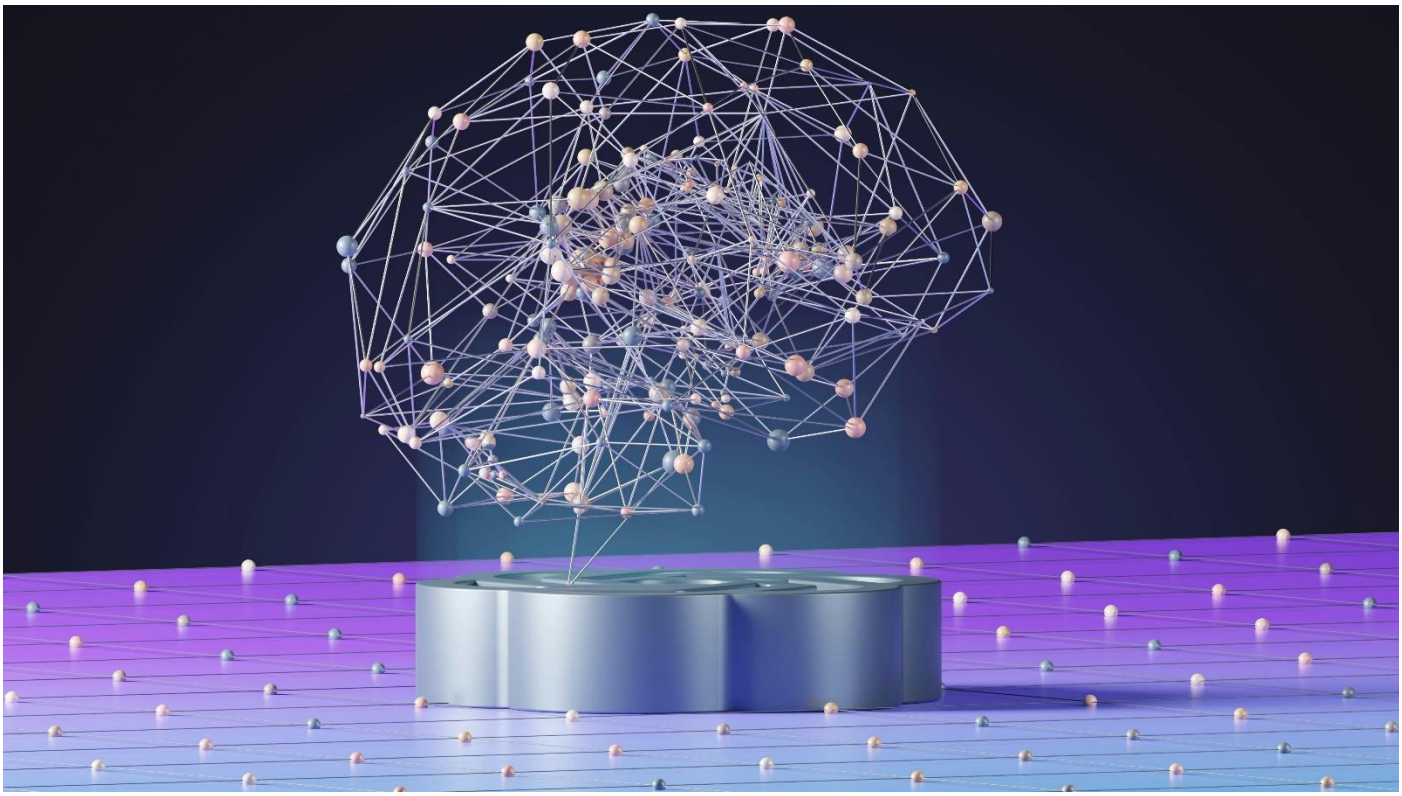
Hewlett Packard Enterprise

Diverta

テクニカルホワイトペーパー

Kuroco RAG on HPE

RAG の精度向上を実現するオンプレミス RAG ソリューション



目次

1 はじめに

Kuroco RAG（Retrieval-Augmented Generation）は、企業が増大するデータを効率的に管理し、正確かつ迅速な情報提供を実現するために開発されたソリューションです。

現代では、膨大な情報の中から正確なデータを迅速に検索し、その情報に基づいた回答を生成することが求められています。しかし、従来のシステムではその限界が明らかになってきました。本ソリューションは、検索技術と自然言語生成を組み合わせることで、企業の情報活用を効率化し、蓄積された情報に基づいて迅速に回答を提供する必要がある場面で、その効果を発揮します。

本ソリューションは、HPE ProLiant サーバー上でオープン LLM を起動し、Kuroco RAG と連携させることで、オンプレミス環境で RAG システムを完結して利用できるアプライアンスとなります。

本ホワイトペーパーでは、本ソリューションの実際の動作、利用方法を解説し、ソリューション概要をご紹介します。

2 協業パートナー情報

本ソリューションは、ディバータ社と共同検証し、共同販売します。

2.1 会社情報

会社名	株式会社ディバータ
住所	東京都新宿区神楽河岸 1-1 セントラルプラザ 6 階
事業内容	・Contents Management System（CMS）開発 ・Web アプリ構築 ・インターネットサイトの企画・構築から運営まで ・インターネットを利用した新サービスの開発研究

2.2 主要ビジネス概要

株式会社ディバータは、CMS（コンテンツ管理システム）を主要サービスとしてクラウド商材を提供する企業です。代表的な製品である「Kuroco」は、API 管理や認証機能を統合したヘッドレス CMS で、企業のデジタルコンテンツの効率的な管理と配信を支援します。2024 年 4 月には、「Kuroco」の機能を拡張し、RAG（Retrieval-Augmented Generation）を実現する「Kuroco RAG」のサービス提供を開始しました。これにより、高度な検索技術と自然言語生成（NLG）を組み合わせた AI 技術の推進にも取り組んでいます。

