3,000時間/月の業務削減を実現する Dify × TiDB Cloud StarterによるAI基盤の裏側

株式会社サイバーエージェント

AIオペレーション室
開発エンジニア

西 悠之

TARGET

01 全社横断的なAI活用普及の取組に興味を持っている方

02 セルフホスト×ベクトルDB選定に関心があるアーキテクト／エンジニア

03 コスト最適化とガバナンスを両立した運用に関心がある方

Agenda

01 部署紹介とミッション（AIオペレーション室の取組み・活動）

02 AIツールDifyとは？（概要と活用事例、競合製品との比較、プラン形態）

03 システム構成（開発体制、想定パフォーマンス要件、AWSアーキテクチャ）

04 全社展開に伴うベクトルDBの選定（TiDB）

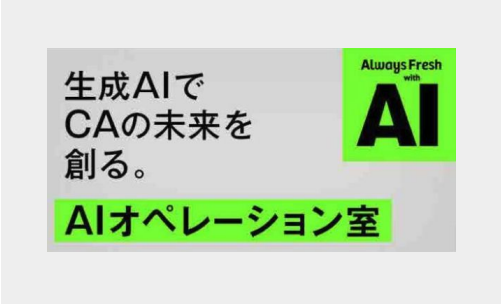
05 まとめ

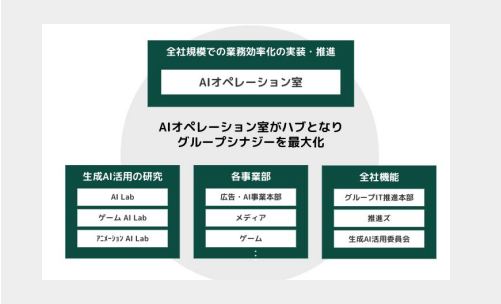
01

部署紹介とミッション

CyberAgent | AIオペレーション室

CyberAgent | AIオペレーション室について





設立背景

2022年11月～ ChatGPTの登場を受け、
「生成AIを徹底的に活用した会社がそうでない会社に大きな差をつける時代になる」
2023年9月に開催された「あした会議」にて決議。同年10月、AIオペレーション室設立

https://www.cyberagent.co.jp/way/list/detail/id=30894

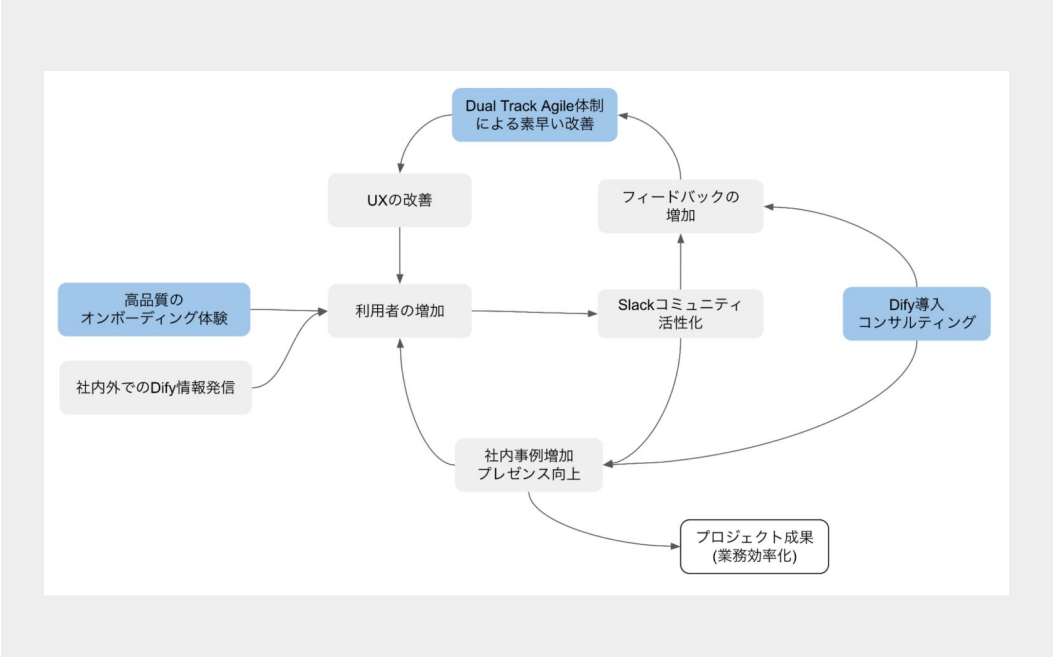
ミッション

全社横断的に生成AIの取組みを加速させ、その活用をサイバーエージェントの競争力へと繋げていき2026年までにオペレーション業務を 6 割削減するミッションを掲げています

