Vertex AI Searchで実現する膨大な資料の活用術

～島津製作所の技術選定から導入までのプロセス公開～

2025/02/14
株式会社grasys

自己紹介

ホソヤ マサノリ

細谷 政徳

株式会社grasys
Cloud Tech div.
Cloud Development sec.

- 東京出身、東京在住、2児の父
- ngmocoのゲームにはまり、株式会社gloopsの前身GMSに入社
  - 開発推進Gのマネージャー、開発部長を経て退職
- スタートアップ・株式会社meleap・フリーランスなど
  - ゲーム・ARスポーツ・メタバース
- 2024年 株式会社grasysに入社
  - Cloud Tech div. Cloud Development sec. の Leader を務める

Certified：

Google Cloud Professional Cloud Architect

自己紹介

ウエマ カンジ
上間 貫司
株式会社grasys
Cloud Tech div.
Cloud Development sec.

- 沖縄県出身、東京在住。
- 2022年に株式会社grasysに入社。主にサーバーサイドを担当。
- 株式会社grasysで初めてクラウドサービスでの構築を行った。Cloud Runをよく使っている。
- 基本的にはPythonでコーディングを行なっているが、今はRustを勉強中。

Certified：

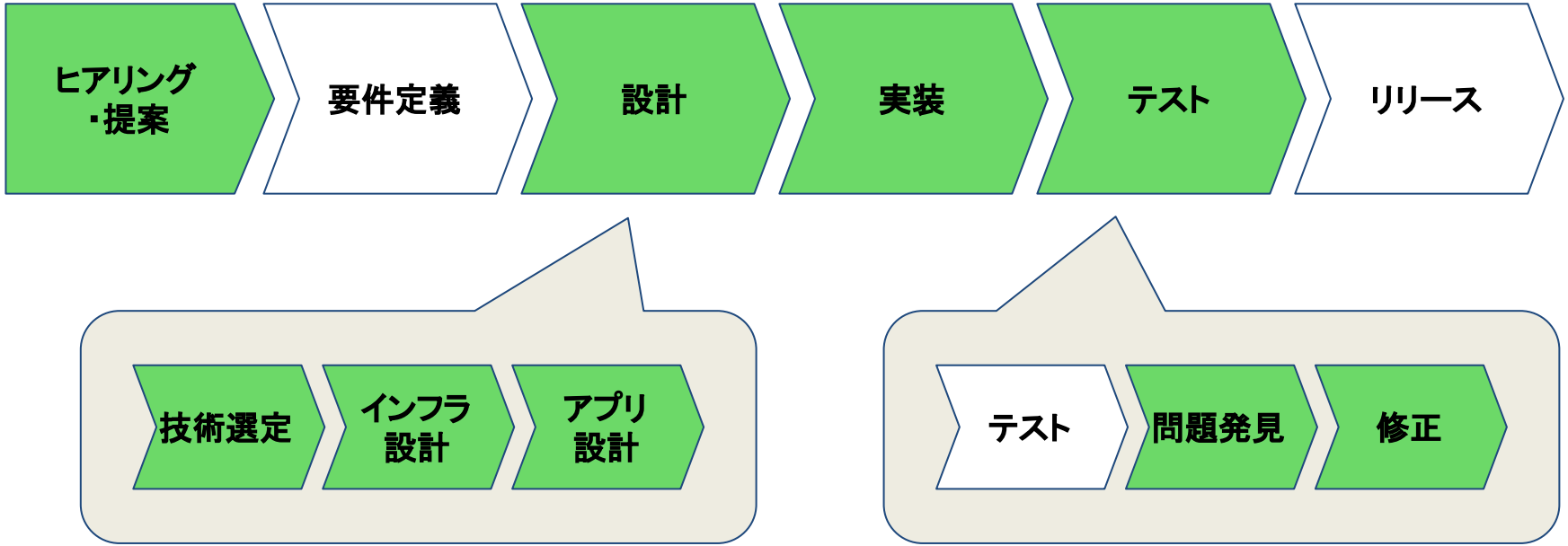
Google Cloud Professional Data Engineer

本日のアジェンダ

- 島津製作所様が抱えていた課題 ～ 提案
- サービス選定
- 構成
- 実装
- 問題 / 要望
- 感想

リリースまでのプロセス

本日お話する内容



島津製作所様が
抱えていた課題～提案

課題

航空業界は安全性の観点から規格が多く、航空機器事業部 品質保証部では業務上確認すべき資料が膨大にあるため、検索 / 確認 に時間を要している

要望

- 手書き資料の検索や確認に時間がかかる課題を解消したい
- セキュリティ面の考慮もしたい
- 費用は最低限にしたい
- 利用ユーザーが増加しても問題ないようにしたい

