

Lakehouse を活用してデータ利活用から AI 活用にシフトしよう！

～データサイエンス / 機械学習の可能性を最大化～

Shunichiro Takeshita

Databricks Japan, Solution Architect

Agenda

- デジタル マチュリティ -> 未来予測型 / Innovation
- DS/ML(AI)のフル ポテンシャルを引き出す
- 次世代データ基盤 & Mission First
- 次世代データ基盤 = Lakehouse
- Mission First(Translatorの育成 by Sol Acc.)
- Call to Action

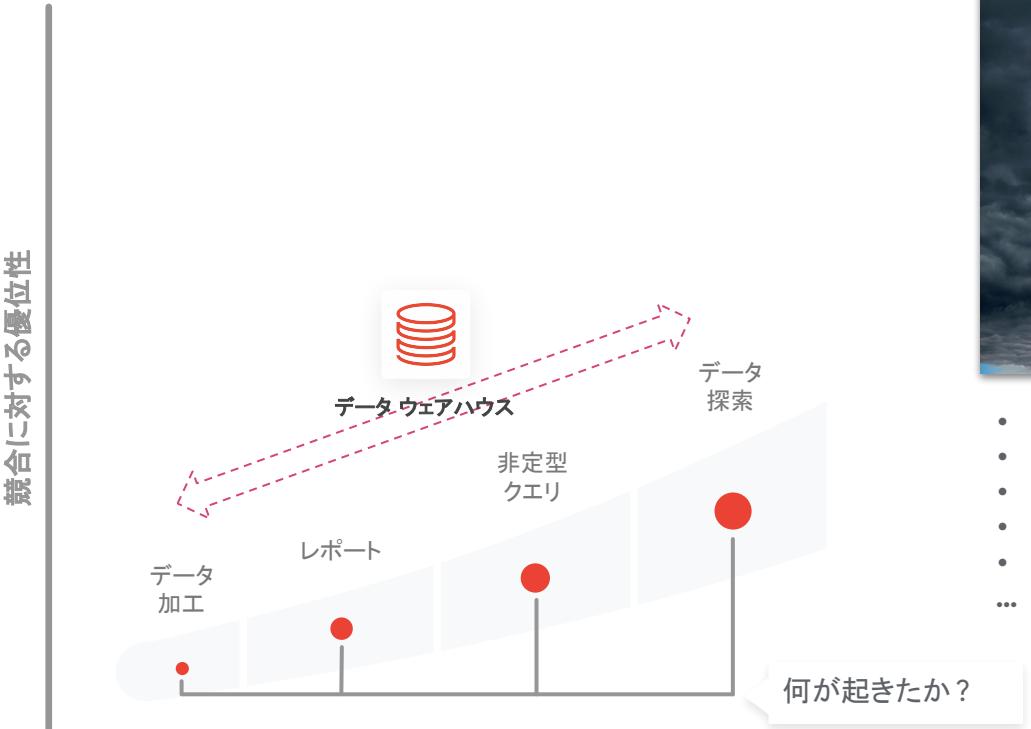


「データと AI の民主化」を推進

データ・エンジニアリング、データ・サイエンス、データ・アナリティクス
の分野において、イノベーションを加速させる SaaS 型統合分析基盤
「レイクハウス・プラットフォーム」を提供