「Kuroco RAG」は、企業内外の膨大なデータから迅速かつ正確に情報を検索し、そのデータに基づいて最適な回答を生成することで、問い合わせ対応や情報提供の効率化を実現します。クラウドベースのソリューションを通じて、迅速なサービス展開と柔軟な運用が可能となり、企業の DX（デジタルトランスフォーメーション）を強力に推進します。

3 ソリューション概要

本ソリューションでは、協業パートナーが提供する「Kuroco RAG」とオープン LLM を同梱したアプライアンスを提供します。

3.1 主要機能一覧

3.1.1 高精度な回答データ抽出

・RAG 技術を活用し、利用者の質問に対して最適な回答を提供。

・クエリ拡張（※1）とベクトル検索を駆使して、登録された回答データから正確な情報を抽出。

・回答データにカテゴリやタグなどの情報を付加することで、回答結果の調整が可能。

3.1.2 ジェネレーティブ UI を活用したユーザー画面での質問回答

・利用者の質問について、回答データを使ってテキストおよび画像を用いた回答画面を自動生成。

・回答画面上で、参照元情報を表示。

・回答画面上で、利用者の質問に対して、AI を使い関連する質問で生成（※２）。

3.1.3 Web 管理画面を使った RAG の精度調整

・カテゴリやタグを使ったデータの構造化が可能。

・エンベッディングモデルを指定し、回答データに合ったモデルを指定可能。

・データ登録時に回答データを AI にて自動で最適化（ベクトルデータの最適化）。

3.1.4 レスポンスの最適化

・利用シーンに応じた 3 種類のレスポンスのパターンを用意し選択可能。

・要約処理を行う場合の生成 AI のモデル指定が可能。

・プロンプトの設定を行うことで、回答内容に関する調整が可能。

3.1.5 データ登録の多様な手法

・管理画面からの入力および CSV アップロードで回答データの登録が可能。

・Web サイトやストレージから html や PDF などのファイルをクロールし、機械的に回答データ取り込み。

・ソリューションが提供する API で、様々なシステムやクラウドサービスとの回答データのデータ連携構築が可能。

3.1.6 様々なシステムと API 連携

・REST API でさまざまなシステムに組み込み可能で、お客様の要望に応じてカスタマイズが可能。
（質問回答、回答データの登録/変更/削除、そのほか）

3.1.7 オープン LLM の柔軟な選択

・日々進歩するオープン LLM を柔軟に設定可能。

3.1.8 必要十分な GPU を搭載

・RAG システムを構成するために必要な LLM を起動するために必要十分なスペックの GPU を搭載。

※1　クエリ拡張

クエリ拡張は、本ソリューションの回答精度を高める独自機能です。

一般的な RAG の動作は、利用者が質問した内容と事前に登録された回答データをベクトル検索し、データの類似性（ベクトル距離の近さ）から情報を抽出した後、その抽出結果を生成 AI によって回答として生成する流れとなります。しかし、利用するデータによっては、ベクトル距離だけに依存することで誤った情報を抽出する場合があります。

本ソリューションでは、利用者の質問内容に対して AI による処理を行い、事前に登録された回答データのカテゴリやタグを自動判定します。これにより、ベクトル検索を行う前に一次的な絞り込みを行い、ベクトル検索のみの場合よりも高精度な情報抽出を可能にしています。

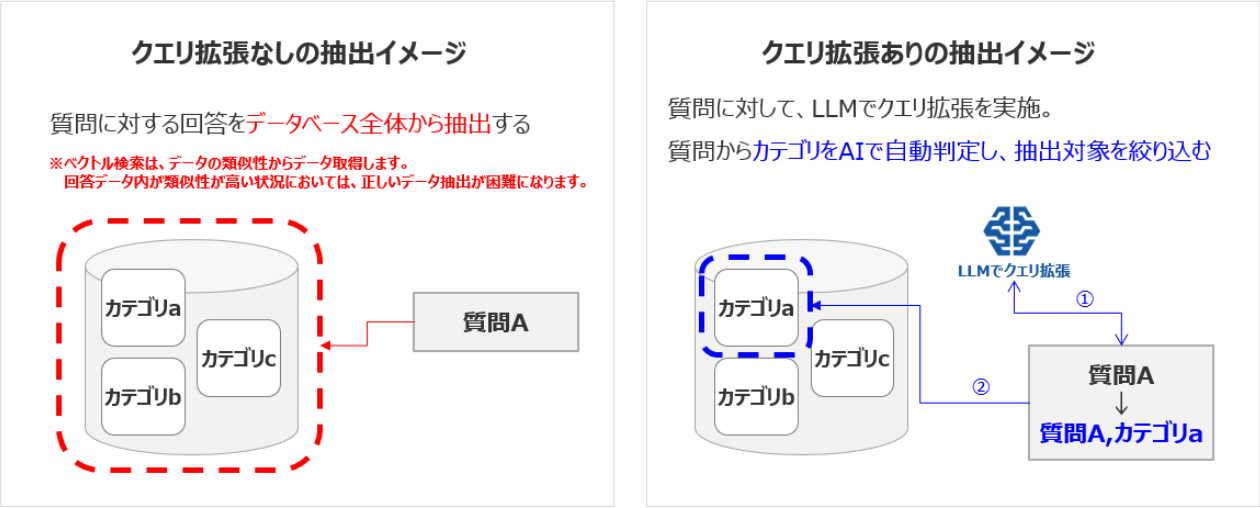


図 1　カテゴリを利用したクエリ拡張の抽出イメージ

※2　ユーザー画面での「関連する質問」の生成

利用者が入力した質問を AI で処理し、関連する質問の候補を自動生成します。

利用者は生成された関連質問から適切なものを選択して画面を進めることで、求める情報にスムーズにアクセスできるようになります。これにより、AI に対する質問が苦手な利用者でも期待する情報に到達しやすくなり、自由入力のチャット形式と比較して利用者のリテラシーに依存することなく、円滑な情報取得が可能となります。

