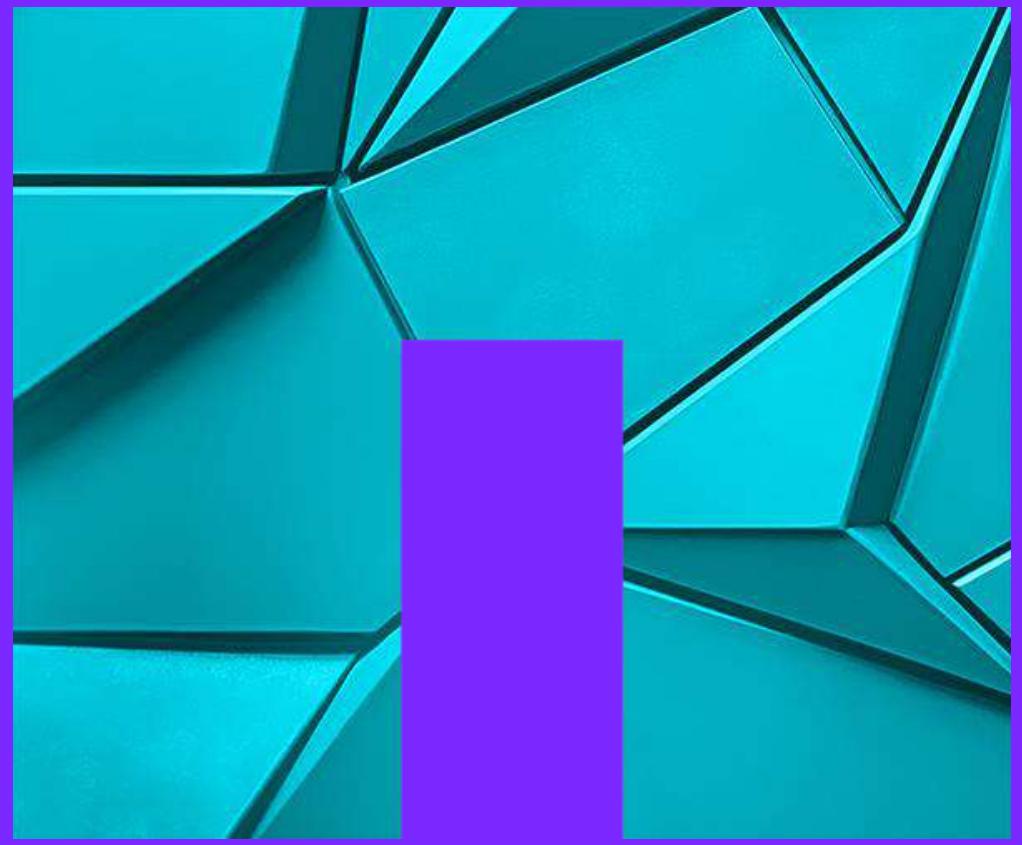
NetApp

SNIA 日本支部 主催
2024年度 第1回 最新技術動向講演会

生成AIの可能性を拡げる インテリジェントデータインフラ



ネットアップ合同会社
Solutions Architect 井上耕平
2024/10/22

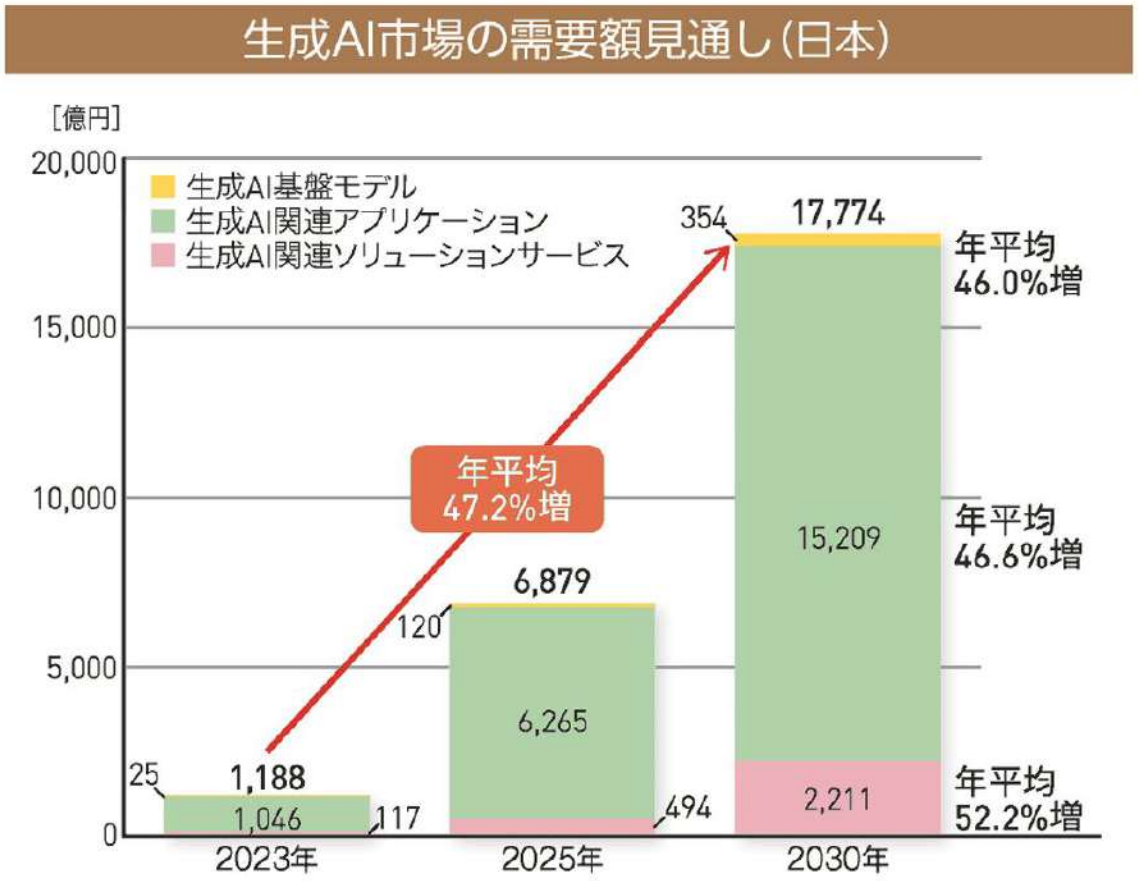
Agenda

- 最近のAI関連動向
- AI向けストレージインフラとは
- AIを作るためのストレージ
- AIを使うためのストレージ
- 可能性を”更に”拡げるインテリジェンスとは

