Customer CASE STUDY

databricks

レイクハウス・プラットフォームを活用し、年間利用会員数7,000万人のデータから独自の「顧客DNAのプロファイリングデータ」を生成。

CCC MARKETING GROUP

CCCマーケティンググループは、カルチュア・コンビニエンス・クラブ(CCC)内において、「UNIQUE DATA , SMALL HAPPY.」をミッションに掲げ、テクノロジーと対話力と提案力で、パートナーの課題を解決するマーケティング・ソリューション事業を展開している。中核となるTポイント事業は、もともとビデオレンタルサービス向けの顧客管理から始まり、顧客のポイントカードの利便性向上のため、日本で初めて共通ポイントサービスを開始した歴史がある。現在、圧倒的に大規模、高品質なデータを日々運用しており、国内最大規模のデータ保有企業の1つと言える。

CCCマーケティング株式会社
IT　シニアマネージャー
松井 太郎 氏

CCCマーケティング株式会社
データベースマーケティング
研究所　技術開発ユニット
シニアMLエンジニア
岸部 友裕 氏

ハイライト

**複雑だった
データ処理全てを
Pythonに統一可能に！**

**ジョブの実行時間が
1/10に短縮！
(340時間→34時間)**

**クラウドコストも
1/5へと
大幅に削減！**

課題

データ分析基盤を次世代アーキテクチャーに進化させるために、複雑化するジョブフローとデータ処理の克服が必要。

現在のTポイント事業の規模としては、年間利用会員数7,000万人を超え、その約6割が月に一回以上利用しており、関連購買トランザクションは年間35億回を有する。扱っているデータの種類としては、会員の属性データや購買データを中心に、ウェブやTV視聴データ、また天候や住居関連などの連携データなどもある。これらの膨大なデータを活用して、1,000以上の項目から構成された独自の「顧客DNAのプロファイリングデータ」を生成している。顧客DNAの例としては、婚姻状況、飲食の嗜好、運転免許の有無などがあり、これらのプロファイリングの適応確率をAI/機械学習を活用して予測をしている。

顧客DNAのスコアは最低でも1ヶ月に1回は更新する必要があり、全体の処理を1ヶ月以内で完了させる必要がある。しかしながら、データブリックス導入前の顧客DNA運用基盤には、大きく3つの課題があった。1つ目は、ジョブフローが複雑な点である。共有DBからの「データ取得」→「データ前処理」→「モデル作成の各処理」が、それぞれ異なる環境で実行されており、各処理においてつまづくポイントが複数あった。また、場合によっては、それに気づけない場合もあった。2つ目は、データ処理が複雑である点である。処理によって、異なる実行環境でpython、SQL、shellで書かれており、メンテンスや変更は難しい。3つ目は、精度向上の限界である。勾配ブースティング系のアルゴリズムの誕生以降、アルゴリズムの大幅な精度向上は難しくなっていた。新しい特徴量の開発が必要となるが、大規模データであるがゆえ、試行錯誤できる環境がない。これら3つの課題により、新しいデータ分析基盤の検証含めて、次世代アーキテクチャへ進化させる必要があると感じていた。

