AWS AI/ML@TOKYO

AWS AI/ML サービスアップデート

2022年の注目アップデートをご紹介

田原 慎也

技術統括本部 ソリューションアーキテクト
アマゾン ウェブ サービス ジャパン合同会社

© 2022, Amazon Web Services, Inc. or its affiliates.

自己紹介

田原 慎也 (Shinya Tahara)

- ソリューションアーキテクト

  AWS のサービスをこれから本格的に活用していく
  フェーズのお客様に対して技術的な支援を実施

- 好きな AWS サービス

機械学習プラットフォーム
Amazon SageMaker

量子コンピューティング
Amazon Braket

© 2022, Amazon Web Services, Inc. or its affiliates.

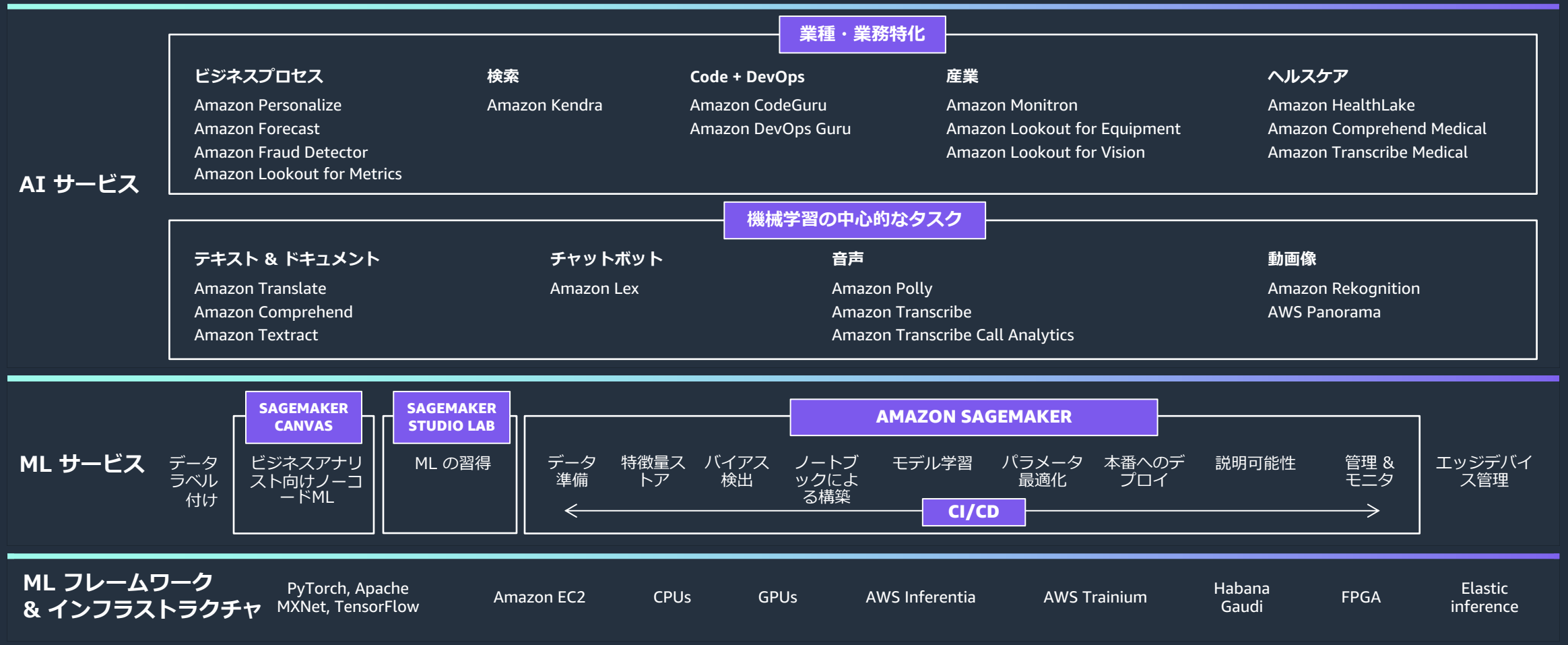
Agenda

1. AWS AI サービスの注目アップデート - 5つ
2. AWS ML サービスの注目アップデート - 7つ

© 2022, Amazon Web Services, Inc. or its affiliates.

AWS の機械学習スタック

Broadest and most complete set of machine learning capabilities



© 2022, Amazon Web Services, Inc. or its affiliates.

