AIC Databricks プロジェクト

西川誠人 [1]，勝又圭 [2]，好田駿成 [3]
[1] 慶應義塾大学大学院理工学研究科
[2] 慶應義塾大学理工学部情報工学科
[3] 慶應義塾大学経済学部

Abstract:
昨今，ビジネスにおいてビッグデータを活用する取り組みが広く行われている．一方で，データ活用のためには相応のデータ基盤が必要になるため，人材やノウハウの無い企業においてはデータ活用が遅れている．この現状に対し，企業にデータ活用プラットフォームを提供するサービスが生まれている．我々は，このようなサービスの 1 つである,Databricks について調査し,AIC の講習会を行うことを目的に活動している．本ポスターではこれまで我々が行ってきた調査内容についての発表を行う．
Keywords: Databricks, Machine Learning, Datawarehouse, AI, Open Source

1. 研究背景・目的

2022 年度の AIC 新設のプロジェクトとして，新しい AI/Data Science Platform である Databricks の調査・評価プロジェクトが 2022 年 6 月からスタートした。Databricks とはデータの収集・蓄積から分析・機械学習モデル開発並びに運用を一気通貫・効率的かつ安価に行うことができるオープンソースの統合プラットフォーム [1] である。機械学習・AI が注目される昨今だが、情報インフラ構築の難易度やノウハウの欠如からデータ活用ができていない企業も多く、機械学習エンジニアの雇用・分析環境の構築の観点から機械学習の大規模な活用は大企業にとどまることがほとんどである。今回の調査対象である Databricks を提供する Datbricks 社はデータと AI の民主化を掲げ、機械学習等のデータ活用を大企業以外にも活用できるようにノウハウを持たない企業におけるデータ活用を推進している。Databricks は Gartner の 2022 年「クラウドデータベース管理システム（CDBMS）部門のマジック・クアドラント」において、2 年連続でリーダーの 1 社 [2] として位置付け今後の De Fact Standard 化が期待されている。そんな Databricks について学生視点から分析を行い、慶應義塾大学内での活用可能性や Databricks を学ぶことの意義を評価することが本プロジェクトの目的である。

2. 方法

本プロジェクト実施にあたり Databricks の提供手法の一つである Databricks On AWS を使用した。他のクラウドプラットフォームとして Google Cloud Platform, Microsoft Azure 等が挙げられるが、実際に Databricks が活用されている企業の中で一番使われていること、Databricks 社が 2022 年 8 月に実施した Datbricks Hands On にて使用されていたことなどから AWS を選択した。

調査方法は主に以下の手順によって実施した。

1. Databricks 社の提供する Hands On への参加
2. Databricks 社が提供する学習コンテンツ Databricks Learning を用いた学習の実施
3. オープンデータを用いた自由な Databricks のデータ分析の実施
4. 他の類似サービスとの比較・優位性の調査

