© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS re:invent recap - インダストリー編

建設・不動産・物流・交通 業界向け re:Invent 2023 キーノート および技術アップデート

馬渕 俊介

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

© 2024, Amazon Web Services, Inc. or its affiliates.

自己紹介

馬渕　俊介

アマゾンウェブサービスジャパン
ソリューションアーキテクト

AWS での役割 :

主に交通業界のお客様を担当、
AWSの導入を支援

好きな AWS サービス :

Amazon CloudWatch

© 2024, Amazon Web Services, Inc. or its affiliates.

Agenda

- re:Invent 2023 概要
- キーノート紹介
- 生成系 AI 関連アップデート紹介
  - 生成系 AI を利活用するアプリケーション (Amazon Q)
  - 生成系 AI を自社アプリケーションに組み込む (Amazon Bedrock)
- データ関連アップデート紹介
- まとめ

© 2024, Amazon Web Services, Inc. or its affiliates.

Agenda

- re:Invent 2023 概要
- キーノート紹介
- 生成系 AI 関連アップデート紹介
  - 生成系 AI を利活用するアプリケーション (Amazon Q)
  - 生成系 AI を自社アプリケーションに組み込む (Amazon Bedrock)
- データ関連アップデート紹介
- まとめ

© 2024, Amazon Web Services, Inc. or its affiliates.

AWS re:Invent 2023 イベント概要

- AWSによるクラウドコンピューティングに関する世界最大規模の「学習型」カンファレンス
  - 2023年11月27日(月) ~ 12月1日(金)
  - 米国ラスベガスにて開催
- 多くのお客様がご参加
  - 現地参加：50,000+ 名のお客様・パートナー様
  - 日本からも 1,700+ 名 が現地でご参加
- 多数のコンテンツ
  - 5テーマの基調講演、17のイノベーショントーク、2000+のブレイクアウトセッション
- AWS の基盤技術から生成系 AI アプリケーションまで、多様な技術アップデートやユーザー事例を発表

Agenda

- re:Invent 2023 概要
- キーノート紹介
- 生成系 AI 関連アップデート紹介
  - 生成系 AI を利活用するアプリケーション (Amazon Q)
  - 生成系 AI を自社アプリケーションに組み込む (Amazon Bedrock)
- データ関連アップデート紹介
- まとめ

© 2024, Amazon Web Services, Inc. or its affiliates.

reInvent

© 2024, Amazon Web Services, Inc. or its affiliates.

基調講演

Peter Desantis
SVP, AWS Utility Computing
基盤技術・サーバーレス

- サーバーレスの技術革新
  - Amazon Aurora Limitless Database
  - Amazon Elasticache Serverless
  - Amazon Redshift Serverless
- 分散システムのための高度な時刻同期
- 量子コンピューティングの取り組み

Adam Selipsky
CEO, AWS
理念・技術革新・生成系 AI

- AWS の技術開発の理念 (Working Backwards)
- 基盤技術の reinvent (ストレージ、コンピューティング)
- 生成系 AI のアップデートと AI 領域のパートナーシップ (NVIDIA, Anthropic)
- 衛星通信の技術革新 (Project Kuiper)

Dr. Swami Sivasubramanian
VP, Data and AI, AWS
データ・機械学習

- 生成系 AI の 3 層スタック
- Amazon Bedrock での基盤モデルの選択肢
- 組織のデータが差別化の要素となる
- 生成系 AI 活用推進支援プログラム

Dr. Werner Vogels
VP and CTO, Amazon.com
コスト・AI・開発者体験

- クラウドのメリットを活かす Cost-aware な基盤
- 可観測性の促進
- AI for good
- 開発者体験を向上するアップデート

© 2024, Amazon Web Services, Inc. or its affiliates.

Keynote by Adam Selipsky

- **技術開発における AWS の理念（Working Backwards）**
  - AWSは、セキュアで、信頼できるプラットフォーム。Working Backwardsが重要で、お客様の問題から逆算して考えている。
  - インフラを 1 から作り直し (reinvent) 、セキュリティやネットワークプロトコルなどのレベルから他のクラウドプロバイダと異なることを行っている。
- **AI・機械学習領域における新技術とパートナーシップ（NVIDIA, Anthropic）**
- Project kuiper（低軌道衛星コンステレーション）の進捗と実現される世界
- **事例登壇**
  - **BMW** : 何百 PB のデータを AWS で活用して乗車体験を変革
  - **Pfizer** : コロナウイルスワクチンを超高速で実現するのに AWS クラウドを活用

© 2024, Amazon Web Services, Inc. or its affiliates.

Agenda

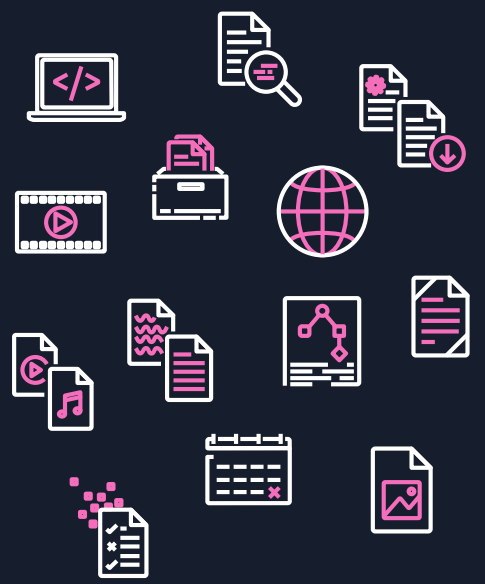
- re:Invent 2023 概要
- キーノート紹介
- 生成系 AI 関連アップデート紹介
  - 生成系 AI を利活用するアプリケーション (Amazon Q)
  - 生成系 AI を自社アプリケーションに組み込む (Amazon Bedrock)
- データ関連アップデート紹介
- まとめ

© 2024, Amazon Web Services, Inc. or its affiliates.

