Google Cloud

より良いユーザー体験を提供するために。悪質コメントを検知するソリューションを解説

Wonha Shin

Google Cloud Japan G.K.
Customer Engineer

スピーカー自己紹介

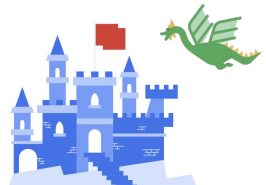
Wonha Shin
グーグル・クラウド・ジャパン合同会社
カスタマーエンジニア

サービス企業でソフトウェア エンジニアとしてキャリアを始め、

- 開発共通基盤の構築・運用
- マイクロサービス設計
- CI/CD パイプラインの構築
- 社内クラウドや OSS を用いたプラットフォームの構築・運用

などの業務を担当していました。

現在はゲーム業界のお客様向けに Google Cloud の
技術的な支援を行っています。

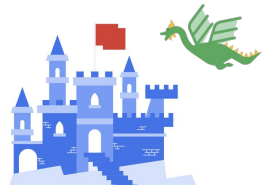


多様なユーザーが楽しめるゲームを支える AI 技術

Google Cloud

様々なユーザーの言語を自動翻訳

- Cloud Translation API
- 100+ の言語を検知・翻訳
- カスタム辞書でゲーム内の用語を登録
- AutoML Translation でモデルカスタマイズ