これからからの企業が向かうべき行き先

過去の振り返り・現状の可視化

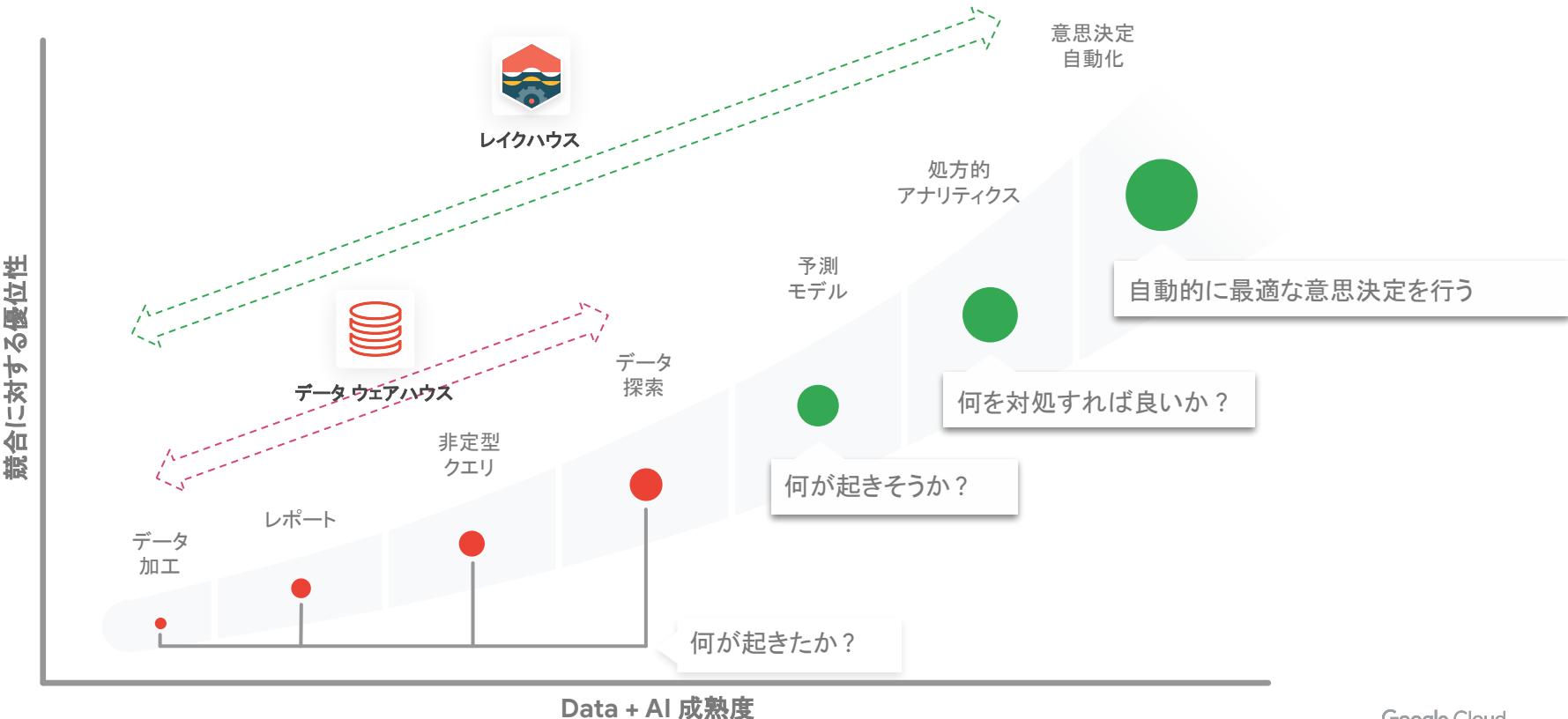


- 米中貿易摩擦(デカップリング)
- ロシアによるウクライナ侵攻
- 技術革新(AIの進化、ブロックチェーン適応拡大)
- パンデミック
- SDGs(気候変動とエネルギー政策 etc)

...