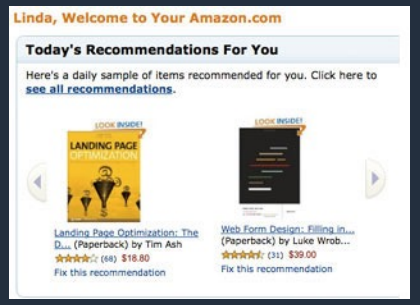
AWS の機械学習スタック

Broadest and most complete set of machine learning capabilities

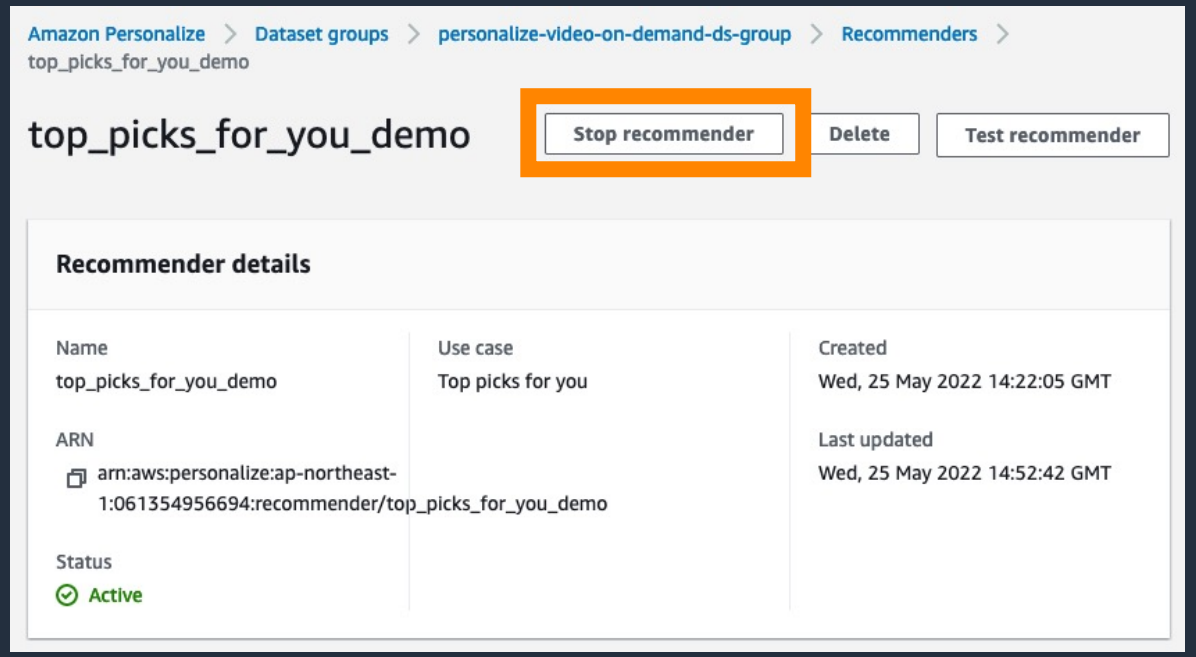
© 2022, Amazon Web Services, Inc. or its affiliates.

Personalize がレコメンダーの停止と開始をサポート

- パーソナライズされたユーザエクスペリエンスを提供する Amazon Personalize において、レコメンダーを停止および開始の制御ができるようになりました。
- 停止中は料金がかからないため、レコメンデーションが不要な場合に停止させることでコストを抑えることができます。



商品推薦エンジン

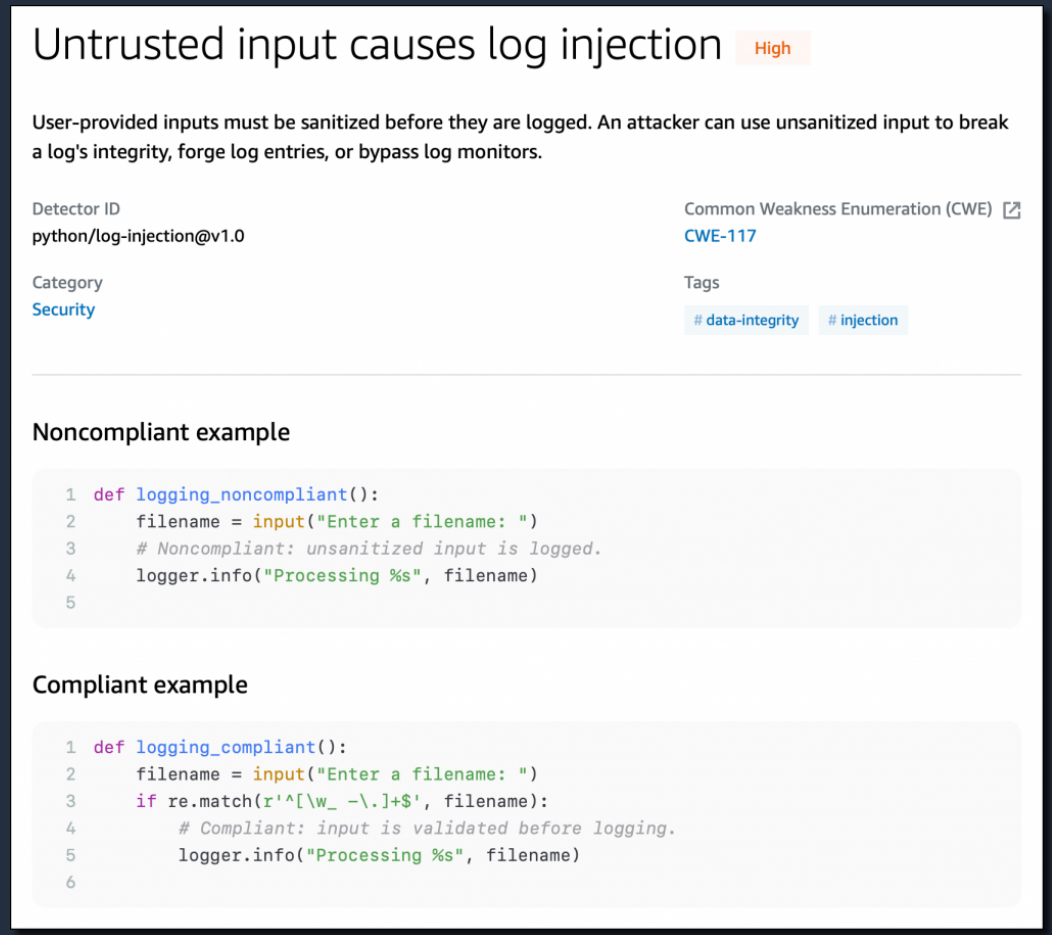


https://aws.amazon.com/jp/about-aws/whats-new/2022/04/amazon-personalize-starting-stopping-recommenders/

© 2022, Amazon Web Services, Inc. or its affiliates.