最近のAI関連動向

国内生成AI市場の見通し

注目分野に関する動向調査2023, JEITA



生成AI市場の国内需要額は

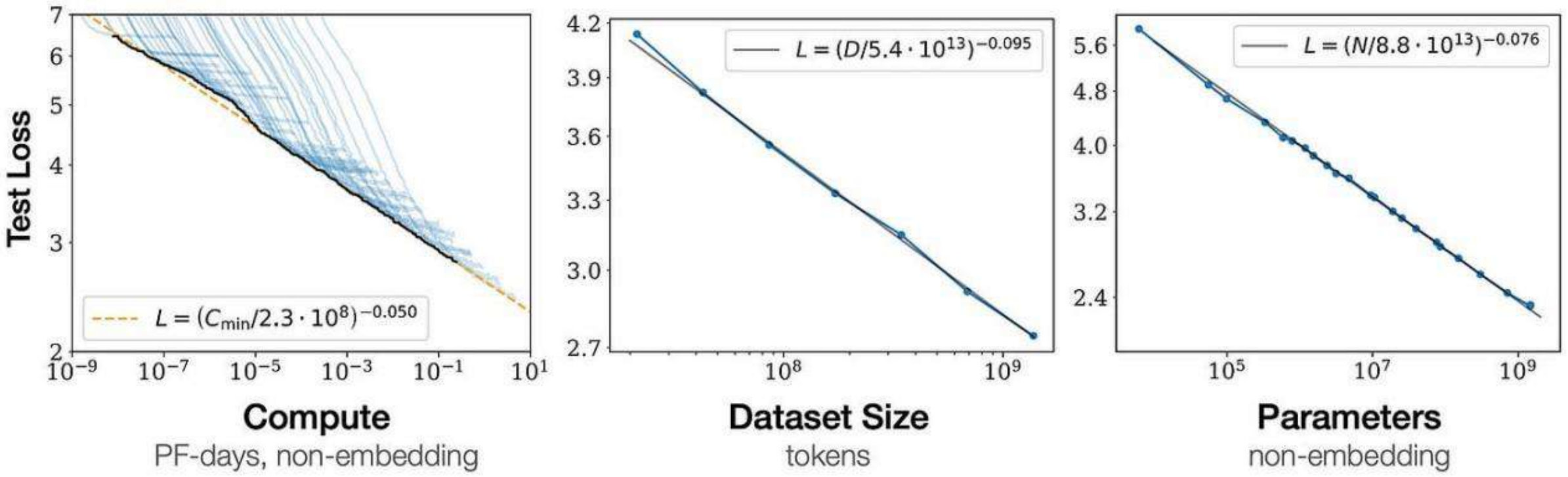
年平均47.2%で成長

2030年には**17,774億円**

2023年の約15倍に到達

生成AIの発展を支えるスケーリング則

ムーアの法則のように、生成AIに用いる大規模言語モデルの性能向上が予測
投資をすればするだけリターンが得られるため、大手IT企業がこぞって大規模言語モデルの開発に着手

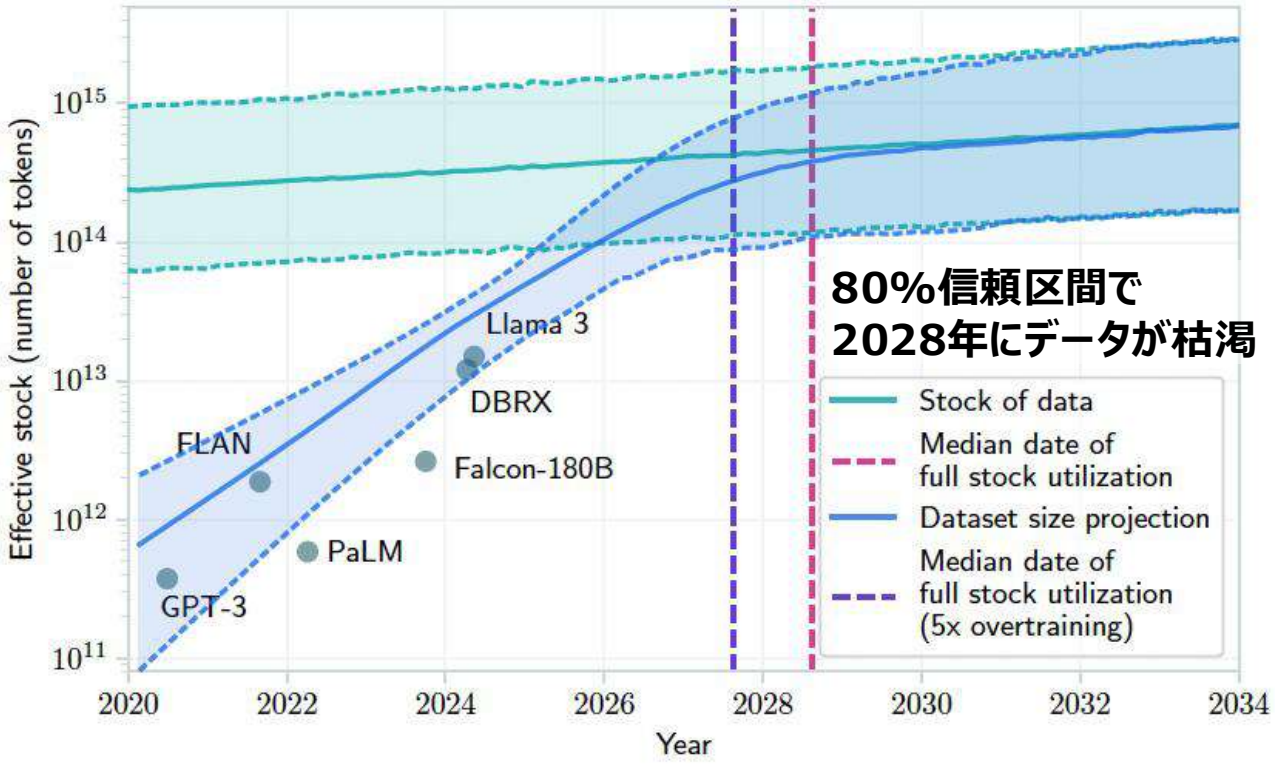


AIモデルの性能は<u>データ量</u>と<u>パラメータ数</u>を増やせば向上

Scaling Laws for Neural Language Models (Kaplan et al., 2020)

学習データの枯渇問題

学習に使うインターネット上の高品質なパブリックテキストデータは2028年に枯渇する可能性



学習データを調達するための
4つの戦略を提案

1. AI生成データ
2. マルチモーダルと転移学習
3. **プライベートデータ**
4. データ効率の向上

AIにおけるプライベートデータの重要度が増す

Will we run out of data? Limits of LLM scaling based on human-generated data(Villalobos et al., 2024)

AI向けストレージインフラとは

2種類のAI と 必要となるインフラ

大規模言語モデルの登場により、AIモデル自体を開発しなくても実現できるタスクの範囲が大きく拡がる

AIを作る

（AIモデル自体の開発）

一般的には…
既存モデルでは実現できない専門的なタスク
大規模GPUクラスター環境
パフォーマンスを限界まで使い切る
Infiniband、~800Gbps、RDMA、水冷

AIを使う

（既存のAIモデルを使った開発）

一般的には…
既存モデルやそのカスタマイズで実現できるタスク
既存インフラに推論環境(GPU/NPU)を追加
サービスに求められる最低限のパフォーマンス
Ethernet、~100Gbps、非機能要件

2種類のAI と ストレージインフラ

企業のユースケースに応じて、求められるストレージインフラも変わる

AIを作る

（AIモデル自体の開発）

NVIDIA SuperPOD/BasePOD アーキテクチャ
並列分散ファイルシステム、NFS
GPU Direct Storage（GDS）
Kubernetes / VMware / Singularity / MLOps
On-premise / DGX Cloud