以上の手順により Databricks の学習コスト、他のプラットフォームを使用した際と比較したデータ分析のしやすさを評価した。

3. Databricks について

3.1 サービスアーキテクチャ

3.1.1 レイクハウスとは

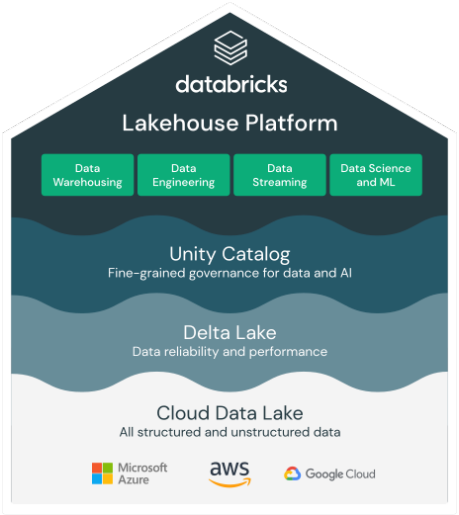

Fig. 1 レイクハウスプラットフォーム概略図

Databricks では新たなデータ活用プラットフォームの形としてレイクハウスを提案している。レイクハウスとはデータウェアハウスと多様なデータを単一のシステムで活用するためのデータレイクそれぞれの利点を組み合わせたアーキテクチャである。一般に古くから用いられるデータウェアハウスは近年の機械学習の発達により活用が促進されている画像や音声などの非構造化データの格納には適しておらず、データレイクはトランザクション・データ品質の保証等に適していないという課題があった。これらの課題を解決するのがレイクハウスであり、データウェアハウスと同様のデータ構造とデータ管理能力を搭載している他、データレイクのメリットである低コストなストレージへの直接アクセスを実現している。構造化データは勿論、非構造化データ等のさまざまなデータタイプが格納できるほか、トランザクションや BI、様々な言語が同一のノートブックで使用でき

るなど他のプラットフォームと比較した優位性があるといえる。Databricks におけるレイクハウスプラットフォームは Fig1 のように構成されており、Databricks でレイクハウスは Delta Lake と呼ばれる OSS で実現されている。

3.1.2 インフラ

Databricks で用いられるインフラ構築は一括で各クラウドプラットフォームをもとに行われている。Databrick では Microsoft Azure、Amazon Web Servbices、Google Cloud Platform の 3 種のクラウドに対応しており、それぞれのインフラ管理についても Databricks 専用の GUI によって管理可能である。ノウハウが少ない中小企業での活用を目的としていることもあり、Databricks に必要な機能に絞った構成となっており、各クラウドから自力で環境構築するより容易であり学習コストも低い。

3.2 主要機能

Databricks では主に Data Science & Engineering、Databricks SQL、Databricks Machine Learning の 3 つの機能から構成されている。

3.2.1 Data Science & Engineering

Data Science & Engineering はデータの取得・加工を目的とした Apache Spark を中心として基づくデータ分析プラットフォームである。HTML 等を用いた機能実装による非エンジニア人材のデータ利用の促進が可能など、データエンジニア・データサイエンティスト等社内のさまざまな人材間のデータコラボレーションが可能なプラットフォームである。使用感としては Jupyter Notebook や Google Colabratory 等と似ており、Databricks を初めて使用するユーザーであっても親和性の高いデザインであることが魅力である。

Fig. 2 Notebook 使用画面

ETL(Extract/Transform/Load) のためのインフラとしてさまざまな OSS から構成されており、特に Spark Core API が導入されていることで R、SQL、Python、Scala、Java を宣言一つで切り替えられるのは他のプラットフォームにはない魅力と言える。各ファイルは Notebook と呼ばれ、この Notebook 単位でジョブを定義でき、定期実行やパイプラインの作成が GUI からも容易に設定が可能で定期的に保存されるデータの継続的な加工に適している。また、各ノートブックの運用環境として Apache Spark クラスターが用いられ、GPU・CPU 性能を容易にスケールアップ/スケールダウンでき、その設定も共有できる点も魅力的だ。

3.2.2 Databricks SQL

Databricks SQL では、SQL 開発者が ETL、分析、ダッシュボードの作成を行うための一連のツールが用意されているコンポーネントである。BI ツールと直結したデータウェアハウスでのダッシュボードが作成可能のため、ダッシュボード反映までのスピードが早い点が魅力。SQL の実行結果をワンクリックで可視化ができ、直感的なダッシュボード構築など非エンジニアなデータアナリストへの配慮が多くなされている。Data Science & Engineering のクラスターと同様に SQL ウェアハウスと呼ばれる実行環境が用意されており、クラスター同様に必要度に応じてスケールアップ/スケールダウンを容易に実現可能である。また、BI ツールの必要に応じて Databricks では Looker をは始めとしてさまざまなサードパーティー製のツールとの連携ができ、可視化のためのダッシュボード作成にはさまざまな手段が用意されている。