CodeGuru Reviewer がログインジェクションの脆弱性検出を開始

- Amazon CodeGuru Reviewer は機械学習を使ってコードのセキュリティの脆弱性を検出、改善をレコメンデーションするサービスです。
- 今回、ログインジェクション攻撃に利用される可能性がある、ロガー回りの処理を発見する新しいログインジェクションディテクタが追加されました。JavaとPythonに対応しています。

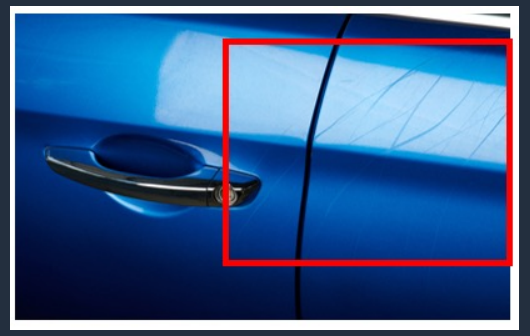


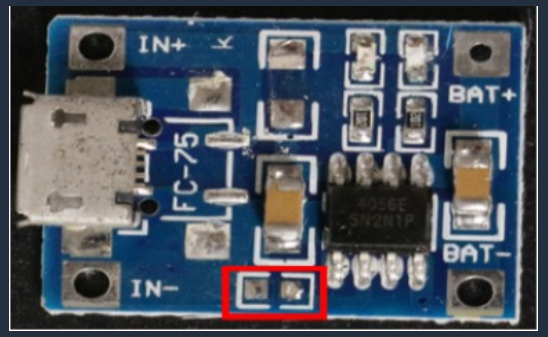
https://aws.amazon.com/jp/about-aws/whats-new/2022/02/amazon-codeguru-reviewer-detects-apache-log4j/

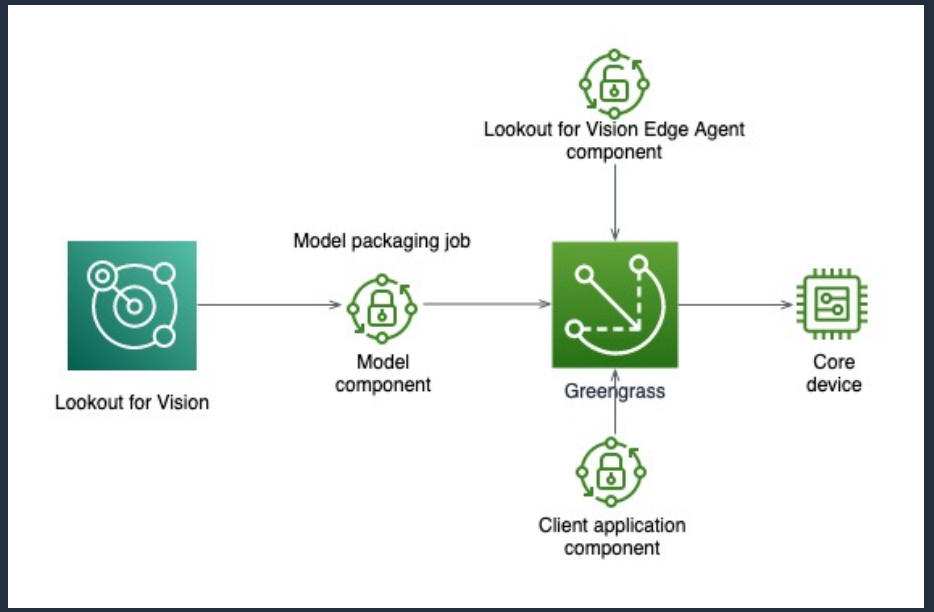
© 2022, Amazon Web Services, Inc. or its affiliates.

Lookout for Vision のエッジでの使用が一般提供開始

- Amazon Lookout for Vision はコンピュータービジョン (CV) を使用して視覚表現（画像）上で欠陥や異常を発見するサービスです。
- Amazon Lookout for Vision のエッジデバイス（AWS IoT Greengrass 等)での異常検知が一般提供開始（GA）になりました。東京リージョンでもご利用いただけます。







https://aws.amazon.com/jp/about-aws/whats-new/2022/03/automate-visual-inspection-product-defects-amazon-lookout-vision/

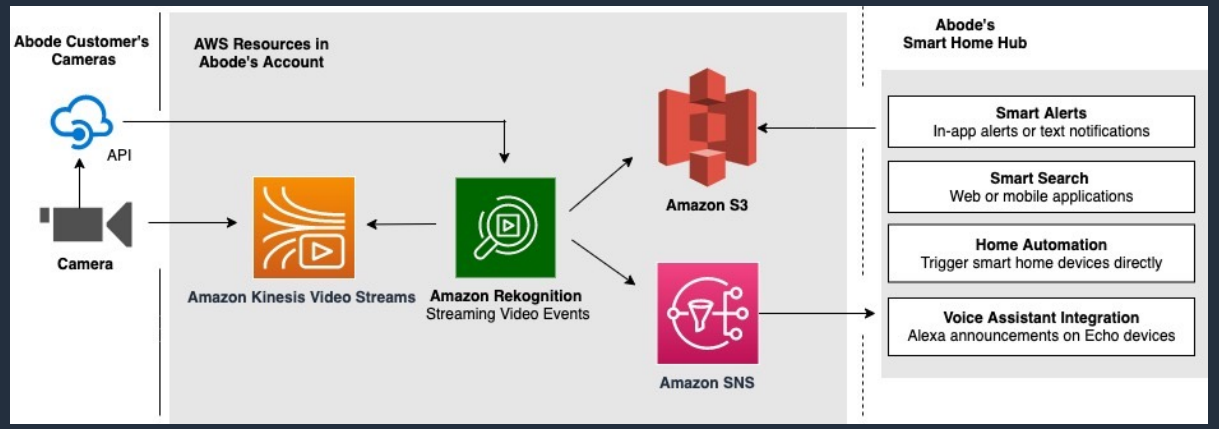
© 2022, Amazon Web Services, Inc. or its affiliates.