提案内容

業務上で利用する膨大な資料の検索について、

AI を活用して迅速に見つけ出して業務の効率化を図る

サービス選定

サービス候補

1. Amazon Kendra（AWS）
    - AWSを利用している
2. Vertex AI Search（Google Cloud）
    - 簡単にできて好感触だった
3. Microsoft 365 Copilot（Microsoft 365）
    - Microsoft365 アカウントの使用率が高い

Amazon Kendra / Vertex AI Search / Microsoft 365 Copilot – サービス比較

項目	Amazon Kendra	Vertex AI Search	Microsoft 365 Copilot
料金	・$1.4/時間 ・$1,008/月	・$4.00/1,000クエリ ・要約機能追加時$4.00 / 1,000 クエリ	・$30 ユーザ/月
OCR	×	○	×
取り込み場所	・S3 ・コネクタを使用して他のサービスと接続可能(Box等)	・Cloud Storage ・BigQuery ・他サードパーティ(Jira他)	・One Drive for Business ・SharePoint Online
ファイルサイズ	最大50MB	最大100MB	※取り込み場所の仕様による
ドキュメント数	100,000	1,000,000	18,000 - 20,000

※ 2024年3月時点の比較

Vertex AI Search / Microsoft 365 Copilot – 検証比較

項目	備考	Vertex AI Search	Microsoft 365 Copilot
根拠ページ	PDFのどのページから抽出したか	○	×
根拠資料	どのPDFから抽出したか	○	○
表（Text Data）	グリッド形式の表から名称を抽出できるか	○	○
図面（Text Data）	指定した工具の図面から横幅を抽出できるか	○	×
手書きデータ	OCR検索できるか	○	△
検索の回答精度	回答の精度は高いか	○	△

※ 2024年3月時点の比較

Vertex AI Search 採用！

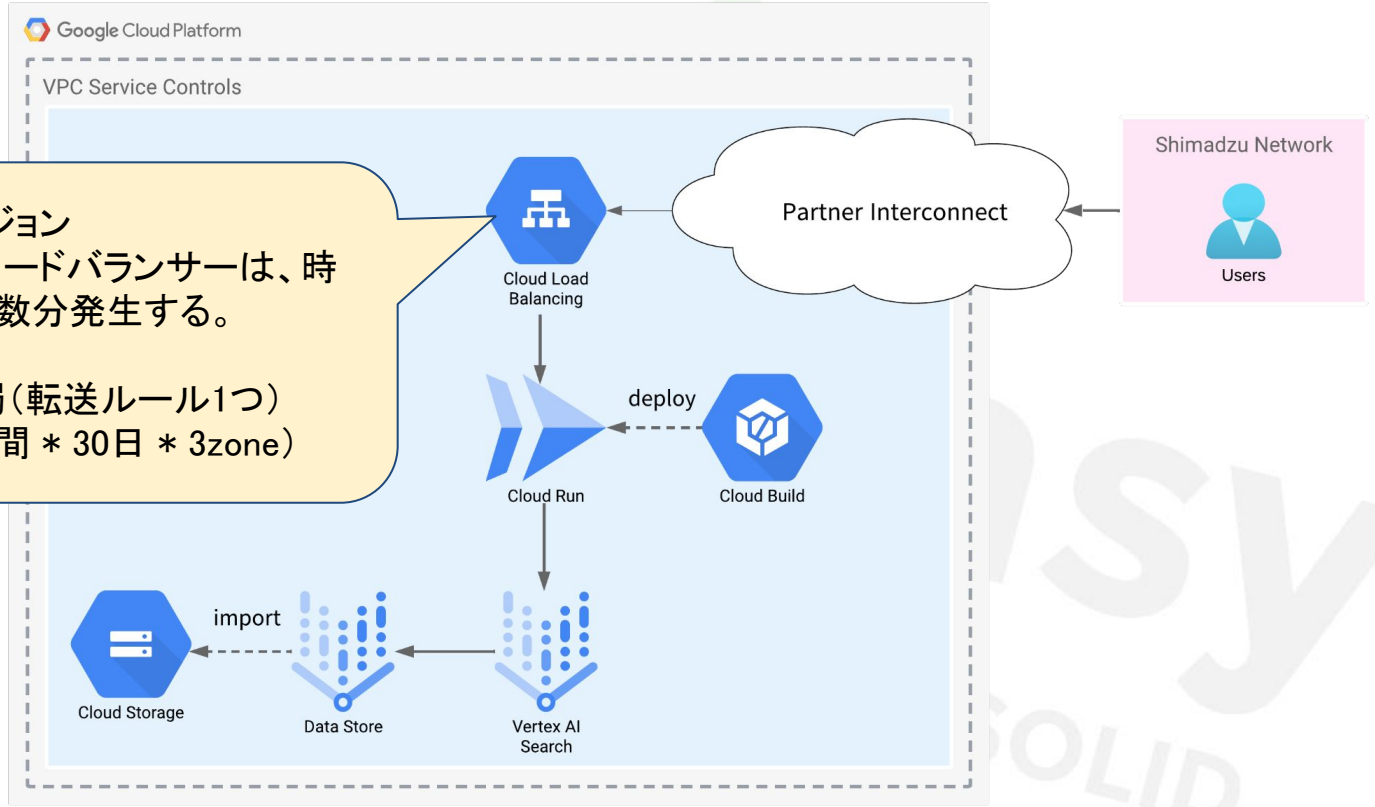
- 料金が安い
  - 見積もりで50クエリ * 20日で月1000クエリの想定
- OCRの設定が簡単にできる
- Gemini（LLM）の回答の精度が高い