Data + AI の成熟度を高める取り組み

過去の振り返りから将来の予測へ



Databricks Lakehouse

Data + AI の民主化を実現をご支援するプラットフォーム

世界 7,000 社を超えるユースケースを実現



金融サービス / 不正検知



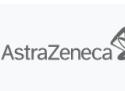
店舗需要予測



レコメンド エンジン / LTV



ゲノム解析 / 論文検索



ESG スコアリング



IOT 故障検知 / 品質管理



databricks レイクハウス・プラットフォーム

データ
ウェアハウス

データ
エンジニアリング

データ
ストリーミング

データサイエンス
機械学習

Unity Catalog

データとAIの為のきめ細やかなガバナンス

Delta Lake

高い信頼性と高い処理性能を担保

クラウド・データレイク

全ての構造化、非構造化データ、ストリーミング データ

Google Cloud

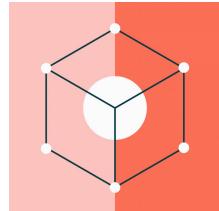
Google Cloud

データサイエンス・機械学習・AIの フル ポテンシャルを引き出すデータブリックス

シンプルな基盤

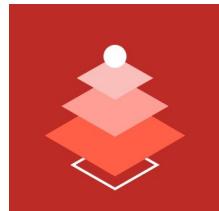


人材育成



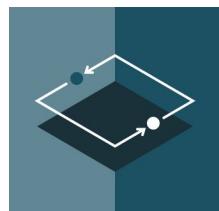
すぐに分析作業をスタート

- クラウドネイティブサービス
- 数クリックで環境準備
- ライブラリがプレインストール



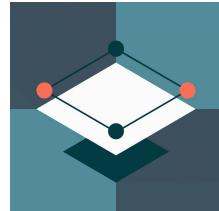
資産をしっかり管理できる

- MLflow でモデル管理
- Delta Lake で多構造データ管理
- Unity Catalog でデータ辞書化



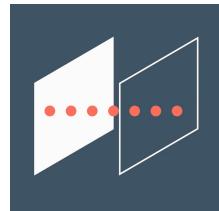
手触り感のある人材育成をご支援

- ソリューション・アクセラレータ
- 業界別の開発済みパッケージ
- 需要予測、不正検知、顧客360、IoT等



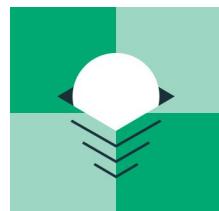
コラボレーションが簡単

- ノートブックで共同作業
- データ、モデル、グラフを共有
- アクセス権限管理が簡単



モデル資産を簡単に本番化

- ペタバイト級へのデータ対応
- リアルタイム処理への対応
- ジョブオーケストレーション



アジャイルにQuick Winをご支援

- Data + AI 施策の実現のサポート
- 人材育成、ユースケース選定支援
- レイクハウス、MLOps の Quickな実現

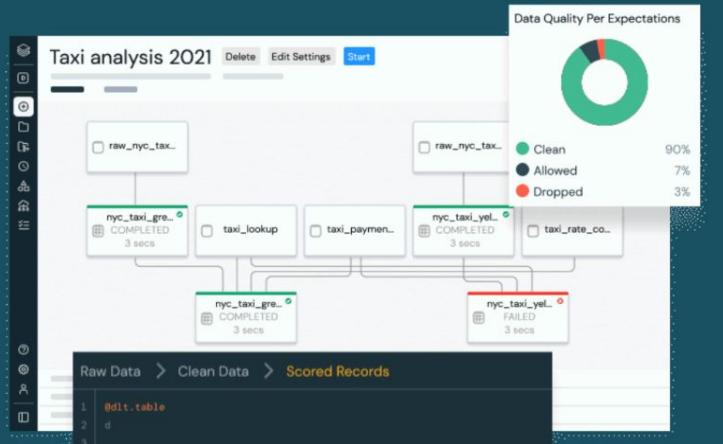
データサイエンスの仕事が捲ります！！

Google Cloud

Databricks SQL : SQL検索やダッシュボード利用を簡単に

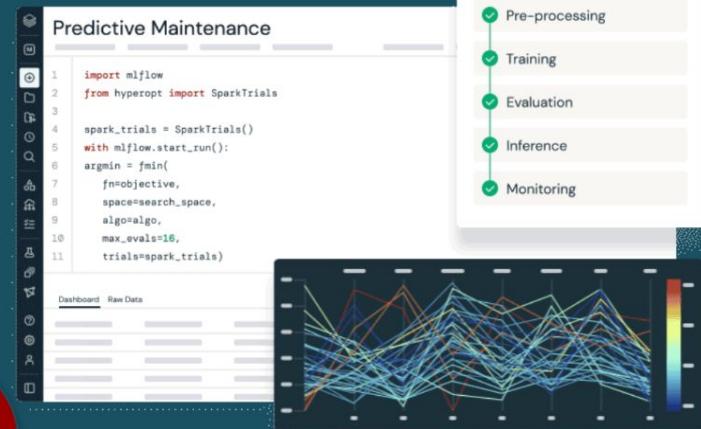


Databricks Delta Live Table : データパイプライン、データ品質管理



Data + AI
民主化を加速
させる新機能

Databricks Notebook & AutoML : SQL/PythonベースのプログラミングからAutoMLまで



Databricks Unity Catalog : データ辞書、データリネージ

The screenshot shows the "unity_catalog/loc_data" table in the "Firehose" database. It displays the schema (ProductID, Status, PricingTier, DistributionTier, AccountInfo), lineage (relationships between tables like "initial", "preprocess", and "inventory"), and privileges (User/Group, Permissions, Modified). Below the table is a "Data Sample" table with sample data rows.

AI First ではなく Mission First が必要

6月開催の Data + AI Summit にて Andrew NG さんが登壇されます

MIT Technology Review

ARTIFICIAL INTELLIGENCE

Subscribe

Andrew Ng: Forget about building an AI-first business. Start with a mission.