生成系 AI とは？

- 会話、ストーリー、画像、動画、音楽など、
  新しいコンテンツやアイデアを創造
- 一般に 基盤モデル（Foundation Model) と呼ばれる
  膨大なデータに基づいて事前にトレーニングされた大規模モデルを利用

基盤モデルの仕組み



膨大なラベル無しデータ

事前学習

幅広い知識

基盤モデル

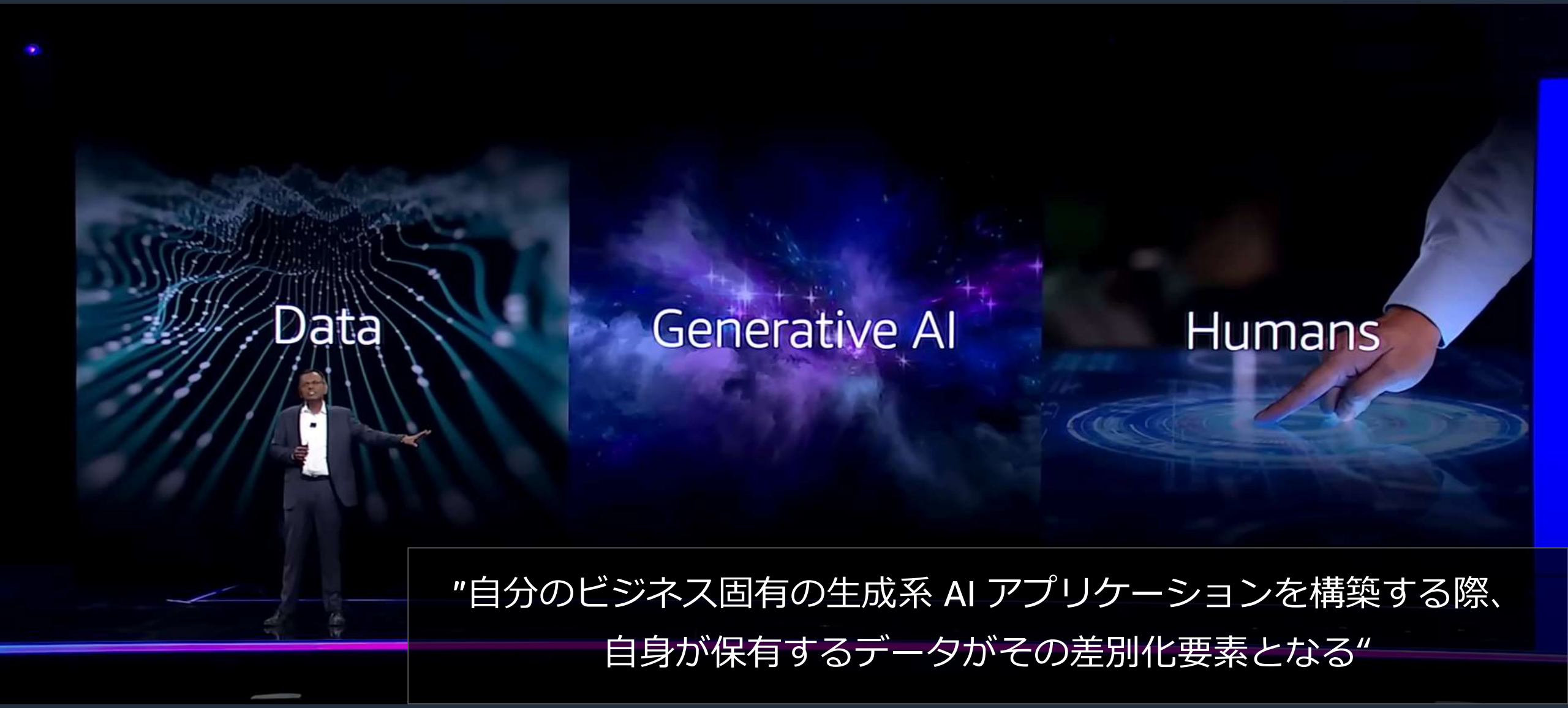
ファインチューン

特定業務の正確性

- テキスト生成
- 要約
- 情報抽出
- Q&A
- チャットボット

少数のラベル付きデータ

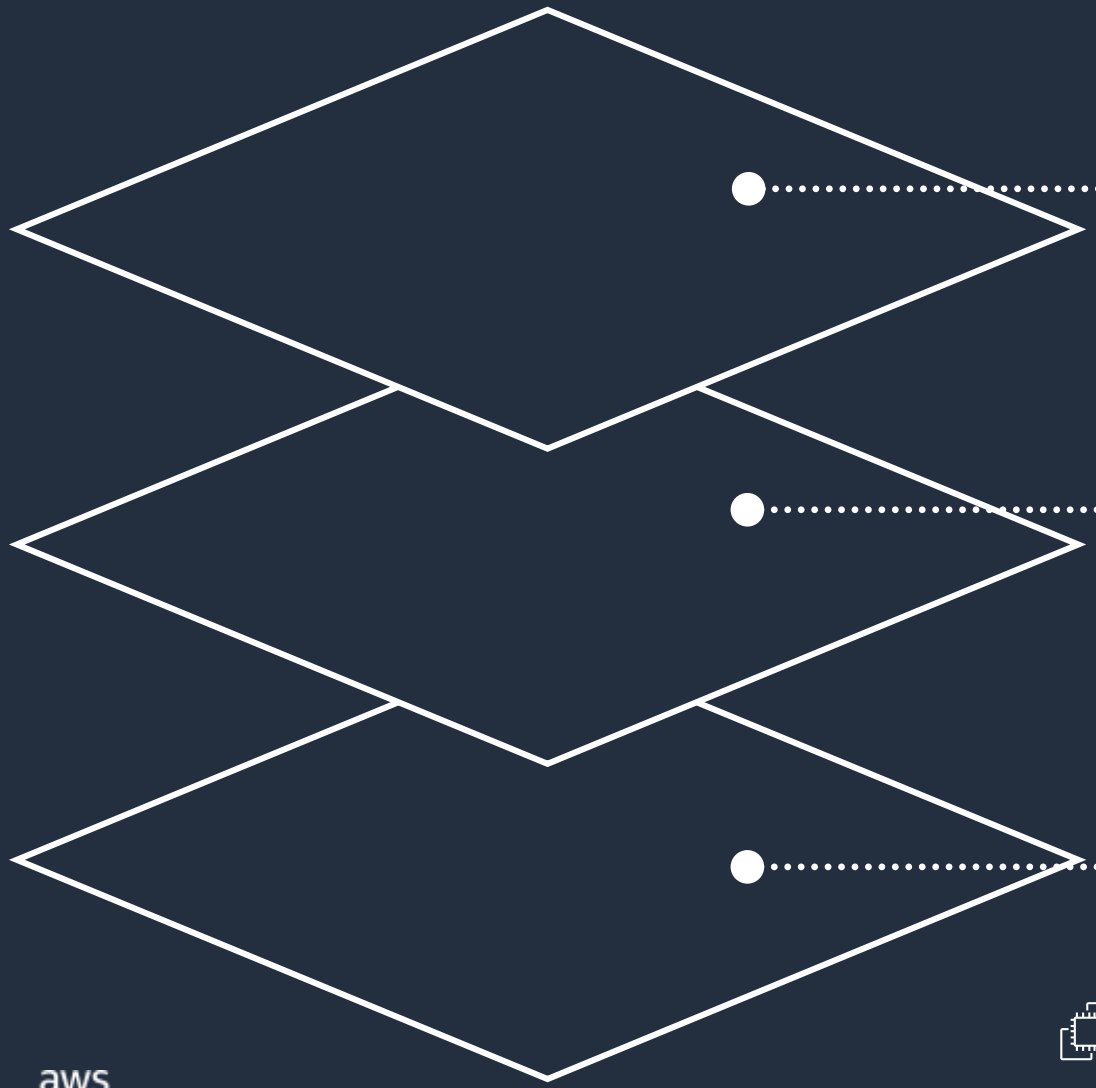
© 2024, Amazon Web Services, Inc. or its affiliates.



