

# 社内ドキュメントに基づく情報検索

シンプレクス株式会社

- シナリオ概要
- アプリUI
- アーキテクチャ図
- 考慮事項
- デプロイ方法
- バリデーション用 検証環境情報
- Appendix

## Azure OpenAI Serviceを用いた高度な社内ドキュメント検索

生成AIの学習データに含まれていない社内ドキュメントなどの情報を扱う場合、検索エンジンと組み合わせたRAG（Retrieval Augmented Generation）を構築する必要があります。

本アーキテクチャにより、社内ドキュメントの情報をを用いた検索、生成AIによる回答が可能になります。また、回答根拠としてドキュメントのリンクを提示することで、ハルシネーションの確認が容易になります。

本アーキテクチャの特徴	説明
生成AIの文字数制限への対策	生成AIには入力可能な文字数の制限があります。サイズの大きいドキュメントも扱えるようにドキュメントを分割する仕組みを導入しています。適切な分割サイズを指定することで検索精度の向上も期待できます。
検索アルゴリズムのカスタマイズ性	Azure Cognitive Searchを検索エンジンとして採用しており、全文検索・セマンティック検索・ベクトル検索といった様々なアルゴリズムの利用やチューニングが可能です。
アクセス権限制御	社内のアクセス権限と同じ設定を本アーキテクチャに転用できる仕組みを構築しています。閲覧可能なドキュメントのみを生成AIにインプットする制御が可能です。
ネットワークの安全性	本アーキテクチャでは、さまざまなAzureリソースを使用しています。ドキュメントの機密性を保護するため、リソース間の通信を閉域化することで最大限セキュアに設計しました。一部、閉域化未対応のAzureリソースに関してはIP制限しています。

- SlackのBot(@schat)に対してメンション付きの質問をします。
- ユーザの閲覧可能なドキュメントに基づく回答と、その根拠となったドキュメントが出力されます。

 [Redacted] 4日前

@schat doc 勉強用に書籍を購入したいのですが、会社から補助はされますか？書籍の種類に制約はありますか？

 schat アプリ 4日前

はい、会社では一定の勉強用書籍の購入を経費精算で認めています。ただし、以下の条件を満たす書籍に限ります。

- 業務遂行上の必要に基づく書籍であること。
- 職務に直接必要な技術や知識を習得させるためのものであること。

最終的な判断は [Redacted] によって行われます。また、一か月に購入できる冊数は [Redacted] までで、一冊の上限価格は [Redacted]、一か月の上限価格は [Redacted] です。

[関連Path]