Rekognition でライブ動画内のオブジェクト検出ができる Streaming Video Events が利用開始

- 画像や動画の分析ができる Amazon Rekognition で、ライブ動画ストリーミング内にある目的のオブジェクトを検出しリアルタイムアラートを出すことのできる Streaming Video Events が利用可能になりました。

ユースケース

- 「裏庭にペットがいます」等のスマート通知
- 人の動きが確認されたときに車庫の照明をつける等のホームオートメーション
- 宅配業者を認識して「荷物が表玄関に配達されました」等の Alexa 音声サポート連動

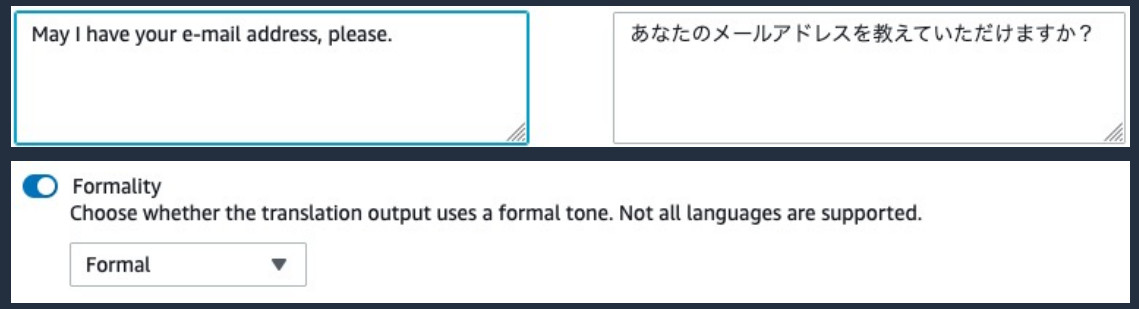


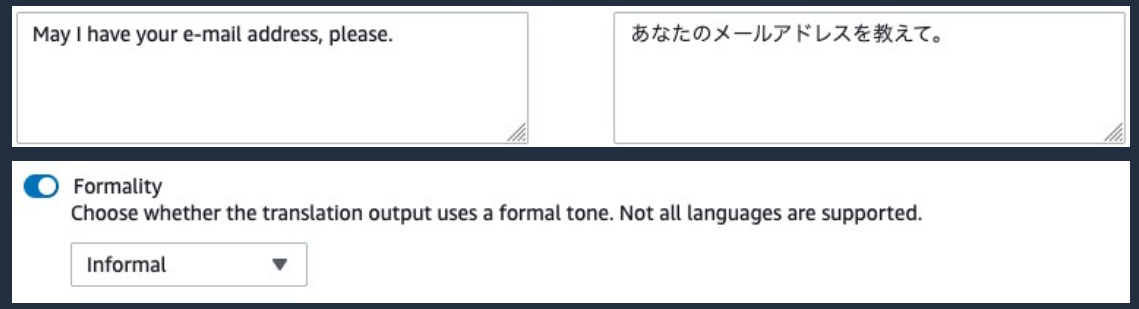
https://aws.amazon.com/jp/about-aws/whats-new/2022/04/amazon-rekognition-streaming-video-events-real-time-alerts-live-video-streams/

© 2022, Amazon Web Services, Inc. or its affiliates.

Translate でフォーマル度のカスタマイズが可能に

- 言語には「表現の丁寧さ」の概念があり、保険やヘルスケア業界ではより丁寧でフォーマル度の高い表現が、ゲームやソーシャルメディアではより親しみやすい、フォーマル度の低い表現が好まれる傾向があります。
- Amazon Translate が出力する訳文についてフォーマル度を選択し、カスタマイズできるように。日本語をはじめ6つの言語でご利用頂けます。