採用のメリット

圧倒的な分析性能に加え、時間的・経済的にも大きなメリットを享受。データ&AIの専門家に適切なアドバイスも得られる。

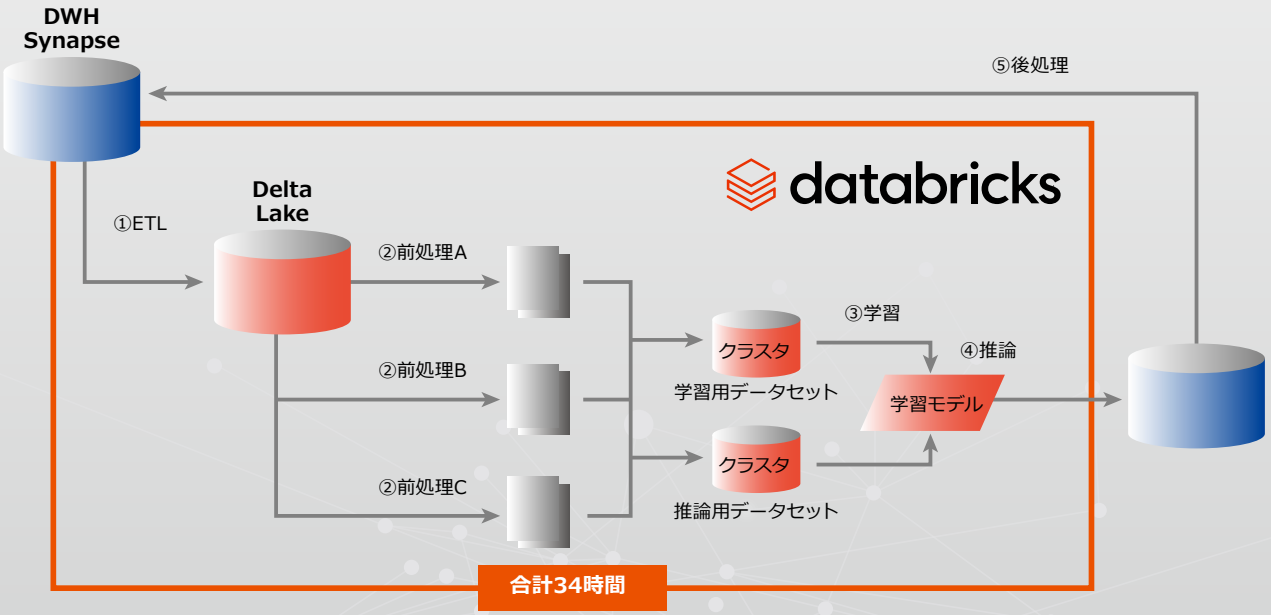
データブリックスの採用により、3つの課題を解決することができた。複雑だったジョブフローに関しては、実行環境がデータブリックスに統一されたため、疎結合を生かしたままジョブが失敗しにくくなり、失敗した場合も気づきやすくなった。また、複雑だったデータ処理に関しては、ほぼ全ての処理はPythonに統一する事ができた。また、ノートブックで書かれているため、メンテナンスや変更も容易になった。加えて、精度向上の限界に関しては、大規模データを活用した試行錯誤を容易にできるため、積極的に精度向上に向けた取り組みを行うことができるようになった。

時間的、かつ経済的にも大きなメリットを享受している。データブリックスを活用することにより、ジョブの実行時間が1/10(340時間→34時間)になり、結果としてクラウドコストも1/5へと大幅に削減することができた。また、運用時間に余裕ができたことにより、その余裕を活用し更に高度なデータ処理にも挑戦することができるようになった。

改めて顧客DNAの運用基盤をデータブリックスに置きかえて振り返って見ると、2つの真の価値があると感じている。まず、データブリックスという汎用的なプラットフォームを利用することで、スクラッチで専用設計しているシステムと比較すると、圧倒的な分析性能が得られること。次に、顧客DNA生成という大規模のデータ処理と同等の処理が、専用のシステムの開発をすることなく通常の分析業務で行えることである。

データブリックス活用に関しては、オープンソース版の活用と比較をされることも多いが、データブリックスには3つの利点があると考えている。1つ目は(SaaSに近い)PaaSであるため、クラスターの作成や立ち上げもGUIで簡単に始めることができ、専用のノートブックUIでそのままコーディングができる点である。データ分析業務が圧倒的に容易に開始できる。2つ目は、クラウドプロバイダーとのアカウントが統合され、サービスの連携が用意されているため、クラウドサービスとの親和性が非常に高い点である。3点目は、データブリックスには、データ&AIの専門家が多数いるため、何か問題が発生したときに適切なアドバイスを求めやすい点である。

高品質、かつ大規模なデータを活用し、データを社会に価値還元すべく、分析業務に引き続き取り組んでいく。



工程	処理内容	実行時間(現行)	実行時間(Db)
①ETL	DBやData Lakeからデータを抽出する	28h	**1.5h**
②前処理	集計や次元縮約を適用し、学習用データセット・推論用データセット作成	140h	**4.5h**
③学習	学習用データセットを使用して、モデル学習	30h	**1.5h**
④推論	推論用データセットを使用して、全モデルで推論	120h	**14.5h**
⑤後処理	スコアの加工をしたり、推論結果を共有DBにアップロード	20h	**12h**
	合計	340h	**34h**