AIC Databricks プロジェクト

西川誠人 [1]，勝又圭 [2]，好田駿成 [3]，石川繁樹 [4]，小林真里 [5]
[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
[4] 慶應義塾高度 AI コンソーシアム特任教授,[5] 慶應義塾高度 AI コンソーシアム特任準教授

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．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 Databricks On AWS を使用した．

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. Databricks を用いたオープンデータ分析
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との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 レイクハウスとは

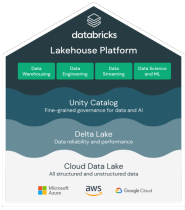

Fig. 1 レイクハウスプラットフォーム概略図

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．データウェアハウスは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．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，データレイクの低コストなストレージへの直接アクセスという特性をデータレイクは実現している．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場合大きな優位性がある．Databricks では新たなデータ活用プラットフォーム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．データウェアハウスは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．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，データレイクの低コストなストレージへの直接アクセスという特性をデータレイクは実現している．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場合大きな優位性がある．

3.2 主要機能

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 に基づくデータ分析プラットフォーム．非エンジニア人材のデータ利用の促進が可能など, 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．Jupyter Notebook 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的.

Fig. 2 Notebook 使用画面

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．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．各ノートブックは Apache Spark クラスターにより運用され,GPU・CPU 性能のスケールアップ/スケールダウンが容易にでき, その設定も共有できる.

3.2.2 Databricks SQL

Databricks SQL は,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．直感的なダッシュボード構築は非エンジニア人材への配慮が多くなされている．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.

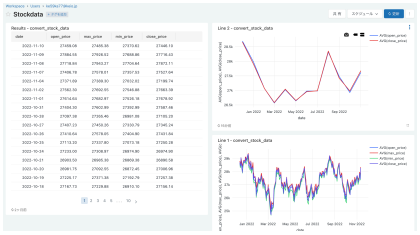

Fig. 3 Databricks SQL Dashboard 使用画面

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.

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.

1. データ活用に不可欠な環境をブラウザ上で使用できる
   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で行うことが可能
2. コスト効率が一般に良い
   メンテナンス・初期導入コスト等が低く維持しやすい。
3. クラウドインフラのリソース (コスト) 管理が容易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, 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く、直感的に使用可能な UI により学習コストも削減できる.

4.2 大学内での使用可能性評価と今後の展開

前述の内容より学生にとって Databricks がオーバースペックであることは否めない．一方で, 本プロジェクトではアカデミアで Databricks は無価値とは考えておらず、以下のような理由から有用であると考える。第一に, 学生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. 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おり、各種技術の学習きっかけにもなり得るツールである．第二に, 授業や研究活動においてデータ解析を手段として用いるような研究や授業で有用である. 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では Databricks の活用で環境構築・維持にかかる負荷の低減が実現できる. 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．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.

5. 結論

本プロジェクトにより Databricks 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] Data Lakehouse Architecture and AI Company. Databricks. (n.d.). Retrieved February 14, 2023, from https://www.databricks.com/.

[2] Databricks、ガートナー クラウドデータベース管理システムのマジック・クアドラントのリーダーの 1 社に. Databricks. (n.d.). Retrieved February 15, 2023, from https://www.databricks.com/jp/resources/analyst-paper/databricks-named-leader-by-gartner

[3] Mssaperla. (n.d.). Azure Databricks のドキュメント. Microsoft Learn. Retrieved February 15, 2023, from https://learn.microsoft.com/ja-jp/azure/databricks/