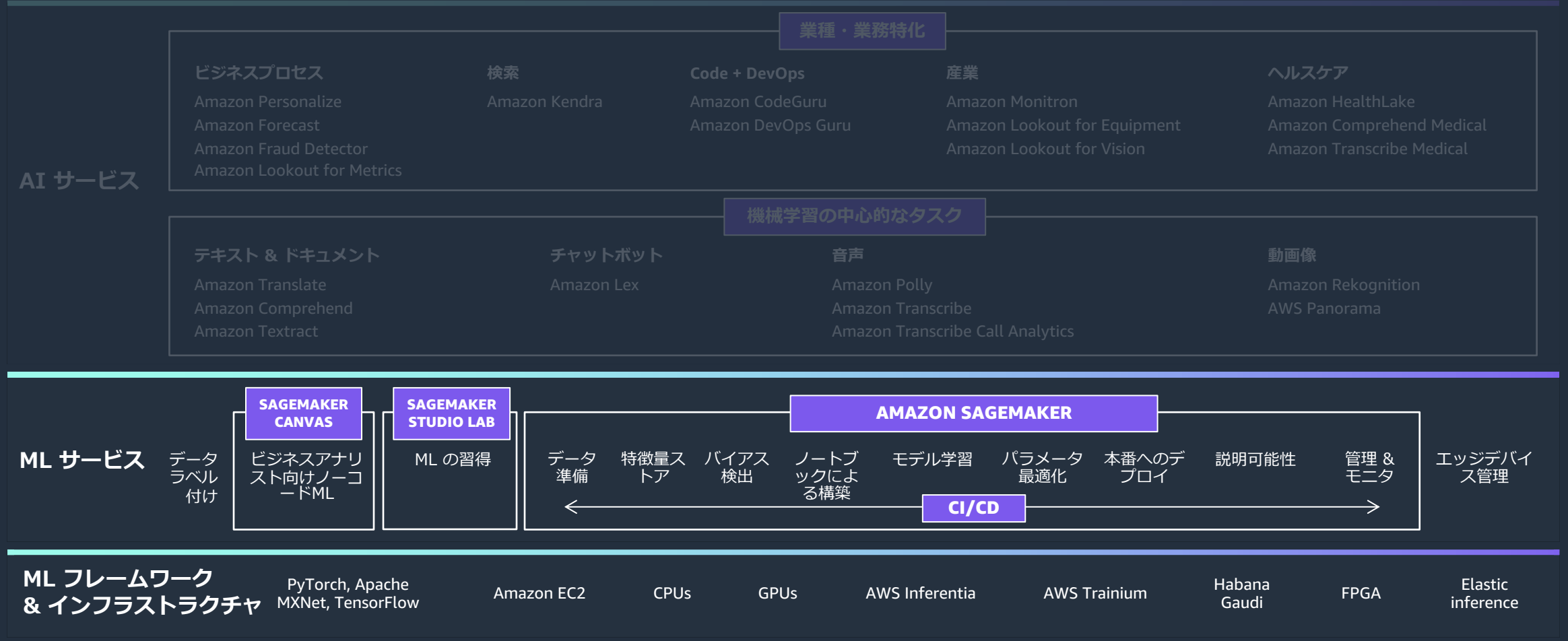


https://aws.amazon.com/jp/about-aws/whats-new/2022/02/amazon-translate-formality-customization/

© 2022, Amazon Web Services, Inc. or its affiliates.

AWS の機械学習スタック

Broadest and most complete set of machine learning capabilities



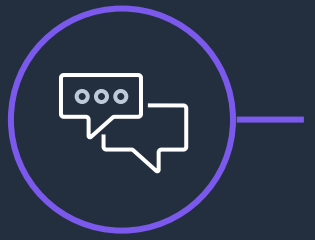
© 2022, Amazon Web Services, Inc. or its affiliates.

Amazon SageMaker Canvas

ML モデルを構築し、正確な予測を生成
ー コーディングは不要

機械学習のためのデータへの迅速なアクセスと準備

組み込み AutoML でモデルを構築し、正確な予測を生成



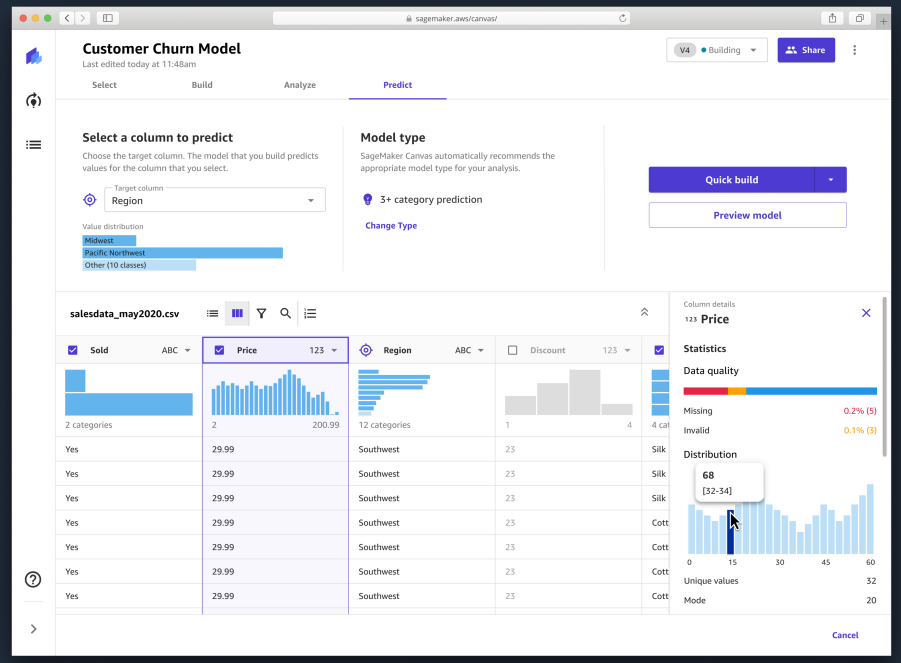
ML モデルの共有とデータサイエンスチームとの連携

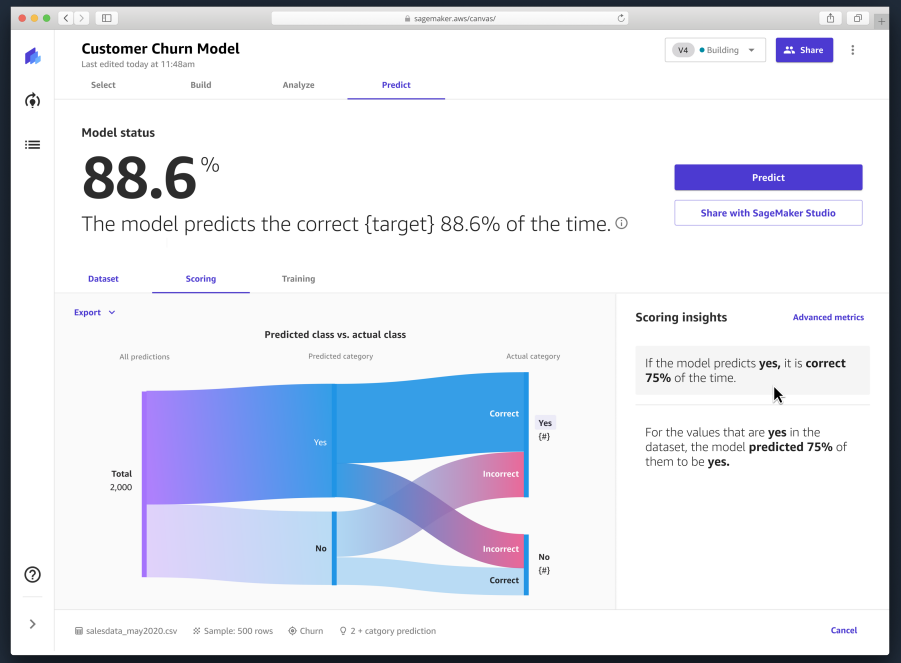
ライセンス料を回避し、TCO を削減するための使用量ベースの価格設定

© 2022, Amazon Web Services, Inc. or its affiliates.