これまでの主な活動

- 生成AI活用ガイドラインの策定・運用
- 全社アイデアコンテストの開催
（アイデア発掘）
- 社内プロダクトの AI開発／導入支援 ― 各事業部の要望からプロジェクト化
- AIファースト開発・ SaaSによる効率化推進
（価値創出、コスト削減、売上増）

今回紹介したい事例：AIファーストのSaaSの推進- Difyについて



当社は約8,000名の従業員が在籍し、インターネットを軸に、メディア&IP事業、インターネット広告事業、ゲーム事業などを中心にビジネスを展開しています。
このような多様な業態に対してDifyを社内SaaSと捉え、業務自動化などの推進を行っています。
開始半年以上で社内ユーザー数は約2,000名弱に達し、そのうちそのうち25%以上が毎週コンスタントに利用するアクティブユーザーとなるなど、比較的速いペースで利用を拡大中

引用: サイバーエージェント社員の20%が使うAIプラットフォーム「Dify」、プロダクト主導で3,000時間/月削減する方法
https://developers.cyberagent.co.jp/blog/archives/56492/

02

Difyとは？

（概要と活用事例、競合製品との比較、導入形態）

POINTS

Difyとは？

1 Difyの概要と活用事例

2 競合製品と選択した理由

3 Difyのプラン形態

Difyの概要

AIアプリのノーコード開発フレームワーク

専門知識がなくても効率的にAIアプリ開発が行えるプラットフォーム

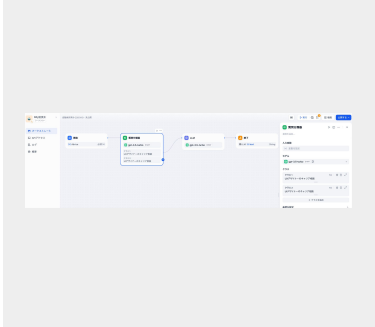
※PythonやNode.jsのコードも実行可能

スプレッドシートやExcel Slackなど外部連携可能

Excelやスプレッドシート、Slackといった外部ツールと連携が可能で既存の業務フローにAIを簡単に組み込める

PDFや、Markdownなどのドキュメントを RAG利用可能

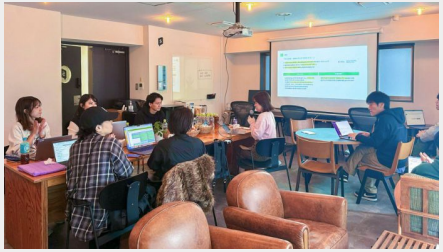
PDFやMarkdownなどの資料を取込み、RAGを用いた検索・回答が可能。知識を活かした高度な応答を実現できる



管理画面から簡単にプロンプトやワークフロー編集可能

管理画面から直感的にプロンプトやワークフローを編集可能で専門的なプログラミングなしで柔軟に調整できる

具体的な活用事例: ABEMA LIFE



1.Dify合宿

AIオペレーション室主体で、AIアプリを“Dify”上で作る合宿を様々な部署に対して実施。AmebaLIFE事業本部では、ビジネス職も含めて社員が「今日はこの業務を効率化して帰る」という目標を持ち取り組んでいただけました

