AIC Databricks

調査・検討内容の報告

西川・勝又・好田

アジェンダ

- Databricks の概要
- 調査結果
  - Databricks の強み
  - 慶應におけるユースケース
  - Databricks に関する講座を AIC で提供することの意義
  - 授業内容案

Databricks の概要

Databricks とは

- クラウド型統合データ分析基盤
  - ノウハウを持たない企業向け
  - ビジネスにおけるデータ分析に必要な機能をオールインワンで提供
- OSS ベースで構築されている
  - Databricks 社も OSS に貢献している

→ Gartner Magic Quadrant における高い評価

databricks

企業がデータ活用において抱える課題

- 情報インフラ構築の難易度
  - 計算機環境
  - データウェアハウス (DWH)
  - データ処理 (ETL) パイプライン
- ノウハウの欠如
  - 機械学習の運用 (MLOps)
  - データの維持管理

これらの課題を乗り越えられる企業は少ない

企業がデータ活用において抱える課題

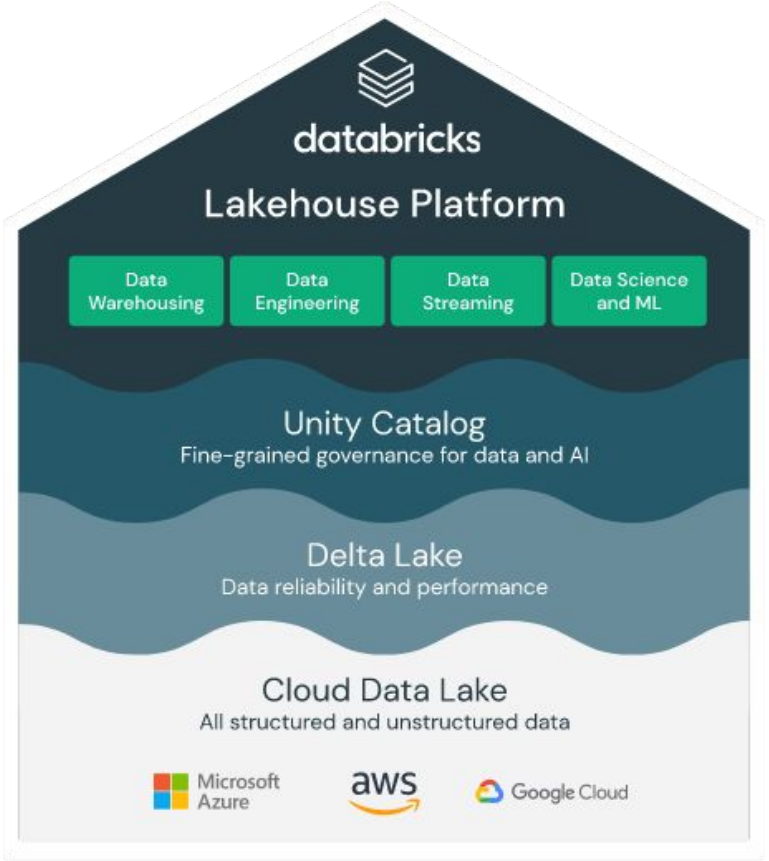
- 情報インフラ構築の難易度
  - 計算機環境
  - データウェアハウス (DWH)
  - データ処理 (ETL) パイプライン
- ノウハウの欠如
  - 機械学習の運用 (MLOps)
  - データの維持管理

これらの課題を乗り越えられる企業は少ない

→ Databricks は**課題を解決するためのプラットフォーム**

レイクハウスプラットフォーム

- レイクハウス
  - DWH + データレイク
  - DeltaLake として OSS 化されている
- 主要機能
  - **Data Engineering** : ETL 環境
  - **Databricks SQL** : データウェアハウス, BI
  - **Databricks ML** : MLOps
- インフラ
  - マルチクラウド
  - Databricks 上でインフラ管理も可能