3.2セキュリティ

本ソリューションでは、以下のセキュリティ対策について設定可能です。

管理画面	完全に暗号化された HTTPS 通信
	TLS 証明書
	ID/PWD によるアクセス制限
	ユーザーのグループ設定による細かな権限の制御
	アクセスログ（監査ログ）・アプリケーションログ
	（オプション） IP アドレス制限
	（オプション） 認証アプリによる 2 要素認証の設定
	（オプション） SAML/OAuth による外部ログイン連携
ユーザー画面	完全に暗号化された HTTPS 通信
	TLS 証明書
	アクセスログ（監査ログ）・アプリケーションログ
	（オプション） ユーザーのグループ設定による細かな権限の制御
	（オプション） IP アドレス制限
	（オプション） 認証アプリによる 2 要素認証の設定
	（オプション） SAML/OAuth による外部ログイン連携
API	完全に暗号化された HTTPS 通信
	TLS 証明書
	独自ドメイン設定
	固定トークン・ダイナミックトークン・Cookie を利用したアクセス制御
	CORS の柔軟な設定
	ユーザーのグループ設定による細かな権限の制御
	アクセスログ（監査ログ）・アプリケーションログ
	（オプション） IP アドレス制限
	（オプション） SAML/OAuth による外部ログイン連携

表 1　Kuroco RAG のセキュリティ対策

※上記は、2024 年 12 月時点の情報となります。

※オプションの項目については、要件確認の上でご提案いたします。

3.3画面イメージ

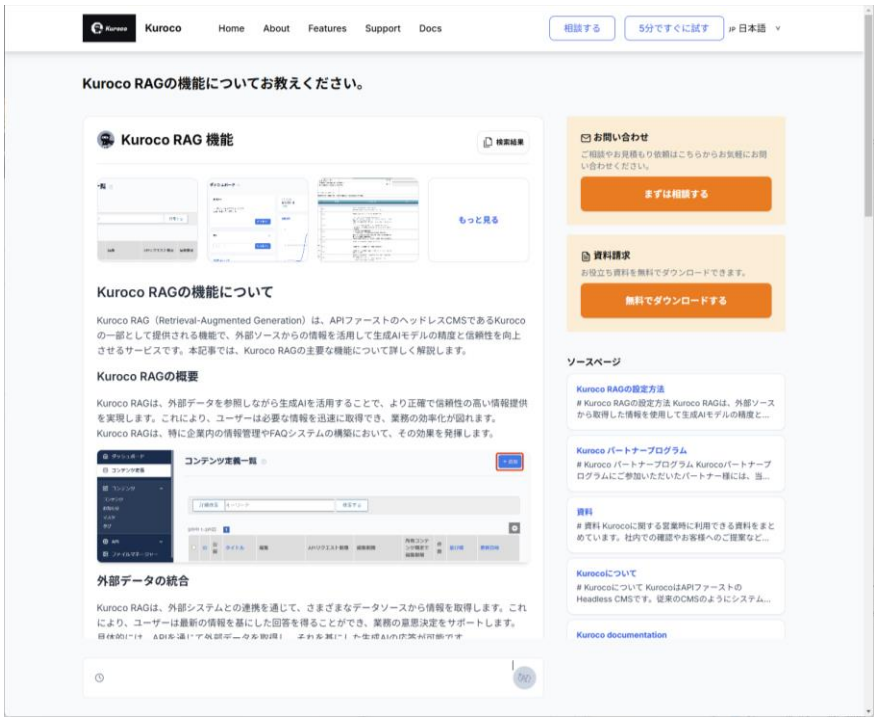


図 2　ユーザー画面（質問に関する回答画面生成）

図 3　ユーザー画面（関連する質問の生成)

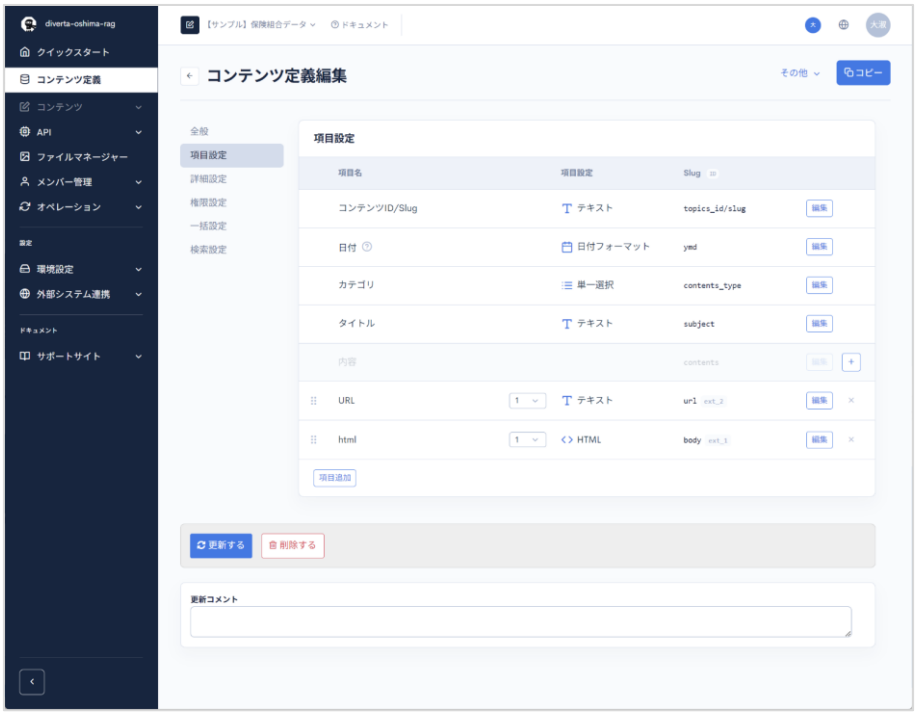


図 4　コンテンツ定義 - 項目設定

図 5　コンテンツ (回答データ)