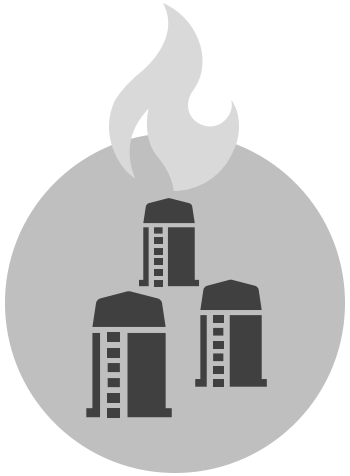
AIを使う

（既存のAIモデルを使った開発）

NVIDIA OVX アーキテクチャ
NFS、Object Storage
運用管理性、セキュリティ
Kubernetes / VMware / DevOps / DataOps
On-premise / Cloud / Hybrid

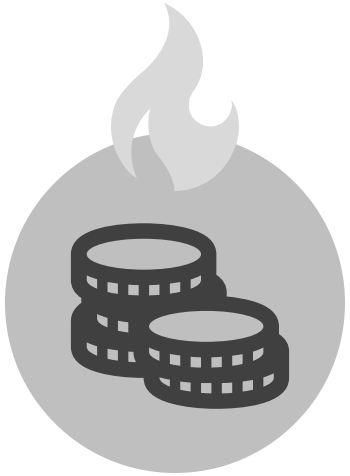
今後重要となる プライベートデータ活用 におけるストレージの課題

プライベートデータを「AIを作る/使う」インフラから”継続的に”活用するために、ストレージが乗り越えるべき課題



既存データインフラとの連携

- プライベートデータをAIから活用するために新たなサイロを構築
- 既存のでータインフラから最新データを継続的に収集し、活用するための仕組み作りで、インフラが複雑化
- 年々進化するAIに対応する事が出来ず、利用が進まず負債化



増加するストレージコスト

- AIモデルの価値を高めるために、プライベートデータの活用が重要に（2026年問題）
- プライベートデータの量が企業の競争力に影響を与えるが、無尽蔵に蓄積するとストレージコストが膨らむ

サイバー攻撃への対応

- AIのプライベートデータ活用が進むと企業の競争力の根源であるナレッジが集まるため、生成AI基盤はサイバー攻撃の格好の標的に
- データポイズニングのような新たな攻撃手法に加え、従来のランサムウェアやデータの窃取に対応しなくてはいけない

ネットアップが考える、AI用データインフラに必要なインテリジェンス

AIに適用できるタスクの範囲が拡大したことで、プライベートデータを活用する「AIを作る/使う」インフラにも、
エンタープライズで求められるインテリジェンスが求められる

既存と同じデータインフラを活用

- ✓ 従来と変わらない運用管理で、既存のデータインフラからデータを収集しやすく、AIインフラから活用しやすい
- ✓ 求めるユースケースに適したインフラ構成検討の助けとなる幅広いリファレンスアーキテクチャが提供されており、AIを用いた価値創造に集中できる

コストの最適化と柔軟性

- ✓ AIで活用するデータの増大に対して、重複排除などのコスト効率機能を用いて効率よくデータを保存できる
- ✓ オンプレミスのハードウェアにおいても、クラウドのようにサブスクリプションで柔軟な利用がが可能

充実したセキュリティ

- ✓ 機密性の高いプライベートデータを格納するためのセキュリティが、認証などで担保されている
- ✓ サイバー攻撃に対応できる様々なセキュリティ機能が組み込まれている
- ✓ データへのアクセス動向からランサムウェアの被害を検知し、被害拡大を防げる

AIを作るための
ストレージ

NVIDIA DGX SuperPOD の 性能要件

どのようなデータを学習させるかによって、求められるパフォーマンスは異なるが、
GPUクラスターの性能を使い切るために、ストレージがボトルネックとならないような点のパフォーマンスを重視

性能レベル	学習対象データ	データセットサイズ
Good	自然言語	データセットは通常、ローカルキャッシュ内に収まる
Better	LLMトレーニングなどの圧縮画像、圧縮音声、テキストデータ	多くのデータセットは、ローカルシステムのキャッシュ内に収まる
Best	大容量のビデオや画像ファイル（AV再生など）、オフライン推論、ETL、安定拡散などの生成ネットワーク、医療用U-Netなどの3D画像、AlphaFoldなどのゲノムワークロードやタンパク質予測	キャッシュに収まらないほどデータセットが大規模、最初のエポックにおける膨大なI/O要件、データセットを一度しか読み込まないワークフロー

パフォーマンス 特性	Good (GBps)	Better (GBps)	Best (GBps)
シングルSU※ アグリゲートシステム 読み込み	15	40	125
シングルSU アグリゲートシステム 書き込み	7	20	62
4 SU アグリゲートシステム 読み取り	60	160	500
4 SU アグリゲートシステム 書き込み	30	80	250

※ SU=Scalable Unit : B200世代の場合、1SU=32node

https://docs.nvidia.com/dgx-superpod/reference-architecture-scalable-infrastructure-b200/latest/storage-architecture.html

ネットアップ の SuperPOD リファレンスアーキテクチャ

並列ファイルシステムのBeeGFS と E-seriesの組み合わせで、DGX H100 クラスターのモデル開発を高速化

検証済みアーキテクチャ
- GPU使用率95%を実現
- BeeGFS on Demandによるキャッシュ効率の向上

シンプルなデプロイ と プロビジョニング
- BeeGFS対応のAnsible PlaybookとDeepOps
- K8S対応のContainer Storage Interface