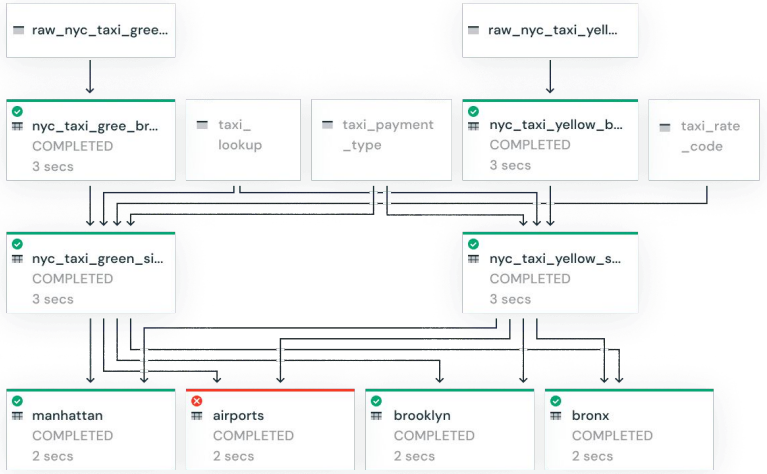


https://www.databricks.com/jp/product/data-lakehouse

機能紹介: Data Engineering

ETL (Extract / Transform / Load) のためのインフラ

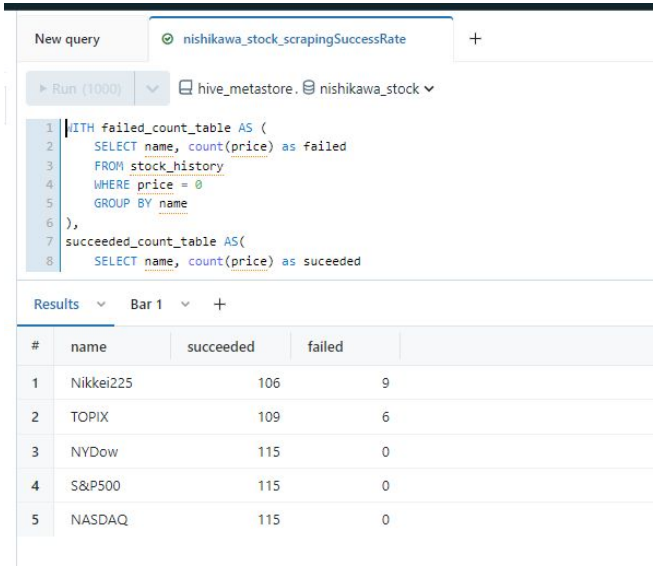
- Jupyter Notebook を使用したプログラミング
  - 複数の言語 (Python, SQL, Spark, R) を使用可能
- ジョブオーケストレーション
  - Notebook 単位でジョブを定義
  - 定期実行やパイプラインの作成が可能

