Google Cloud

# Google Cloud ではじめる リアルタイム分析

西村 哲徳

グーグル クラウド ジャパン合同会社

Data Analytics Specialist

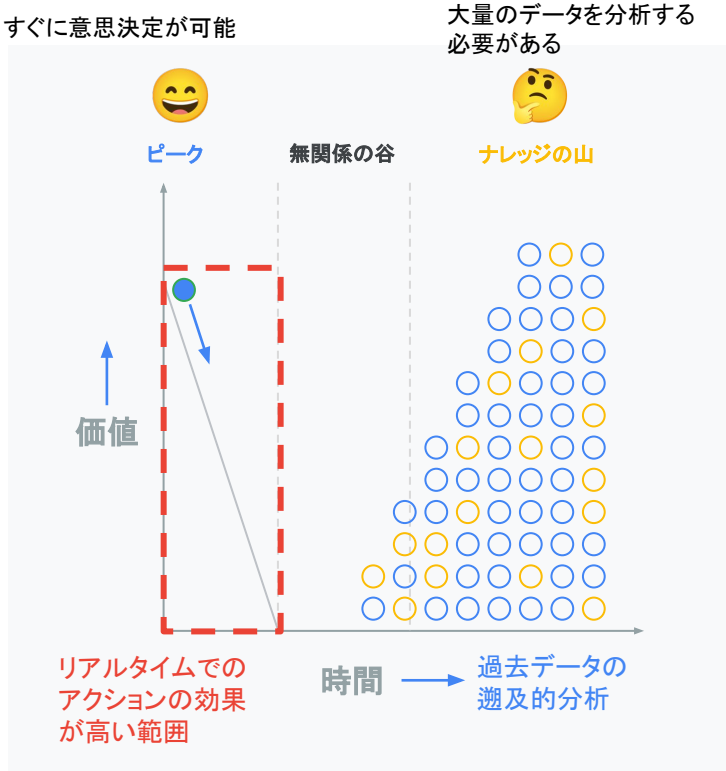
# アジェンダ

- リアルタイム分析のすすめ
- ユースケースから見るデザイン  パターン

# リアルタイム分析のすすめ

# リアルタイム分析の必要性

- 機会損失を防ぐ
  - 競争優位を獲得
  - 変化の早い状況への適応
  - より適切なデータポイントを収集
- 誤った判断を防ぐ
  - 鮮度の高いデータでの意思決定
- タイムラグによる選択肢の減少を防ぐ
  - 低コストの打ち手が取れる
  - 例）輸送コスト： 飛行機 vs トラック
  - 余裕をもった変更への適用
  - 意思決定の精度を高める



# 全ての業界に変革をもたらすリアルタイム分析

## ヘルスケア

- リアルタイムの患者モニタリング
- 医療機器 IoT でのパーソナライズケア
- 保険金詐欺、廃棄物、薬の乱用の警告
- リアルタイムの医薬品供給と在庫の最適化

## 金融サービス

- 不正検出と防止
- 将来のコンプライアンスを計画
- リスクの予測と分析
- 顧客へのパーソナライズされたオファーを促進

## 小売

- リアルタイムのパーソナライズされたマーケティング
- ダイナミックプライシングとレコメンド
- リアルタイムの可視化で購入未完のアラート
- オムニチャネルでの顧客を即座に可視化