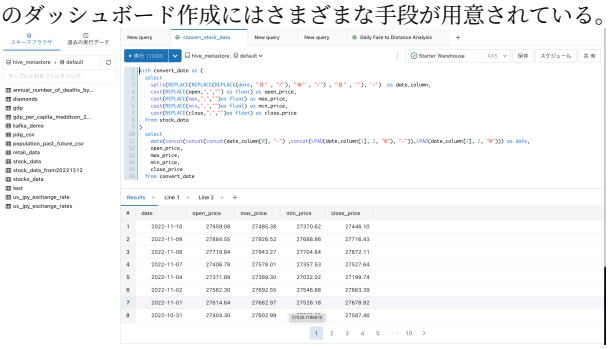

Fig. 3 Databricks SQL Query 使用画面

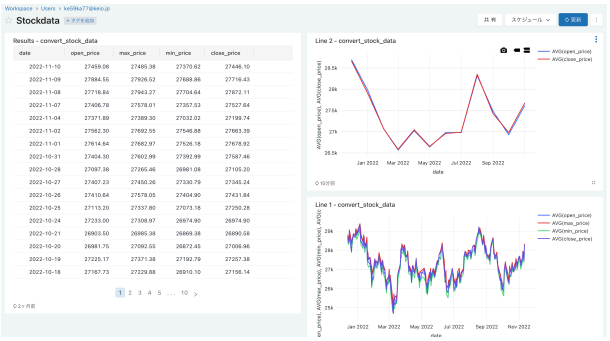

Fig. 4 Databricks SQL Dashboard 使用画面

3.2.3 Databricks Machine Learning

Databricks Machine Learning とは機械学習のモデル作成においてそのトレーニングの過程の追跡や管理を行うための機械学習プラットフォームである。主に先述の Data Science & Engineering の Notebook を用いてデータ加工、学習を行うおこなう。機械学習モデルの作成のため、Databricks Macine Learning ではモデルの自動作成のための Auto ML、ハイパーパラメーターチューニングを行うためのモデルのトレーニング状況を追跡を行う ML Flow を用いた実験 (Experiment) 機能、モデルの共有・管理・提供のためのモデルレジストリがこのコンポーネントでは提供されており、自

前の環境では環境構築の難易度が高い機械学習向けインフラを容易に使用できるエンドツーエンドのプラットフォームである。

			Metrics	
	Run Name	Created	Duration	accuracy
	capricious-hen-793	1 month ago	6.5min	0.446
	skittish-mule-893	1 month ago	5.5min	0.457
	classy-calf-715	1 month ago	5.6min	0.404
	resilient-pug-521	1 month ago	5.7min	0.261
	unequaled-sloth-565	1 month ago	8.8s	-
	kindly-stork-742	1 month ago	2.5min	0.265

Fig. 5 Databricks Machine Learning モデルレジストリ

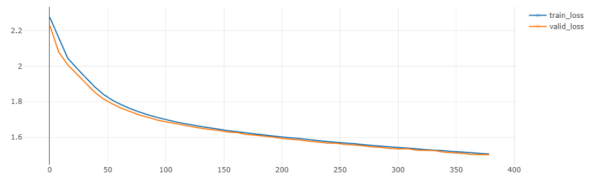

Fig. 6 Databricks Machine Learning 実験 (Experiment)

4. 調査結果

4.1 Databricks の強み

前節までの調査をもとに Databricks の強みをまとめる。まず第一にデータ活用に必要な機能が一通り揃っている環境をブラウザ上で容易に使用できる点が挙げられる。Databricks を用いれば先述の諸機能を利用してデータ収集から ML モデル運用までを一気通貫に同一プラットフォーム上で行うことが可能である。また、GUI で容易に操作が可能な UI 設計がされており、機能間での連携に詳しい知識がいらない点も魅力的と言える。

第二にコスト面である。Databricks は OSS のアプリケーションであるため、個々の機能がマネージドサービスとして提供されており、分析基盤の維持管理が不要である。従って、大きく事業者にのしかかるメンテナンスコストの削減が見込める。また、クラウドインフラのリソース (コスト) 管理が容易であり、オンプレミスで環境を用意する場合と比較して事業の成長度合いやプロジェクトの進捗により規模のスケールアップ/スケールダウンが容易に変更できる k とおは大きな魅力だ。第三に非エンジニアでも使える UI 設計がある。Notebook、SQL Query、Dashboard に至るまで非エンジニアが利用することを想定した UI 設計がされており、エンジニア側でも機能拡張の伸び代も大きいなどチームでの利用にとても親和性が高いプラットフォームと言える。煩雑なクラスターの設定など低レイヤの技術が隠蔽されていることで初心者や非エンジニアにとっても利用しやすく、データ分析に対するネガティブな印象を減らすことができるほか、直感的に使用可能な UI により学習コストも削減できる。以上のような理由から組織でのビジネスには非常に強力なツールであると言える。