"自分のビジネス固有の生成系 AI アプリケーションを構築する際、
自身が保有するデータがその差別化要素となる"

© 2024, Amazon Web Services, Inc. or its affiliates.

生成系 AI の技術スタック



基盤モデルを利活用するアプリケーション

Amazon Q
Amazon Q in Amazon QuickSight
Amazon Q in Amazon Connect
Amazon CodeWhisperer

基盤モデルと大規模言語モデルを活用したアプリケーションを構築するツール群

Amazon Bedrock

Guardrails | Agents | カスタマイズ機能

基盤モデルの学習と推論用の基盤

GPUs
Tranium
Inferencia
SageMaker

Ultra Clusters
EFA
EC2 Capacity Blocks
Nitro
Neuron

© 2024, Amazon Web Services, Inc. or its affiliates.

生成系 AI の技術スタック

アプリケーションとしての UI 等を含めて
Ready-to-use な形で提供されている
生成系 AI ベースのサービス群。
生成系 AI での生産性向上に利用可能

基盤モデルを利活用するアプリケーション

Amazon Q
Amazon Q in Amazon QuickSight
Amazon Q in Amazon Connect
Amazon CodeWhisperer

生成系 AI の基盤モデルのホスティングと、
関連する機能を提供するサービス。
API を通じて AI モデルを使えるため、
自社アプリケーションに
生成系 AI を組み込むのに利用可能

基盤モデルと大規模言語モデルを活用したアプリケーションを構築するツール群

Amazon Bedrock
Guardrails | Agents | カスタマイズ機能

AI の学習・推論に使われるインフラ層。
AI モデル・サービスを開発する会社はこの層を利用して
AI を開発・提供しているほか、
AWS が AI サービスを安定提供するのにも利用。
独自の AI を構築・運用したい場合に利用可能

基盤モデルの学習と推論用の基盤

GPUs
Tranium
Inferencia
SageMaker
Ultra Clusters
EFA
EC2 Capacity Blocks
Nitro
Neuron

© 2024, Amazon Web Services, Inc. or its affiliates.

生成系 AI の技術スタック

アプリケーションとしての UI 等を含めて
Ready-to-use な形で提供されている
生成系 AI ベースのサービス群。
生成系 AI での生産性向上に利用可能

基盤モデルを利活用するアプリケーション

Amazon Q
Amazon Q in Amazon QuickSight
Amazon Q in Amazon Connect
Amazon CodeWhisperer

生成系 AI の基盤モデルのホスティングと、
関連する機能を提供するサービス。
API を通じて AI モデルを使えるため、
自社アプリケーションに
生成系 AI を組み込むのに利用可能

基盤モデルと大規模言語モデルを活用したアプリケーションを構築するツール群

Amazon Bedrock
Guardrails | Agents | カスタマイズ機能

AI の学習・推論に使われるインフラ層。
AI モデル・サービスを開発する会社はこの層を利用して
AI を開発・提供しているほか、
AWS が AI サービスを安定提供するのにも利用。
独自の AI を構築・運用したい場合に利用可能

基盤モデルの学習と推論用の基盤

GPUs
Tranium
Inferencia
SageMaker
Ultra Clusters
EFA
EC2 Capacity Blocks
Nitro
Neuron

© 2024, Amazon Web Services, Inc. or its affiliates.

Agenda

- re:Invent 2023 概要
- キーノート紹介
- 生成系 AI 関連アップデート紹介
  - 生成系 AI を利活用するアプリケーション (Amazon Q)
  - 生成系 AI を自社アプリケーションに組み込む (Amazon Bedrock)
- データ関連アップデート紹介
- まとめ

© 2024, Amazon Web Services, Inc. or its affiliates.

Amazon Q

© 2024, Amazon Web Services, Inc. or its affiliates.

“生成系 AI はデジタルワークプレイスにおいて
これまでで最も破壊的なイノベーションとなるだろう。

我々は、生成系 AI が **80%** の仕事にある程度の影響を与え、**知的労働者の仕事が迅速かつ劇的に変化する**と予想している。”

“生成系 AI がサポートする仕事は、サポートのない人間の作り出す仕事よりも**効率的**で質が高い傾向がある。

平均して、**生産性は 30% 以上向上**する。”

Gartner, Four GenAI Use Cases for the Digital Workplace – 10 October 2023

GARTNER is a registered trademark an service mark of Gartner, Inc. and/or its affiliates in the U.S. and internationally and is used herein with permission. All rights reserved.

aws © 2024, Amazon Web Services, Inc. or its affiliates.

生成系 AI チャットアプリケーションのリスクと課題

誤った回答
(ハルシネーション)

