

リクルート横断データ プラットフォームがGoogle Cloud 移行によってパフォーマンス・安定性ともに大幅向上



株式会社リクルートテクノロジーズ

<https://recruit-tech.co.jp>
〒100-6640 東京都千代田区丸の内 1-9-2
グラントウキョウサウスタワー

リクルートグループのビジネスにおける IT・ネットマーケティング テクノロジーの開発・提供を行う機能会社として 2012 年、株式会社リクルートを会社分割する形で設立。グループ横断で利用するデータ基盤やインフラなどの構築を行う。従業員数は 2019 年 4 月 1 日時点で 701 名。

※ 本記事に登場する組織名および人物の所属組織は 2020 年 3 月時点のものです。

求人広告から販売促進メディア、人材派遣まで、さまざまな分野で業界トップクラスのサービスを提供しているリクルートグループ。当然ながら、そこで扱われる情報は膨大です。今回お話を伺った、株式会社リクルートテクノロジーズ データプラットフォーム部は、2016 年 10 月から約 2 年半の歳月をかけて、段階的に同グループの情報基盤に Google Cloud を導入していきました。その理由と成果について聞いてきました。

(Google Cloud ソリューション)

データ ウェアハウス モダナイゼーション、データレイク モダナイゼーション

(利用している Google Cloud サービス)

BigQuery、Google Kubernetes Engine、Cloud Dataproc など

データプラットフォーム部
データプラットフォーム 4 グループ
グループマネジャー
鶴谷 誠文

顕在化してきたオンプレミス環境の問題を Google Cloud 活用で段階的に解消

「私達のグループは、グループ共通で使われている『リクルートID』を軸とした全社横断のデータ プラットフォームを運営しており、2016 年 10 月までは多くのリソースが、オンプレミス環境で動作していました。規模感としてはサーバーが 300 台超、そこで取り扱うデータのボリュームが 2,000 TB 超、月あたりのクエリ数が 1 億超、そして、その運用に関わる人員が 90 人超となります。」(鶴谷さん)

このデータ プラットフォームでは、膨大なデータマートの生成や機械学習パッチジョブの実行、各種マーケティング施策、アドホック分析を行ってききましたが、昨今の取り扱いデータの増大や処理の複雑化などによって、さまざまな問題が発生。当初はサーバーを追加し続ける形で対応してきましたが、それを迅速に実行し続けるのは難しく、クラウドを活用した根本的な対策が求められるようになっていたそうです。

まず最初に顕在化したのが、ユーザーに対しておすすめの商品をレコメンドするレコメンデーション API が膨大なリクエストをさばききれなくなるという問題です。当初はこれを App Engine で解決しようとしたのですが、

App Engine は当時の仕様では、オンプレミス上の HBase とのプライベート ネットワーク構築ができず断念(*)。一方で、Compute Engine では VM の管理が必要となり、管理が煩雑となることが想定されたため、最終的に Google Kubernetes Engine (GKE) を採用することになりました。2016 年 10 月から約半年ほどかけてこの移行を行い、課題となっていた同時アクセス性能の大幅な改善のほか、スケーラビリティの確保なども実現しています。

そして、その次に問題となったのが、オンプレミス上のデータ ウェアハウスにおいて、アドホックな分析処理が、重要度の高い定期実行処理と混在するような状況になっていたこと。これが原因で、落ちてはならない重要なバッチ処理が止まってしまうという問題が頻発していたと、当時を振り返ります。そこで、まずはアドホックな分析処理を BigQuery に移行させることにしました。この際、工夫をしたのが、BigQuery で扱うデータソースの移行方法。

*現在は VPC やインターコネクトと通信できる "Serverless VPC Access" 機能が発表されている

半分以上がオンプレミス上にあるものなので、これと Google Cloud を高速な専用線で接続する Cloud Interconnect(Dedicated Interconnect)を導入しましたが、データ量が膨大なため、何も考えずに転送すると、それだけで帯域を使い切ってしまう。そのためデータの特性に応じてデータロードの頻度を変えるなどといった工夫で転送量を制限。他のアプリケーションの通信に影響を及ぼさないようにしています。データ転送のリードタイム長期化によって、データ鮮度の低下が懸念されましたが、今のところそういった問題は発生していないとのこと。また、Cloud Interconnect も、今はさらに広帯域なものが選べるようになっており、現在はそちらが採用されているとのこと。



限界が見えつつあった大規模 Hadoop クラスタを Google Cloud に全面移行

そして、2018 年 4 月、レコメンデーション API の GKE 移行と、アドホックな分析基盤の BigQuery 移行、この 2 つの成功を受けて、オンプレミス上の Hadoop クラスタを Google Cloud に全面移行するというプロジェクトがスタートします。