図 6　API 設定

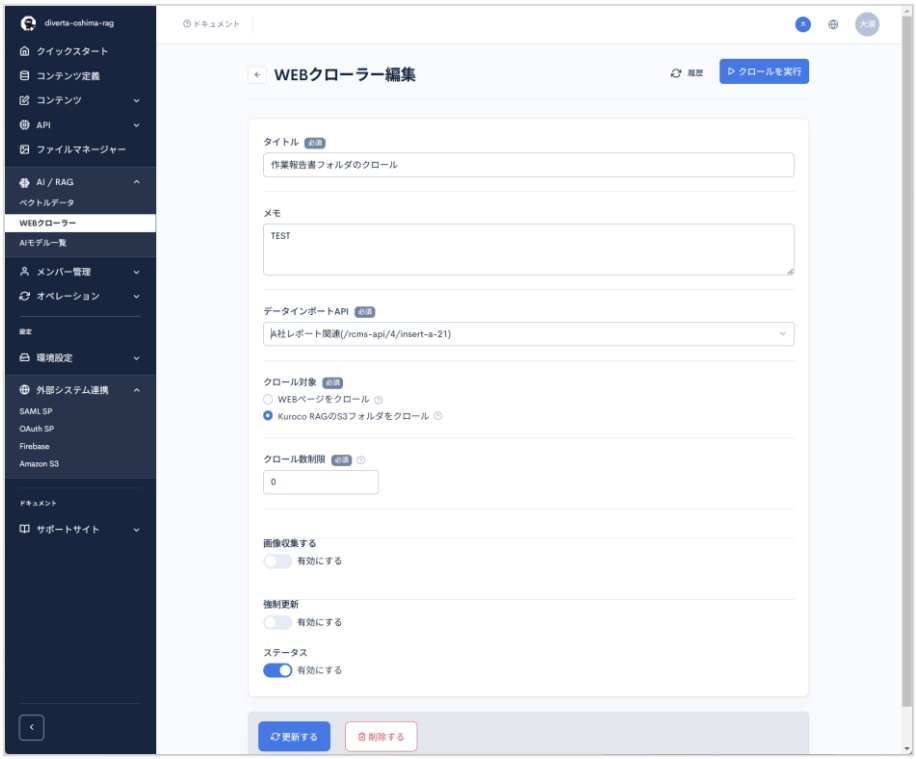


図 7　WEB クローラー

3.4 使用している言語モデル

使用している言語モデルは、下記となりますが、下記以外のオープン LLM も設定可能です。LLM を起動するフレームワークのサポート状況により対応していない場合もあります。

Completions models	Llama-3-ELYZA-JP-8B
Embedding models	multilingual-e5-large-instruct

※2024 年 10 月時点の情報です。

Completions models	Qwen/Qwen2.5-14B-Instruct
Embedding models	BAAI/bge-m3

※2024 年 12 月時点の情報です。

表 2 使用している言語モデル

3.5 アーキテクチャ

Kuroco RAG on HPE のアーキテクチャ図は以下となります。

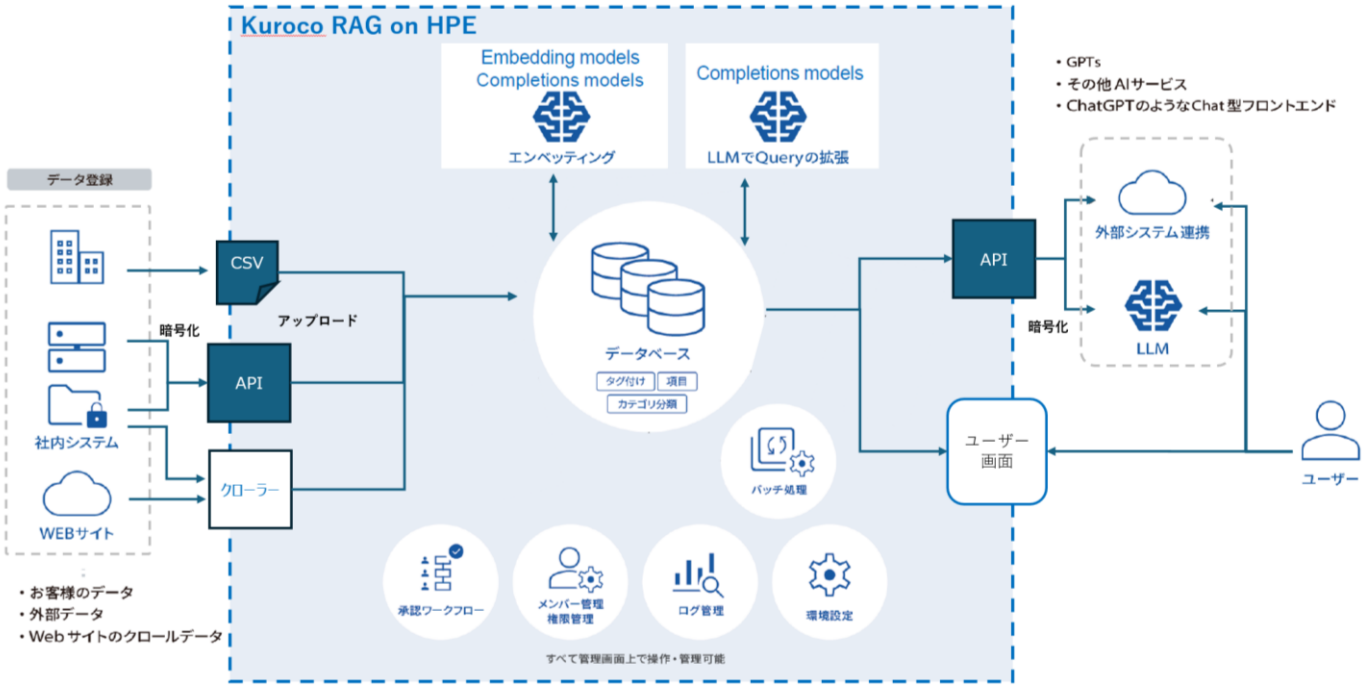


図 8　Kuroco RAG on HPE アーキテクチャ図

注記

2024 年 12 月時点でのアーキテクチャ図となります。あくまで参考例としてご参照ください。予告なく変更する可能性がございます。

4 ソリューションの利用方法

本ソリューションの利用方法をご紹介します。

4.1 基本的な利用

本ソリューションの利用方法としては下記となります。

4.1.1 ユーザー画面での質問・回答

利用者は、本ソリューションで提供されるユーザー画面へ質問を入力することで回答を表示することができます。回答には、RAG の回答で利用した引用元情報や、入力された質問を AI で処理し関連する質問を作成し表示します。回答画面からの再質問については、テキストでの質問だけでなく、関連する質問を選択することで、利用者のリテラシーに依存せずに、求めている情報を入手できるユーザーインタフェイスを提供します。