2.ワークフローで自動化

コメント監視やSEO最適化など人手依存の高かった業務をAIで自動化。AIオペレーション室の支援により、モデル選定やコスト最適化も実現

3.成果

コストは従来の10分の1に圧縮。社員はコンサルや企画業務など付加価値の高い領域へシフト

※ABEMA LIFE: ABEMAブログなど扱う部署

→ こうした成功の裏側の 1つの要因には“ツール選定の判断基準 ”があると考えています

数ある選択肢の中からなぜDifyを選んだ背景

Dify

直感的なUI/UXでAIアプリを素早く構築できるローコードなフレームワーク。ビジネス職のユーザーも利用しやすく、開発と運用の両面をカバーできる。

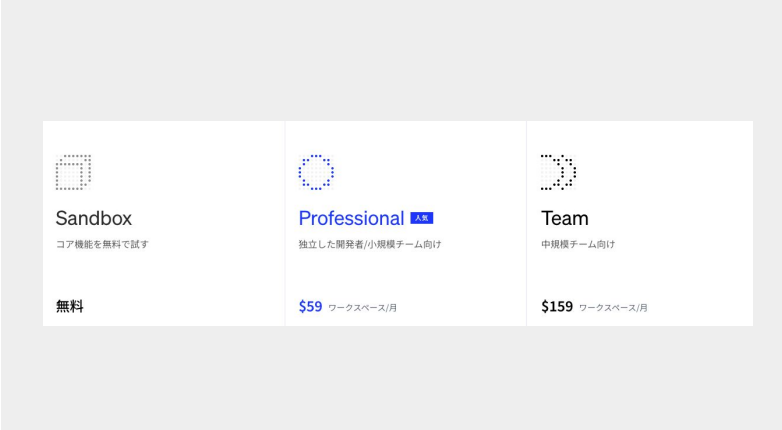
n8n

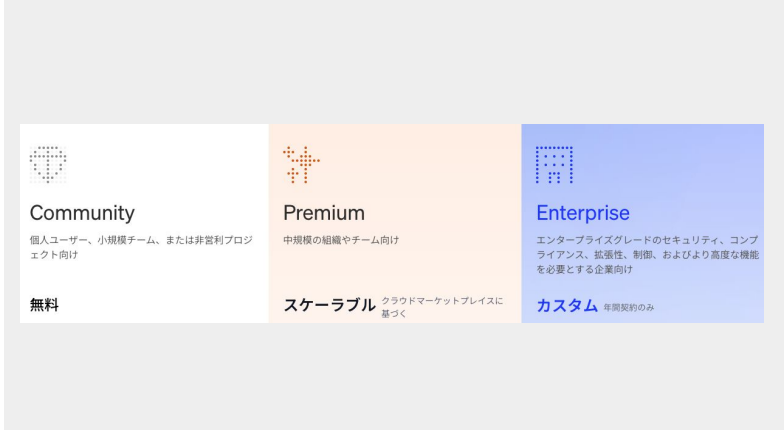
ワークフロー自動化に強みを持つツール。多彩な外部サービスと連携し、ノーコードで業務効率化を実現します。

LangChain

LLM（大規模言語モデル）の利用に特化した開発ライブラリ。柔軟な拡張性で高度な AIアプリ開発を支援します。

Difyのプラン形態





クラウド

Dify が提供する SaaS 形式で、環境構築や運用管理が不要。スケーラビリティやアップデートが自動で行われるため、すぐに利用開始できる。

プラン: 無料、Pro(約9000円/月)、Team(約2万円/月)

制限: ユーザー数、アプリ数、リクエスト数、ストレージに制約あり

例:Teamの場合 ユーザー数50人、アプリ数200、ストレージ: 20GB

https://dify.ai/jp/pricing

セルフホスト

カスタマイズ性やデータ管理の自由度が高く、オンプレミスやAWS MarketPlaceなどで購入し独自クラウドに展開可能。

プラン: Community(無料)、Premium、Enterprise(120万円/月～ 会社規模による)

制限: カスタム

セルフホスト（Community）を選んだ3つの理由

1.コスト

全社展開を見据えた際にEnterprise版は大企業向けのサポートなどがついているが、月額約100万円以上と高額なため、初期投資としてスケールと運用コストのバランスを考慮した

2.カスタマイズ性

OSSであるため拡張可能

例：ReBACによるユーザー間アプリ共有機能の追加やNotionデータの連携拡張など機能カスタマイズができる

※ 自由度が高い一方で、導入・運用には環境構築やバージョン管理などの継続的な技術的メンテナンスメンバーの開発力が求められる

3.データガバナンス

機密情報を扱う前提から、自社で信頼できるストレージを選定できる

社内SSGの要件を満たすには追加のセキュリティ対応が必要であることが多いため、OSSであればBasicやOAuthの認証追加が必要

※ 例：DifyのアプリはMCPサーバーとしてURLを公開することができます。ただ、現行機能として認証機能がないためURLが第3者にもれると機密データにアクセスされるリスクがあるなど