性能 と スケーラビリティ
- 32台から数百台まで対応するスケーラビリティ
- GPU Direct Storageによる超高速データ入出力

データパイプライン と セキュリティ
- DCでのトレーニングから クラウドへ柔軟にバースト
- NetApp DataOps Toolkit による容易な管理

ネットアップ の BasePOD リファレンスアーキテクチャ

インテリジェント機能が搭載されたストレージとDGX H100により、
小/中規模なモデル開発から推論まで対応

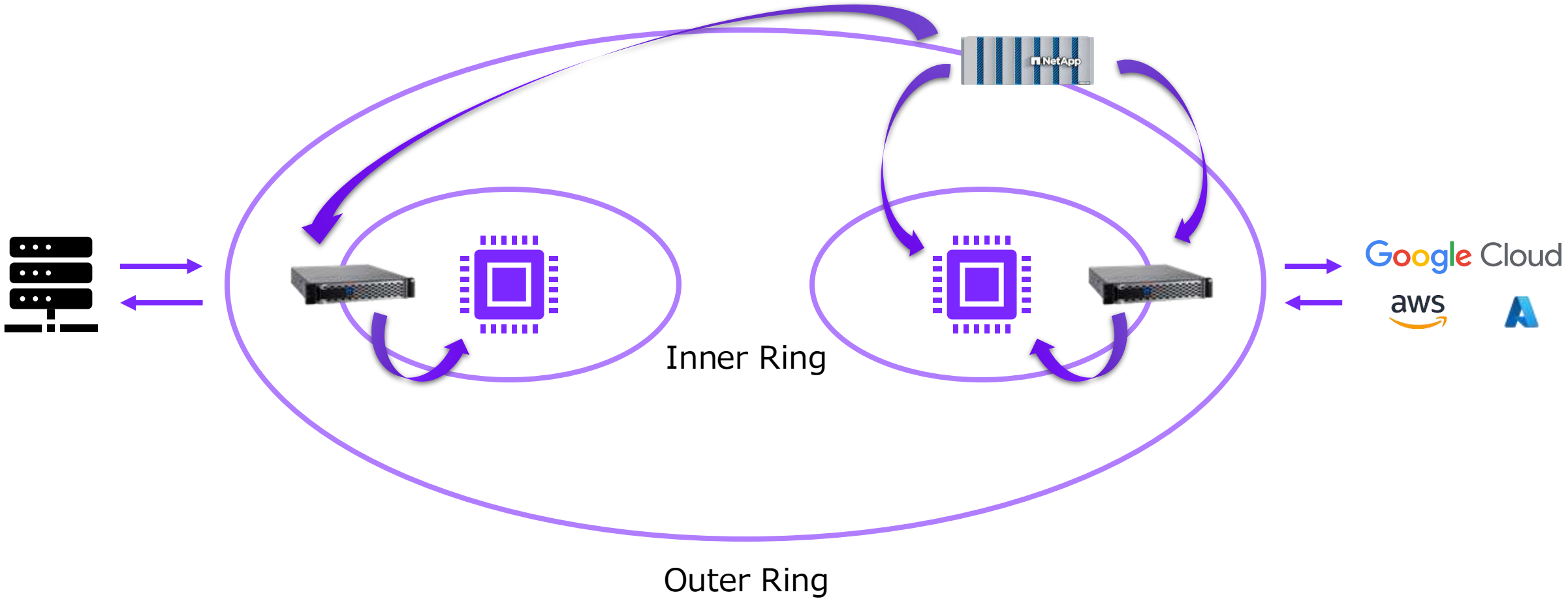
**データセンターで実行される分析、トレーニング、推論を
一つのAIインフラに統合**

- NVIDIAが検証した柔軟なソリューションで、MLとDL用のデータパイプラインを簡素化、高速化、統合
- 適切なパフォーマンスと無停止のスケーラビリティを提供
- インフラのサイロをなくし、AIワークロードを統合
- NVIDIA DGX BasePOD設計を活用
- GPU Direct Storage（GDS）サポート
- ネットアップの最新世代A90とDGX H100構成に対応

NetApp AIPod for NVIDIA DGX BasePOD

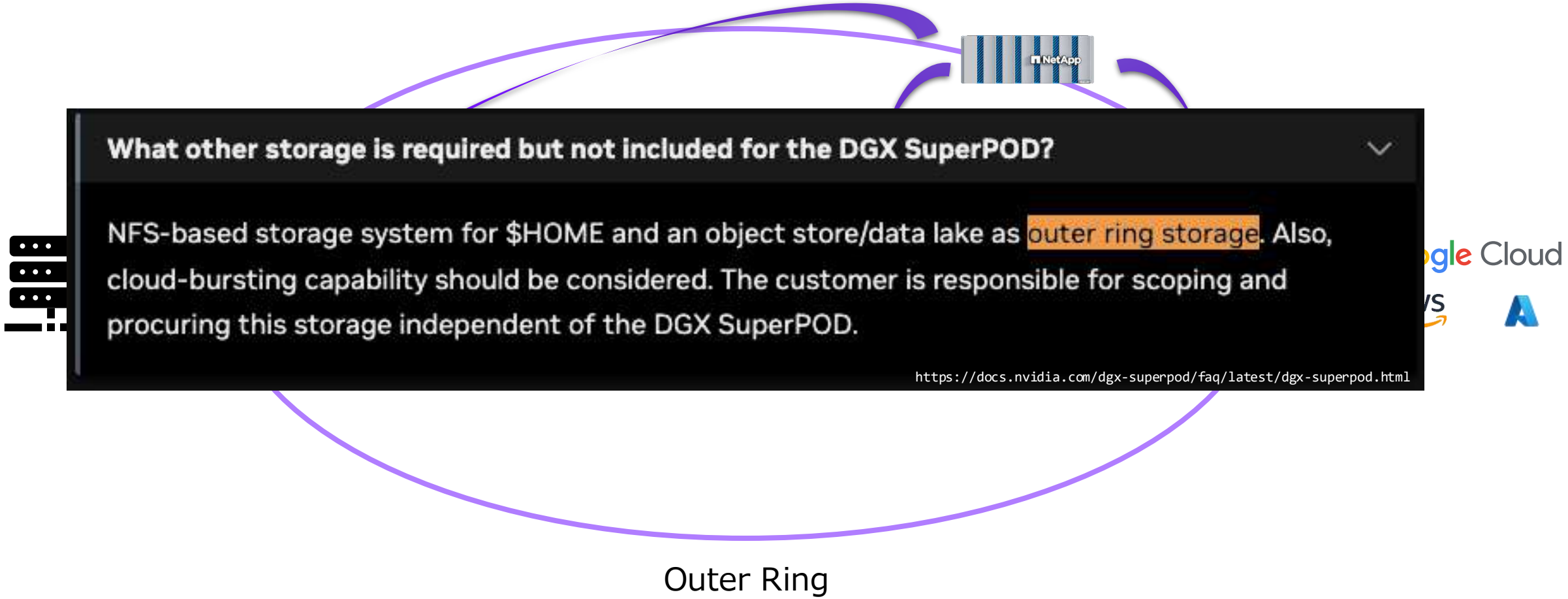
インナーリング と アウターリング

SuperPOD リファレンスアーキテクチャには、GPUクラスターから直接扱う インナーストレージ の要件のみ定義
データの性質によって、インナーリングにないインテリジェンスを備えたアウターリングストレージが必要