4.2 大学内での使用可能性評価

前述の内容を考えると学生にとって Databricks がオーバースペックであることは否めないことは確かだ。一方で本プロジェクトメンバーではアカデミアで Databricks は無価値とは考えていない。慶應義塾大学におけるユースケースを考察すると学生の個人利用と授業や研究活動の主に二分されると考える。まず、学生の Databricks の個人利用だが、純粋にデータ解析のみを行いたい学生には有用なツールと言える。今後 De Facto Standard になりうるツールを学ぶことは社会に出てから役立つといえ、学生向け無償プランも提供されている点からも導入する価値は高い。Databricks には先述の通り多様な OSS によりデータ分析における様々な機能が格納されている。Databricks をデータ分析のプラットフォームとして利用するもよし、参考に OSS への知見を深めるのもよし、各種技術の学習きっかけにもなり得るツールである。授業や研究活動において「非情報系だがデータ解析を行う」ような研究や授業では有用であると言える。もちろん情報系の研究室でデータ解析自体を研究対象としている研究室であればオンプレミスの開発環境を所持していることが多く必要性は薄い。一方で機械学習を用いて別の分野の研究を行う研究室やプロジェクト、授業において Databricks を活用すれば環境構築・維持にかかる負荷の低減が実現でき、より本質的な議論へ時間を割くことができる。その時間の削減により新たな発展が生まれることを期待する。また、Databricks を用いることで最先端のデータ解析ツールを容易に用いることができ、個人の開発環境では用いることのできない多様な機能によりより精度が高く、社会的価値の高い研究が多発生しうる機会を提供できるプラットフォームである。

5. 今後の展開

本プロジェクトを通じて Databricks の慶應義塾大学における利用可能性が高いことが評価できた。様々な職種、特に非エンジニアでも使いやすい Databricks というプラットフォームを利用して学習を進めることで、同様の内容を環境構築を含めて行なった場合と比較してデータ分析に挑戦することへの抵抗がなくなる可能性が期待できる。また、すでにデータ分析等を既存のツールを使用しているエンジニアの学生に対しても手軽にデータ活用する方法を紹介し、将来の De Facto Standard となり得る Databricks の学習を行うことは将来役立つ経験となることは確かだろう。そして、OSS ベースで構築される ML インフラの全体像を掴むことで、データ分析基盤の全体像を知ることができ、現代社会で盛んに利用されている OSS という方式によるアプリケーション開発・活用の実態を知り、背後で稼動する OSS の挙動を考えながらシステムを触るきっかけとしたい。以上のような理由から Databricks を用いたハンズオンを中心として、Git, SQL 等の関連技能の講習を含めた講習会をの実施を 2023 年春に企画している。

6. 結論

本プロジェクトにより Databricks 社により提供されている Databricks というデータ活用プラットフォームの有用性を評価することができた。ビジネス文脈で使用される機能を多く備えたツールであるという側面は理解した上で、機械学習を使用することを目的として研究や学習用途で学生が使う有用性は大いに認められるだろう。AIC Databricks プロジェクトとして Databricks の慶應義塾大学における普及・De Falcto Standard 化を促進するため、Databricks に関する授業提供を実施する。学生に機械学習が容易に活用できるという選択肢を提示することにより今後の慶應義塾大学での研究・学習における機械学習の活用が促進されることで未来の研究にポジティブに寄与してゆきたい。

参考文献

[1]
[2]
[3] クイックスタートガイド
[4] https://learn.microsoft.com/ja-jp/azure/databricks/scenarios/what-is-azure-databricks-ws