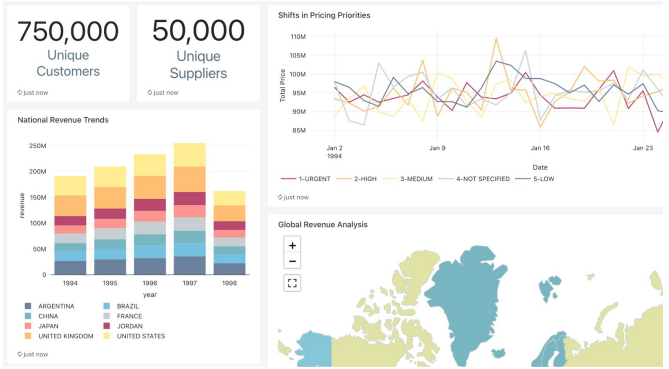


機能紹介: Databricks SQL

BI ツールと直結したデータウェアハウス

- スムーズな BI
  - SQL の実行結果をワンクリックで可視化
- 直感的なダッシュボード構築
  - 非エンジニアなデータアナリストへの配慮



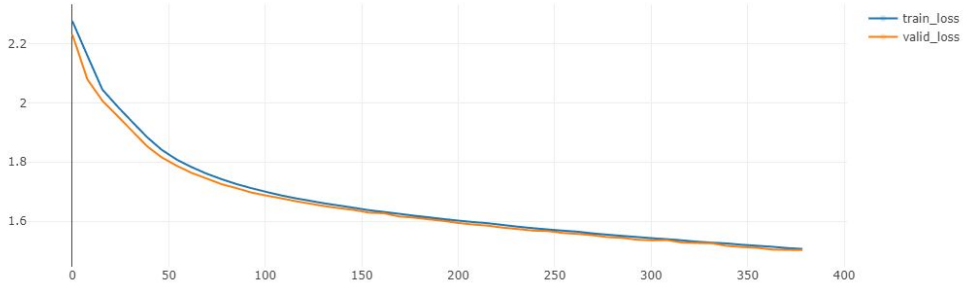


機能紹介: Databricks ML

機械学習向けインフラ

- ML 向け Notebook
  - 環境構築が不要
  - データ前処理・モデル学習のパイプライン化
- データのバージョン管理
  - モデル構築の再現性を保証
- マネージド ML Flow
  - 実験管理が容易

				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



Databricks の調査結果

Databricks の強み

- データ活用に必要な機能が一通り揃っている
  - ETL 環境，DWH，BI，MLモデル運用
- SaaS である
  - 分析基盤の維持管理が不要
    - 個々の機能がマネージドサービスとして提供されている
    - 機能間での連携に詳しい知識がいらない
  - クラウドインフラのリソース (コスト) 管理が容易
- 非エンジニアへの配慮がある
  - 直感的に使用可能な UI
  - 低レイヤの技術が隠蔽されている

→ 組織での**ビジネスには非常に強力**

慶應におけるユースケース

- 学生の個人利用
  - 純粋にデータ解析のみを行いたい学生には有用
  - 今後デファクトスタンダードになりうる（？）ツールを学ぶことは社会に出てから役立つ
- 授業や研究活動
  - 「手段としてデータ解析を行う」ような研究や授業では有用
    - 環境構築・維持にかかる負荷の低減
  - 研究室という組織として活用できる可能性がある

→ **アカデミア向けプラン**も提供されている

講座を通して学生に提供できる価値

- データ分析に挑戦することへの抵抗がなくなる
  - 手軽に挑戦できる方法の紹介
- データ分析基盤の全体像を知ることができる
  - Databricks を例にデータ基盤についての一般的な講義
- 社会でも役立つ経験
  - データ分析の基本的な流れについてのハンズオン
  - デファクトスタンダードになりうるツールの使用

講座の内容案

- 概要の説明（1コマ）
  - そもそもデータ活用とは
  - Databricks の紹介
    - Databricks 以外の方法も紹介する
- Databricks を使用するに当たって必要な関連技術の入門（1コマ？）
  - Git, SQL などを扱う予定
  - ハンズオンに含めてしまう可能性もある
- ハンズオン（計3コマ）
  - Data Engineering　　: インターネットからデータをスクレイピング・ DB に格納
  - Databricks SQL　　: ↑のデータをSQLで取得してダッシュボード化
  - Machine Learning　　: DNN モデルの構築・実験管理

結論

- Databricks は強力なツールである
  - 特にビジネスではデファクトスタンダードになりうる
  - 慶應においても活用可能な場面は存在する
- Databricks に関しての授業を AIC で提供することに価値はある
  - データ分析を手軽に始める方法
  - データ分析基盤に関しての知見
  - 社会でも役立つ経験

Appendix

データ活用とは？

ビジネスプロセスで生成されるデータを収集・蓄積し，得られたデータを分析することで得られた知見をサービスに反映すること [1]

データ活用の流れ

1. データの収集
2. データの蓄積
3. データの分析
4. 得られた知見の活用

[1] https://www.soumu.go.jp/johotsusintokei/linkdata/r02_05_houkoku.pdf

ETLとは？

データウェアハウスを構築するためのプロセスの総称

- Extract　　: データソースからのデータ抽出
- Transform　: データの加工
- Load　　: データウェアハウスへの格納

Databricks の弱点

- そもそもクラウドインフラ代が高い
  - Databricks にはアカデミア向けプランがあるが，クラウドインフラにかかる費用は別
  - オンプレ Databricks などが実現可能ならば嬉しい
    - 慶應 JupyterHub のような提供が可能になる可能性
    - Databricks が OSS ベースなので出来ないことはなさそう
- 個人利用するには高機能すぎる
  - Databricks 自体は組織での使用が前提とされている
    - それでも使用したい個人は存在するかもしれない
  - 個々のコンポーネントには代替方法が多数ある
    - Google Colabratory, WandB など