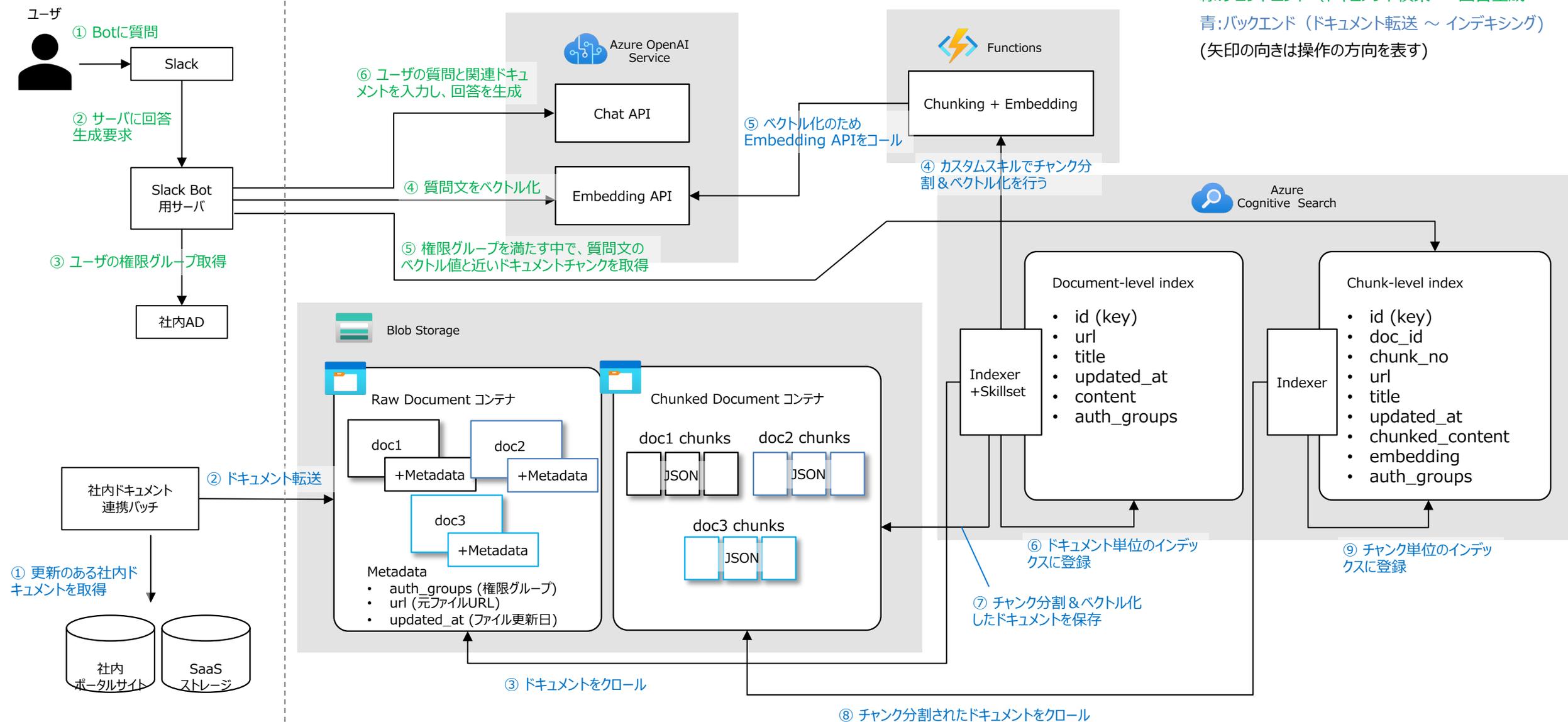
- 資料A
- 資料B
- 資料C

- 質問文の文頭に付ける文字列（コマンド）によって様々な出力を得ることができます。

コマンド	機能
(なし)	関連ドキュメントを検索せず、モデル学習時の知識のみで回答を生成
doc	関連ドキュメントを検索した上で、その内容に基づく回答を生成
path	関連ドキュメントのパスのみを出力
pathv	関連ドキュメントのパスと、該当部分のテキストを出力

# アーキテクチャ図 (リソースとデータフロー)

緑:フロントエンド (ドキュメント検索 ~ 回答生成)  
青:バックエンド (ドキュメント転送 ~ インデキシング)  
(矢印の向きは操作の方向を表す)



- 本アーキテクチャで重要な要素をいくつかピックアップして解説します。

- インデキシング

本システムでは検索エンジンとしてAzure Cognitive Search を採用しており、生成AIに的確な情報をフィードするためにAzure Cognitive Searchへのデータ登録（インデキシング）が重要です。

