Google Cloudの大規模障害についてGoogleが原因と対策を説明

22161412　髙橋 和真

原因はAPI管理システムの新機能のコード不備

5月29日にService Controlに追加された新しいクォータポリシーチェック機能のコードに、適切なエラーハンドリングが欠けており、また重要な「機能フラグ」による保護もされておらず、6月12日にService Controlが使用する地域ごとのSpannerテーブルに、意図しない空のフィールドを含むポリシー変更が挿入された。

この誤ったポリシーデータが処理される際、不備のあるコードパスが実行され、nullポインタエラーが発生。これによりService Controlのバイナリが世界中でクラッシュループに陥った。

障害発生から2分以内に対処開始

Site Reliability Engineeringチームは、障害発生から2分以内に対応を開始し、10分以内に根本原因を特定した。

問題のあるAPI提供パスを無効化する「赤ボタン」を導入し、障害発生から40分以内にサービスの回復が始まった。サービス回復後すぐに、Service Conrolシステムのすべての変更と手動でのポリシー更新を一時的に停止した。

今後の改善アプローチ

- Service Controlのアーキテクチャのモジュール化
- グローバルに複製されるデータの監査
- 機能フラグによる保護の徹底
- エラーハンドリングとテストの改善
- 指数関数的バックオフの徹底
- 外部コミュニケーションの改善
- 監視および通信インフラストラクチャの継続的な運用
- API管理プラットフォームの堅牢化
- メタデータの伝播保護

コメント

松本さん

エラー処理部分の障害は事前検証では見つかりにくいところです

エラー発生後2分後から対策開始しているところを見るとかなり適切に処理に移っている様です

クラウドサービスはこうした裏の人たちが安定稼働に向け24時間365日対応することで安定化をしています

デジタルだから安定しているという訳ではなく、如何に人が苦労しながら止まらないシステムを支えているのか、こういう障害時に垣間見えます

感想

- Androidスマホでは、保存している画像の大半はドライブ保存になっています。文書や画像をドライブに保存した方がなくならず安全という売り文句があるからこそ利用している人もとても多いサービスです。

- コメントでもありましたが、24時間監視して、もしバグが起こってもすぐには反映されないように仕組み化することはいいですが、その代わりの膨大な労働力にはどのような考慮がされるのかが働き方改革の一環として考えられると思います。