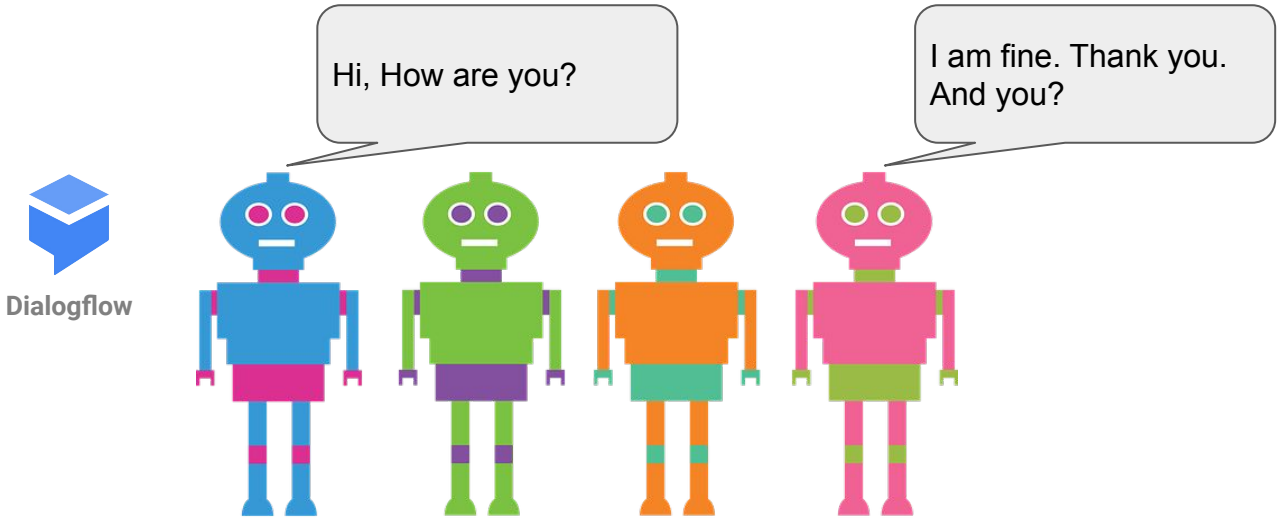


ユーザーの感情分析

ゲーム内及びコミュニティー掲示板などのテキストから、ユーザーの感情を分析

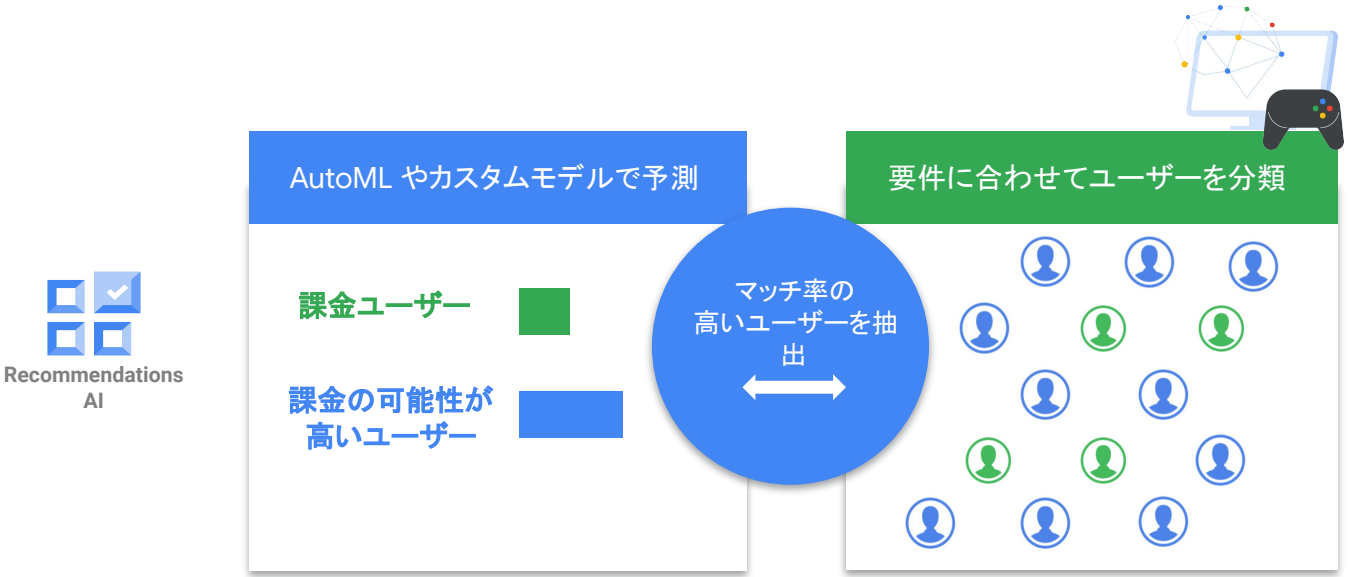
会話ができる NPC を実現

Dialogflow で単純なロジックやスクリプトで話す以上の会話ができる NPC を実現



ゲーム内のアイテムや広告をレコメンド

Recommendations AI で、ユーザーに合わせたアイテムや広告を提案



マッチングのバランス調整

学習したモデルで、プレイヤー マッチングのバランスを調整（デック・アイテム・スキル など）

Vertex AI

AutoML

Source: Leveraging Machine Learning for Game Development | Google AI Blog, 2021

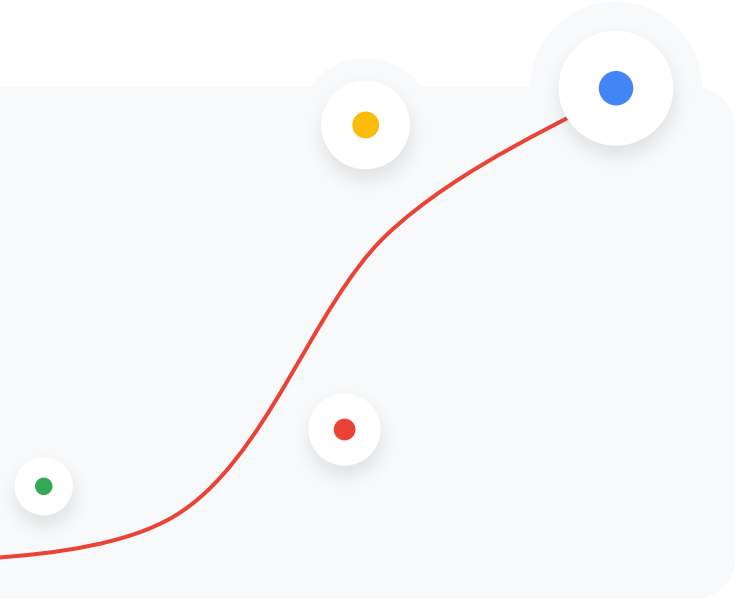
マルチプレイゲームにおける嫌がらせの問題

Google Cloud

マルチプレイ ゲーマーが感じる

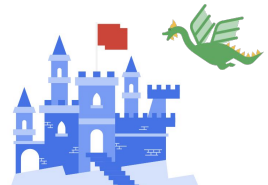
嫌がらせの増加

83%



Source: Hate is No Game: Harassment and Positive Social Experiences in Online Games 2021