生成AIの入力文字数制限も踏まえ、ドキュメントを細かく分割したチャンク単位のインデックスを設計しました。

- 権限制御

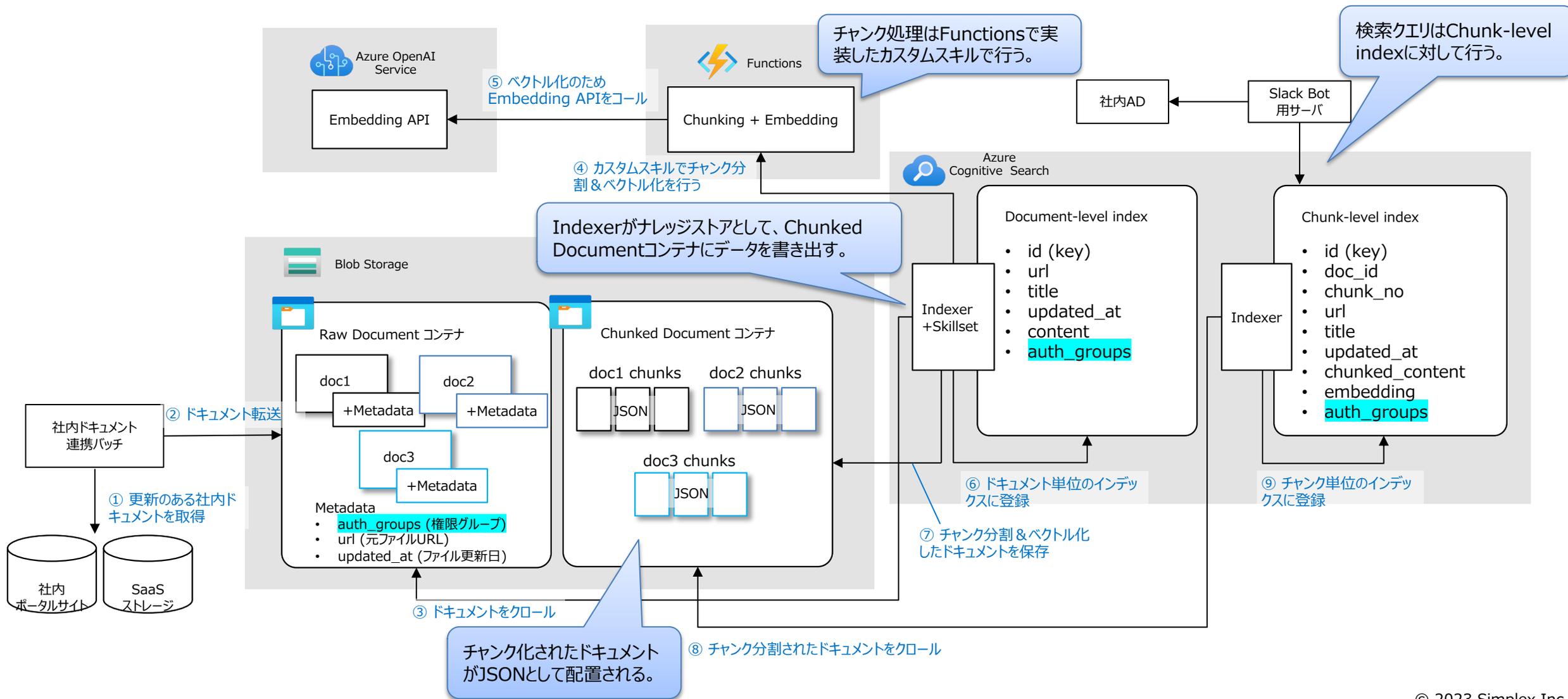
ユーザがアクセスできない情報を含んだ回答を生成しないようなアクセス権限制御は必須の機能です。

ユーザの権限に応じてAzure Cognitive Searchに対するクエリをフィルタリングできる構成を設計しました。

- 通信経路

ビジネスドキュメントという機密性の高い情報を扱うことから、リソース間の通信を可能な限り閉域化またはIP制限し、最大限セキュアとなるようにネットワークを設計しました。

- ドキュメント全体に対してインデックスを構築します (Document-level index)。
- ドキュメントを所定の長さ分割したチャンクを作成し、インデックスを構築します (Chunk-level index)。
- 「ドキュメント」と「チャンク毎のインデックス」は、1:Nの関係になります。これによりチャンク粒度の検索、ドキュメントの引用が可能になります。



- インデックスのスキーマを示します。

## Document-level index

Field	Type	Retrievable	Filterable	Sortable	Facetable	Searchable	備考
id (Key)	Edm.String	Yes	No	No	No	No	
url	Edm.String	Yes	No	No	No	No	
title	Edm.String	Yes	No	No	No	Yes	
updated_at	Edm.DateTimeOffset	Yes	No	No	No	No	
content	Edm.String	Yes	No	No	No	Yes	
auth_groups	Collection(Edm.String)	Yes	Yes	No	No	No	

