



1. イントロダクション

Azure Machine Learning – 構築・運用編

Keita Onabuta

FastTrack for Azure
Senior Customer Engineer for AI/ML

Agenda

機械学習サービスの選択
Azure Machine Learning 基本

機械学習サービスの選択

基礎知識

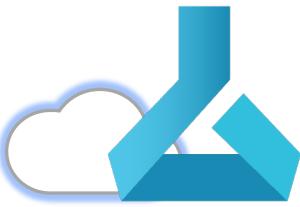
- Azure Machine Learning
- Azure Databricks
- Azure Data Science VM

ガイドライン・実装手順

- Azure Machine Learning と Azure Databricks

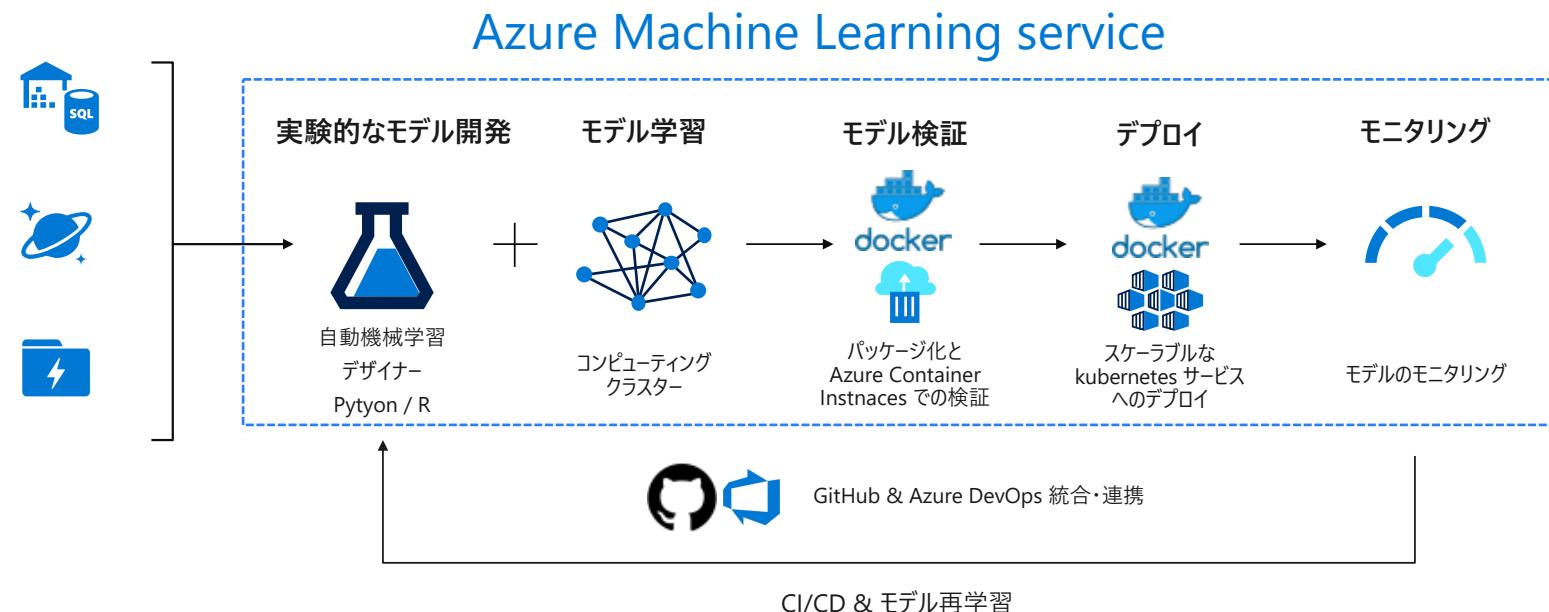
検討事項

- 利用する機械学習サービスの検討