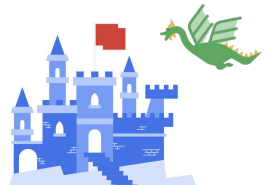
Google Cloud



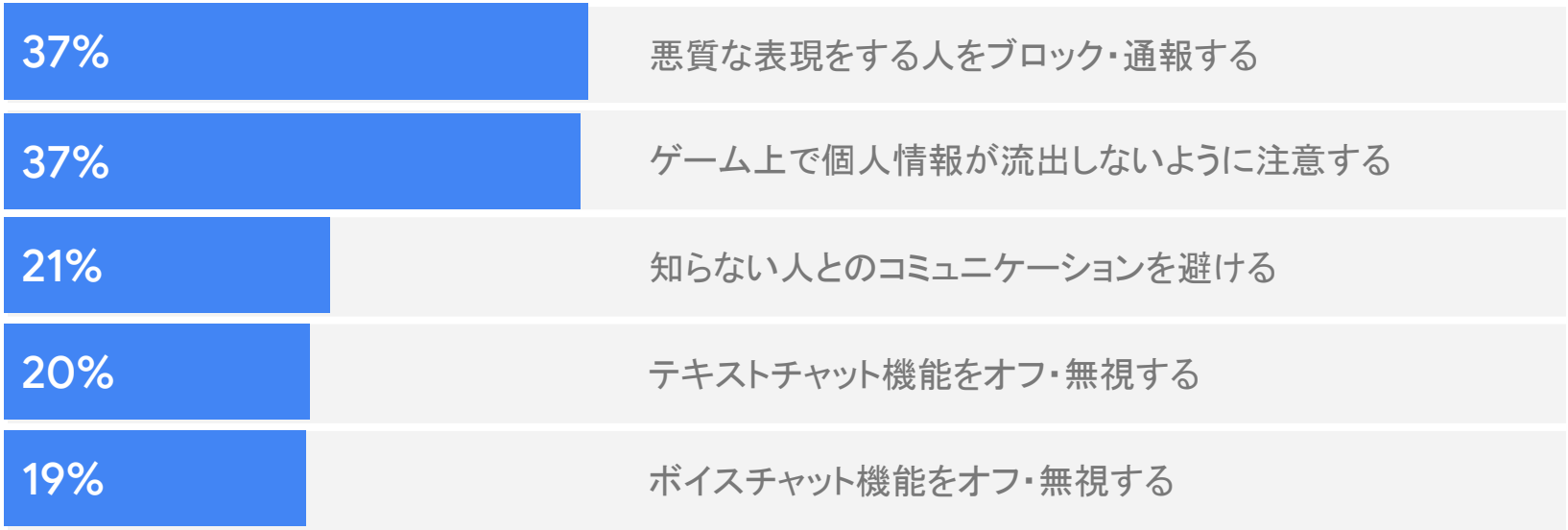
ゲーム内嫌がらせの種類

25%	不快に感じる表現
21%	個人を特定した攻撃
19%	差別
18%	脅迫
17%	セクハラ

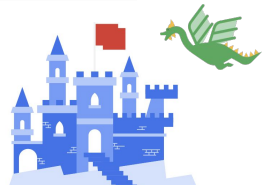
Source: Google for Games Dev Summit 2021



ゲーム内の嫌がらせから自らを守るために...



Source: Google for Games Dev Summit 2021



悪質な発言によるユーザーの離脱

64% のゲーマーは悪質な発言がゲーム経験に悪影響を及ぼすと感じています

5 人に 1 人 以上は、嫌がらせ・悪質な発言を経験し、ゲームを辞めることがあります

Source: Anti-Defamation League/NewZoo, 2020 -- Representative Survey (1,000 P18-45)

Google Cloud

ユーザーの離脱を防ぐための戦略

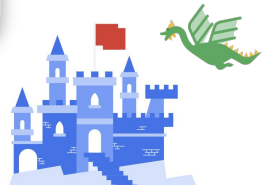
事前の予防

嫌がらせの発生率が低い
健全なコミュニティーを作る

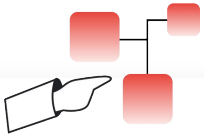
＋

事後の対応

悪質な発言がもたらす
影響を抑える
&
再発を防止する



嫌がらせ判定の難しさとその理由



ユーザーからの通報

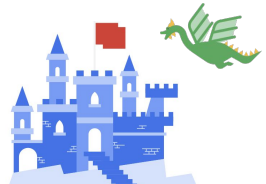
- システムが悪質な表現を通報するプレイヤーに依存
- 通報された内容を運営がレビューするための労力を要す