4.1.2 API を使った他システムへの組み込み

本ソリューションでは、RAG の質問・回答機能を API で提供します。既存のグループウェアや業務システムに API を組み込むことで、ご利用中のシステム上で RAG を追加できます。社内システムにも容易に RAG の機能を統合できるため、内部業務の効率化にも活用できます。

4.2 チューニング

本ソリューションでの主な調整項目は下記となります。

4.2.1 カテゴリ・タグの調整

クエリ拡張機能で利用するためのカテゴリとタグを設定可能です。カテゴリとタグで利用者の質問内容に応じて、ベクトル検索によるデータ抽出の前に一次的な絞り込みを行い、不要な情報を検索範囲から除外することができます。

4.2.2 チャンキングおよびエンベッティングの調整

管理画面上でチャンキングおよびエンベディングに関する設定が可能です。回答データに適したエンベディングモデルを選択することで、ベクトル検索の精度が向上します。また、AI を使ったベクトルデータの最適化やベクトルテンプレートの編集を行うことで、ベクトル化を行う項目の調整も可能です。

4.2.3 レスポンスの調整

管理画面上で RAG の回答に関する調整が可能です。生成 AI のモデル指定、生成 AI のプロンプト設定、レスポンスの種類などを設定することができます。API を使って AI チャット連携や業務システムへの組み込みを行う場合は、レスポンスの種類を選択することで、回答時間を最適化することもできます。（ベクトル検索のみ、検索拡張＋ベクトル検索、検索拡張＋ベトル検索＋要約処理）

4.3 導入検討時のサービス確認

本ソリューションの導入検討時のサービス確認について、フリートライアル環境のご提供可能です。
フリートライアルは、オンライン上での簡単な申し込み手続きでご提供可能で、システム要件や画面の操作イメージ、回答精度などのご確認を素早く行うことができます。フリートライアル環境を使った PoC の実施もご提案可能です。

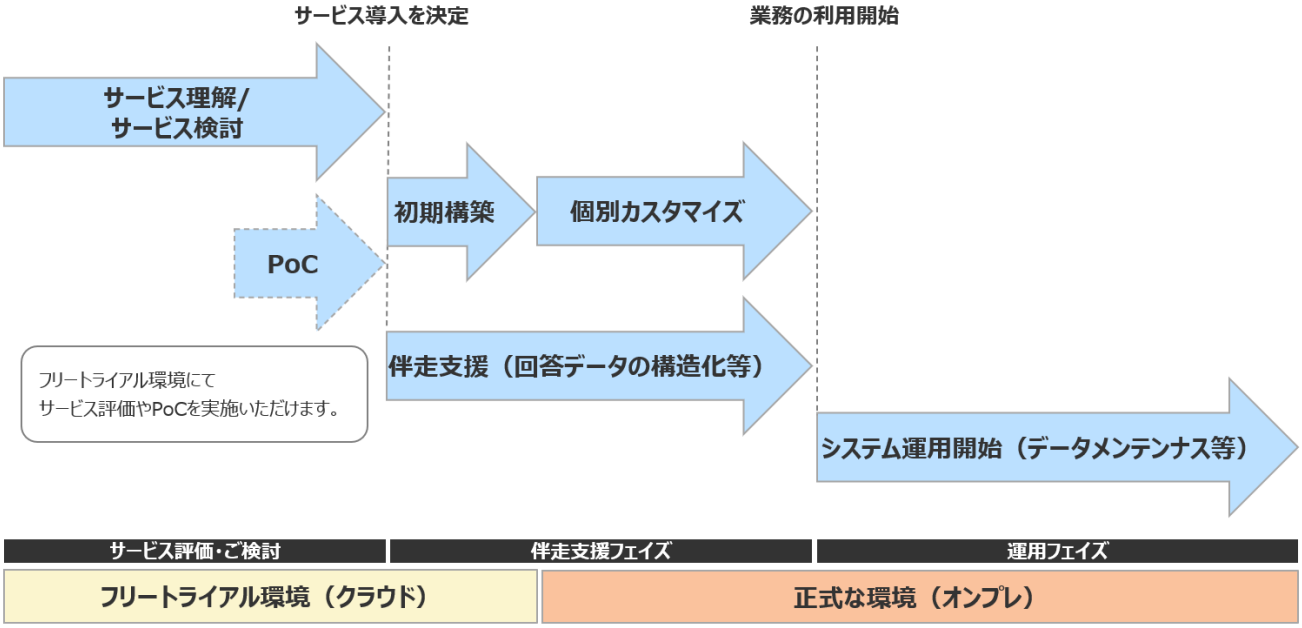


図 9　導入検討の流れ（想定モデルケース）

<ご注意>

フリートライアル環境は、本ソリューション（RAG および LLM）についてクラウドサービスとしてご提供します。正式導入時の環境とは異なる環境であることをご留意ください。