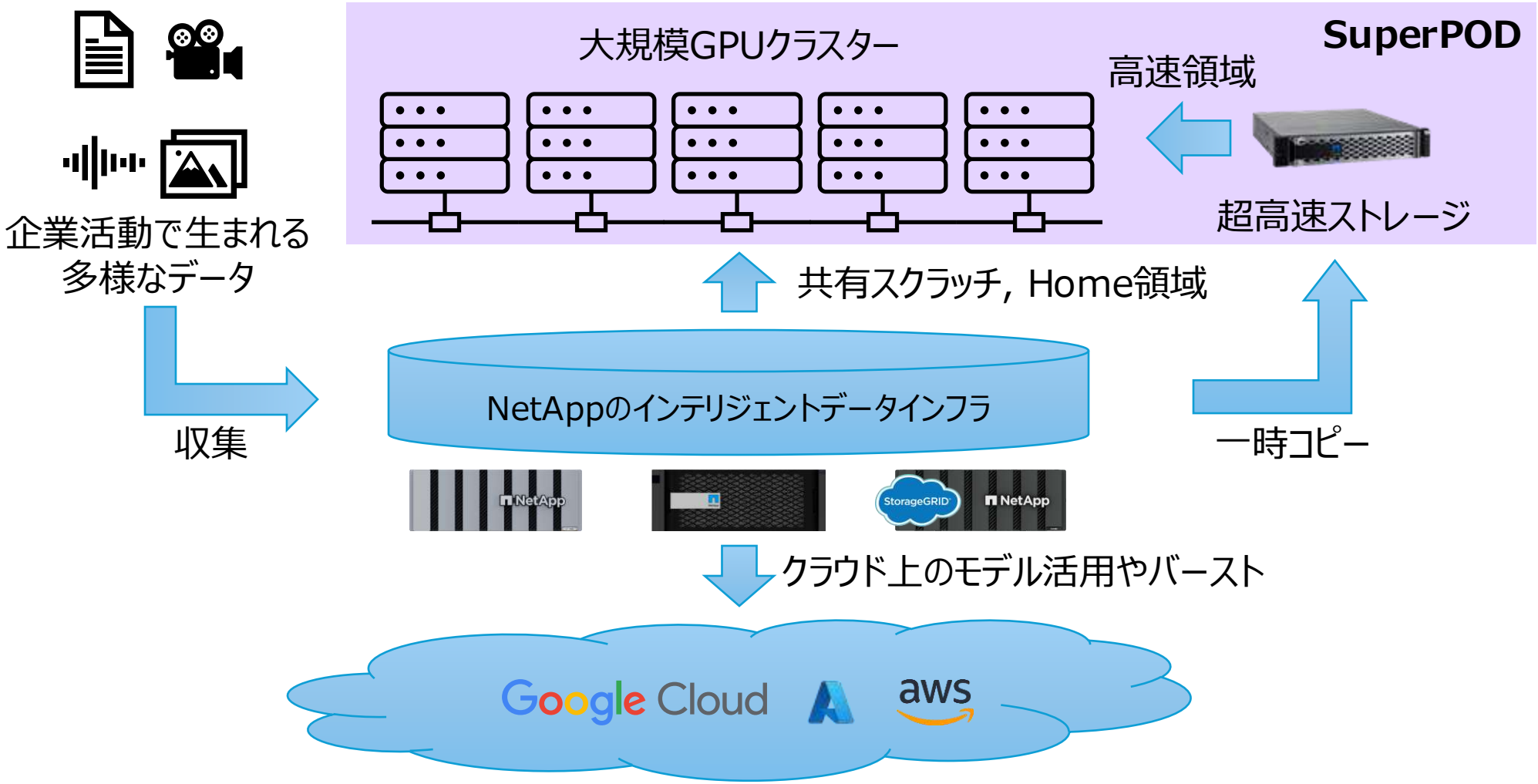
インナーリング と アウターリング

SuperPOD リファレンスアーキテクチャには、GPUクラスターから直接扱う インナーストレージ の要件のみ定義
データセットの性質によって、インナーリングにないインテリジェンスを備えたアウターリングストレージが必要



ネットアップ の アウターリング ストレージ

超高速ストレージとインテリジェント機能が搭載されたストレージを組み合わせることで、
パフォーマンスを維持したまま本番運用へ耐えれる構成を実現



【補足】DGX SuperPOD with ONTAP

ネットアップAFF A90におけるNVIDIA DGX SuperPOD サポートのテストがまもなく開始

最も強力なモデルトレーニングが、最も強力なデータ管理とともにまもなく利用可能に

- インテリジェントなデータ管理による安全で効率的なストレージ
- ハイブリッドクラウド全体で業界最高水準のデータ管理を、最も要求の厳しい次世代AIワークロードにもたらす
- パフォーマンスを犠牲にすることなく、データパイプライン全体で一貫性のある機能豊富なデータ管理を実現するように設計

TEST STARTING SOON

NVIDIA

+

NetApp

AIを使うための
ストレージ

各ベンダーから提供される様々なAIを使うための環境

自社開発、パートナーやオープンなモデルを使用できる生成AI環境がベンダー各社から提供

ベンダー	生成AIサービス	使用可能なモデル例
Amazon Web Service	Amazon Bedrock	Amazon Titan Text G1, Anthropic Claude 3.5 Sonnet
Google Cloud Platform	Vertex AI	Google Gemini 1.0 Pro
Microsoft Azure	Azure OpenAI Service	OpenAI GPT-4o
Nvidia	AI Enterprise	VILA, llama-3