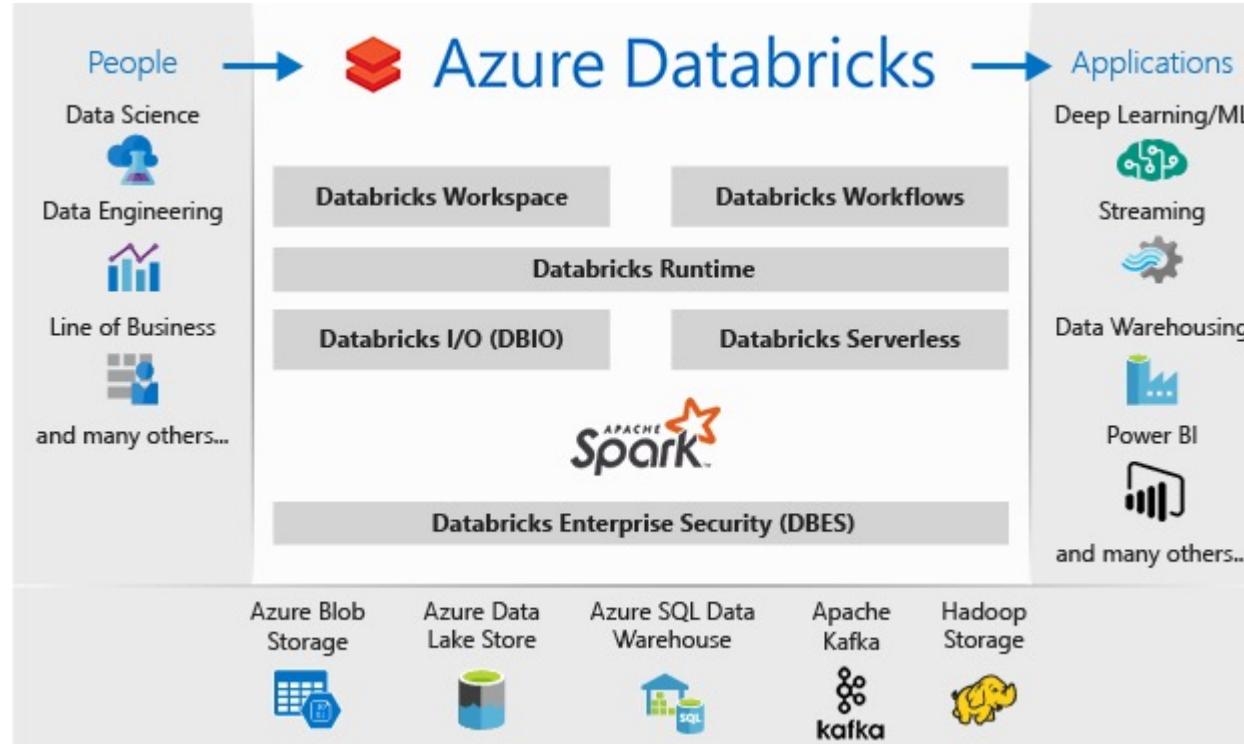
Azure Machine Learning

- ・ 機械学習プロセスをエンドツーエンドでサポートするマネージドサービス
 - ・ 必要なシステムモジュールをあらかじめビルトインしている
- ・ **自動機械学習やパラメータチューニング機能**による効率的なモデル開発
- ・ 繙続的なモデルのデプロイ & 運用管理をサポート
- ・ スケーラブルな計算環境による並列分散処理 etc



Azure Databricks

迅速、簡単で協調的な Apache Spark ベースの分析プラットフォーム



できること

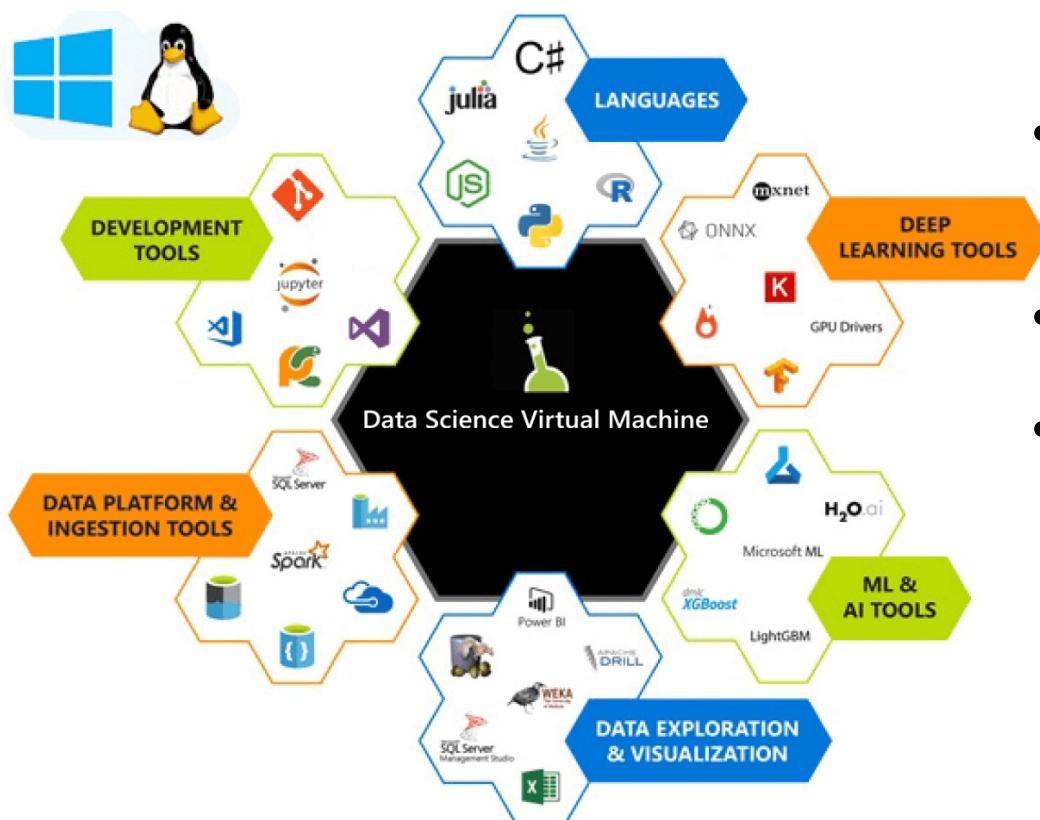
Azure Active Directoryと連携した認証・アクセス制御により、セキュアな環境での機械学習をインメモリ分散テクノロジー Spark をベースに高速に実現できます