PoC の実施については、実施内容により有償対応となる場合がございます。

4.4カスタマイズ

本ソリューションは、利用シーンや業務に応じたカスタマイズが可能で、企業の業務効率向上に貢献します。

※　カスタマイズについては、都度要件を確認させていただき、個別見積りさせていただきます。

4.4.1　ユーザーインタフェイスの自由なカスタマイズ（ユーザー画面の個別構築）

本ソリューションは、ユーザーごとに異なるニーズに応じてインターフェースを柔軟に設計、ご提供いたします。パーソナライズされた操作画面や、業務内容に応じた入力項目の追加など、各業務に最適なユーザー体験を提供します。

4.4.2　イントラシステムへのシームレスな組み込み（既存システムへの組み込み）

本ソリューションは、イントラシステム内に RAG の質問応答機能の UI を直接組み込むことが可能です。利用シーンに合わせたカスタマイズ可能なインターフェースにより、直感的で使いやすい質問応答機能を提供し、社内業務の操作性と効率を大幅に向上させます。

4.4.3　多彩な API 連携で幅広い社内サービスとデータ連携

本ソリューションには、社内の各種サービスやシステムと容易に統合できる API が備わっています。データの一元管理や情報共有が円滑になり、各部署での情報アクセスが強化され、業務スピードの向上を支援します。また、WEB クローラーや CSV アップロードを活用し、多様なデータソースを取り込み、AI 応答の精度を最適化します。

5 構成のベストプラクティス

このセクションでは HPE のサーバーポートフォリオとその特徴、および Kuroco RAG on HPE の推奨構成を提示します。

5.1 HPEサーバーポートフォリオとその特徴

HPE では、汎用的なラックマウント、タワー型のサーバーシリーズである HPE ProLiant サーバーや、ブレード型の HPE Synergy、エッジ型の HPE Edgeline、High Performance Computing の Cray シリーズ、ミッションクリティカルの Superdome シリーズ、無停止コンピュータシリーズの Non S top、仮想化専用機のハイパーコンバージドインフラなど、幅広いラインナップをご用意しています。