SageMaker Canvas が東京リージョンで利用可能に

- ビジネスアナリスト向けのノーコードの機械学習環境である Amazon SageMaker Canvas が東京リージョンで利用できるようになりました。





https://aws.amazon.com/jp/about-aws/whats-new/2022/04/amazon-sagemaker-canvas-available-asia-pacific-tokyo/

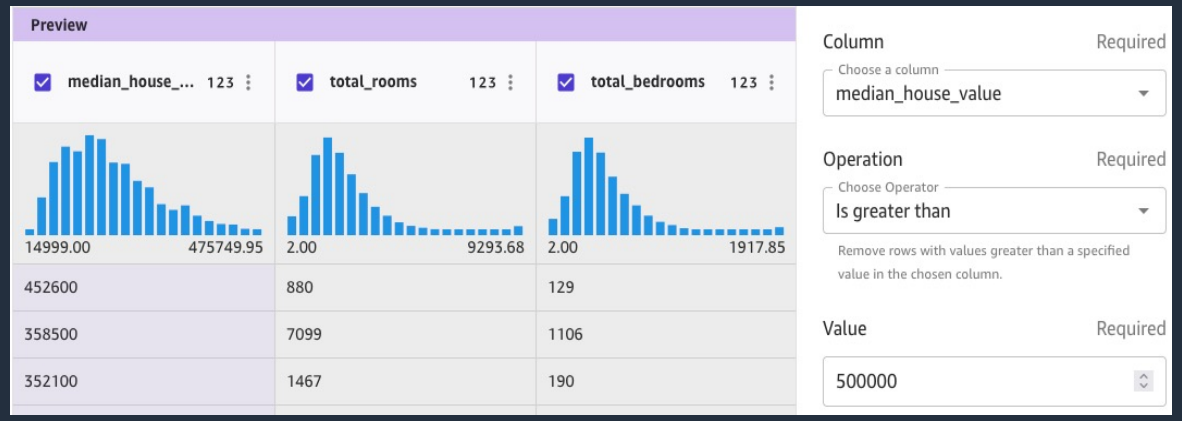
© 2022, Amazon Web Services, Inc. or its affiliates.

SageMaker Canvas の新しいデータ準備機能が追加

- Amazon SageMaker Canvas でモデルを構築する前にデータセットに対して新しいデータ準備機能が追加されました。

行のフィルタリング

欠損値や外れ値をプレビューして削除することができます。また、条件を指定した行の削除もできるようになりました。



タイムスタンプの変換と抽出

タイムスタンプ列で複数フォーマットをサポートし、その列から日付と時刻の情報の抽出し新しい列の作成ができます。

https://aws.amazon.com/jp/about-aws/whats-new/2022/05/amazon-sagemaker-canvas-adds-new-data-capabilities-usability-updates/

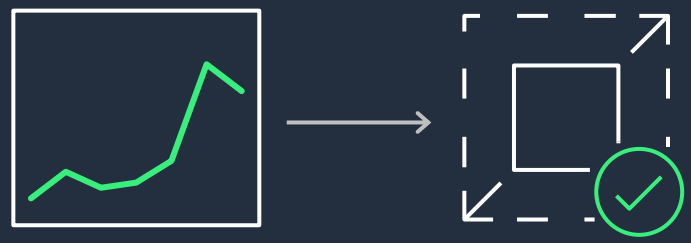
© 2022, Amazon Web Services, Inc. or its affiliates.

SageMaker Serverless Inference の一般提供開始

- インフラの構成・管理をする手間なく、リアルタイム推論を実行できる Amazon SageMaker Serverless Inference が一般利用開始になりました。
- 推論コードの実行時間と処理されたデータ量に応じた料金が発生します。推論リクエストがない場合は、料金は発生しません。

さらに、以下の機能強化がされています。

- ✓ SageMaker Python SDK サポート
- ✓ SageMaker Model Registry サポート
- ✓ 最大同時実行数の引き上げ（50→200）



推論リクエストに応じたスケーリングを実現

https://aws.amazon.com/jp/about-aws/whats-new/2022/04/amazon-sagemaker-serverless-inference/

aws

© 2022, Amazon Web Services, Inc. or its affiliates.

（補足）推論処理の選択肢

Real-time inference
リアルタイム推論

低レイテンシー

高スループット

マルチモデルエンドポイント

A/B テスト

Batch transform
バッチ変換

大量の推論データの処理

ジョブ形式の実行

New

Serverless Inference
サーバーレス推論

マネージドな推論処理基盤

サーバーレス

管理不要なスケーリングポリシー

Asynchronous inference
非同期的推論

ニアリアルタイム

1 GB までのペイロード

15 分までの実行時間

© 2022, Amazon Web Services, Inc. or its affiliates.