お客様のメリット

Azure Active Directory と連携したクラウド内の完全に管理された Apache Spark クラスターを迅速に立ち上げ、インフラを意識することなく分析者は共同分析作業に集中できます。また、処理に応じて柔軟にスケールを変更することができ、これまで分析に要していた時間を短縮することにより様々なアプローチに取り組むことができます。

Azure Data Science VM

データ分析、機械学習、AI トレーニングでよく使用されるツールを事前インストール、構成、テストされた Azure 仮想マシンイメージ

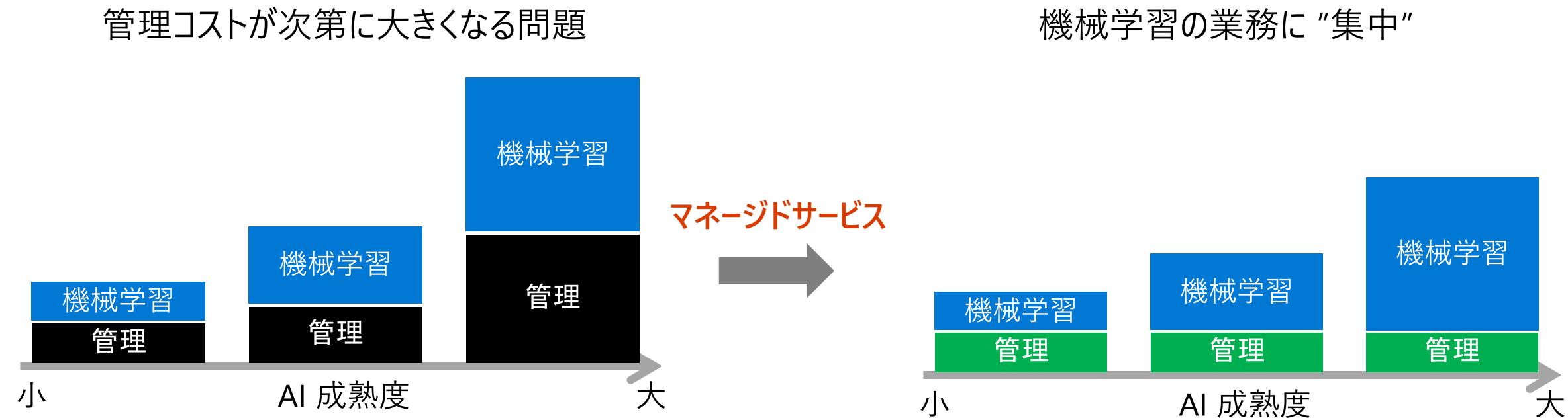


- 任意の VM インスタンスで稼働
 - ✓ GPU のドライバも同梱
- OS は Windows / Linux 選択から可能
- PyTorch / TensorFlow などの主要 Deep Learning フレームワークも事前インストール済み

<https://docs.microsoft.com/ja-jp/azure/machine-learning/data-science-virtual-machine/overview>

