプロイメントの ttl パラメータの最大可能値 (サポートされている場合)
minimumFrequencyHour	実行のスケジュールに使用可能な最小値 (サポートされている場合)

一般的な AI API 仕様に加えて、追加のユースケースに対応する多数の拡張もあります。これらは、すべてのランタイムエンジンに実装されるわけではありません。

拡張は以下のとおりです。

内線	詳細
analytics	分析拡張には、リソースグループまたはテナントの分析情報をフェッチするためのエンドポイントが含まれています。
metrics	メトリック拡張には、実行時に生成されたメトリックを保存および取得するために、メトリックエンドポイントへの書込およびメトリックエンドポイントからの読込を行うためのエンドポイントが含まれています。
resourceGroups	リソースグループ拡張には、リソースグループを管理するためのエンドポイントが含まれています。
dataset	データセット拡張には、ファイルをアップロードおよびダウンロードするためのエンドポイントが含まれています。

関連情報

サービステンプレート
ワークフローテンプレート
AI API 仕様
Meta API を使用したカスタムランタイム機能
分析拡張
リソースグループ拡張
インテリジェントシナリオライフサイクル管理

リソースグループ

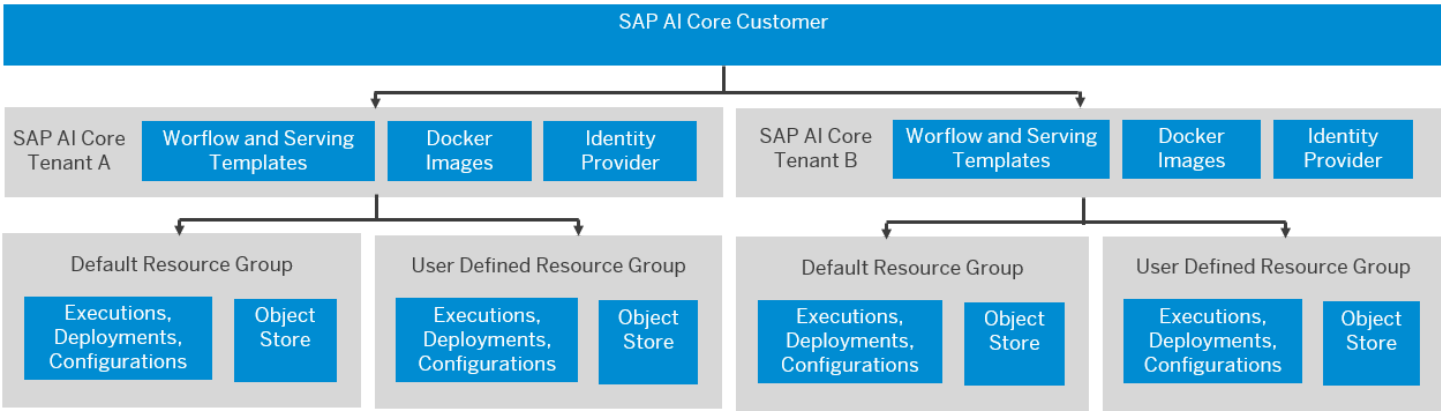
SAP AI Core テナントでは、リソースグループを使用して、関連する ML リソースおよびワークロードが分離されます。シナリオ、実行可能ファイル、および Docker レジストリシークレットは、すべてのリソースグループで共有されます。

リソースグループは、1 つの SAP AI Core テナントの範囲内の関連リソースの仮想コレクションを表します。テナントがオンボーディングされると、デフォルトのリソースグループが即時に作成されます。AI API を使用して、テナント管理者が追加のリソースグループを作成または削除することができます。テナントでは、対応する使用シナリオに基づいてリソースグループをマッピングできます。

SAP AI Core テナントでリソースグループを使用してシナリオコンシューマテナントが分離され、リソースグループが後で削除されると、シナリオコンシューマのプロビジョニングが解除されます。SAP AI Core では、テナントのシナリオコンシューマが認識されません。標準の XUSAA マルチテナンシーモデルに従います。