秘匿情報の共有

不十分、もしくは
古くなったデータ

あなたの業務や顧客に
ついての知識がない

あなたのロールや権限を
知らない

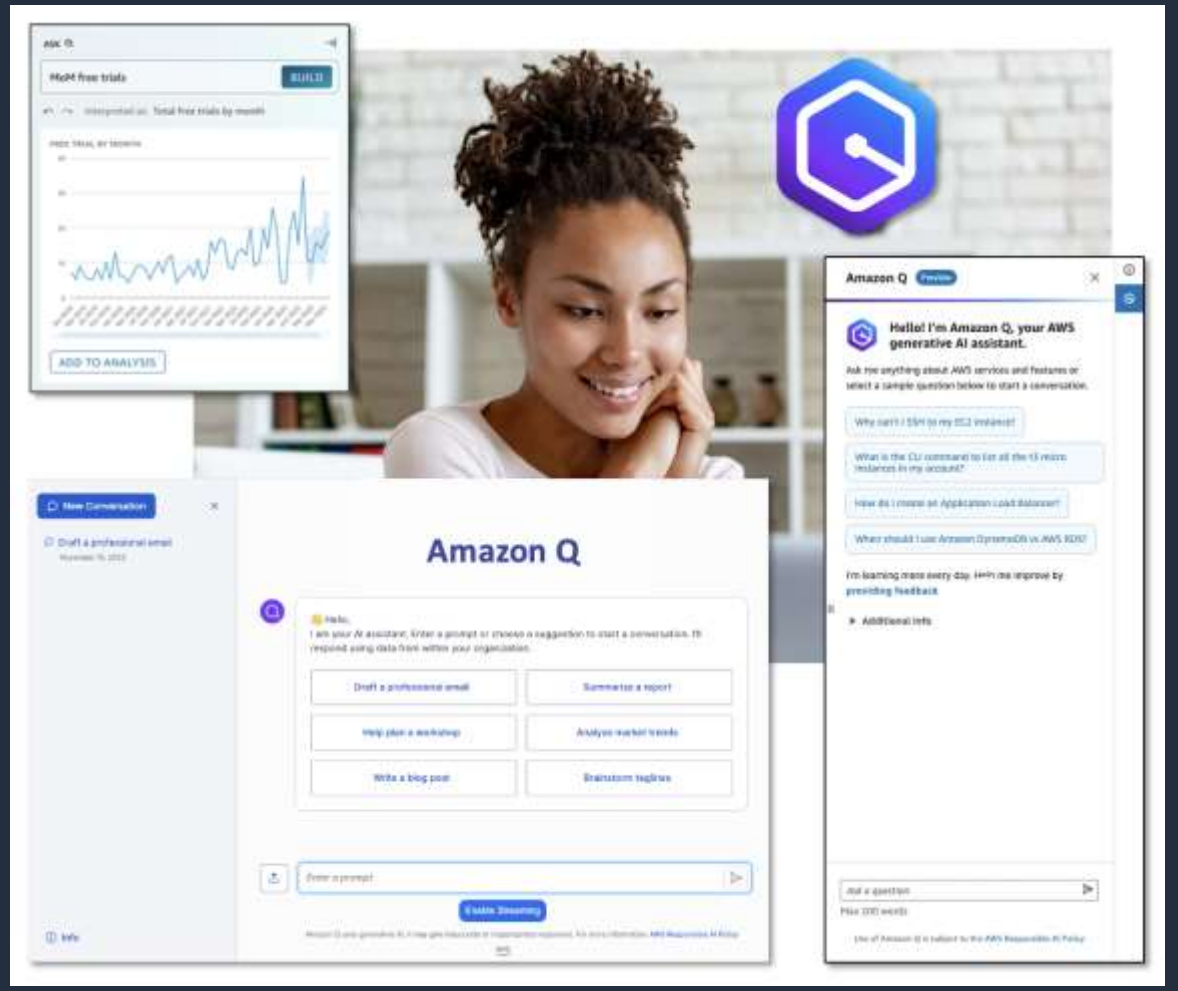
組織内のデータに接続
できない

© 2024, Amazon Web Services, Inc. or its affiliates.

Amazon Q

業務用に設計されたお客様専用の生成系 AI アシスタント

- 問題の解決、コンテンツの生成、アクションを起こすための会話を実施
- お客様の企業の情報、コード、システムを理解
- お客様のロールや権限に基づいてやり取りをパーソナライズ
- セキュアかつプライベートに構築



© 2024, Amazon Web Services, Inc. or its affiliates.

Amazon Q の専門分野

お客様のビジネス

AWS 上での構築

Amazon QuickSight

Amazon Connect

AWS Supply Chain

© 2024, Amazon Web Services, Inc. or its affiliates.

Amazon Q の専門分野

お客様のビジネス

ビジネス向け AI エージェント

- ビジネス上の質問に対して、適切な回答をセキュアでプライベートに提供
- 多様なデータソースに容易に接続し、参考文献や引用を含む回答で事実確認を容易に
- ファイルのアップロードと分析も可能

ユースケース：
検索体験合理化、文書からのインサイト抽出、要約生成、コンテンツ制作加速

AWS 上での構築

開発者向け AI エージェント

- AWS マネジメントコンソール上で、AWS の知識を学習したチャットボットが開発/運用をサポート
- 統合開発環境上で、AI が開発を支援
  - 文法や AWS サービスに関する質問に回答
  - コードの生成・修正・説明・最適化・変換

ユースケース：
アプリケーション開発高速化、アーキテクチャ検討相談、トラブルシューティングの容易化

Amazon QuickSight

BI ツール内での生成系 AI

- 自然言語でユーザーが指示すると、その内容に基づきデータをビジュアル化
- 「無料ユーザを有料ユーザに引き上げるためのストーリーを作って」と指示すれば、ビジュアルとテキストからなるレポートを自動生成。修正も容易

ユースケース：
ビジネスレポーティングの効率化、データ活用民主化

© 2024, Amazon Web Services, Inc. or its affiliates.

まとめ : Amazon Q

- **Amazon Q** は、業務用に設計されたお客様専用の生成系 AI アシスタント。
  ハルシネーションを抑制しつつセキュアに利用でき、生産性を向上
- **ビジネス向け AI エージェント** : 容易にデータソースと連携して正確な回答が
  可能。アップロードしたファイルを分析してタスクを実行することも可能
- **開発者向け AI エージェント** : AWS について質問に正確に回答。
  さらに、統合開発環境上でコードの生成・修正・説明・変換などが可能
- **Amazon Q in Quicksight** : BI ツール上で自然言語で指示すると、
  データのビジュアル化やデータからのレポート作成を自動で実行

© 2024, Amazon Web Services, Inc. or its affiliates.

Agenda

- re:Invent 2023 概要
- キーノート紹介
- 生成系 AI 関連アップデート紹介
  - 生成系 AI を利活用するアプリケーション (Amazon Q)
  - 生成系 AI を自社アプリケーションに組み込む (Amazon Bedrock)
- データ関連アップデート紹介
- まとめ

© 2024, Amazon Web Services, Inc. or its affiliates.

生成系 AI の技術スタック

アプリケーションとしての UI 等を含めて
Ready-to-use な形で提供されている
生成系 AI ベースのサービス群。
生成系 AI での生産性向上に利用可能

基盤モデルを利活用するアプリケーション

Amazon Q
Amazon Q in Amazon QuickSight
Amazon Q in Amazon Connect
Amazon CodeWhisperer

生成系 AI の基盤モデルのホスティングと、
関連する機能を提供するサービス。
API を通じて AI モデルを使えるため、
自社アプリケーションに
生成系 AI を組み込むのに利用可能