厳しすぎる規制

- 安全すぎる空間では楽しむ要素が減る
- ゲームを盛り上げる戦いや対立もある

ルールベースでの判定

- 文脈を考慮した分類ができない
- 言語の相違による受け取り方の違い



嫌がらせ問題の防止策

ユーザーの離脱を防ぐための戦略

事前の予防

嫌がらせの発生率が低い
健全なコミュニティーを作る

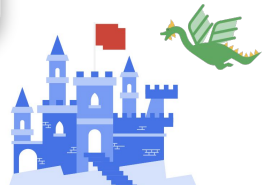
JIGSAW

＋

事後の対応

悪質な発言がもたらす
影響を抑える
&
再発を防止する

clean chat



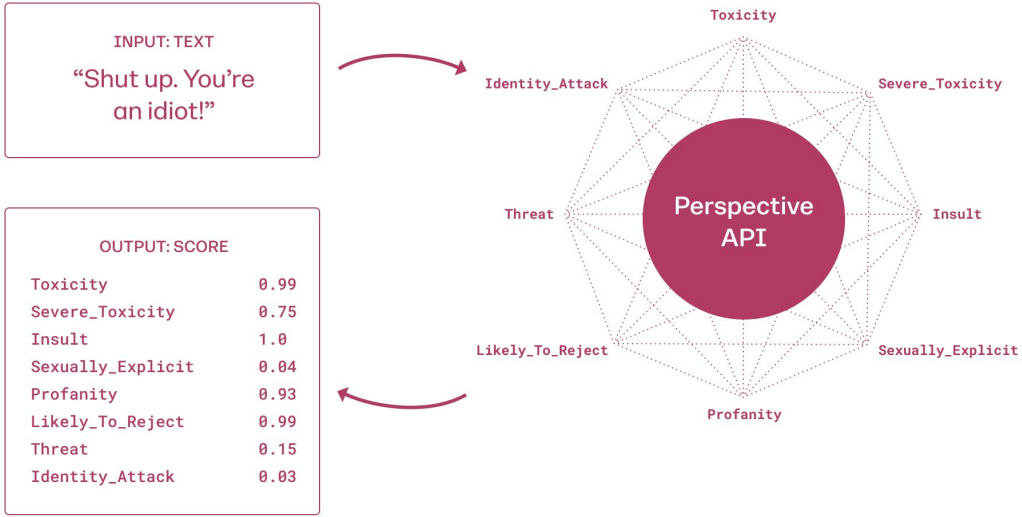
嫌がらせの検知

Perspective は、より安全な会話を
補助するため、Google 親会社 Alphabet 傘下
の Jigsaw が開発した API です。

入力テキストに対し、悪質な発言かどうかの
判定を 0 から 1 のスコアで返します。

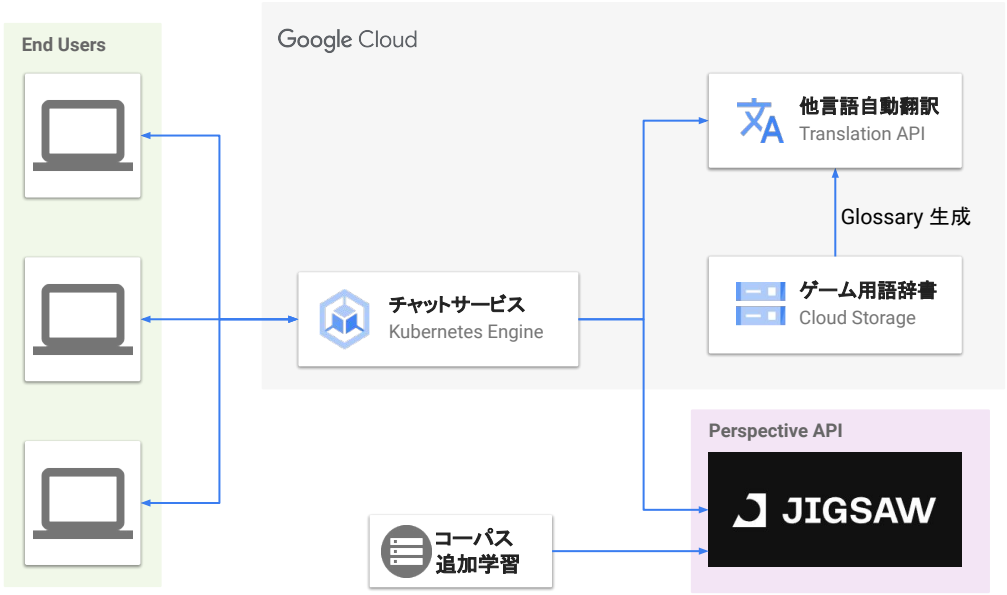
文脈と言語ごとの特徴を考慮して
各属性ごとにスコアリングを行い、
最終的な Toxicity スコアを算出します。