→ 実際にどのようなシステム構成で運用しているのかをご紹介します。

03

システム構成

（開発体制、アーキテクチャ概要、AWS構成）

POINTS

システム構成

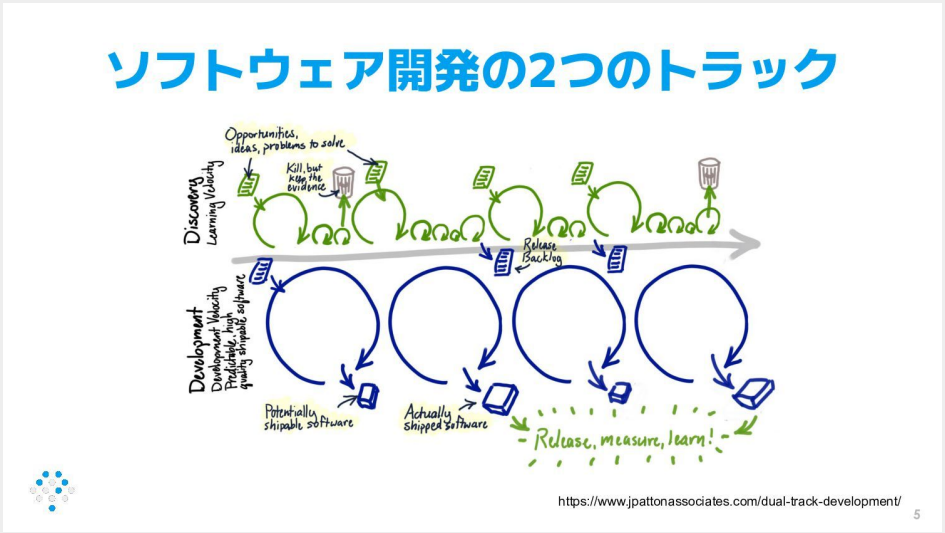
1 開発体制

2 Difyアーキテクチャ概要

3 想定ユーザー規模とパフォーマンス要件

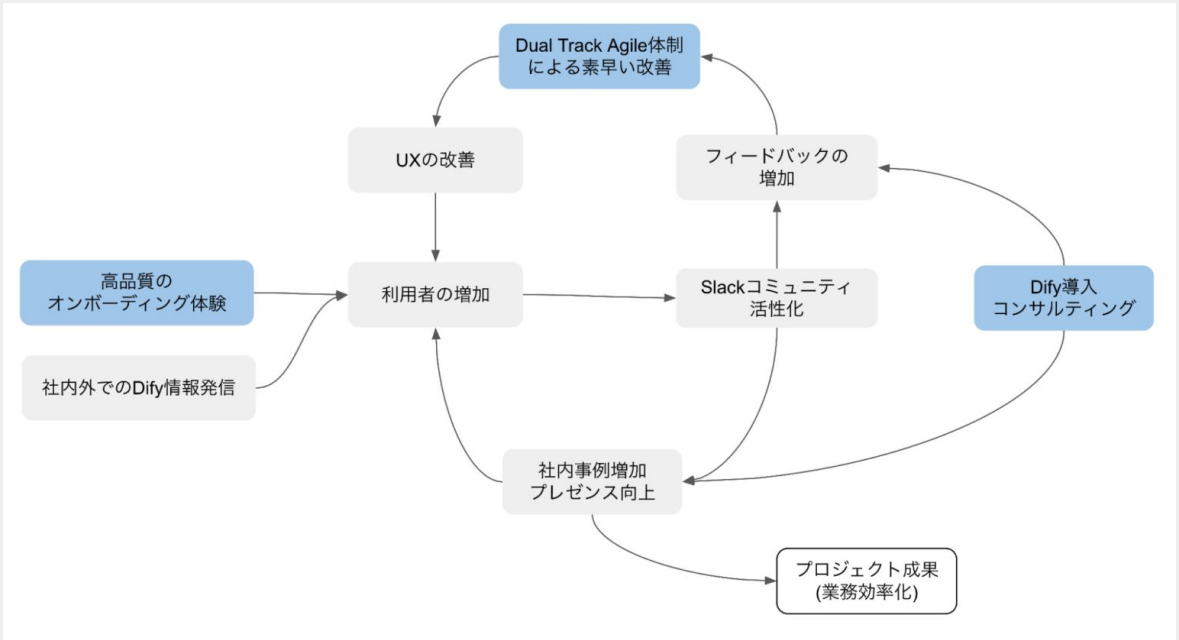
4 AWSアーキテクチャ

開発体制: デュアルトラックアジャイルによる社内プロダクト開発



トラック	概要	本件での担当
Discoveryトラック	社内コンサルティングやSlackサポートを通じたフィードバック収集、仮説の構築・検証	筆者を含む社内コンサルタント・Slack運営担当の2名
Deliveryトラック	実際の機能開発	AWS/開発エンジニア2名