基盤モデルと大規模言語モデルを活用したアプリケーションを構築するツール群

Amazon Bedrock
Guardrails | Agents | カスタマイズ機能

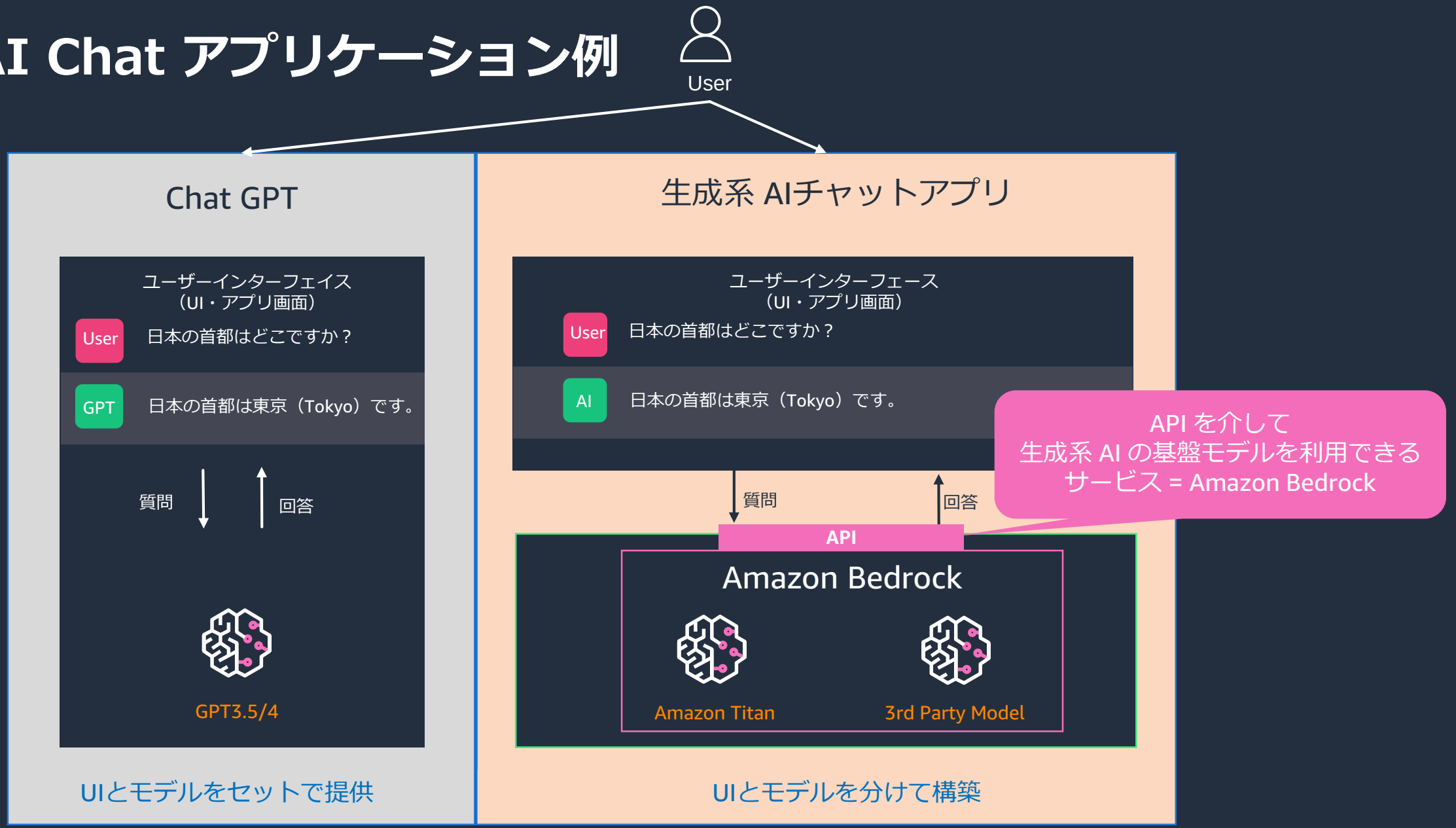
AI の学習・推論に使われるインフラ層。
AI モデル・サービスを開発する会社はこの層を利用して
AI を開発・提供しているほか、
AWS が AI サービスを安定提供するのにも利用。
独自の AI を構築・運用したい場合に利用可能

基盤モデルの学習と推論用の基盤

GPUs
Tranium
Inferencia
SageMaker
Ultra Clusters
EFA
EC2 Capacity Blocks
Nitro
Neuron

© 2024, Amazon Web Services, Inc. or its affiliates.

生成系AI Chat アプリケーション例



© 2024, Amazon Web Services, Inc. or its affiliates.

Amazon Bedrock

サーバーレスの API サービスを介して基盤モデルを活用した生成系 AI でアプリケーションを構築

東京リージョンで利用可能

厳選された基盤モデルから業務に最適な基盤モデルを選択・活用

自社データを使用し基盤モデルをプライベートな環境でカスタマイズ

実績ある AWS のセキュリティ機能によりデータ保護を強化

aws

© 2024, Amazon Web Services, Inc. or its affiliates.

基盤モデルの選択肢を拡大

AI21labs　ANTHROP\C　co:here　Meta AI　stability.ai　amazon

JURASSIC	CLAUDE	COMMAND + Embed	LLAMA 2	Stable Diffusion	AMAZON TITAN
テキスト生成用 多言語対応 LLM 英語、西語、仏語 独語、葡語、伊語 蘭語	誠実かつ責任あるAI システム研究に基づく 会話、質問応答 ワークフロー自動化用 LLM	テキスト要約 コピーライティング 対話、情報抽出 質問応答等 業務アプリ用 テキスト生成モデル	質問応答、文書読解等 自然言語処理タスク 用に事前学習、 ファインチューニン グ済LLM	ユニークかつリアルな 高品質画像生成 アート、ロゴ、デザイン 生成	テキスト要約 テキスト生成、 分類質問応答、 情報抽出、埋め込み 検索

New Models in re:Invent 2023

- Claude 2.1
- Llama2 70B
- Stable Diffusion XL 1.0 (GA)
- Amazon Titan Image Generator (Preview)
- Amazon Titan Multimodal Embeddings
- Amazon Titan Text Lite and Express (GA)

AWS は生成 AI の用途にあった基盤モデルの選択肢を提供します。

LLM: Large Language Model (大規模言語モデル)

aws
© 2024, Amazon Web Services, Inc. or its affiliates.

amazon | ANTHROPIC

Dario Amodei
Co-Founder and CEO of Anthropic

- ✓ 主要クラウドプロバイダーにAWS を選択
- ✓ ワークロードの大部分をAWSで実行する
- ✓ 学習用のTrainium と推論用のInferentia を使用し、将来の基盤モデルを構築、学習、展開する
- ✓ アクセラレータの開発でも協業

- Bedrock で、20万のトークンをサポートし、ハルシネーションを軽減した Claude2.1 が利用可能に
- ファインチューニングにも今後対応予定であることが発表された