リソースの範囲

テナントおよびリソースグループで使用可能なリソースは、利用可能な範囲によって異なります。



テナントレベルのリソース

テナントレベルのリソースには、以下が含まれます。

- ワークフローテンプレート
- 対応テンプレート
- Docker レジストリ (Docker イメージを含む)
- ユーザ認証および権限 (UAA)

ユーザ認証および権限は、SAP AI Core テナントに基づきます。テナントは、SAP AI Core サービスキーを使用して取得されたアクセストークンの所有者です。SAP AI Core テナントでは、実行時に、または AI API を使用してライフサイクル管理中に、要求ヘッダにリ

ソースグループを設定することができます。リソースグループが設定されていない場合は、デフォルトのリソースグループが使用されます。

リソースグループレベルリソース

テナントレベルの実行可能ファイルは、すべてのリソースグループで共有されます。リソースグループレベルでは、リソースグループヘッダを設定することでオブジェクトストアが登録されます。

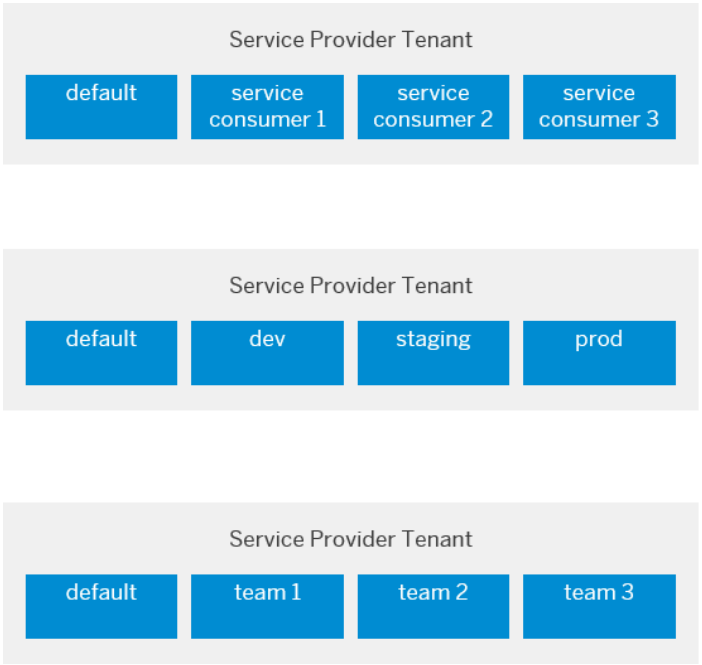
SAP AI Core テナントは、AI 機能の設計においてセキュリティの側面を考慮する必要があります。