## 通信

- ネットワークの信頼性の改善
- 解約の最小化
- 加入者の不正を未然防止
- ネットワーク容量の分析と最適化

## 製造

- リアルタイムの在庫最適化
- 予知保全とアラート
- スマートサプライチェーン
- フリート管理と分析

# 従来のプラットフォームでのリアルタイム処理の難しさ

増加するデータソースの取り込み

データ処理の拡張性や柔軟性の不足

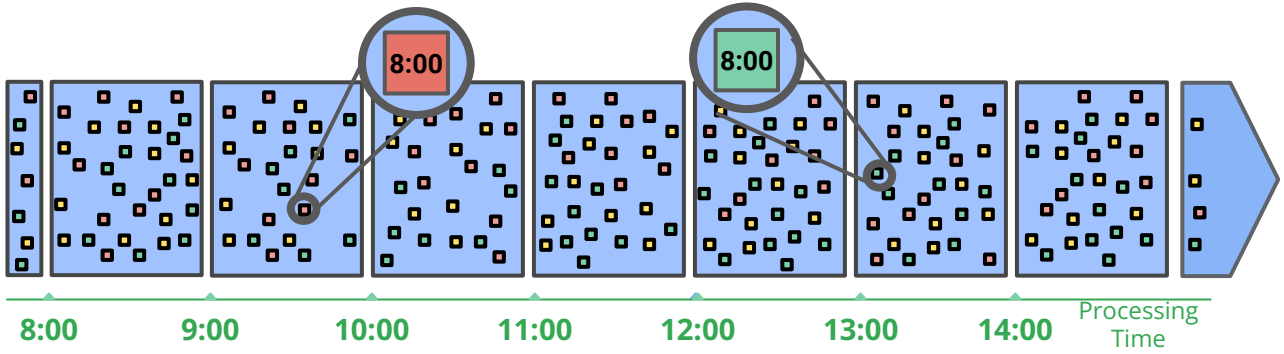
すぐに利用可能なスキルの不足

分断されたツール間の連携

# リアルタイム分析の難しさ

- 無制限なデータに対するウィンドウ集計
  - 集計期間
  - ウィンドウの種類
  - トリガー
  - イベント発生時間と処理時間の違い
  - 遅延データの取り扱い

イベント発生時間は8:00 だが
パイプラインへ到着時間がずれると処理時間は別々になる



# Google Cloud が提供する リアルタイム分析 ソリューション

## 堅牢なデータ収集サービス

- データの冗長性を抑えながら複数の サブスクライバにデータをパブリッシュする信頼性 の高いデータ/ イベントの取り込み

## 統合されたストリームとバッチ処理

- ストリーム処理によりデータとイベントが アクション可能な洞察に変わります。
- バッチとストリームの統合により ラムダ アーキテクチャの負荷の軽減
- 様々なウィンドウ集計とウォーターマークによる遅 延データのハンドリング