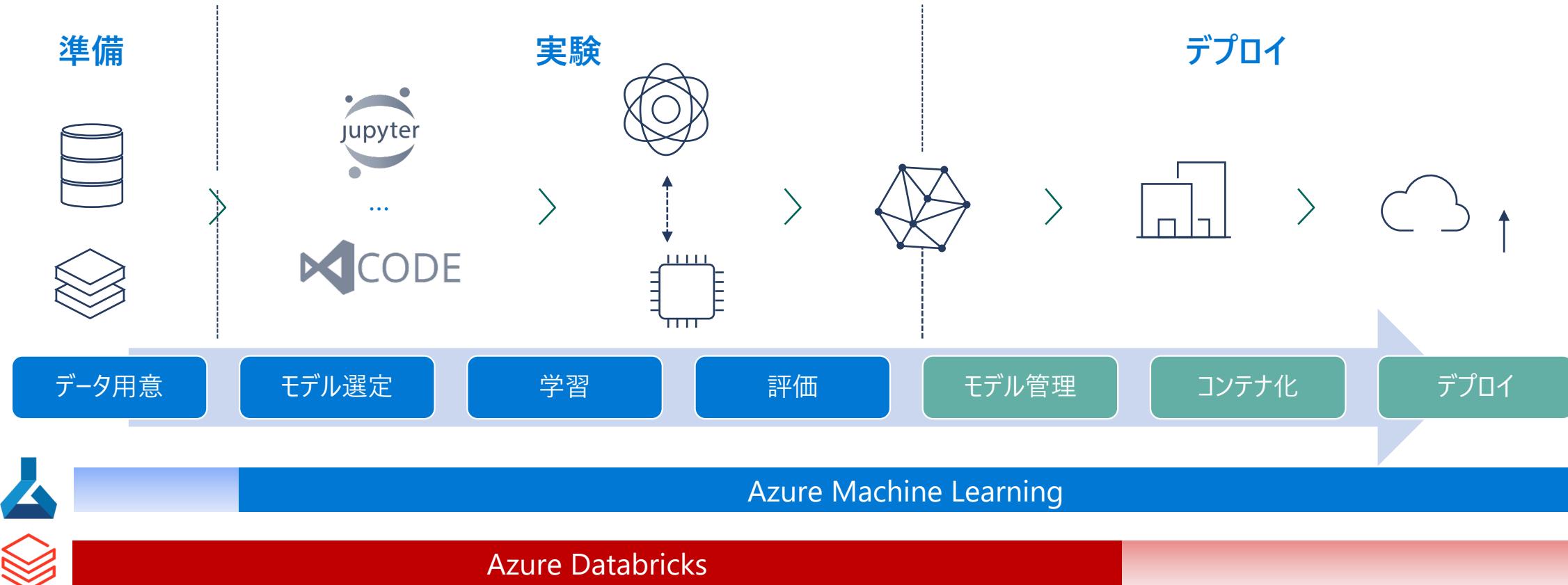
利用する機械学習サービスの検討

機械学習プラットフォームを構成する方法は多数あります。社内のエンジニアリソースが豊富にある場合はスクラッチで開発することも選択肢に上がりますが、機械学習を構成する要素は多岐に渡っており簡単ではありません。極力マネージドサービス (PaaS) を利用することで管理コストを軽減することができます。



Azure Machine Learning と Azure Databricks