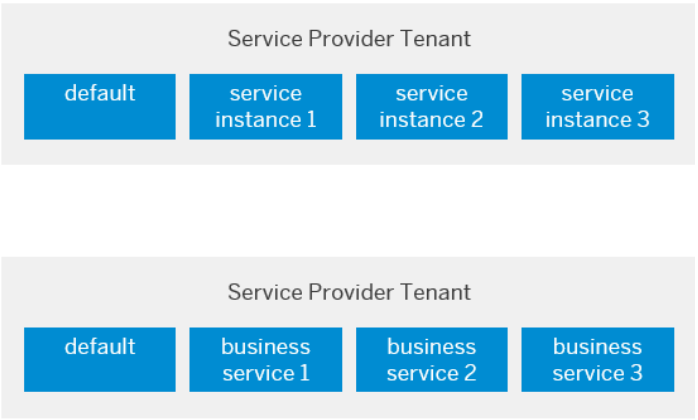
→ 推奨事項

複数のリソースグループに対して、AWS IAM ユーザが同じオブジェクトストアバケットを使用しないでください。

実行、デプロイメント、設定、アーティファクトなどのランタイムエンティティは、特定のリソースグループに属し、リソースグループ間で共有することはできません。

リソースグループマッピングの例





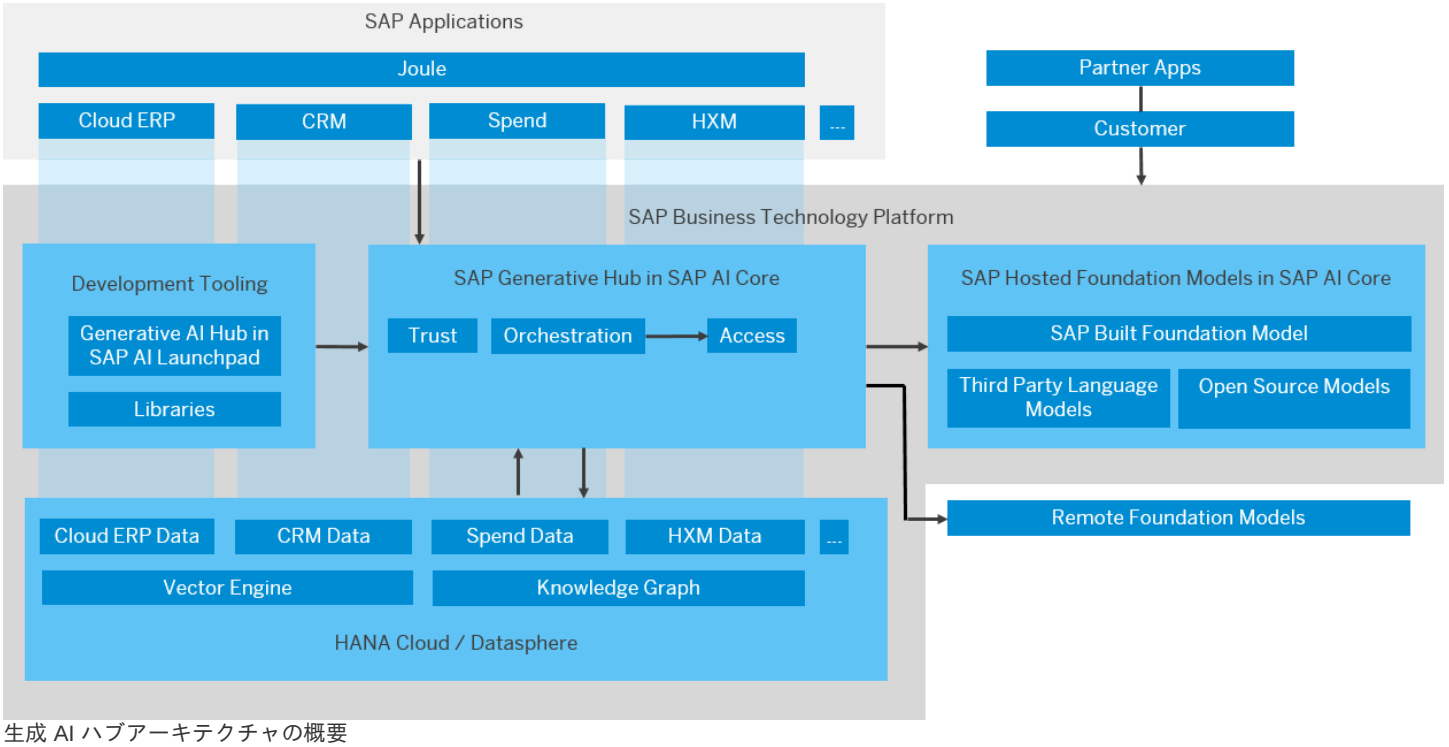
SAP AI Core の生成 AI ハブの概要

生成 AI ハブでは、SAP AI Core および SAP AI Launchpad の AI 活動に生成 AI が組み込まれています。

生成 AI モデルは、ラベルのない大量のデータでトレーニングされた自己教師ありのディープラーニングモデルです。AI テクノロジーと業界規模の計算リソースを使用して、複雑なパターンとセマンティックナレッジベースを学びます。これらのモデルは、自然言語処理 (NLP) などのタスクで優れています。プロンプトなどの入力を解析し、ターゲット単語を予測することで、コンテキストに関連する応答を自然言語で返します。1 つのモデルで、さまざまな入力書式と出力モードを使用して複数のタスクを処理することができます。

生成 AI モデルは設計上一般的ですが、追加の埋め込みを使用して微調整できます。これにより、専用またはドメイン固有のユースケースに適しています。

SAP AI Core と生成 AI ハブは、LLM と AI を新しいビジネスプロセスにコスト効率の高い方法で統合するのに役立ちます。



生成 AI ハブアーキテクチャの概要