## サーバレス アーキテクチャ

- リアルタイムソリューションのキーである大量 データに対応する拡張性
- スパイクの調整やプロビジョニングからの解放

## 包括的な分析ツール

- DWH、機械学習、オンライン アプリケーションとの親和性

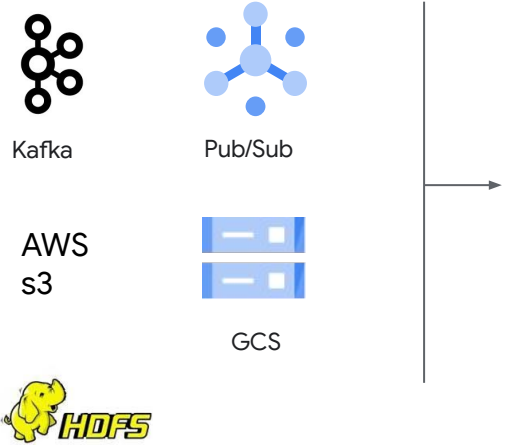
## 柔軟性の高さ

- 既存のスキルセットにあわせて移行可能なプ ラットフォーム

# オープンでインテリジェントかつ柔軟な基盤の上に構築されたエンド ツー エンドのソリューション

## 01 収集

信頼性の高い
データ取り込み、配布




## 02 処理

高速で正確な計算を迅速かつ簡単に


## 03 格納と分析

機械学習、データ ウェアハウス、オンライン処理

# Pub/Sub

## アプリケーションとマシンがあらゆるものからインサイトを得るための ハイパー スケール、サブスクリプション メッセージング

**Pub/Sub**

リアルタイム分析のための メッセージングとイベントの取り込み

**拡張性、耐久性のあるイベントの取込みと配信**
サーバーレス、自動スケーリング、自動プロビジョニング。一貫したパフォーマンス。データレプリケーション。最大 31 日間、99.95%以上の SLA

**拡張性のある パブリッシュ / サブスクライブ パターン**
トピックごとに最大10,000 のサブスクライバー; Push と Pull の配信モデル

**グローバル ルーティング**
地理に関係なく、イベントをパブリッシュおよびサブスクライブ

**深いインテグレーション**
Dataflow を使用したスケーラブルな分析、Functions を使用したサーバーレスアクション、組み込みの監視、監査ログ、コンプライアンス

Architecture

# Dataflow

## データ パイプラインの最適化と自動化、
## ストリーミング インテリジェンスの複雑さを取り除く

**Dataflow**

ストリーミングおよびバッチデータ処理のための高度に自動化された自己修復データ パイプライン実行

**非常にシンプル**
サーバーレス、自動プロビジョニング、および自己修復。自動化されたパフォーマンス、ワークバランシング、統合されたバッチとストリーミング

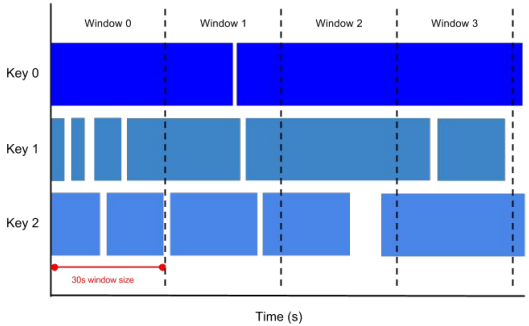
**ML へのアクセス性**
すぐに利用可能な MLOps、ML 推論、GPU と大規模データ処理