検索拡張生成（RAG）とは？

生成AIモデルは、知らないことを正確に回答できない

Retrieval **検索**で、知識を

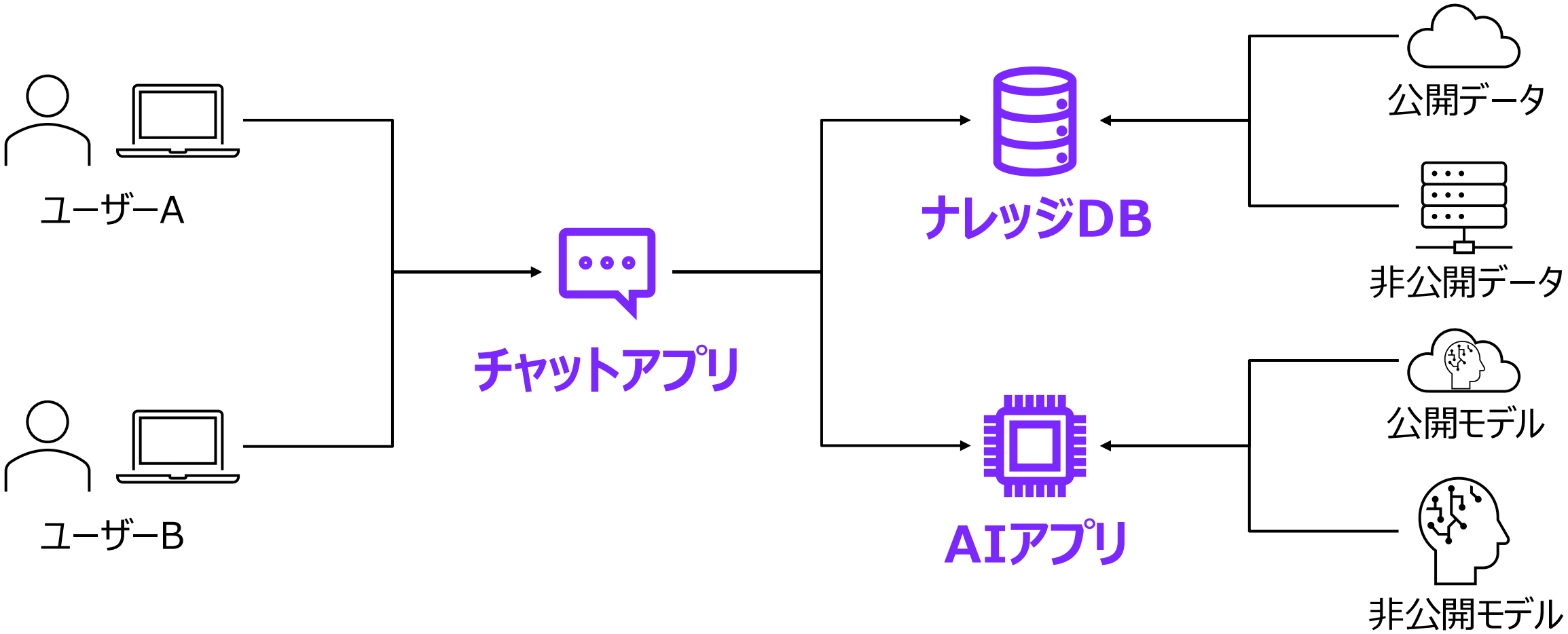
Augmented **拡張**して、文章を

Generation **生成**するAI

教科書持ち込みOKのテストのようなもの

RAGを活用した生成AIシステムの基本アーキテクチャ

大きく3つのコンポーネント（チャットアプリ、AIアプリ、ナレッジDB）とデータ・モデルを組み合わせて構築



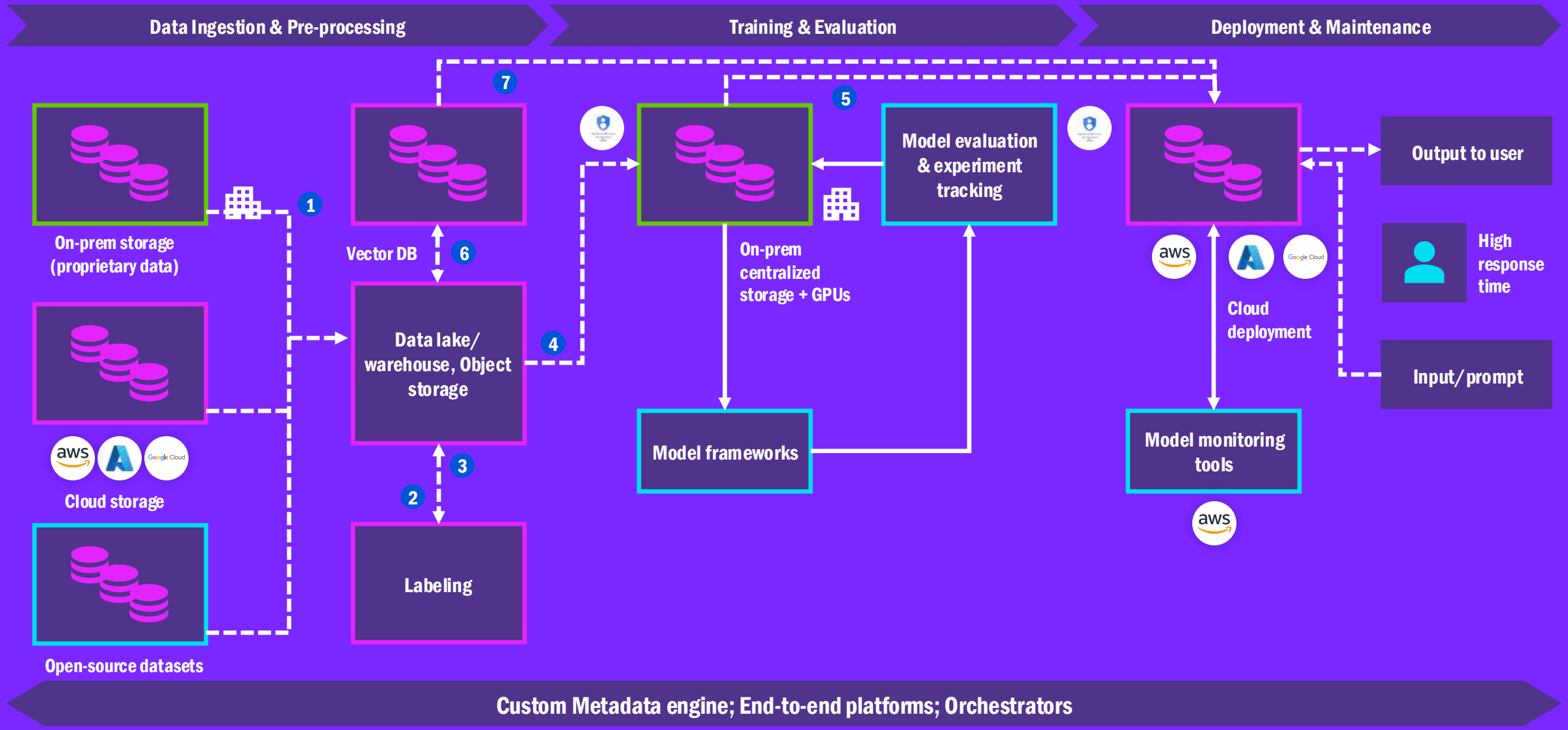
各ベンダーのAI環境上でのRAGによる非公開データの活用例

オブジェクトストレージをはじめとした各社のデータサービス上のデータや
SalesforceやSharepointのようなサードパーティ上のデータの活用が可能

ベンダー	非公開データ活用サービス	利用可能なデータソース例
Amazon Web Service	Amazon Bedrock Knowledge Bases	Amazon S3, Confluence, Microsoft Sharepoint, Salesforce, Web Crawler
Google Cloud Platform	Vertex AI Agent Builder	BigQuery, Cloud Storage, Google Drive, Cloud SQL, Spanner, Firestore, Bigtable, AlloyDB, Third-party(Confluence, Jira, Salesforce, ServiceNow, Sharepoint, Slack, Dropbox, Box, OneDrive)
Microsoft Azure	Azure OpenAI On Your Data	Azure AI Search(Blob Storage, CosmosDB, Data Lake Storage, SQL Database, Table Storage, Files, MySQL, Sharepoint), Web Crawler, Blob Storage
Nvidia	NeMo Framework	Web Crawler, RRS News, S3, Slack, TwitterなどのOSSデータローダー（LlamaIndexを活用）

RAGによる生成AI活用 を 既存インフラとどう組み合わせて実現するか

既存のデータインフラ と RAG を組み合わせた サンプルアーキテクチャ



ネットアップ の クラウドAI リファレンスアーキテクチャ & サービス

クラウドベンダー各社が提供している生成AIサービスと、ネットアップ技術が組み込まれたストレージサービスを
組み合わせて利用するためのリファレンスアーキテクチャと構築運用サービスを提供