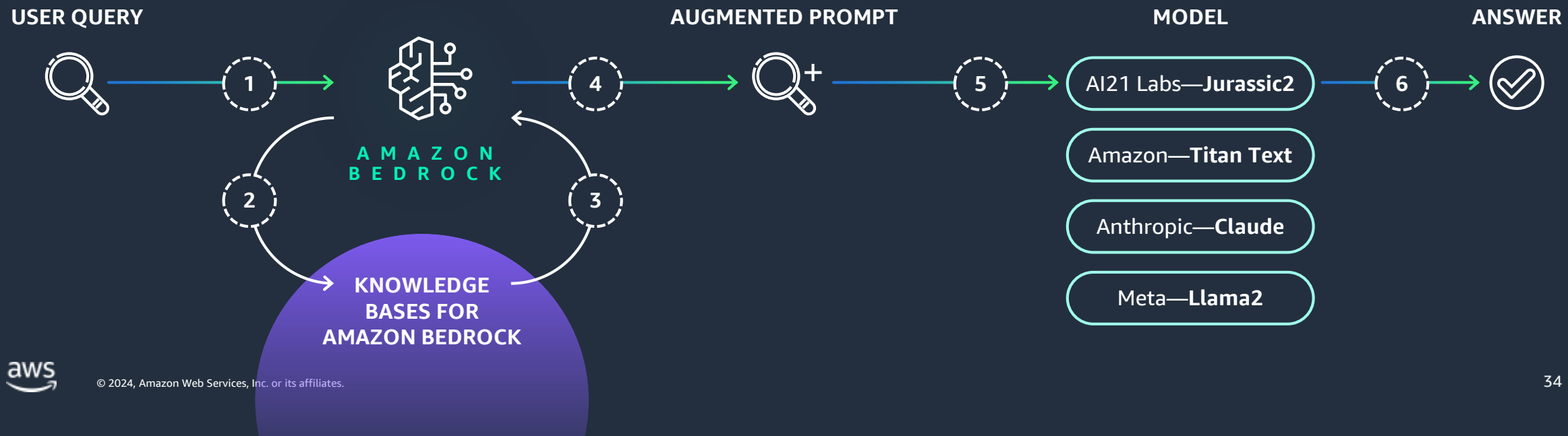
© 2024, Amazon Web Services, Inc. or its affiliates.

検索拡張生成 Retrieval augmented generation (RAG)

Knowledge Bases for Amazon Bedrock

基盤モデルを組織のデータソースと統合して、より正確で関連性の高い回答を提供

ユーザーのクエリに関連する企業データをプロンプトに束ねて基盤モデルに渡すことで、ハルシネーション (誤った発言) のリスクを軽減



© 2024, Amazon Web Services, Inc. or its affiliates.

Delta Airlines : 生成系 AI を顧客対応に活用

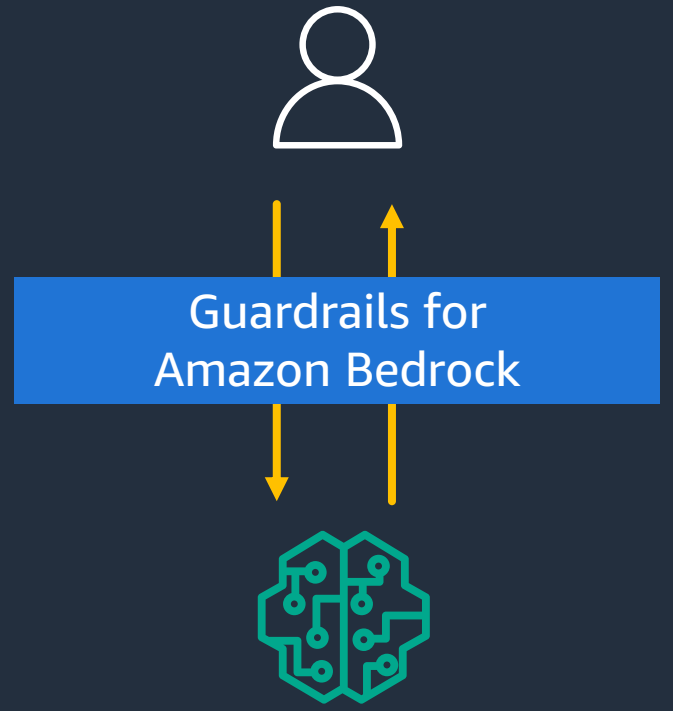
- Amazon Bedrock 上で**複数の基盤モデル**を活用し
  カスタマーサービスを改善
  - **自社の様々なデータを統合**し、
    顧客の質問に対話で回答可能なシステムを構築
    （データの例）
    - 規約
    - リアルタイムのフライトスケジュール
    - 再予約オプション
    - 空港の状況
  - 例えば、「何個の手荷物を預けられますか？」
    「猫をキャビンに連れてジャマイカにいけますか？」などに回答可能

© 2024, Amazon Web Services, Inc. or its affiliates.

プレビュー

Guardrails for Amazon Bedrockを発表

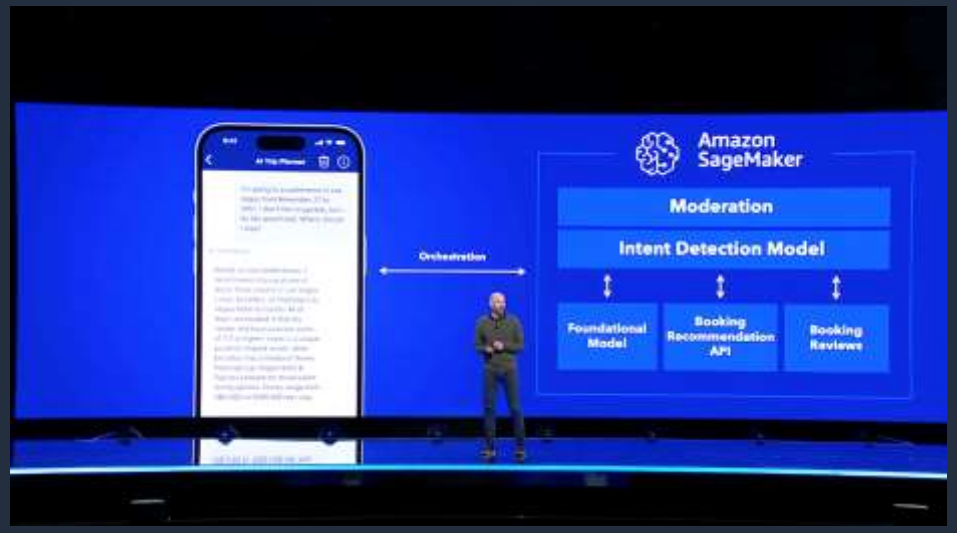
- ユーザのクエリや基盤モデルの応答を評価し、不適切なコンテンツを評価・フィルタする機能
  - 多くの基盤モデルは保護機能を有しているが、独自の基準やルールを適用したい場合に便利な機能
  - ユースケースに応じた「責任あるAI利用」を実現するために利用できる
- 現時点ではテキストベースのモデルで、英語に対応。Agents for Amazon Bedrockとカスタマイズ済みモデルにも対応
- バージニア、オレゴンでプレビューを開始



© 2024, Amazon Web Services, Inc. or its affiliates.