An AI pioneer reflects on how companies can use machine learning to transform their operations and solve critical problems.

by Karen Hao March 26, 2021



JEREMY PORTJE

<https://www.technologyreview.com/2021/03/26/1021258/ai-pioneer-andrew-ng-machine-learning-business/>

”私は顧客主導またはミッション主導になる傾向があり、テクノロジー主導になることはほとんどありません”

”業界が AI に対応しているパターンのひとつは、デジタルトランスフォーメーションが行われ、データが存在するかどうかです”

”前へ進みましょう！今日、地球上のどの企業も、ハイテクの巨人でさえも、自社データが完全にクリーンで完璧であるとは考えていない”

”私は自分でビジネスについて少し学び、ビジネスリーダーが AI について少し学ぶのを助けるように努めます。”



Databricks ソリューション・アクセラレータ

Databricks 上で AI の迅速な導入を支援する、開発済みのツール群

<https://databricks.com/solutions/accelerators>



10%+

需要予測を Databricks に移行した際の予測精度の向上率平均

- 精度の向上**: 細かい粒度(日次、店舗、SKU)レベルの予測を限られたサービス時間内でスケーリングして実行
- 大幅な処理性能向上**: サービス ウィンドウの期間中、1 日に膨大なモデル反復を実行
- スケーラブルな分析環境**: テクノロジー制限による分析の深さ幅の妥協を排除

需要予測



商品推奨



在庫管理



動体検知

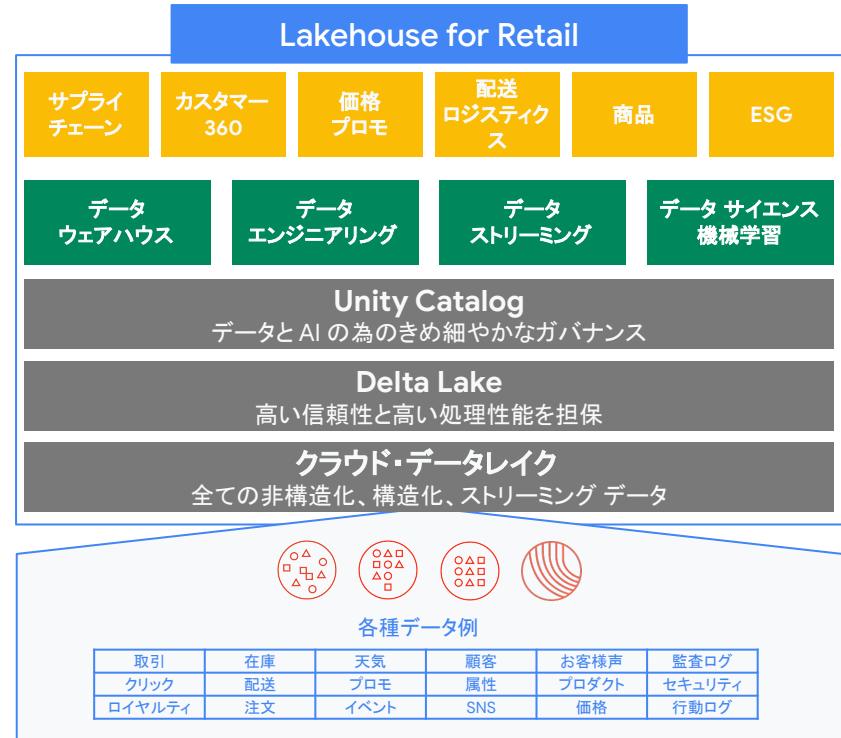


Google Cloud

Lakehouse for Retail の発表

データ + AI 領域で B2C 企業が勝つためのテクノロジー、新しいデータ共有イニシアチブ、ビジネスアクセラレーター、パートナー エコシステム、投資の加速を統合的に提供するものです。

- 業界における重要なニーズを統一プラットフォームでサポートします。
- ビジネスソリューションアクセラレータにより POC や POV を迅速に遂行します。
- オープンなデータ共有とコラボレーションにより、より多くのパートナーにデータコラボレーションを拡大し、サービスレベルを向上させます。
- 経験豊富なパートナー・エコシステムが Lakehouse for Retail に組込まれています。
- データ+AI のコミュニティを各インダストリーで実施しています。