Azure Databricks は Spark を利用した強力なデータの前処理をスムーズに行うことができます。Azure Machine Learning はデプロイに強みを持っており、様々なリソースとの連携が可能です。



Azure Machine Learning と Azure Databricks (cont'd)

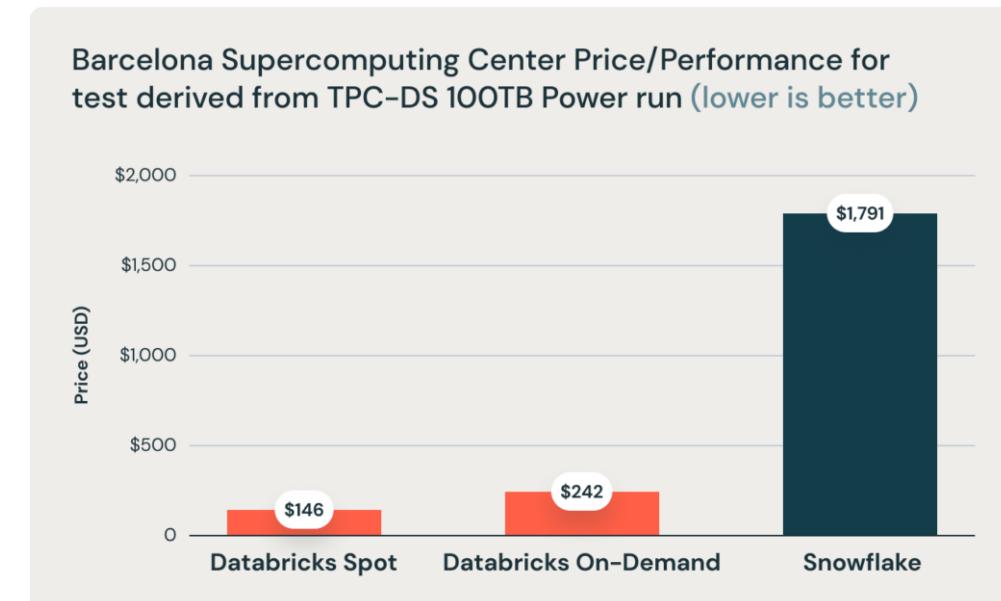
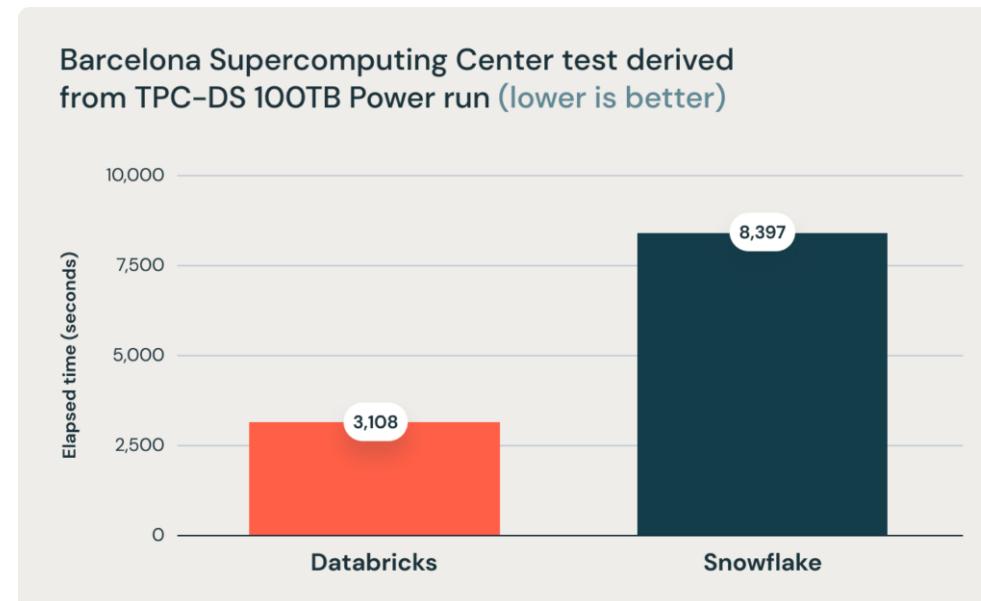
Azure ML の強み