フィードバックサイクル

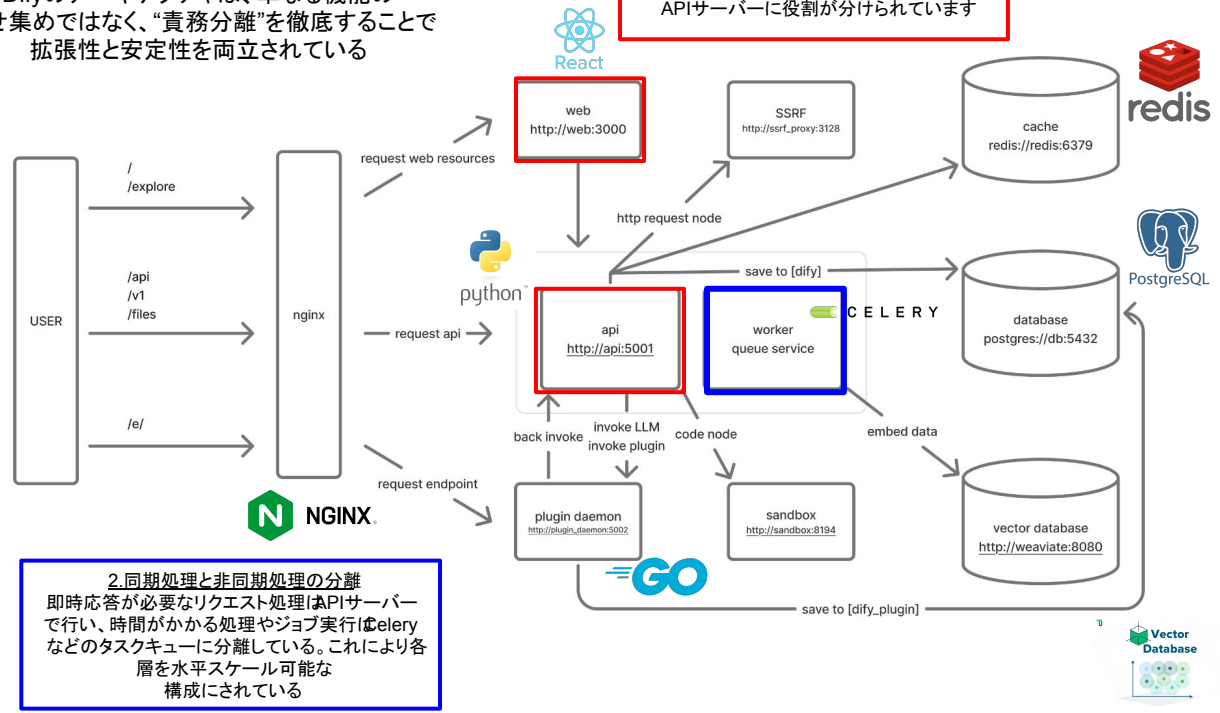


引用: サイバーエージェント社員の20%が使うAIプラットフォーム「Dify」、プロダクト主導で3,000時間/月削減する方法
https://developers.cyberagent.co.jp/blog/archives/56492/

Difyアーキテクチャ概要

Difyのアーキテクチャは、単なる機能の
寄せ集めではなく、“責務分離”を徹底することで
拡張性と安定性を両立されている

1.フロントエンドとバックエンドの分離
Reactベースのフロントエンドと、
Python/Golangなどで構築された
APIサーバーに役割が分けられています



2.同期処理と非同期処理の分離
即時応答が必要なリクエスト処理はAPIサーバーで行い、時間がかかる処理やジョブ実行はCeleryなどのタスクキューに分離している。これにより各層を水平スケール可能な
構成にされている