事例： Booking.com

- 生成系AIで新しい体験を提供している。
- 2800万の宿泊施設、54カ国の飛行機の予約、5200のレンタカーの予約など、これら、150PB以上のデータを活用するため、AWSを活用した。
- トリッププランナーでは、Llama2を使い、会話型で、旅行の予約について相談できる。Intent Detection Modelを使っている。会話が旅行に関連するのかを把握し、プライバシーの保護にも考慮している。また、リコメンドにおいても、Amazon SagemakerやAmazon Bedrockを利用している。
- (セッションの構成も、Amazon Titanに確認してみている)

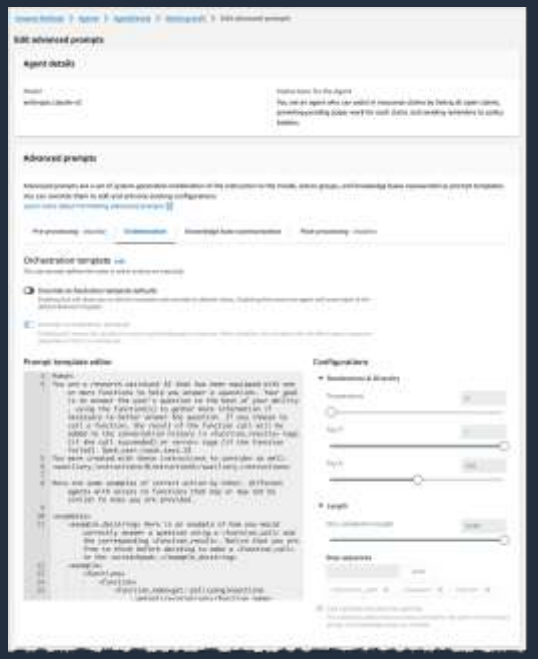


© 2024, Amazon Web Services, Inc. or its affiliates.

一般利用開始

Agents for Amazon Bedrockが一般利用開始に

- 基盤モデルだけで完結しないタスクを実行することを容易にするフルマネージドサービス
  - 自然言語で指示を記述し、組織内の別システムへのアクセス方法を提供し、Lambda関数を定義することでエージェントを作成
  - エージェントはリクエストを分析。基盤モデルの推論能力を活用し、リクエスト完了までのステップを決定・実行する
  - データは全て暗号化され、安全
- バージニアとオレゴンで一般利用開始に



© 2024, Amazon Web Services, Inc. or its affiliates.

一般利用開始

Agent ができること– タスク分解

- ユーザーの入力を複数の小さなタスクに分割し、タスクごとに適切な API を呼び出すことで回答を生成する

ユーザー

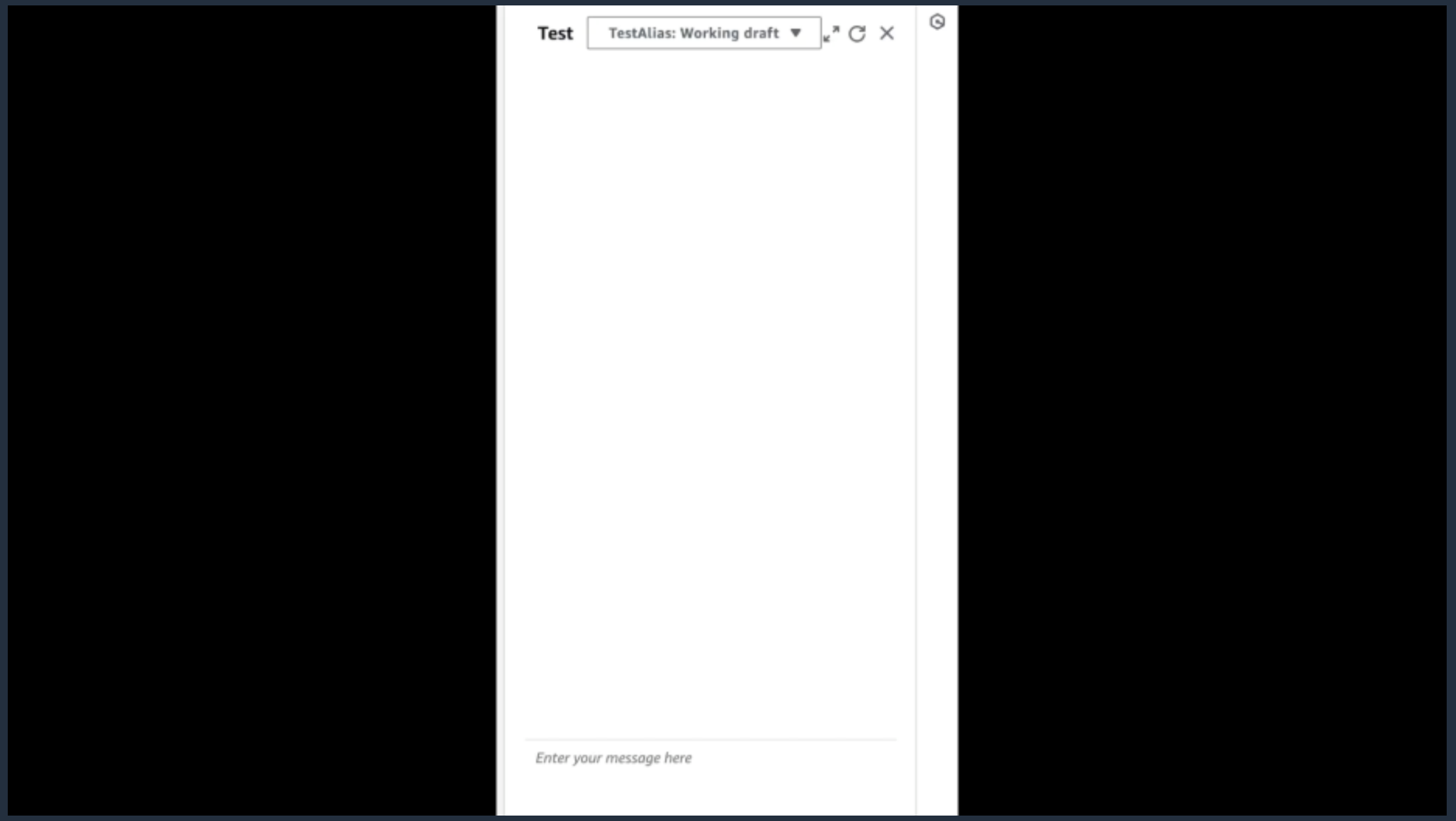
XXX の現在の株価を教えてください。あと、XXXに関する最近のニュースを教えてください。

Agent (LLM)

文脈によると、**XXX は企業の名前**だな。
銘柄コードを取得して、銘柄コードから**株価を取得**することはできそうだ。
最近のニュースは **Web から検索**してくる必要がありそうだ。
これらの情報をまとめたら、回答が生成できそうだ。

© 2024, Amazon Web Services, Inc. or its affiliates.

demo



https://www.youtube.com/watch?v=f1AfRngvSU8

© 2024, Amazon Web Services, Inc. or its affiliates.