構成

構成図

リージョン
内部アプリケーションロードバランサーは、時間課金でゾーン数分発生する。

Tokyo: 約8,000円弱（転送ルール1つ）
（0.025円/時 * 24時間 * 30日 * 3zone）



アプリケーション構成

- Python + Flask
  - 開発のスピードを出せるため
- React + Next.js
  - 拡張性が高いため
- Docker
  - Cloud Runを使用するため

実装

実装機能一覧（一部）

1. データストア切り替え
2. 要約言語モデル切り替え
3. 参照ドキュメントを開く
4. 監査ログ

1. データストア切り替え

- データストアは、質問や検索に対する回答を見つけるために使用
- 要望
  - 規定集や手書き文章などドキュメントの種類で検索の範囲を切り分けたい
- 実装
  - データストアを種類ごとに作成
  - 検索時にデータストアを選択するように実装
  - 必要に応じて簡単に追加・変更ができる

2. 要約言語モデル切り替え

- 言語モデルごとの回答の精度が変化に対応するため
- 要望
  - 下記の要約言語モデルを選択したい
    - Text-Bison2(2025/4/9に廃止予定)
    - Gemini 1.5 Flash V2
- 実装
  - 検索時に要約言語モデルを選択するように実装
  - 必要に応じて簡単に追加・変更ができる

3. 参照ドキュメントを開く

- 検索結果に、参照したgsutil URI(gs://)が返るが、閲覧できない
- 要望
  - 誰がどのドキュメントを見たのかをログに残したい
  - 回答に使用した参照ドキュメントの確認をしたい
- 実装
  - クエリパラメータでgsutil URIを受け取る
  - ユーザーとgsutil URIをログに残す
  - gsutil URIから署名付きURLを生成、リダイレクト

4. 監査ログ

- 要望
  - 誰がログイン / 検索 / ドキュメント参照をしたのか残したい
- 実装
  - Cloud Logging に必要な情報を書き込む
  - ログシンク で 定期的に Cloud Strorage にログを保存
    - BigQuery にインポートして分析も可能

実装完了！

問題 / 要望

問題 / 要望 一覧(一部)

1. 漢字だけの名詞で質問すると、要約が中国語で生成されることがある
2. 産廃と略すと産業廃棄物として認識してくれない
3. OCRで「φ(ファイ)」を「4」と誤認した
4. Vertex AI searchの要約のクォータ制限

1. 漢字だけの名詞で質問すると、要約が中国語で生成されることがある

「特工申請」 → 中国語

「特工申請の方法は？」 → 日本語

解決策

Auto / 日本語 / English から要約言語を選択できるようにした

2. 産廃と略すと産業廃棄物として認識してくれない

解決策

検索モデルをチューニング

- クエリファイル
  - 業界用語が含まれた質問のサンプル
- コーパスファイル
  - 業界用語が含まれた質問に対する回答のサンプル
- トレーニングラベル
  - クエリとコーパスのペアに対して重みをつける