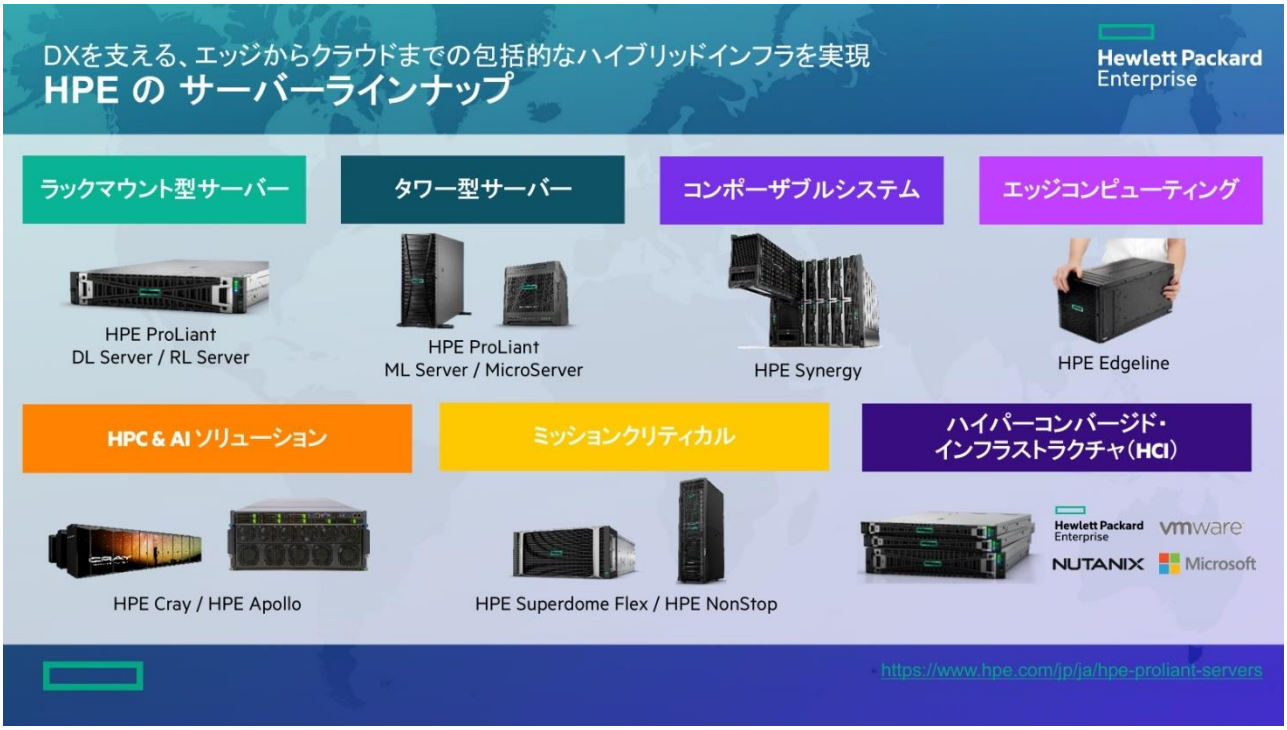


図 10　HPE サーバーラインナップ

エンタープライズ向けの汎用サーバーとしては、ラック型、タワー型を中心に、搭載できる CPU タイプ、CPU 数および、筐体サイズによって分かれます。

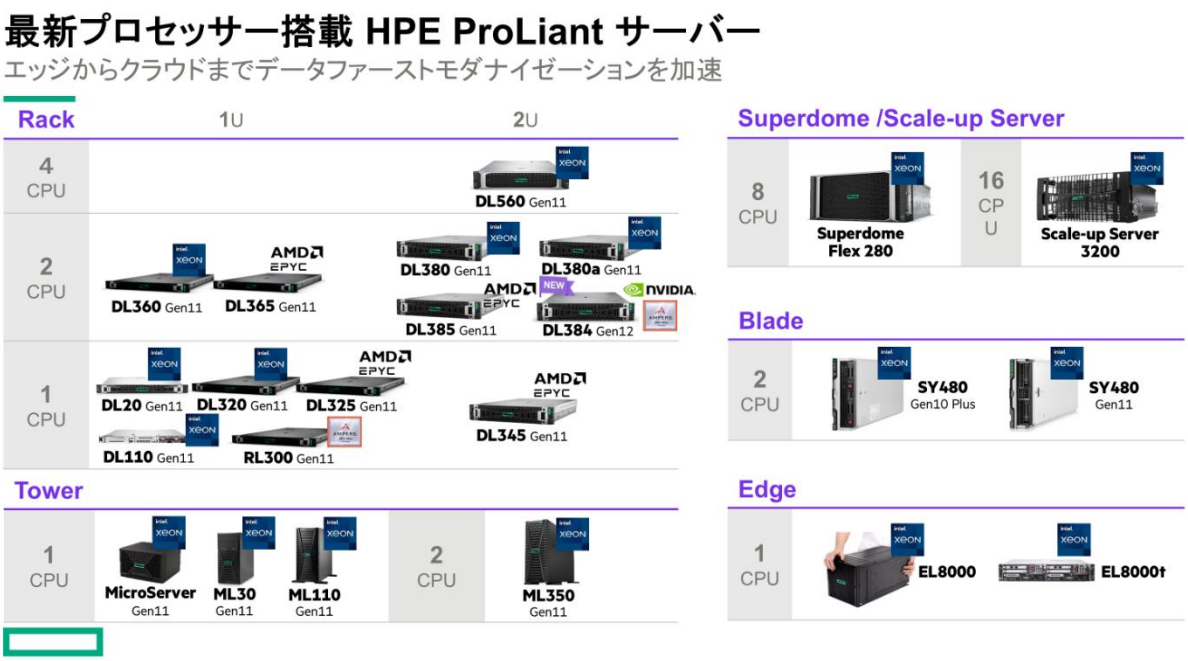


図 11　最新プロセッサー搭載 HPE ProLiant サーバーシリーズ

HPE サーバーの特徴として、“直感的”、“安心”、“最適化”を謳っており、オンプレでもクラウド型の監視・管理が可能な SaaS サービスの提供や、ファームウェア改ざん対策をはじめとしたセキュリティ機能、サービスの提供、中でも AI サーバーとしてご利用いただけるよう最適なサーバー設計として GPU の搭載枚数の拡充や、排熱量低減、省電力化を意識したデザインとなっています。