まとめ : Amazon Bedrock

- 自社アプリケーションに生成系 AI を組み込むのに利用できるサービス

- API 経由で、**複数の基盤モデルをセキュアに利用**することができる
  - Anthropic 社の **Claude** はハルシネーションに強く、20 万トークン入力可能

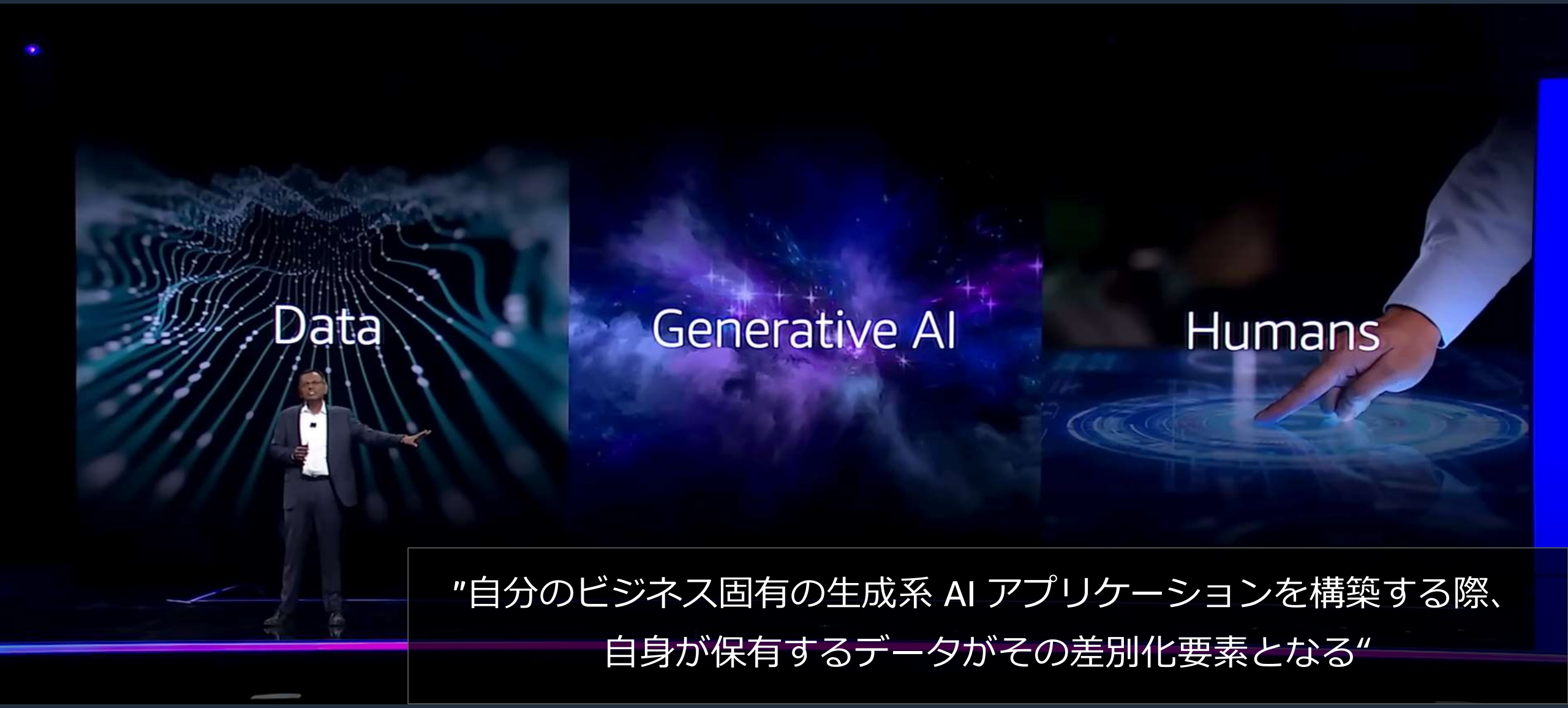
- 基盤モデルをアプリケーションに組み込む際の追加機能あり
  - **Knowledge Base** : データソースと連携することで回答の正確性を向上
  - **Guardrails** : 不適切な回答を評価・フィルタリングし、コンプライアンスリスクを抑制
  - **Agents** : ユーザのリクエストをタスクに分解し、API 等を利用して完遂

© 2024, Amazon Web Services, Inc. or its affiliates.

Agenda

- re:Invent 2023 概要
- キーノート紹介
- 生成系 AI 関連アップデート紹介
  - 生成系 AI を利活用するアプリケーション (Amazon Q)
  - 生成系 AI を自社アプリケーションに組み込む (Amazon Bedrock)
- データ関連アップデート紹介
- まとめ

© 2024, Amazon Web Services, Inc. or its affiliates.



"自分のビジネス固有の生成系 AI アプリケーションを構築する際、
自身が保有するデータがその差別化要素となる"

© 2024, Amazon Web Services, Inc. or its affiliates.

re:Invent 2023 で発表された、データ基盤関連アップデート

- **データ統合 :**
  - **Zero-ETL サービス拡充によるデータ統合の容易化**
    - Amazon Aurora PostgreSQL zero-ETL integrations with Amazon Redshift
    - Amazon RDS for MySQL integrations with Amazon Redshift
    - Amazon DynamoDB integrations with Amazon Redshift
    - Amazon DynamoDB zero-ETL integrationwith Amazon OpenSearch Service
  - **生成系 AI による ETL 処理の作成支援** (Amazon Q Data integration in AWS Glue)
  - Amazon Redshift が Multi-data warehouse writeに対応
- **データカタログ :**
  - **生成系AIが、データカタログにビジネスコンテキストを自動的に追加**
    (Amazon DataZone AI recommendations)
- **データ活用 :**
  - **生成系 AI による、自然言語からの SQL クエリ自動生成**
    (Amazon Q generative SQL in Amazon Redshift)

© 2024, Amazon Web Services, Inc. or its affiliates.

Agenda

- re:Invent 2023 概要
- キーノート紹介
- 生成系 AI 関連アップデート紹介
  - 生成系 AI を利活用するアプリケーション (Amazon Q)
  - 生成系 AI を自社アプリケーションに組み込む (Amazon Bedrock)
- データ関連アップデート紹介
- まとめ

© 2024, Amazon Web Services, Inc. or its affiliates.

まとめ

- AWS はお客様の問題から逆算し、セキュアで信頼できるプラットフォームを作ってきた。そのために、インフラのあり方を 1 から “reInvent” してきた
- re:Invent 2023 では、生成系 AI に関するアップデートを多数発表した
  - 生成系 AI を利活用して、自社の生産性向上に利用できる **Amazon Q**
  - 生成系 AI を自社アプリケーションに組み込むのに利用できる **Amazon Bedrock**
    - 幅広いモデルの選択肢と、アプリケーションに組み込むのに不可欠な機能を提供
- 生成系 AI を利用する際、**自身の保有するデータがその差別化要素となる**

© 2024, Amazon Web Services, Inc. or its affiliates.

Thank you!

© 2024, Amazon Web Services, Inc. or its affiliates.