- 開発環境の充実
 - 市場での認知度が高く実績が豊富な Jupyter, JupyterLab, R Studio, Visual Studio Code がシームレスに利用可能
 - オープンソーステクノロジーとの親和性が高くカスタマイズが可能
 - ローカル環境をスムーズに移行
- Azure Kubernetes Service との連携
 - Auto ML モデルを直接本番水準のクラスターにデプロイして API として利用可能
 - エントリスクリプトや設定ファイルの用意のみで任意のモデルを API としてデプロイ可
- パイプライン
 - Azure Data Factory パイプラインから Azure ML パイプラインを呼び出し可能
 - GUI を使用したパイプライン構築のサポート
 - デプロイまで含めたモデル作成の全工程のカバー
 - Python スクリプトやノートブックをパイプラインに組み込み可能

Azure Machine Learning と Azure Databricks (cont'd)

Azure Databricks の強み

- 大規模なデータに対する処理能力
 - Spark ベースのチューニングされた分散処理の仕組みによる圧倒的処理能力
 - ノートブックを使用してインタラクティブに分散処理の実行が可能



Azure Machine Learning 基本

基礎知識

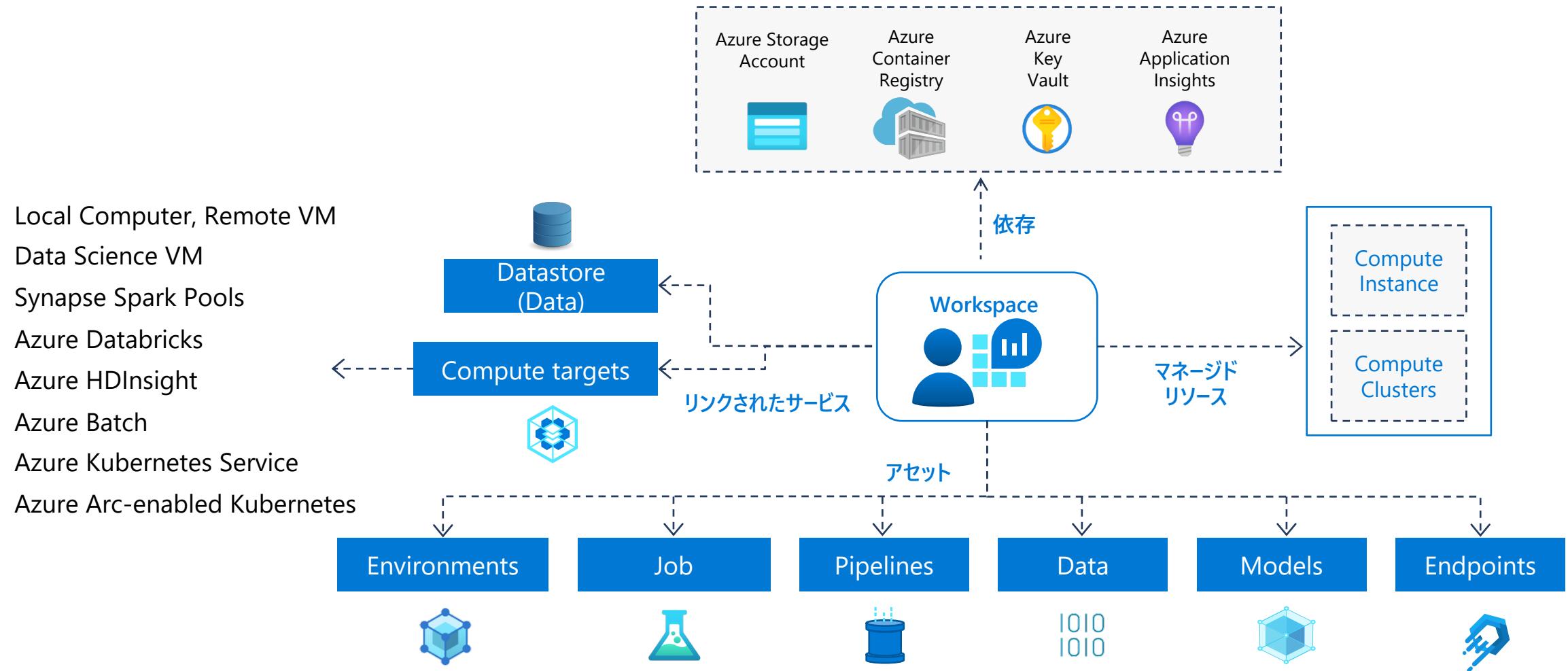
- Azure Machine Learning 基本構成

ガイドライン・実装手順

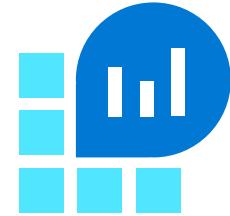
- ワークスペース構成のガイドライン
- クイックスタートテンプレート

Azure Machine Learning 基本構成

Azure Machine Learning Workspace は、最上位のリソースであり、あらゆるリソースを操作・運用管理するための一元的な環境を提供します。



Azure Machine Learning 基本構成 (cont'd)



Workspace



データ (Data)

モデル学習で利用するデータの運用管理



モデル (Models)

学習済み機械学習モデルを運用管理



コンピューティング (Computes)

モデル学習や推論で利用する計算環境



パイプライン (Pipelines)

モデル学習や推論のプロセスをパイプライン化



環境 (Environments)

Python パッケージ、Docker イメージを管理



エンドポイント (Endpoints)

クラウドやエッジデバイスにモデルをデプロイ



ジョブ (Job)

ジョブ実行とメトリックやログの管理



データラベリング (Data Labelling)

画像やテキストへラベルを付与



Workspace

- Workspace は Azure Machine Learning の最上位リソースである。Azure Machine Learning 使用時に作成したあらゆる成果物を集約した一元的な作業スペースを提供する。
- Workspace はモデル学習で利用した Compute Targets の情報やログ、メトリック、出力、スクリプトのスナップショットを含む学習の実行履歴を保持する。

- モデルは Workspace に登録される。
 - 複数の Workspace を作成することができ、それぞれの Workspace は複数人で共有することができる。
 - 新しく Workspace を作成したとき、自動的に以下 Azure リソースが作成される:
 - [Azure Container Registry](#)
 - 学習とデプロイに使用する Docker コンテナの登録に利用される
 - [Azure Storage](#)
 - Workspace のデフォルトの Datastore やログ管理などに使用される
 - [Azure Application Insights](#)
 - 推論エンドポイントをモニタリングした情報が保存される
 - [Azure Key Vault](#)
 - Compute Targets によって使用されるシークレットとその他 Workspace が使用するセンシティブな情報が保存される



Data

- Datastores は Azure のストレージサービスに対する接続情報を保持するために使用されます。
- Data asset、MLTable は Datastores を使用して接続された各種データソース内に保存されたデータを参照する。

Datastores

- Datastores は認証情報や元のデータソースの完全性を危険にさらすことなく接続に必要な情報を保持する。サブスクリプション ID や認証トークンといった接続に必要な情報は Workspace に紐づけられた Key Vault に保存しているため、スクリプトに機密情報を直接実装することなくストレージに対してセキュアにアクセスできる。

Data asset, MLTable

- Azure Machine Learning で利用するデータへのアクセスが容易になる。Data asset を作成するとデータへの参照とメタデータのコピーが作成される。データは元々データがあった位置に維持されるため、余計なストレージコストをかけず、データソースの一貫性を維持することができる。
- MLTable は Data asset の中でも表形式データに対応している。