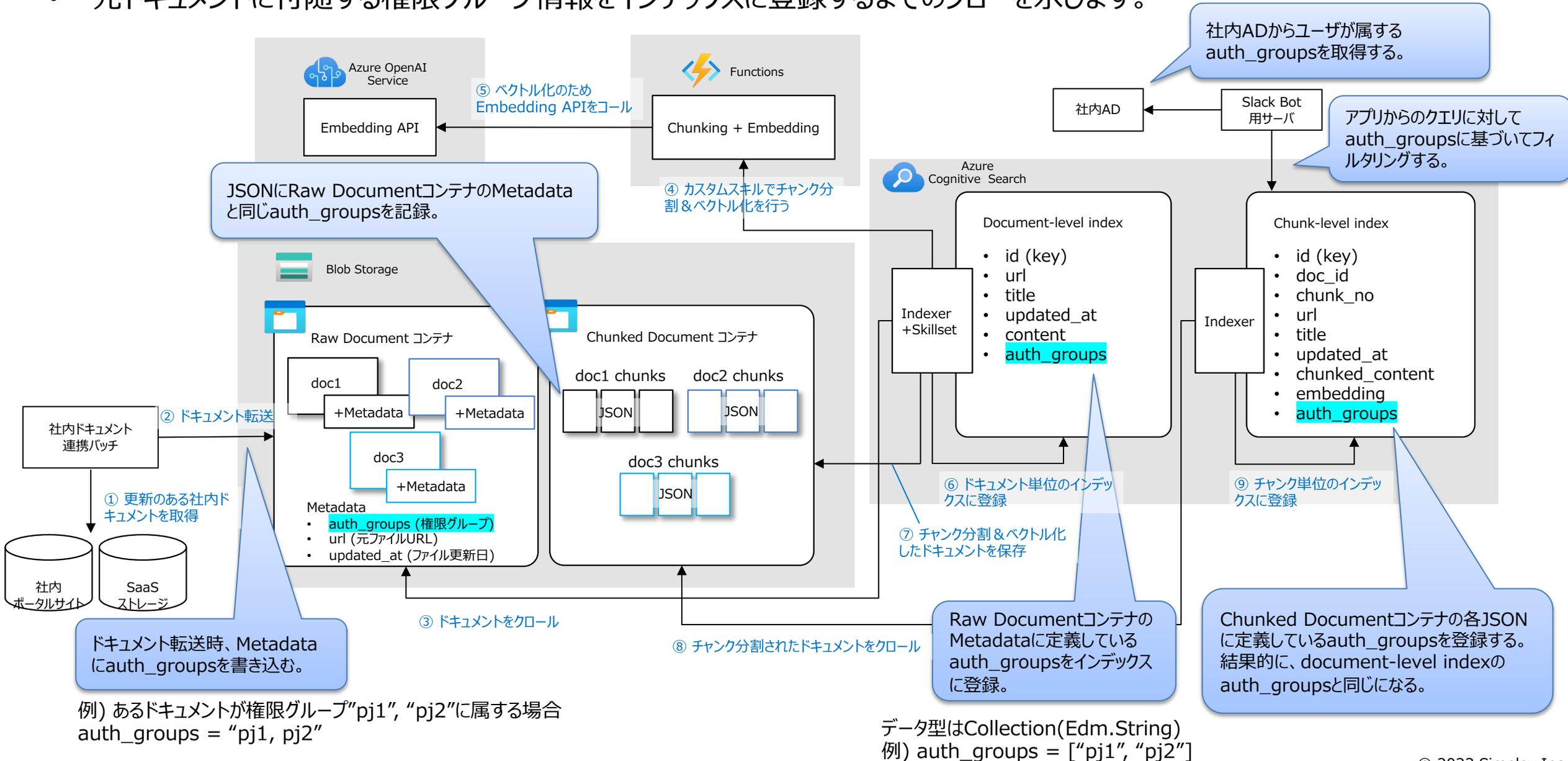
## Chunk-level index

Field	Type	Retrievable	Filterable	Sortable	Facetable	Searchable	備考
id (Key)	Edm.String	Yes	No	No	No	No	
doc_id	Edm.String	Yes	No	No	No	No	document indexのidに対応
chunk_no	Edm.Int64	Yes	No	No	No	No	
url	Edm.String	Yes	No	No	No	No	
title	Edm.String	Yes	No	No	No	Yes	document indexの対応レコードと同一
updated_at	Edm.DateTimeOffset	Yes	No	No	No	No	
chunked_content	Edm.String	Yes	No	No	No	Yes	
embedding	Edm.SingleCollection	Yes	No	No	No	Yes	次元=1536
auth_groups	Collection(Edm.String)	Yes	Yes	No	No	No	document indexの対応レコードと同一