（補足）Serverless Inference のデプロイ方法

1

推論コードが含まれている ECR image の場所を指定

2

モデルアーティファクトが置かれている S3 の場所を指定

3

メモリ量を指定

Serverless Inference Endpoint

自動的にリソースを構築し管理する

負荷の応じて、自動的にスケール

モニタリングやロギングはサービスが管理

リクエスト ⟵

レスポンス ⟶

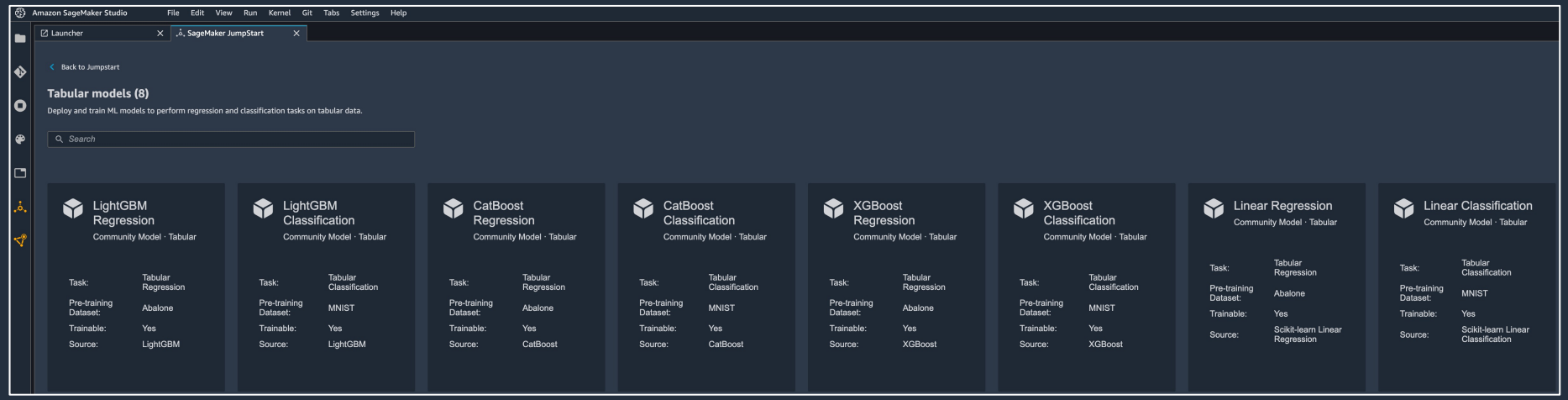
クライアントアプリケーションや他の AWS サービスからトリガー

エンドユーザー

© 2022, Amazon Web Services, Inc. or its affiliates.

SageMaker JumpStart にて表形式データに対するモデルとして LightGBM と CatBoost が追加

- Amazon SageMaker JumpStart は300 以上の定評のあるモデルコレクションと17のエンドツーエンドソリューションにワンクリックでアクセスし、すぐに機械学習を利用開始できるように支援するサービス。
- 今回、表形式のデータに対してLightGBM、CatBoost、XGBoost、Linear Learnerを含むよく知られたモデルとアルゴリズムが追加されました。



https://aws.amazon.com/jp/about-aws/whats-new/2022/01/amazon-sagemaker-jumpstart-lightgbm-catboost-tabular/

© 2022, Amazon Web Services, Inc. or its affiliates.

SageMaker Autopilot がサポートするデータセットやファイル形式の拡大

- Amazon SageMaker Autopilot は、データに基づいて最適化機械学習モデルを自動的なアプローチ（AutoML）で構築、トレーニング、調整します。
- アップデートにより、以下のデータセットに対応するようになりました。
  - ✓ Apache Parquet ファイル形式
  - ✓ 最大100GB までのサイズ
    ※こちらデフォルトのサイズ制限のため引き上げリクエストは可能です。
    以前はデフォルト最大10GBまで。

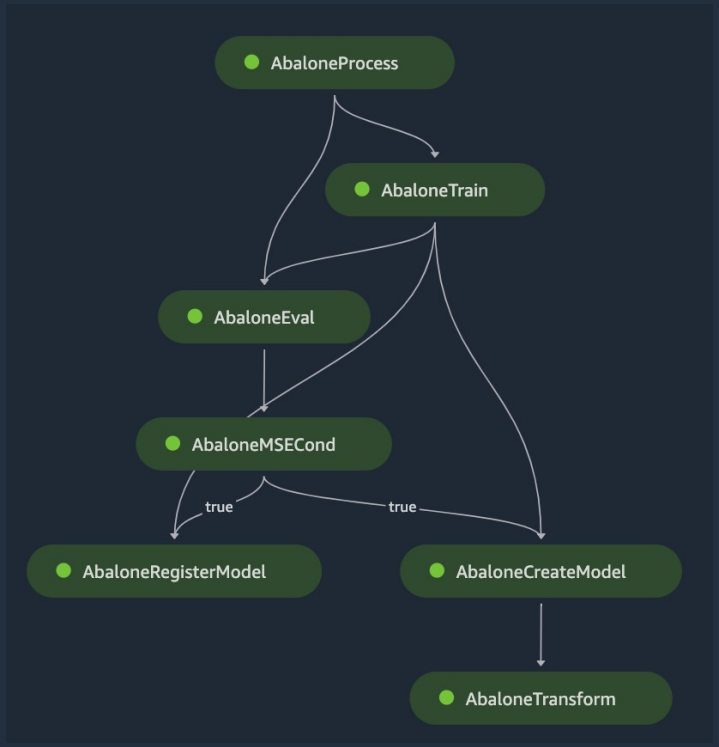
```
"InputDataConfig": [
    {
        "ContentType": "x-application/vnd.amazon+parquet",
        "CompressionType": "None",
        "DataSource": {
            "S3DataSource": {
                "S3DataType": "S3Prefix",
                "S3Uri": f"s3://{bucket}/{prefix}/abalone.parquet"
            }
        },
        "TargetAttributeName": "Rings",
    }
],
```

https://aws.amazon.com/jp/about-aws/whats-new/2022/01/amazon-sagemaker-autopilot-apache-parquet/
https://aws.amazon.com/jp/about-aws/whats-new/2022/01/amazon-sagemaker-autopilot-datasets-up-to-100-gb/

© 2022, Amazon Web Services, Inc. or its affiliates.