Google Cloud

Google Cloud
Vertex AI

＋

NetApp GenAI Toolkit

＋

Google Cloud
NetApp Volumes

Microsoft Azure

Azure OpenAI
Service

＋

NetApp GenAI Toolkit

＋

Azure
NetApp Files

aws

Amazon
Bedrock

＋

NetApp BlueXP Workload Factory

＋

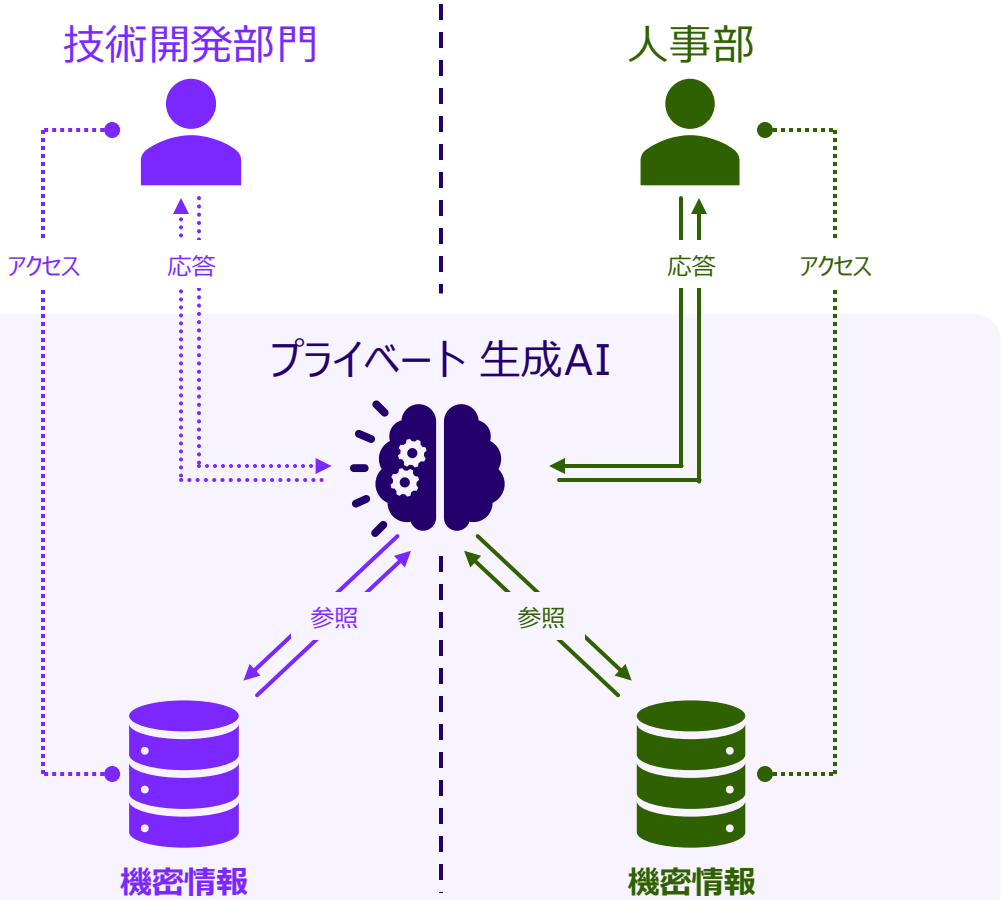
Amazon FSx for
NetApp ONTAP

データの中立性がもたらす オンプレ/クラウド に跨がるデータ基盤の実現

オンプレ、パブリッククラウドにハイブリッド/マルチに対応できるインテリジェントデータインフラによって、
生成AIを実行する環境まで、データを柔軟に移動

生成AI活用するナレッジのアクセス管理

適切なナレッジを 適切な人物にだけ提供 ※



- 企業組織内部でも、情報へのアクセス権限は異なる
- 社内のプライベート AI環境において、どのようにして プライベート 生成AI が 参照するデータの範囲と、適切に回答すべき相手を設定するか?
- コンピュータの基本：「情報」は「ファイル」の中に存在
- これまでの エンタープライズ システム環境
  - データ ファイルに対する ユーザーのアクセス権限 を管理
  - 生成AI 環境も、このアクセス権限を継承すれば、一貫した 情報へのアクセス管理が継続できる

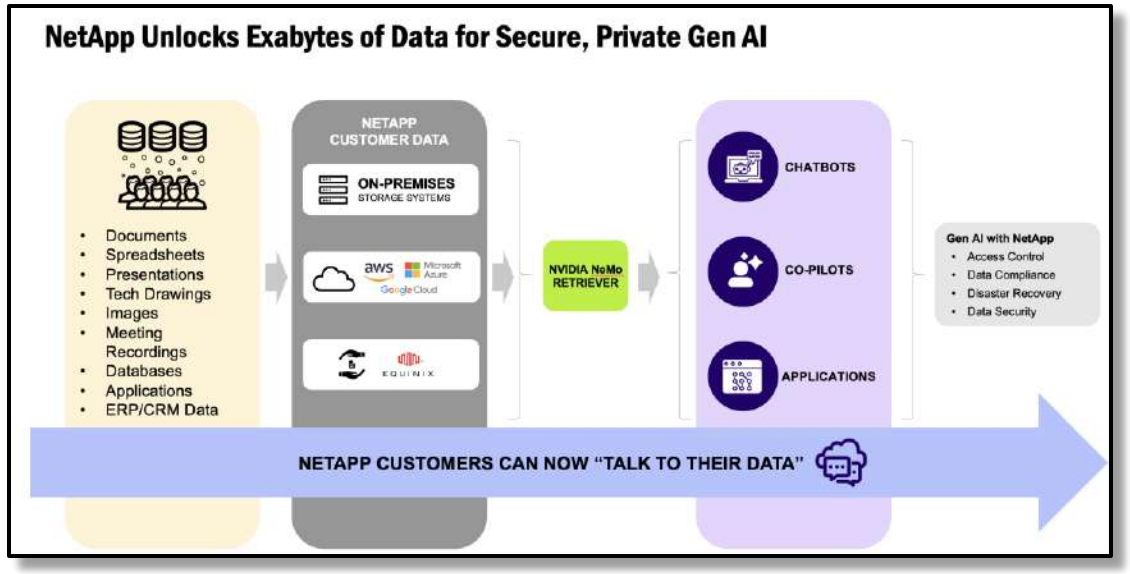
※ アプリケーションも含めた参照実装は BlueXP Workload Factoryのみ提供 (2024/10月時点)

【補足】NVIDIA NeMo Retriever

オンプレミスで生成AIによるRAG活用を支援する
NVIDIA NeMo Retrieverとの連携もサポート

TECH PREVIEW in CY24

3月に開催されたNVIDIA GTCで発表された、
NVIDIA NeMo Retriever マイクロサービスとの連携と接続が、



ただの対話だけで無く、生成AIの為のデータサービスに…

データの発見と管理

- 顧客はNVIDIA NeMo Retrieverマイクロサービス用のデータコレクションをキュレーションし、会話型AIを強化できます
- オンプレミスまたはクラウドを問わず、GenAIアプリケーションのデータ利用をプライベートかつセキュアに確保

戦略的統合

- LenovoとのNetApp AIPod™でNVIDIA OVXとの統合を計画中。
- 企業における生成型AIのデータアクセス/制御の課題に対応

セキュリティとコンプライアンスの重視

- 厳格な企業基準向けに設計。
- セキュアな生成AIの導入を促進します。

可能性を”更に”拡げるインテリジェンスとは

～ネットアップの取り組みから見えてきた今後の課題とビジョン～

可能性を “更に” 拡げるために、AI用ストレージインフラに足りないインテリジェンス

ネットアップのAIソリューションを用いたデモやハンズオン、PoC支援を実施してきた経験から見えてきた、
今後”更に”可能性を拡げるために、ストレージインフラが備えるべきインテリジェンス

データの最新性の維持

- 生成AIを「検索」のユースケースで使う場合、データの最新性を維持することが困難
- 例えば、1日ごとにナレッジを最新化するような定期実行を組み込むことはできるが、ユーザーが知りたいのは最新の情報

データの分類と埋込効率化

- ナレッジ化される情報に機微な情報が無いかを判別し、場合によっては匿名化等の処理を行う義務が生じる
- ナレッジ化する際に、文章を一塊のチャンクに分割して処理を行うが、適切なサイズや、ベクトルの次元を特定しないと、求められる精度が得られない