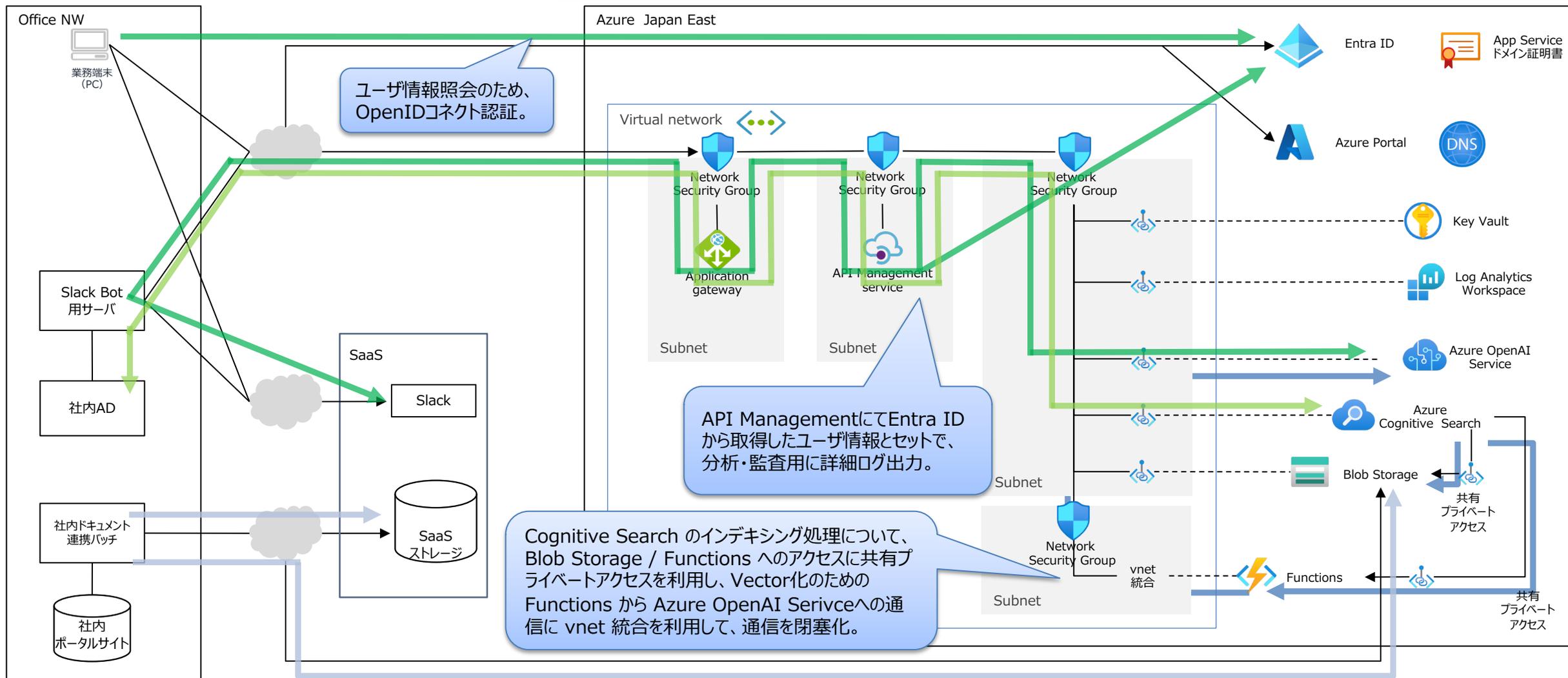
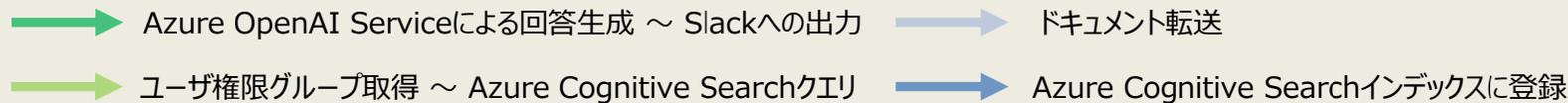
検索時は、auth\_groupsで絞りつつ、embedding値で関連度の高いチャンクをに基づいて回答します。