インフラを構築する上で求められるパフォーマンス要件を定義

ユーザー数・アプリ数

– 全社展開するため数千名規模に拡大想定（最大10000名以上）
– アプリ数（一人当たり10個ほど作成すると少なくとも数十万以上のアプリ数が作成される可能性）

同時接続・負荷

数千、数万のアプリが定期的に実行されたとしてもレイテンシーが落ちないようにしたい

データ量

部署によっては、数万以上のNotion記事やドキュメント情報を活用されているため、コストパフォーマンスがいいものを利用したい

※Enterprise版の月額100万円を超えたら本末転倒

セキュリティ・運用

社内情報などを扱うため、日本国内でデータが閉じている
バックアップ・リストアが容易、監査ログの提供

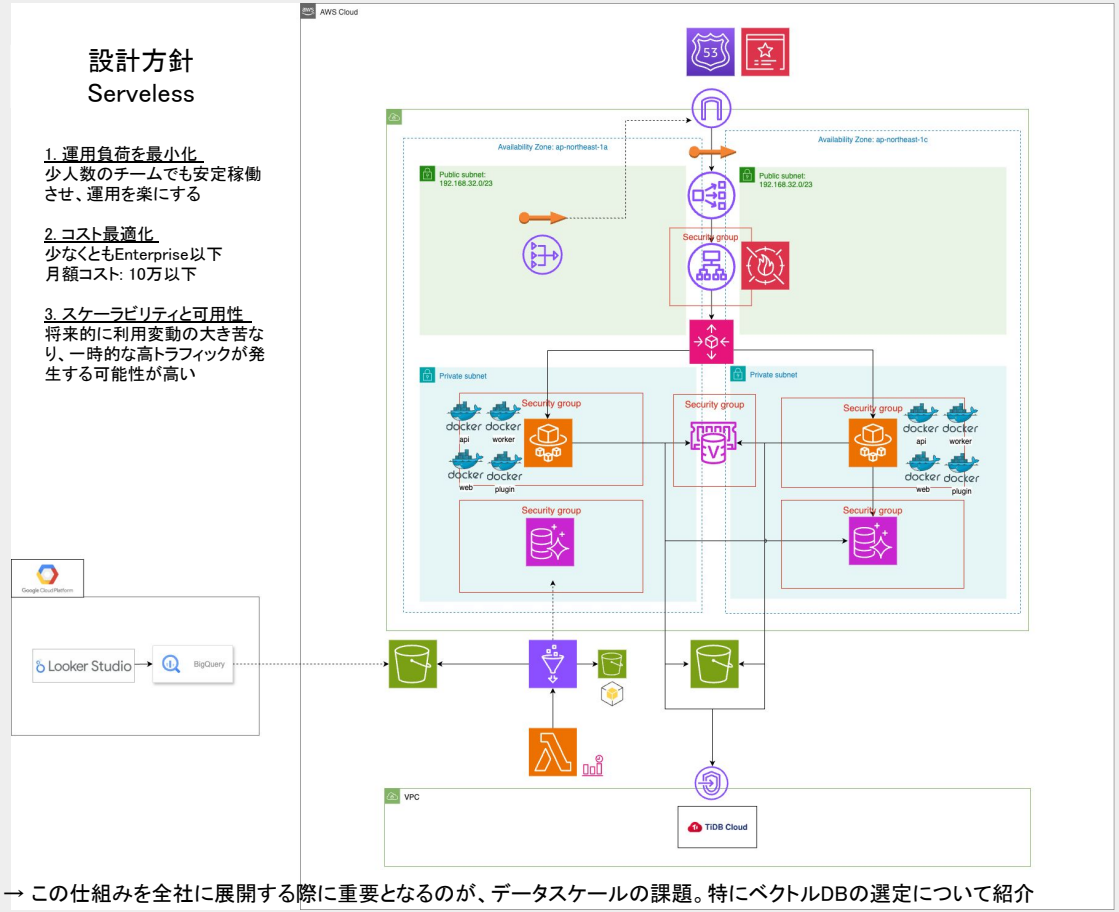
AWSアーキテクチャ

設計方針
Serveless

1. 運用負荷を最小化
少人数のチームでも安定稼働させ、運用を楽にする

2. コスト最適化
少なくともEnterprise以下
月額コスト: 10万以下

3. スケーラビリティと可用性
将来的に利用変動の大きな苦なり、一時的な高トラフィックが発生する可能性が高い



→ この仕組みを全社に展開する際に重要となるのが、データスケールの課題。特にベクトルDBの選定について紹介

04

全社展開に伴う
ベクトルDBの選定

全社展開で見えてきた特徴・性能について

利用者の特徴

ビジネス職の方が主要なユーザーを占める
– 訴求ポイントもビジネスメンバーや人事の業務をDifyを使って効率化提案が現在多い

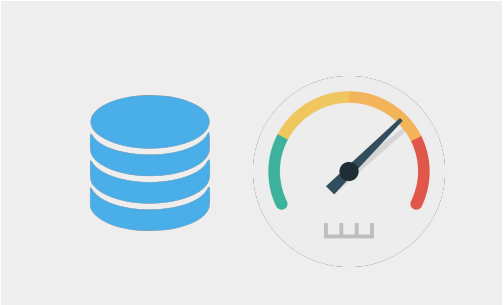
活用パターン

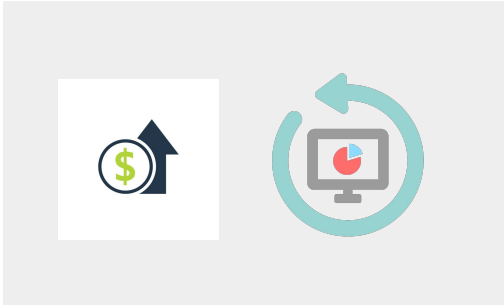
– 日常の定期的な繰り返し作業での活用
– 質問応答システムとしての活用
– Notion、社内Wikiの検索・要約(RAG)

ベクトルDBに求められる性能

– データ量の急激な増加
（数千～数万ファイル追加 / 日）
– RAG検索精度と速度の両立が必須
– コスト効率のバランス

ベクトルDBに求められる要件を定義





1.セキュリティ

社内情報などを扱うため、認証やアクセス制御、暗号化対応など厳格なセキュリティ基準を満たせること
– データが国内に置いてあるか、データベースとの通信が閉域網で閉じられているか
…etc

2.大量データ対応・QPS

RAG活用では応答速度が直接ユーザー体験に影響するので、分散ストレージやシャーディングによる水平スケール、インデックスを活用して検索速度を維持できるか。データ増加時にも一貫した低レイテンシを保てるようにしたい

3.コスト・運用効率

安定した性能を維持しつつ、コストパフォーマンスが高いこと。
運用・監視が容易で、少人数でも回せること
– SQLライクでデータ検索ができる
– ダッシュボードや監視サービスと連携できる

Dify対応ベクトルDB一覧

専用ベクトル DB

Weaviate

Qdrant

Zilliz/Milvus

Chroma

検索エンジン /RDS

OpenSearch、ElasticSearch

Pgvector (PostgreSQL)

NewSQL

TiDB

OceanBase

AWS外のクラウドなど

Tencent Cloud VectorDB, Oracle, Analyticdb, Tablestore, Lindorm, Relyt, Couchbase, OpenGauss

引用: Vector Databases Supported
https://docs.dify.ai/en/getting-started/readme/features-and-specifications