**ベストな OSS と最適化されたプラットフォーム**
ApacheBeamSDK によるオープン イノベーション。複数の言語が利用可能。組込みの監視と観測でワークロードを最適化

**規模に合わせて構築**
水平スケーリングと垂直スケーリングの両方に対応するように設計および構築。あらゆる規模でパイプラインのパフォーマンスを最適化。使用率を最大化してコストを節約し、価値実現までの時間を短縮

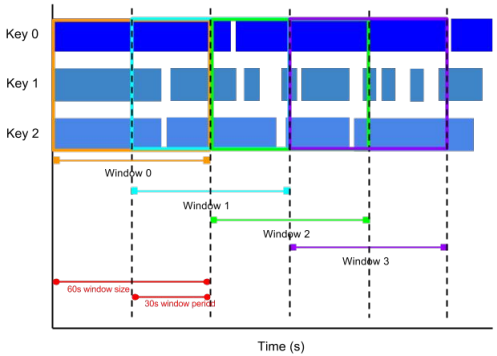
# Dataflow の提供するウィンドウの概念

### タンブリング ウィンドウ



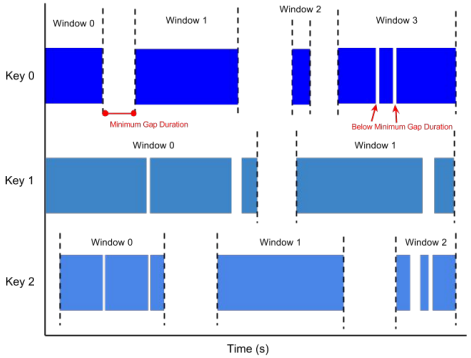

固定ウィンドウのこと。データストリームを重なりなく分ける一定の時間間隔

### ホッピング ウィンドウ



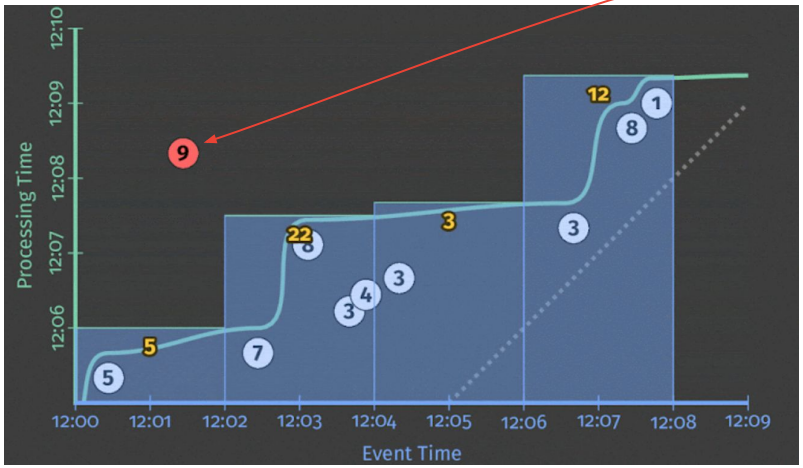

スライディング ウィンドウのこと。一定の時間間隔のウィンドウが指定した期間で重なりあう

### セッション ウィンドウ




あるキーの次の新しいデータが来るまでのギャップ期間を指定し、その間隔を超えると新しいセッションとしてウィンドを計算する

# ウォーターマークによるイベント時間と処理時間の違いを考慮した集計




ウォーターマークとは特定の**ウィンドウ内のすべてのデータがパイプラインに到着したと予想**されるシステム的な概念でこれを常に追跡

ウォーターマークを超えて到着した**遅延データの取り扱いも制御することが可能**

# ユースケースから見る
# デザイン パターン

# ユースケース

シンプルなアーキテクチャで様々なユースケースをサポート

- クレジット カード不正利用のリアルタイム検出
- リアルタイムのクリック ストリーム分析
- 動画クリップ内のオブジェクトの検出
- 個人情報の匿名化と再識別