auth\_groupsの登録フローは次ページの通り

- 元ドキュメントに付随する権限グループ情報をインデックスに登録するまでのフローを示します。



- 通信経路を右記4種類に分け、セキュリティ観点の設計ポイントをまとめます。



## [Microsoft Azure Well-Architected Framework](#)

### 信頼性(可用性)

- Azureのサービスでは「可用性ゾーン」や「リージョン」といった単位で可用性を設計しており、これらを適切に組み合わせることで、ビジネス クリティカルなワークロードの信頼性を実現するように設計することが可能です。詳細は「[Azure リージョンと可用性ゾーンとは](#)」をご参照ください。
- このシナリオで用いられているAzure OpenAI Service、Azure Cognitive Search 等のコンポーネントはそれぞれゾーン冗長、リージョン冗長、geoレプリケーションなど高可用性のオプションや構成を利用可能です。必要となる可用性に応じて導入を検討してください。複数リージョン間/複数ゾーン間でAct-Act構成を取る場合にはAzure Front Door、Azure Application Gatewayなどの利用をご推奨します。

### 信頼性(回復性)

- アプリケーションの正常性を監視するために、Application Insights を使用すると、カスタマー エクスペリエンスや可用性に影響を及ぼすパフォーマンスの問題についてアラートを生成し、対応することができます。詳細については、「[Application Insights とは何か?](#)」を参照してください。
- 回復性に関するその他の記事については、「[信頼性の高い Azure アプリケーションを設計する](#)」を参照してください。

### セキュリティ

- セキュリティは、重要なデータやシステムの意図的な攻撃や悪用に対する保証を提供します。詳細については、「[セキュリティの重要な要素の概要](#)」を参照してください。
- セキュリティで保護されたソリューションの設計に関する一般的なガイダンスについては、「[Azure のセキュリティのドキュメント](#)」を参照してください。
- 本シナリオは企業内ユーザを前提としているので、認証認可にはAzure ADのご利用をご検討ください。