SageMaker Pipelines で同時実行制御をサポート

- Amazon SageMaker Pipelines は、機械学習モデル構築のために継続的インテグレーション及び継続的デリバリー（CI/CD）のフルマネージドサービス。
- モデル構築のワークフローにおける処理（ステップ）に対して、並列で実行できる同時実行制限を設定できるようになりました。



例）最大同時実行ステップ数を50に制限する

```
pipeline.create(role,
                parallelism_config={"MaxParallelExecutionSteps": 50})
```

https://aws.amazon.com/jp/about-aws/whats-new/2022/01/amazon-sagemaker-pipelines-concurrency-control/

© 2022, Amazon Web Services, Inc. or its affiliates.

プレビュー

Amazon EC2 Trn1 インスタンスがローンチ予定

- AWS による第２世代 独自設計機械学習推論チップ **AWS Trainium** を搭載
  - 最大16の Trainium アクセラレータ、512GB の HBM2 アクセラレータメモリを搭載
  - 最大800Gbps の EFA ネットワーク帯域、 768GB/s の高速 NeuronLink チップ間接続
  - EC2 UltraCluster にデプロイされ、最大で数万の Trainium にスケール可能
- Trn1上で学習してデプロイ先は自由、PyTorch 及び TensorFlow をネイティブにサポート
- 自然言語処理、画像認識、レコメンデーション、インテリジェント検索といった用途で、クラウドでの深層学習モデルのトレーニングにおいて最高の価格性能を提供します。
- 2022年6月現在プレビュー中で、2022年にローンチ予定となっています。

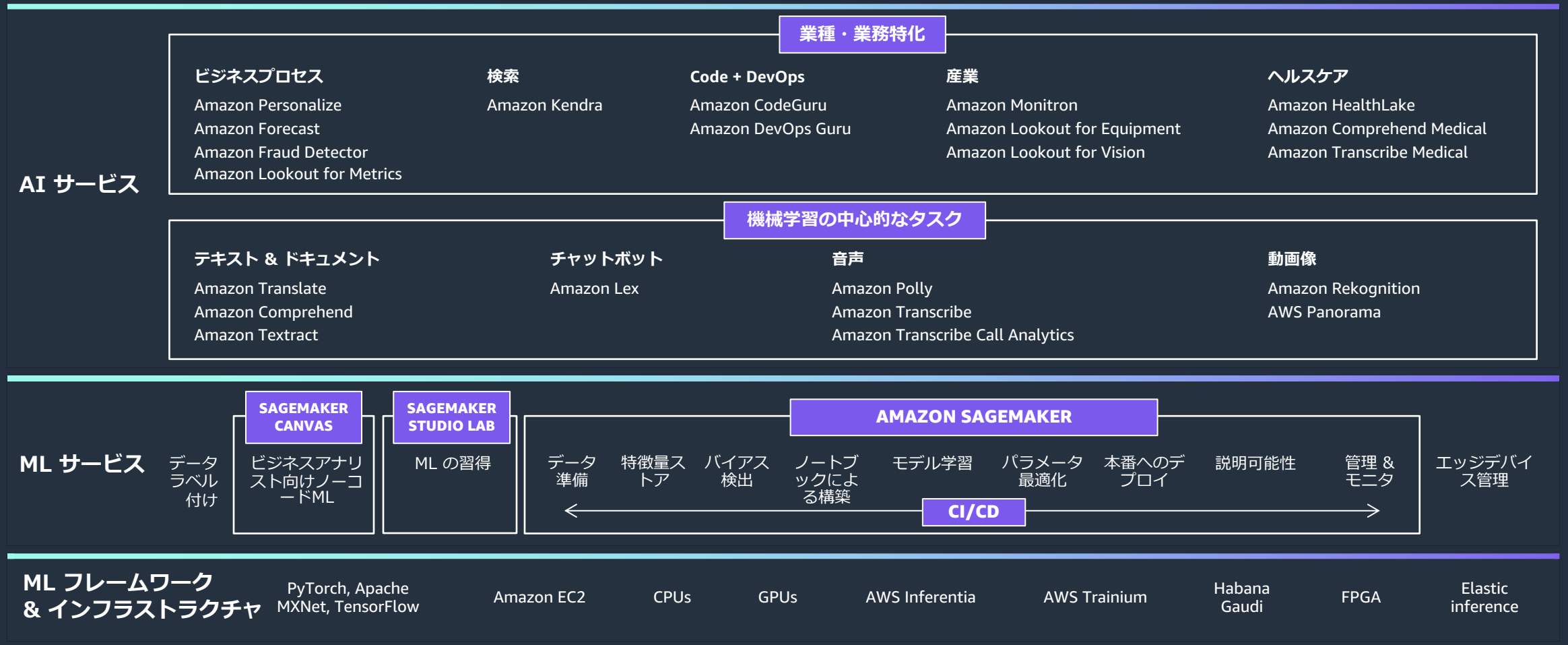
インスタンスサイズ	Trainium チップ数	アクセラレータメモリ (GB)	vCPU	メモリ (GB)	NeuronLink (GB/s)	NW帯域 (Gbps)	NVMe SSD (TB)
Trn1.2xlarge	1	32	8	32	N/A	最大 10	0.5
Trn1.32xlarge	16	512	128	512	768	800	8

https://aws.amazon.com/jp/ec2/instance-types/trn1/

© 2022, Amazon Web Services, Inc. or its affiliates.

AWS の機械学習スタック

Broadest and most complete set of machine learning capabilities



© 2022, Amazon Web Services, Inc. or its affiliates.

Thank you!

Shinya Tahara
Solutions Architect
Amazon Web Services Japan

© 2022, Amazon Web Services, Inc. or its affiliates.