3.OCRで「φ（ファイ）」を「4」と誤認した

※「φ（ファイ）」＝円の直径

「Aの穴径は」の質問に、「φ10」が「410」となっていた

解決策

- Vertex AI Gemini APIを併用して2回クエリ実行する
  1. Vertex AI Searchでクエリを実行してドキュメントを取得する
  2. 1. で取得した結果ドキュメントと1. で使った質問を使って、改めてVertex AI Gemini APIで回答を生成する

※ この課題は検証・解決の確認まで行ったが未実装

4. Vertex AI searchの要約のクォータ制限

デフォルトが1分間に15回まで

解決策

- HTTPステータス: 429 の時、少し時間を空けて再リクエストするよう画面に表示するようにした
  - 現状の規模では、1分間に15回の同時アクセスは滅多にないと判断
- 規模が大きくなり頻発するようであれば、クォータ上限値アップの申請をGoogle Cloudに出す

リリース！

感想

感想

- 上間（メイン開発者）
  - 生成AIやセマンティック検索の深い知識がなくても、スピード感を持ってドキュメント検索のシステムを構築することができた。
  - Vertex AI Searchだけでなく、Kendra、Microsoft 365 Copilotなどを触って検証できた。
- 細谷（リーダー）
  - Vertex AI Search のカスタマイズ性が高い。
  - 時間があまりない中で、無事にリリースまでできたのは非常に良かった。
  - 料金が安い。
    - 1回の検索で約1.4円（LB などの料金は除く）

Vertex AI Search の その他の機能

- 要約をカスタマイズ
  - プロンプトを用意してカスタマイズ
- フォローアップ検索
- Cloud Storage の ACL もインポートすることでドキュメント単位でアクセス制御
- データソースは BigQuery, Cloud SQL も可能
- [preview] サードパーティのデータソースを Vertex AI Search に接続

最後に

エンジニア
募集しています

会社概要

会社名	株式会社grasys
設立	2014年11月13日
本社	東京都渋谷区恵比寿4-20-3 ガーデンプレイスタワー 11F
代表者	代表取締役 長谷川 祐介
従業員数	49名
URL	https://www.grasys.io/

grasys 組織　サービス領域

Tech Team

- AI
- IoT

- PoC
- MLOps
- 技術支援

Cloud Development Team

- KPI分析
- データパイプライン
- システム間連携

- アプリ開発
  - 情報分析基盤構築
    （データ収集・集約・加工）
  - BI
  - WEBアプリ開発

Cloud Infrastructure Team

- Lift&Sift
- ゲームインフラ
- SaaS

- インフラ設計・構築
- 運用・監視
  - 負荷試験
  - 監視
  - チューニング

パートナー

aws

Google Cloud Partner

Microsoft Partner

第三者から認定を受けた技術力・実績

Google Cloud プレミア SELL パートナーである grasys は
インフラ領域において「高い技術力」と「実績」が認められています。

受賞歴

クラウドを活用した最新技術に挑戦しませんか？

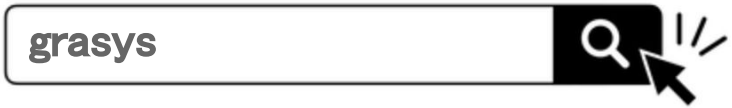
grasys では、のべ3億人以上のユーザーを支えるクラウド基盤や、
数百万人が利用するオンラインゲーム基盤などを手掛けています。

直近では、長年クラウドインフラで培った技術をもとに
Google Cloud や AWS などのクラウド技術を活用しながら、
AI技術を活用した 最前線のプロジェクトも多数手がけています。

大規模システム設計からAIプロジェクトのリードまで、
ここでしかできない経験が待っています。

grasys はこんな会社です！

- 技術を心から楽しめる方が多数
- ゲーム、Web、エンタープライズ案件と幅広く携われる
- クラウド検証し放題（ルールはあるけどね）
- AI技術など最先端の技術を扱うことができる
- デスク環境良い（バロンチェア、27インチ4Kディスプレイ）
- 風通し良い方。やることやってれば比較的自由もきく。
- 恵比寿、美味しいランチが多い（ちょっとお高め）



ご清聴ありがとうございました