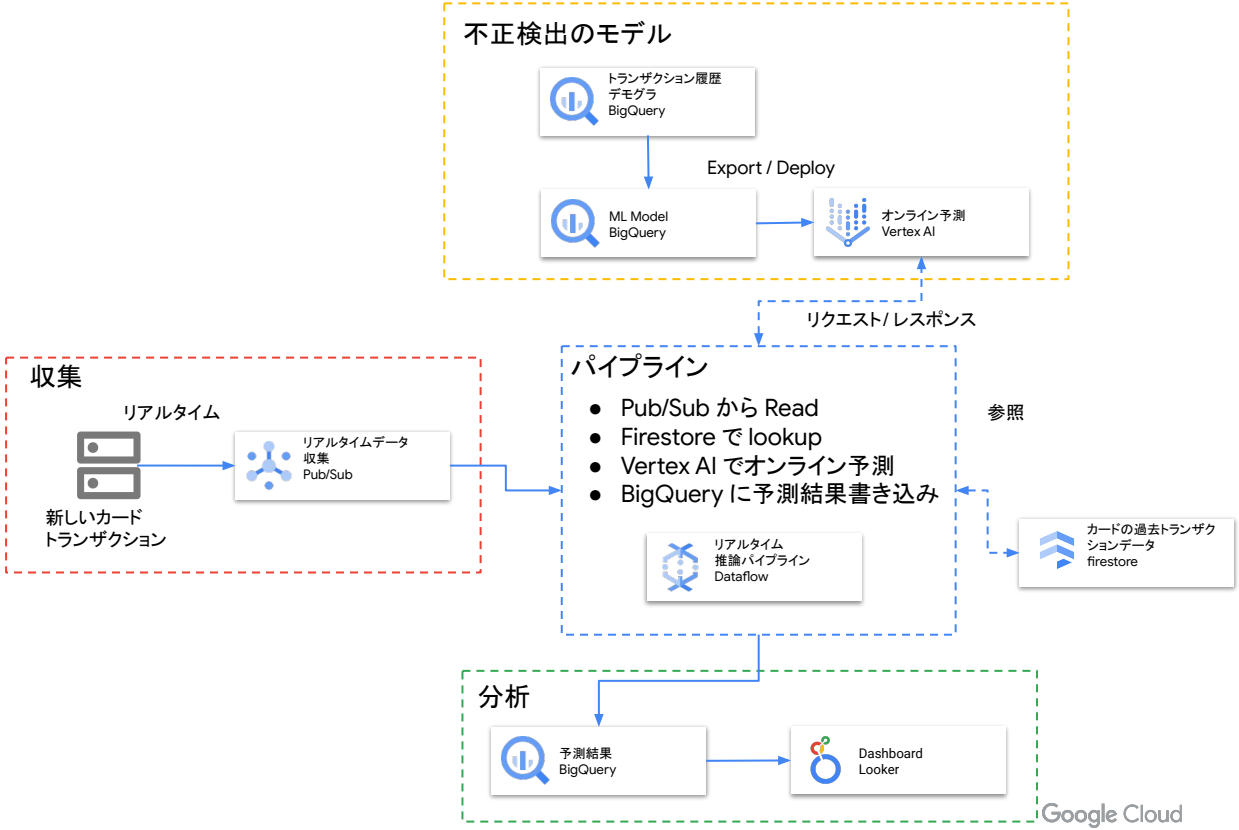
# クレジット カード不正のリアルタイム検出

**BigQuery ML でモデル作成**
履歴データから XGBoost で
モデルを作成

**Vertex AI でモデルのホスト**
オンライン予測用に Vertex AI に作
成したモデルをデプロイ

**Pub/Sub -> Dataflow による
オンライン不正検出**
Dataflow から Vertex AI に
リクエストを送りオンライン予測

**BigQuery / Looker で分析**
不正検出の状況を Looker で
リアルタイムに可視化



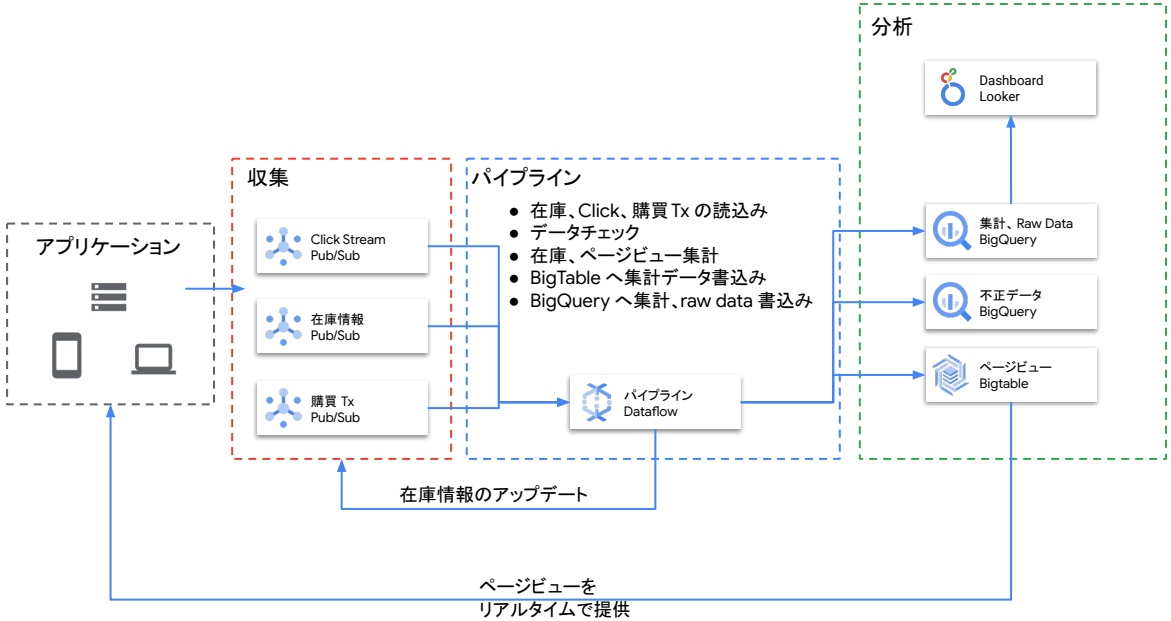
# リアルタイムのクリック ストリーム分析

**マルチソース / マルチシンク**
Dataflow で複数のソースから読み、複数の宛先に書き込む

**リアルタイム集計、データチェック、変換**

**ページビューのリアルタイム表示**
集計データを Bigtable に格納し アプリに提供

**BigQuery / Looker で分析**
在庫、ページビュー、売上を Looker で可視化

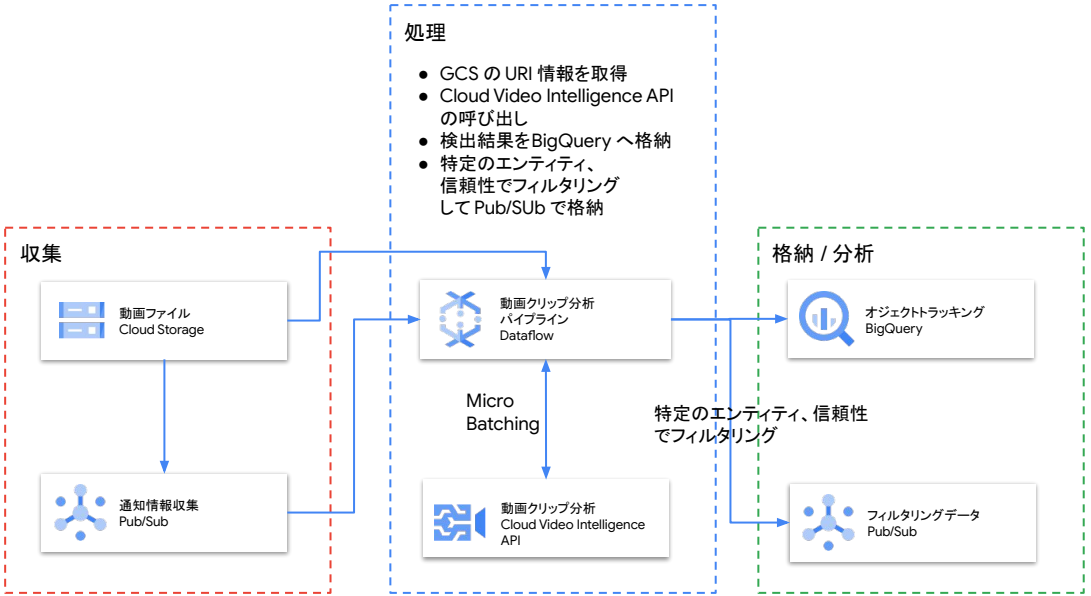


# 動画クリップ内のオブジェクトの検出

**Cloud Video Intelligence**
との連携でオブジェクト検出

**Micro Batching で API Call**
Dataflow では、クリップの
処理時間が短縮されます

**ファイルベースのリアルタイム**
ファイルのアップロードの通知が
Pub/Sub に送られ Dataflow が
自動的に処理

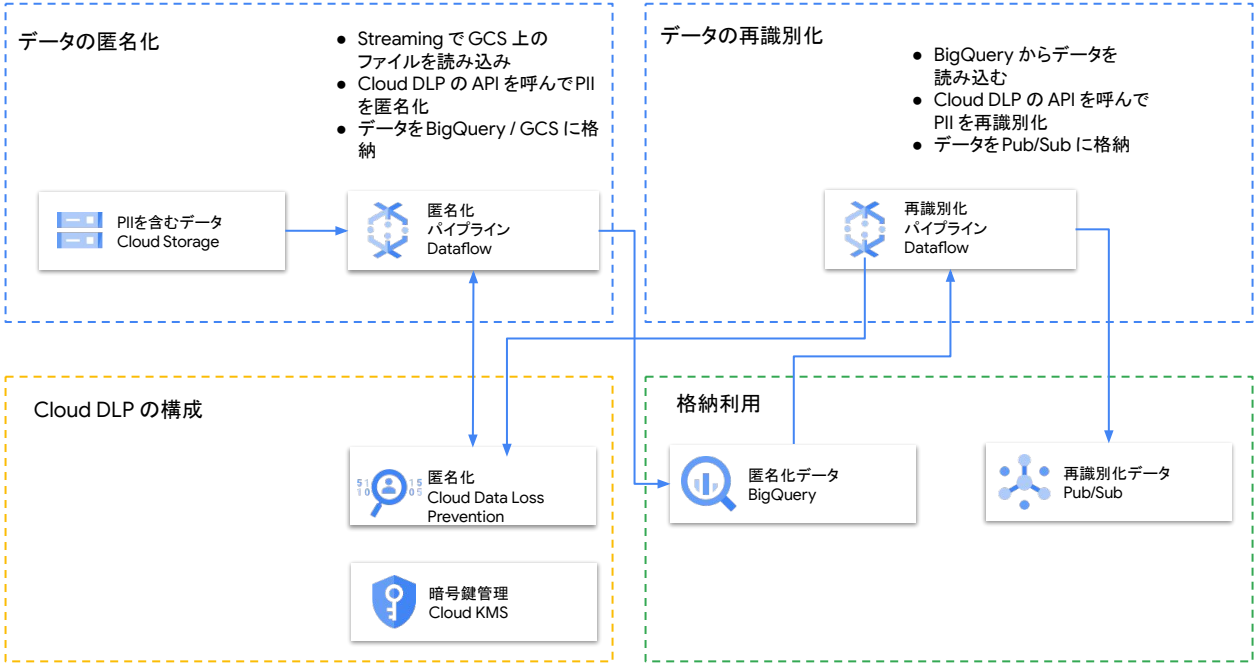


# 個人情報の匿名化と再識別

**Cloud DLP**
との連携でトークン化、
PII データの匿名化

**スループットの最大化**
Dataflow の State と Timer の利用