Environments

Environments は機械学習を実行するソフトウェア環境の構築や運用管理の機能を提供する。

学習や推論スクリプトで利用する Python パッケージ、環境変数、ソフトウェアの設定、ランタイム（Python、Spark や Docker 等）で定義され、最終的には Docker コンテナとして実行される。

Workspace 上で管理・バージョニングされ、再現性、監査可能性、機械学習ワークフローの異なる計算リソース間での相互利用性を確保する

Environment はローカルコンピューター上で以下のように使用することが可能

- 学習スクリプトの開発
- モデルの学習をスケーリングする際、Azure Machine Learning の計算リソース上で同一環境の再利用
- 同一環境を使用したモデルのデプロイ
- 既存モデルの学習時に使用された環境の再現
- Workspace のデフォルト環境として利用可能な curated environments （代表的なシナリオに沿ってプリセットされた environment）の提供



Job

Job は Experiment と Run から構成され、モデル学習などのプロセスを実行したり、結果を管理します。

Experiment は1つのスクリプトに由来する複数の Run をグループ化したもので、Azure ML Workspace に所属する。各 Run の情報は関連付けられた Experiment の配下に保存される。

Experiment

スクリプトの実行によって生成された複数の Run をグルーピング

- Azure ML Workspace 直下に所属
- 各 Run の情報を保存し、横断的に管理

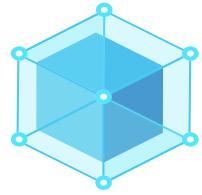
Run

Job を submit することで生成され、以下の情報を保持

- 実行のメタデータ (実行した日時、実行に要した時間等)
- スクリプトから出力したメトリック
- 実験に関する自動収集された、もしくは明示的にアップロードされたファイル
- スクリプトを含むディレクトリの実行直前のスナップショット

Run configuration

計算リソース上でスクリプトがどのように実行されるべきかの定義



Models

Model は学習を経て生成される artifact* である。Model は新しいデータに対する予測結果を得るために使用する。

Model Registry は Azure Machine Learning Workspace 上の全ての Model 追跡する。

* 機械学習の過程で生じる中間生成物、画像、モデル等のあらゆるファイルを artifact と呼称する

Model

Model は Azure Machine Learning 上の Run から生成される

- Note: Azure Machine Learning の外部で学習したモデルを使用することも可能

Azure Machine Learning はフレームワーク非依存であるため Model を作成する際にあらゆる機械学習フレームワークを使用可能

Model は Azure Machine Learning の Workspace 配下に登録可能

Azure Machine Learning に作成された Model はバージョニングされ、モデルのバージョンを比較・選択するためのツールが組み込みで付属

一度 Model が作成されれば、Dockerfile や Docker イメージの生成とあらゆる場所へのモデルのデプロイが可能

Model Registry

Azure Machine Learning Workspace 上の全モデルを追跡

Model は名前とバージョンによって一意に特定される

Model を登録するときに追加でメタデータタグを付与することができ、タグは Model の検索時に利用可能

Docker イメージに使用されている Model は削除できない



Pipelines

1つの Azure Machine Learning Pipeline は 完成した機械学習タスクのワークフローを表し、 機械学習の各サブタスクはパイプラインを構成 数する一連の手順として内包される

Azure Machine Learning Pipeline には、Python スクリプトを呼び出すだけのシンプルなものから、あらゆることを実行するものまで存在する。

タスク

Pipelines は以下のようないくつかの機械学習タスクにフォーカス:

- インポート、検証とクリーニング、変換、正規化、ステージングを含む「データの準備」
- パラメーター化された引数、ファイルのパス、ログとレポート出力の設定を含む「学習の設定」
- 効率的かつ繰り返し実行できる「学習と検証」。特定のデータサブセット、異なるハードウェアからなる計算リソース、分散処理、進捗状況の監視を指定することで、効率性を確保することが可能
- バージョン管理、スケーリング、プロビジョニング、アクセス制御等の要素を含む「デプロイ」



Computes

Computes は学習スクリプトを実行したりモデルを推論用途でホストするための計算リソースです。

Azure Machine Learning Python SDK, CLI, Studio から作成し運用管理することができます。

既存のリソースをアタッチすることもできます。

ローカル環境で実行したのちに、Compute Targets 上でスケールアップ・スケールアウトすることができます。

現在サポートされている Compute Target

Compute Targets	学習	デプロイ
Local Computer	✓	
A Linux VM in Azure (such as the Data Science Virtual Machine)	✓	
Azure ML Compute Clusters	✓	✓
Azure ML Compute Instance		
Azure Databricks	✓	
Azure Data Lake Analytics	✓	
Azure HDInsight	✓	
Azure Container Instance		✓
Azure Kubernetes Service		✓
Azure IoT Edge	✓	
Field-programmable gate array (FPGA)		✓
Azure Functions (preview)		✓
Azure App Service (preview)		✓
Azure Synapse Spark Pool (preview)	✓	
Azure Arc enable Kubernetes (preview)	✓	✓