ベクトルDBの比較

Weaviate

⭐⭐⭐ スケール：水平スケーリング可能
⭐ コスト（60GB/月 保管）:約 230,000円/月
⭐⭐⭐ 検索機能：セマンティック、全文、ハイブリット
⭐ 運用監視：Prometheus/Grafanaなどと統合可能

OpenSearch、ElasticSearch

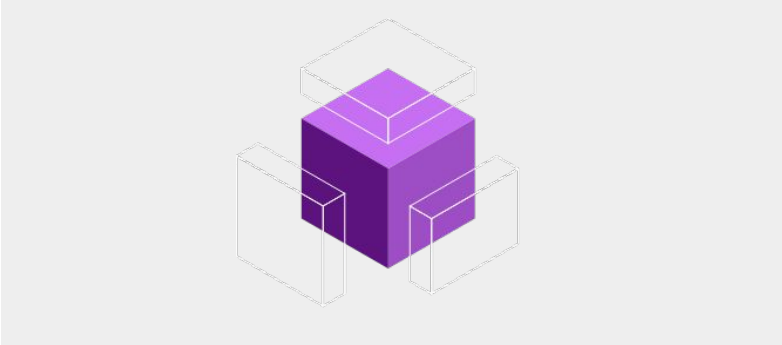
⭐⭐⭐スケール：水平スケール可能
⭐⭐ コスト（60GB/月保管+0.5OCU）:約37,000円/月
⭐⭐⭐ 検索機能：セマンティック、全文、ハイブリット
⭐⭐⭐ 運用監視：Kibana を中心とした可視化や監視ツールが揃っており、既存のオペレーション監視基盤と統合しやすい

TiDB Cloud Starter

⭐⭐⭐スケール：水平スケール可能
⭐⭐⭐ コスト（60GB/月 保管）:約1000円/月
※25GiBの行ストレージ、25GiBの列ストレージ、250Mのリクエストユニットを毎月無料で利用可能
⭐⭐ 検索機能：セマンティック
⭐⭐⭐ 運用監視：標準で馴染みのあるSQLコマンドラインでベクトルデータを検索できる

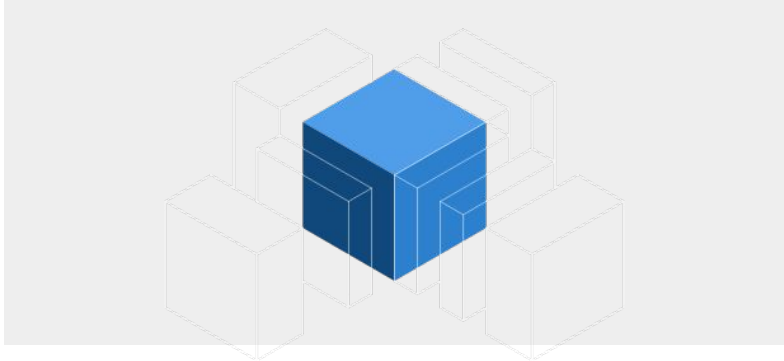
サービス選定をする際に大切にしているポイント
需要変動が激しい現代のサービスにおいて、トラフィックやデータ量の急増が当たり前に発生すると考えている
イベントや負荷に応じてリソースを自動的にスケールし、高負荷でも止まらず、閑散期は無駄なコストが発生しないようにする
Serverlessの思想、無駄遣いしない

TiDBの展開オプション



TiDB Cloud Starter

無料枠（毎月）	各組織で **25 GiB の行ベースストレージ＋25 GiB の列ベースストレージ** および **250M（2億5000万）Request Units (RUs)**。 TiDB +2
課金単価（無料枠を超えた分）	- RUs：1M RU あたり $0.10 TiDB +1 - 行ベースストレージ：1 GiB 当たり $0.20／月 TiDB +1 - 列ベースストレージ：1 GiB 当たり $0.05／月 TiDB
クラスタ数上限・無料クラスタ制限	各組織で最初の **5 クラスタ** は無料／無料枠対象。それ以上クラスタを作るにはクレジットカード登録や予算上限（Spending Limit）の設定が必要。 TiDB +2