**ファイルベースのリアルタイム**
GCS 上のファイルを
リアルタイムに匿名化し
BigQuery や GCS へ連携



# データ分析の設計パターン

本日紹介したユースケースの
サンプルコード以外にも様々なユースケースに
あわせたデータ分析のサンプルコードや技術
リファレンスガイドがございます。

Cloud アーキテクチャ センター › データ分析の設計パターン

この情報は役に立ちましたか？

## データ分析の設計パターン

フィードバックを送信

このページでは、ビジネス ユースケース、サンプルコード、業界データ分析ユースケースの技術リファレンス ガイドへのリンクを紹介します。これらのリソースを使用して、ワークロードの実装を加速するためのベスト プラクティスを学習します。

**注:** データ分析の設計パターンを実装している場合は、ぜひご連絡ください。こちらの短いアンケートにお答えいただき、ご感想をお聞かせいただければ幸いです。

ここに示す設計パターンはコード指向であり、短時間で実装できます。幅広い分析ソリューションを確認するには、ビッグデータに関する技術リファレンス ガイドのリストをご覧ください。

### 異常検出

| 解決策 | 説明 | プロダクト | リンク |
|---|---|---|---|
| K 平均法クラスタリングを使用したデータ通信ネットワーク異常検出アプリケーションの構築 | このソリューションでは、Dataflow、BigQuery ML、および Cloud Data Loss Prevention を使用して、データ通信ネットワークに ML ベースのネットワーク異常検出アプリケーションを構築し、サイバー セキュリティの脅威を特定する方法を説明します。 | • BigQuery<br>• Cloud Build<br>• Cloud Data Loss Prevention<br>• Cloud Storage<br>• Dataflow<br>• Pub/Sub | 技術リファレンス ガイド: Dataflow、BigQuery ML、および Cloud Data Loss Prevention を使用した安全な異常検出ソリューションの構築<br>サンプルコード: Netflow ログの異常検出<br>ブログ投稿: ストリーミング分析と AI を使用した異常検出<br>概要の動画: 安全な異常検出ソリューションの構築 |