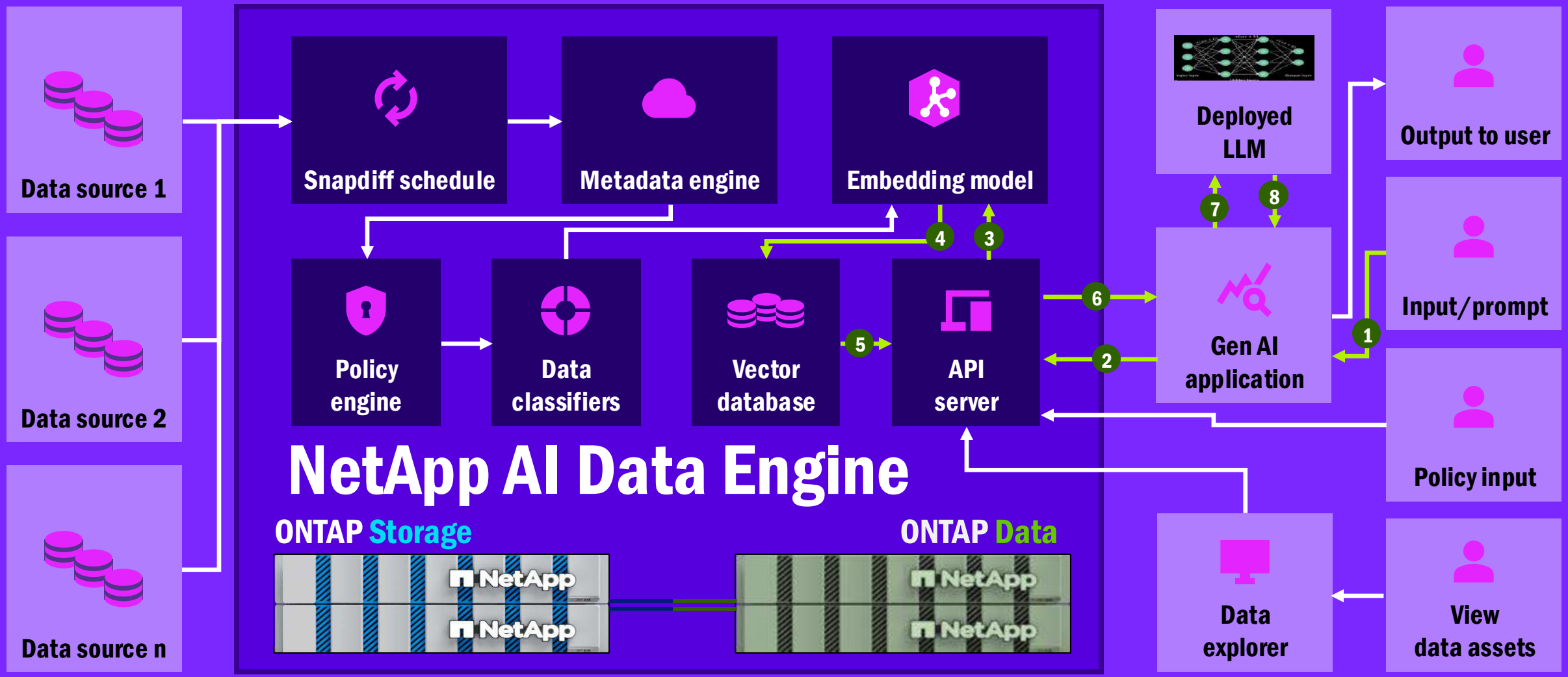
ナレッジベースの運用管理

- RAGフレームワークでは、一般的に運用実績のあまり無いベクターデータベースや全文検索データベースが必要
- 運用が進むと肥大化※してしまい、運用管理に負荷がかかってしまう恐れも
  （※元のテキストの10倍にサイズが増加する事も）

VISION

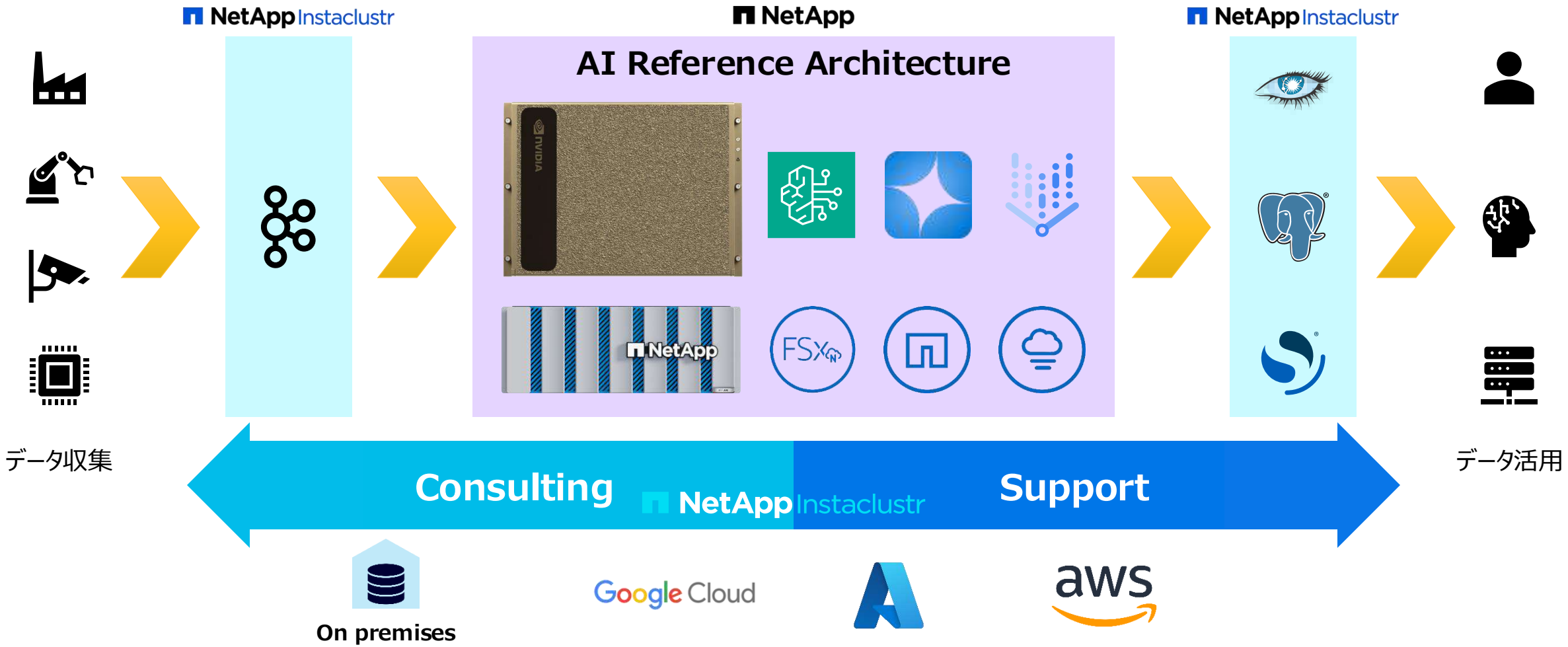
NetApp AI Data Engine Vision

プライベートデータのAI活用の可能性を”更に”拡げるためのデータサービスをONTAP上でも提供予定



【補足】Open Source Software が拡げるAIの可能性

今すぐにストレージへのデータの収集やAI活用の運用を楽にしたい方には、ネットアップの別ブランドであるInstaclustrにて、
Open Source Softwareのコンサルティングやサポートも提供



KEY TAKEAWAYS

01

生成AIは国内でも盛り上がっているが、学習データの枯渇（2026年問題）により、
プライベートデータをどのように活用するのかが重要に

02

AIを作る、AIを使うという2種類のAIインフラでプライベートデータを活用するためには、
既存データインフラとの連携や**増加するストレージコスト**、**サイバー攻撃への対応**が課題

企業で使われる既存データインフラには、そのようなインテリジェンスが搭載されており、
AIにおいても価値を発揮する

03

ネットアップのAIに対する取り組みから見えてきた、既存のデータインフラに足りない
最新性の維持や**データの分類と埋込効率**、**ナレッジベースの運用管理**に関しても、
今後、ストレージの機能として提供予定

THANK YOU