ワークスペース構成のガイドライン

チームやプロジェクトで Azure ML を利用する場合、企業で 1 つの Azure ML Workspace では無く、下記の観点から複数の Workspace を構築・運用管理することが推奨です。

ポイント

□ セキュリティ

- Azure ML Workspace 内部で細かいアクセス制御ができない (2022年4月時点)。
- データソースに対しては Identity ベースのアクセス設定が可能。
- Notebooks は File Share にコード・ノートブックを保持しており、アクセス制御に対応していない。

□ リージョン

- チームやプロジェクトの拠点が地理的に散在しているケースにおいてはアクセス速度や法的な理由から Workspace をリージョン毎に構築することが望ましいケースがある。

□ ワークロード

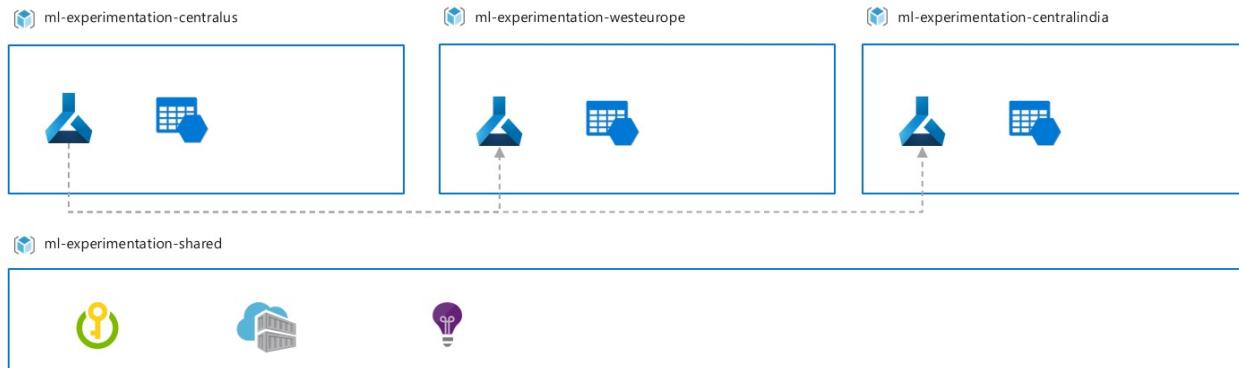
- 開発環境、テスト環境、本番環境などワーカロードに応じて Workspace を分ける。

ワークスペース構成のガイドライン (cont'd)

複数ワークスペースの運用イメージ



- ・ 開発環境、QA環境、本番環境
- ・ 個人 or プロジェクト単位



- ・ 個人 or プロジェクト単位

クイックスタートテンプレートの活用

Bicep もしくは Terraform のテンプレートが準備されており、
クイックにセキュアなリソースを作成することが可能です。

- ベストプラクティスが実装されたパイロット環境がクイックに作成可能
- テンプレートは要件に応じてパラメータ変更が可能
- 利用しているクラウド環境が Azure のみの場合は、Bicep テンプレートの利用を推奨
- マルチクラウド環境の場合は、Terraform テンプレートを利用するケースが多い

Bicep → [Bicep/ARM quickstart](#)

Terraform → [Terraform quickstart](#)

The screenshot shows the Azure portal interface for a 'Machine Learning end-to-end secure setup' quickstart template. At the top, there's a navigation bar with 'Microsoft Azure (Preview)', a search bar, and a dashboard link. The main title is 'Azure Machine Learning end-to-end secure setup' with a subtitle 'Azure quickstart template'. Below the title, there are tabs for 'Basics' (which is selected) and 'Review + create'. A 'Template' section shows a preview icon and 'machine-learning-end-to-end-secure' with 9 resources. There are buttons for 'Edit template', 'Edit parameters', and 'Visualize'. The 'Project details' section allows selecting a subscription ('FTA keonabut - Azure CXP Internal') and a resource group ('Create new'). The 'Instance details' section contains fields for Region ('East Asia'), Prefix (''), Location ('[resourceGroup().location]'), Tags (''), Vnet Address Prefix ('192.168.0.0/16'), Training Subnet Prefix ('192.168.0.0/24'), Scoring Subnet Prefix ('192.168.1.0/24'), Azure Bastion Subnet Prefix ('192.168.250.0/27'), Deploy Jumphost ('true'), Dsvm Jumpbox Username (''), and Dsvm Jumpbox Password ('*****'). The 'Aml Compute Public Ip' field also has 'true' selected.

[テンプレートを使用してセキュリティで保護されたワークスペースを作成する - Azure Machine Learning | Microsoft Docs](#)



Microsoft AI





© Copyright Microsoft Corporation. All rights reserved.