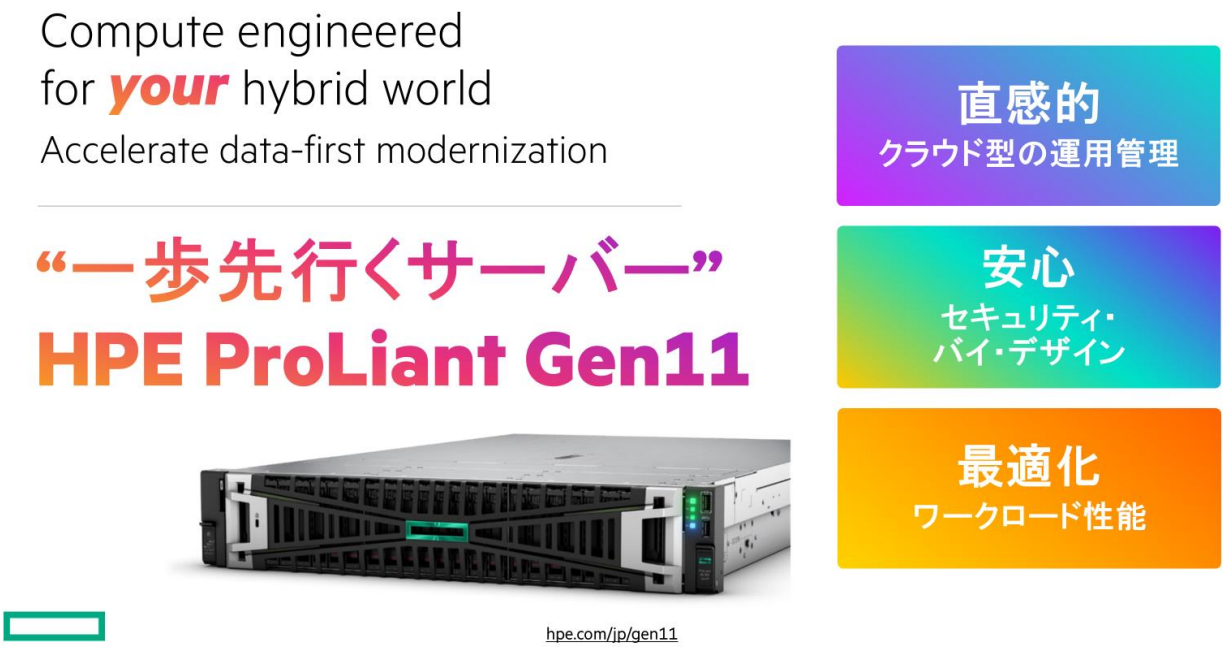


図 12　HPE ProLiant サーバーの特徴

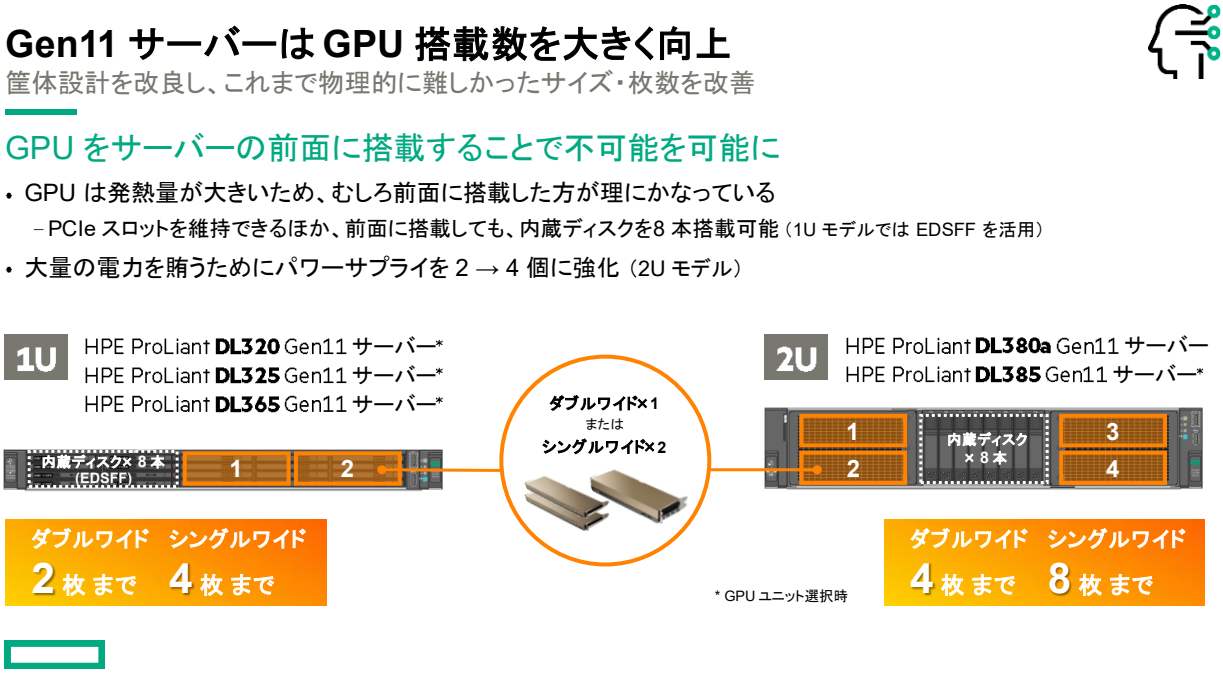


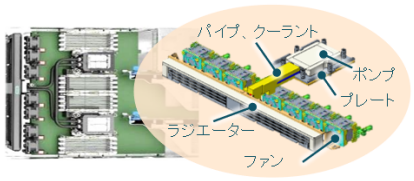
図 13　GPU 搭載枚数の拡充

最新サーバーの高性能がゆえの発熱量・消費電力を抑える工夫

電力効率のよい機種や筐体、パーツの選択

液冷ソリューションの“民主化”

液冷＋空冷で**省電力**



HPE スマートリキッドクーリング

- ✓ ハイブリッド型内部冷却方式（Closed-loop Liquid Cooling）
- ✓ クルマのように、サーバー内部に小型ラジエーターを搭載し、空冷ファンによる風で冷却水を冷やす構造
- ✓ 特別な工事や外部ユニット不要

電源周りの効率化

PlatinumからTitaniumに変更することで、**省電力**

80 PLUS PLATINUM 94%　80 PLUS TITANIUM 96%

- ✓ 最大96%の高効率電源ユニット

- ✓ 500w ～ 2200wまでの出力

1000w　2200w

- ✓ x4電源ユニット搭載可能（x2 main board + x2 GPUs）
- ✓ Ex. DL380a / DL385

筐体サイズの最適化

1Uから2Uに変更することで、**省電力**

1Uサーバー

2Uサーバー

1Uサーバーのヒートシンク

2Uサーバーのヒートシンク

図 14　発熱量・消費電力を抑える工夫

5.2 システム構成

5.2.1 サーバースペック

ディバータ社の Kuroco RAG 及び、オープン LLM をホストするサーバーとして、必要なサーバースペックを図 15 に示します。

	Model	Qty	Spec, Options, Notes
GPU サーバー	 HPE ProLiant **DL380** Gen11	1	**Model:** HPE ProLiant DL380a Gen11 4 Double Wide **CPU:** Intel Xeon-Gold 6426Y 2.5GHz 16-core 185W Processor x2 **MEM:** 512GB / HPE 64GB (1x64GB) Dual Rank x4 DDR5-4800 x8 **Disk:** HPE 1.92TB SATA 6G Mixed Use SFF BC Multi Vendor SSD x2 **Array:** HPE MR216i-o Gen11 x16 Lanes without Cache OCP x1 **OS Boot**: HPE NS204i-u Gen11 NVMe Hot Plug Boot Optimized Storage Device **Network**: Broadcom BCM57416 Ethernet 10Gb 2-port BASE-T OCP3 Adapter x1 **PS:** HPE 1800W-2200W Flex Slot Titanium Hot Plug Power Supply Kit x4 **iLO:** iLO Advanced ライセンス
GPU	 NVIDIA **L40S**	1	**GPU:** NVIDIA L40S 48GB PCIe Accelerator

図 15　Kuroco RAG on HPE 推奨構成例

2U サイズの 2CPU サーバーで、世界で最も人気の HPE ProLiant DL380 Gen11 サーバーに、Ada Lovelace アーキテクチャの汎用万能型 GPU の NVIDIA L40S を 1 枚搭載しています。CPU、GPU をはじめとするマシンスペックを増強することで、さらに高性能な AI 環境をご利用いただくことも可能です。

5.2.2 ソフトウェア要件

Kuroco RAG on HPE のソフトウェア要件は以下となります。

OS	Ubuntu 22.04.4 LTS
GPU ドライバー	NVIDIA Driver 550
コンテナ ランタイム	Docker 27
コンテナ ランタイム (GPU)	NVIDIA Container Toolkit 1.17
推論フレームワーク (LLM, Embedding)	vLLM 0.6.4

表 3　Kuroco RAG on HPE ソフトウェア要件

6 お問い合わせ先

HPE Services

計画から運用、またその後においても、HPE エキスパートはエッジからクラウドに至るまで変革を加速し、運用を最適化するとともに、IT 投資を最大限活用できるようお手伝いします。

https://www.hpe.com/jp/ja/services.html

7 その他のリソース

HPE ProLiant DL380 Gen11 サーバー

お客様のニーズに最適な製品をお選びください。

HPE 営業にお問い合わせください。

Get updates

2025 年 1 月

Hewlett Packard Enterprise