# AIC カンファレンス ポスター発表 要旨

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

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。レイクハウスとは一般的に企業などで構造化データを用いる際に活用される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 Fig.1 のように構成されている。Databricks でレイクハウスは Delta Lake と呼ばれる OSS で実現されており、Databricks の各種サービス・機能は Delta Lake をベースとして提供されている。

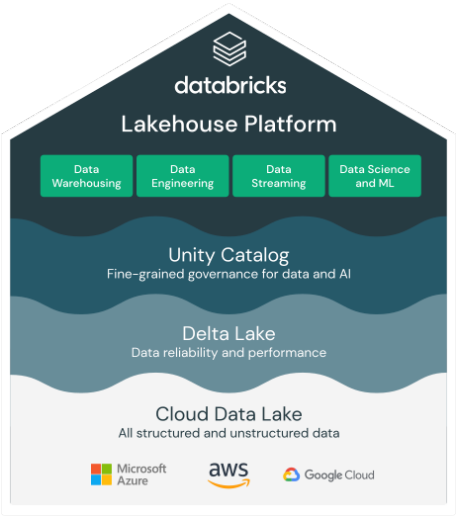


Fig. 1 レイクハウスプラットフォーム概略図

#### 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 ML 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

#### 3.2.2 Databricks SQL

#### 3.2.3 Databricks ML

## 4. 調査結果

### 4.1 Databricks の強み

### 4.2 他プラットフォームとの比較

### 4.3 大学内での使用可能性評価

## 5. 今後の展開

## 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] S. Yagami, Science, **393**, 113–117 (1998).
[2] S. Yagami *et al.*, Proc. IEEE **52**, 284–290 (2013).
[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`
[5]