きっかけはオンプレミス上に構築された Hadoop 環境が EOSL(End of Service Life)を迎えるためでした。この際、移行先として Google Cloud を選定したのは、これまでの検証実績で BigQuery が高速であることが分かっていたためだと言います。例えば、日次で実行されるバッチ処理の例では、毎日 11 時間かかっていた統合データマートの作成処理が約 14 分の 1 の 47 分で終わるようになっていました。これにより、従来環境で慢性的に起きていた処理の長期化という問題が解決できました。そのほか、ユーザー アカウントが、Google アカウントと統合されており、ユーザー管理が非常に楽であること、Web ブラウザから簡単にクエリを書ける環境が提供されていること、事業ごとに分けられている Google Cloud のプロジェクト間におけるデータ共有が簡単に行えることなども、Google Cloud 導入の理由となったとのこと。なお、移行に際しては、Google Cloud のカスタマーエンジニアからの移行におけるベストプラクティスの提案や類似ケースの紹介、助言なども、とても助けになったとのこと。

プロジェクトは 2018 年 4 月にスタートし、多くの処理を BigQuery へ移行。ただし、一部の BigQuery にすぐに移行出来ない処理については、Cloud Dataproc を利用しています。

Cloud Dataproc は、クラウド上で Hadoop 環境をフルマネージドで提供してくれるサービスであり、オンプレミスと比べて、Hadoop クラスタの物理的な運用から解放されるメリットがあります。ここでは、既存のコンポーネントを利用して、ユーザーをレコメンドするためのロジックを生成する機械学習の処理をおこなっていたので、それを移行期間中に Hadoop 以外の処理としてリファクタリングするのは難易度が高いと判断し、まずは処理をリファクタリングすることなく、Cloud Dataproc で動かすことに。現在は、これをさらに進化させるために、BigQuery ML や Cloud Dataflow などを用いて、よりクラウドネイティブに実現できないかの検証が行われています。

大規模な移行でしたが、2019 年 1 月には、無事、移行が完了。その効果

は満足いくものだったと言います。

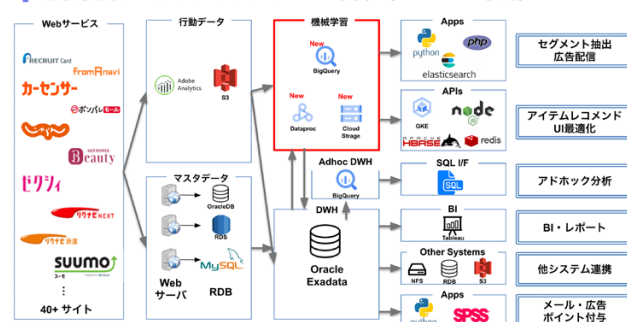
「パフォーマンスと安定性という、解決したかった 2 つの課題が Google Cloud 移行によって無事解決しました。リソース不足などに起因したエラーが減り、データの加工処理がより確実に行えるようになったことで、データの精度・鮮度が向上し、結果としてデータ利用者に、より適切なデータを提供できるようになりました。今後も、Google Cloud を活用してアジリティ、スケーラビリティ、運用効率を高めていきたいです。フルマネージド、サーバーレスなど、よりクラウドネイティブなアーキテクチャーに進化させて行きたいです。今後はこれまで以上に収集できるデータが増えていくので、柔軟なデータ拡充・活用も大きな課題の 1 つ。これらにしっかり対応し、新たなビジネス価値の創出に寄与していきたいと考えています。」(鶴谷さん)

今後のプロジェクトについてはまだ話せる段階にないとのことですが、現在進行形の取り組みとしては、フルマネージドなメタデータ管理サービス Data Catalog(ベータ版)を検証中とのこと。

2014 年ごろから、全社のメタデータを集約して可視化するという取り組みも進められており、それと Data Catalog を協調させることで、不足の機能をおぎなうなどができたらいいとのアイデアも。

「Data Catalog は、Google Cloud の機能と統合されているのが大きなメリット。BigQuery での分析時に、その GUI 上で連携できることで、分析の効率化に寄与してくれそうです。」(鶴谷さん)

■ 全社横断データプラットフォームの変革 (～2018.12予定)



Google Cloud を活用することで、ビジネスの将来に注力できるようになります。インフラストラクチャの管理やサーバーのプロビジョニング、ネットワークの構成などに起因する負担を軽減することができます。つまり、イノベーターもプログラマーも、自分の本来の仕事に集中することができます。

お問い合わせはこちら
<https://goo.gl/CCZL78>