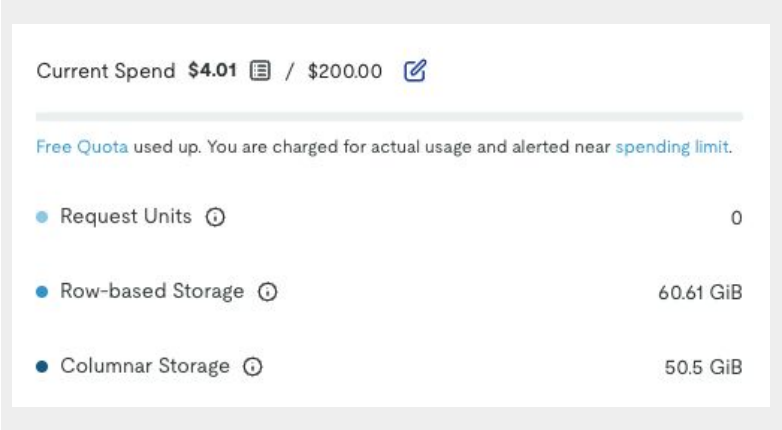


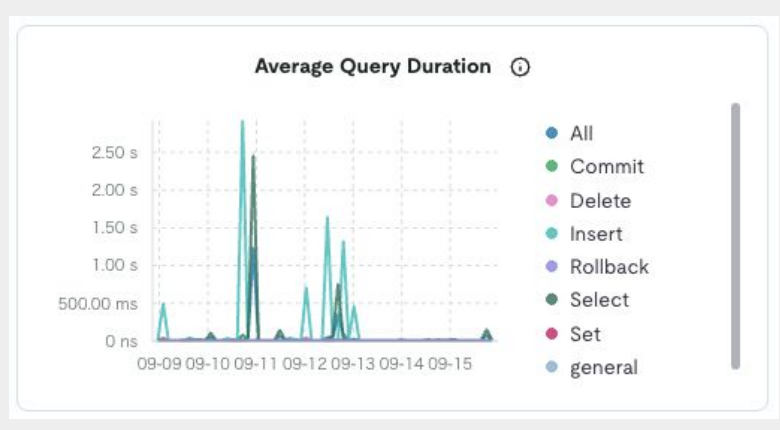
TiDB Cloud Dedicated

最低料金／開始規模	月額 $1,376〜（クラスタのサイズ等による）から。 TiDB +2
ノード構成・vCPU 装備	ノードあたり 4 vCPU 〜 32 vCPU の構成が可能。TiDB／TiKV／TiFlash 各ロールで選択可能。 TiDB +2
ノードごとの課金方式	計算ノード（コンピュート）コスト、ストレージコスト、IOPS／スループットの追加コストなど、使用したノード数や容量・性能で従量制。時間単位で課金。 TiDB +1

https://docs.pingcap.com/tidbcloud/tidb-cloud-intro/

TiDBの運用実績





コスト・ストレージ

ストレージ利用量としてはそこそこの規模（約60GB）ですが利用料は数ドル程度にとどまっておりコスト効率は非常に良好
必要な容量を確保しつつ無駄なコストが発生していない

パフォーマンス

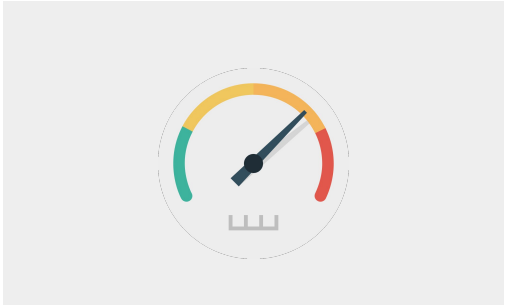
同期間におけるクエリ応答時間に顕著な低下は見られず、ピーク負荷時でも遅延は許容範囲内。例えば、スロークエリの発生も少なく、平均応答時間は（数秒前後）で安定

TiDBを選んでよかったポイント

1.コスト効率のよさ

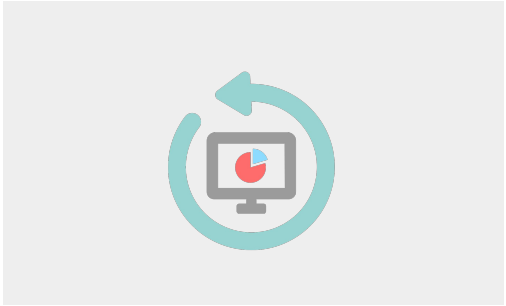
最低課金なしの完全従量課金

TiDBは利用量に応じた従量課金で、ストレージを数十GB使っても数ドル程度と低コスト。性能を維持しながらコストを最小限に抑えられるため、全社展開を見据えた運用でも費用対効果が高い



2.パフォーマンス

ストレージ／コンピュートの両方が動的にスケールされるため性能低下が今のところ少ない



3.運用性

高いScalabilityによる耐障害性、バージョンアップによる構成変更による計画停止がないなど、少人数のチームでも安定運用が実現できている

SQLライクでデータ調査が可能なので調査が楽なのもポイント高い

05

まとめ

まとめ

全社展開にあたり、急増するデータを扱う基盤としてTiDBを選択しました。

コスト効率が非常に高く、数十GB規模の利用でも数ドル程度に抑えられるなど、実運用で“コスパの良さ”を実感しています。

結果として、安定性と費用対効果の両立が可能となり、全社利用に耐えうる基盤を築くことができております。

みなさまの会社の中で、AIツールの導入や技術選定の参考になれば幸いです。


AIオペ室