2021年6月の時点で、総合スコアを表す属性の
Toxicity が日本語対応になっています。



Source: https://developers.perspectiveapi.com/s/about-the-api

嫌がらせを多言語対応で予防 - 実装例



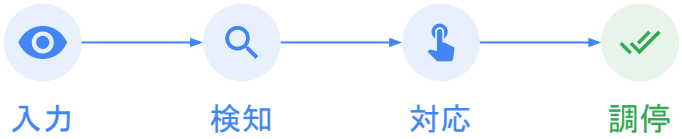
検証時のポイント

- リアルタイム性
- 翻訳の精度
- 嫌がらせ検出の精度
- 負荷耐久性

嫌がらせに対する事後の対応 - Clean Chat

clean chat

Clean Chat は、すべてのユーザーにとって
ゲームをより楽しく安全にするための
分析フレームワークです。
Google Cloud のプロダクトと組み合わせた
パイプライン実装例が OSS として公開されています。

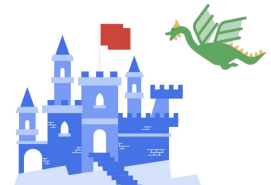


https://github.com/googleforgames/clean-chat

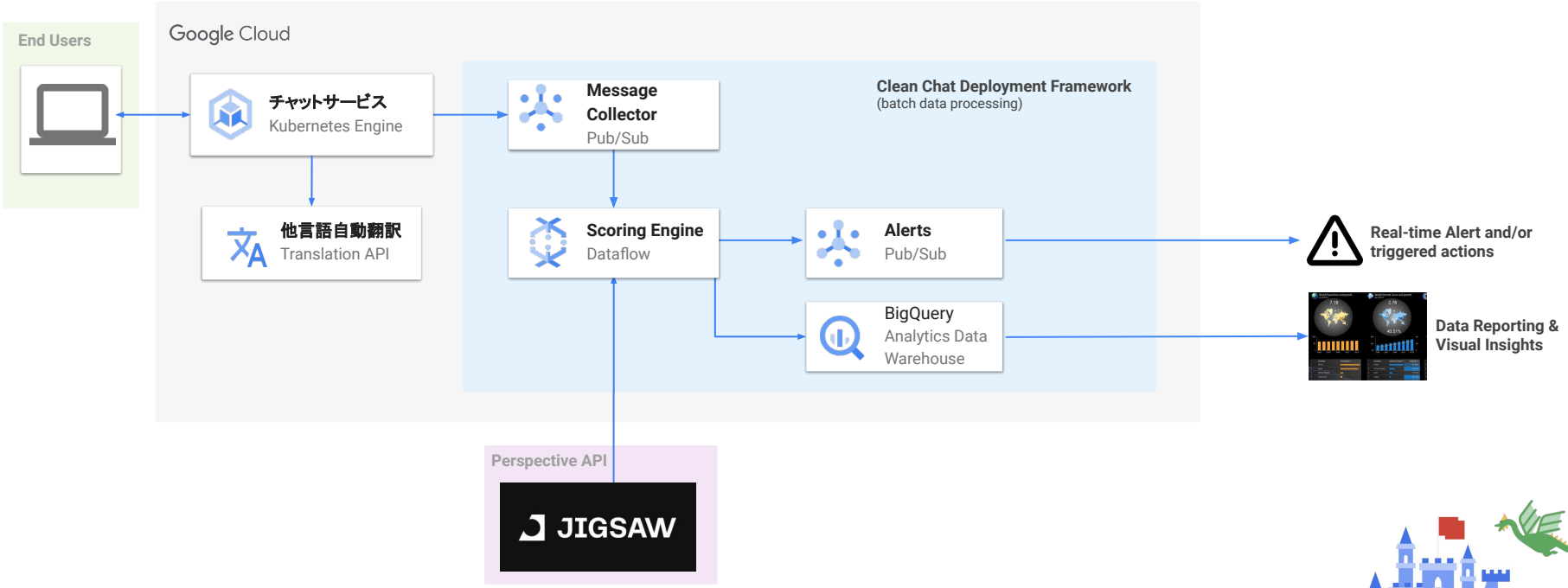
Clean Chat の特徴

- Model training と deployment framework
- イベントドリブンなアーキテクチャ
- Google Cloud のインフラストラクチャ