## コスト最適化

- 不要な費用を削減し、運用効率を向上させる方法を検討することです。詳しくは、[コスト最適化の柱の概要](#)に関する記事をご覧ください

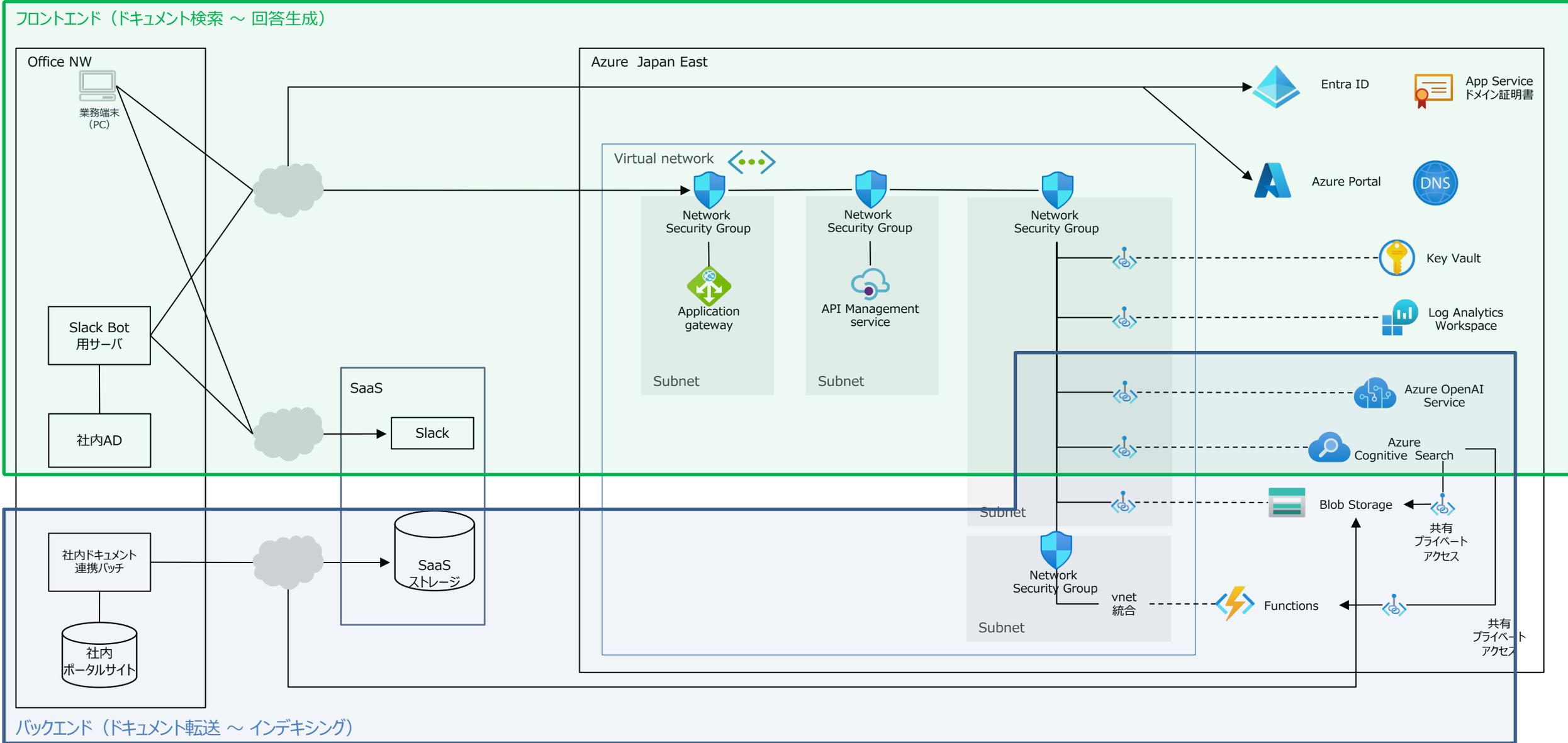
## オペレーショナルエクセレンス

- システムの健全性の担保、トラブルの解決、利用動向の監視を行うためには適切な監視とログ収集が必要となります。詳細は「[ワークロードの監視](#)」をご参照ください。API Managementを利用することで、API利用の監視やトレースを行うことが容易になります。
- ソフトウェアのアップデートや脆弱性への対応など、ソフトウェア/インフラ設計の改修を円滑に進められるよう、DevOpsプロセスを確立してください。詳細は「[リリース エンジニアリングの継続的インテグレーション](#)」をご参照ください。

## パフォーマンス効率

- アプリケーションの負荷が高まることを見越し、スケーラビリティの確保をあらかじめ検討することは重要です。詳細は「[スケーリング用のアプリケーションを設計する](#)」をご参照ください
- Azure Functionsは負荷に応じて自動的にスケールイン/アウトさせることが可能です。詳細については「[Azure Functions でのイベント ドリブン スケーリング](#)」をご参照ください。
- 頻出のクエリについてはアプリ側でキャッシュする等のキャッシュ戦略もご検討ください。詳細は「[キャッシュを使用する](#)」をご参照ください。
- また、特定のユーザーにAzure OpenAI Serviceの利用が集中することを避けたい場合にはAPI Managementによるスロットリング導入などをご検討ください。詳細は「[Azure API Management を使用した高度な要求スロットル](#)」をご参照ください。

- 本アーキテクチャの構築コード（IaC）は非公開とさせていただきます。



本資料は情報提供のみを目的としており、本資料に記載されている情報は、本資料作成時点でのシンプレクス株式会社の見解を示したものです。状況等の変化により、内容は変更される場合があります。

シンプレクス株式会社は、本資料の情報に対して明示的、黙示的または法的な、いかなる保証も行いません。

© 2023 Simplex Inc. All rights reserved.