| BoostedTrees を使用して金融取引の異常をリアルタイムで見つける | このリファレンス実装では、Dataflow と AI Platform で TensorFlow ブーストツリー モデルを使用して不正なトランザクションを識別する方法について説明します。 | • AI Platform<br>• BigQuery<br>• Cloud Storage<br>• Dataflow<br>• Pub/Sub | 技術リファレンス ガイド: AI Platform、Dataflow、BigQuery を使用した金融取引の異常検出<br>サンプルコード: 金融取引における異常検出 |
| LSTM オートエンコーダを使用した時系列データの異常の検出 | このリファレンス実装を使用して、時系列データを前処理し、ソースデータのギャップを埋める方法を学習します。その後、LSTM オートエンコーダを使用してデータを実行して異常を特定します。オートエンコーダは、LSTM ニューラル ネットワークを実装する Keras モデルとして構築されています。 | • BigQuery<br>• Dataflow<br>• Pub/Sub | サンプルコード: 時系列データの処理 |

データ分析の設計パターン

Google Cloud

# まとめ

- リアルタイム分析は競争優位性を高める
  - 機会損失を防ぐ、打ち手を増やす、意思決定の精度をあげる
- リアルタイム分析実現にともなう困難は Google Cloud で軽減することが可能
  - 拡張性、オープンでインテリジェントな統合されたソリューション
- シンプルなアーキテクチャで様々なユースケースをサポート
  - データ分析の設計パターン

# Thank you.

Google Cloud