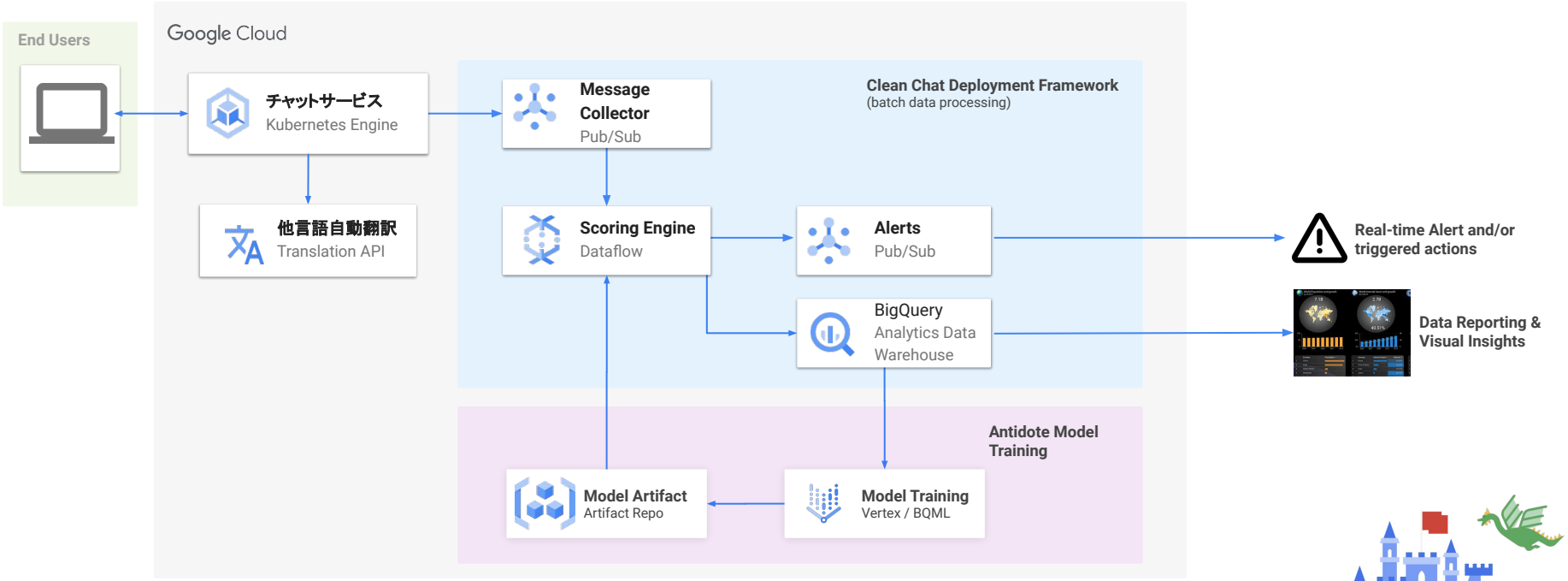
Google Cloud Players



嫌がらせに対する事後の対応 - Clean Chat

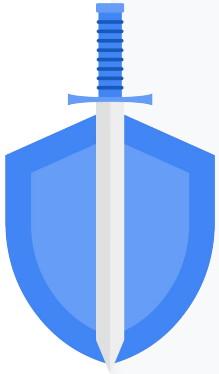


嫌がらせに対する事後の対応 - Clean Chat



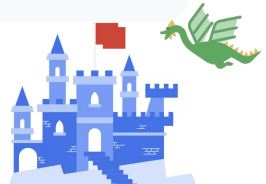
まとめ

- 嫌がらせによりユーザーの離脱
- ユーザー離脱を防ぐための
  事前の予防・事後の対応
- ベンダー ロックインのない
  Google Cloud のソリューション



Google Cloud では多様な目的に
合わせた ML ソリューションを
提供しています。

目的に合わせたソリューションの導入
で、ビジネスをアジャイルに展開できます。



Google Cloud

Thank you