某小売業様での Lakehouse for Retail 利用状況

データ + AI 領域において大きな成果を実現

- **1000+** Spark ETL パイプラインを本番稼働
- **20+** ユースケースを実装

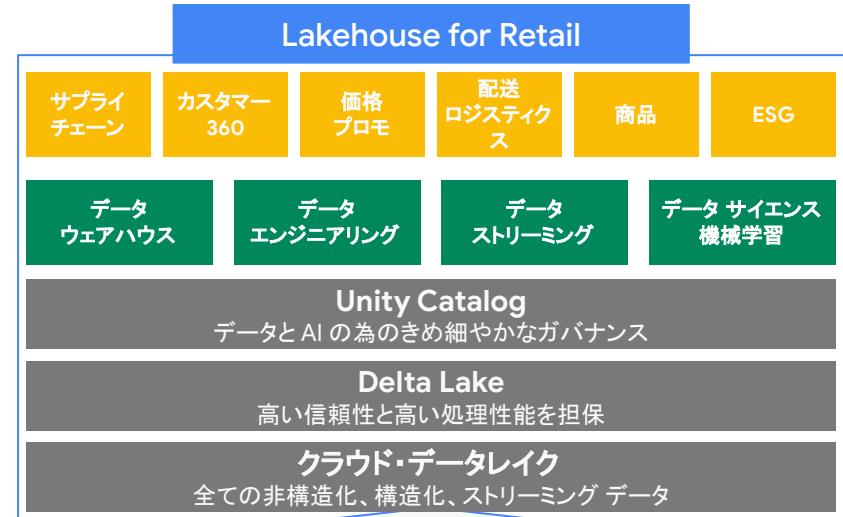
Customer 360、POS(バスケット分析)、ロイヤルティ(クーポン、プロモーション)、商品階層分析、店舗オペレーション最適化、労働生産性最適化等

- **10+** ペタバイトデータを Delta Lake で利用
- **30倍** Delta テーブルへのクエリーが高速
従前はオンプレ DWH やデータレイクを利用していた
- **1500+** Lakehouse 月間アクティブ ユーザ

当初は数名からスタートし、データ ウェアハウス機能、機械学習機能を利用する人財がどんどん増加

- **100+** アクティブなワークスペース数

プロジェクト別、部門別等でワークスペース環境を構築

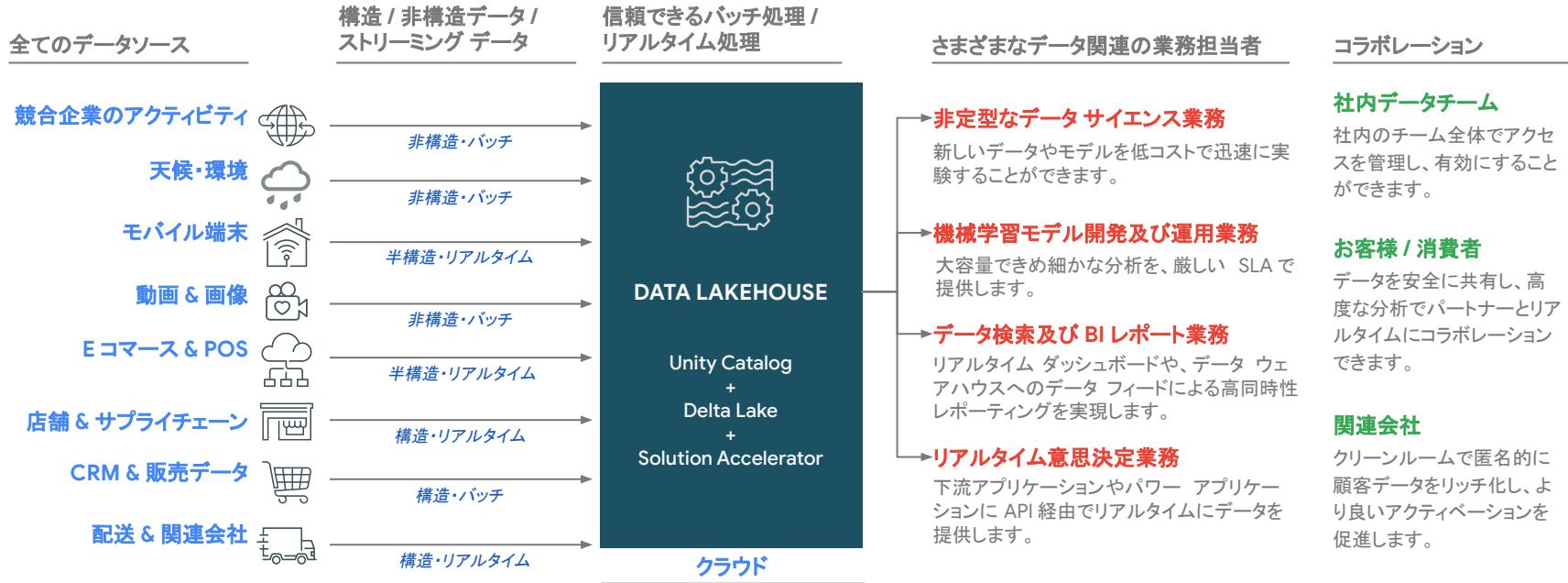


各種データ例

取引	在庫	天気	顧客	お客様声	監査ログ
クリック	配送	プロモ	属性	プロダクト	セキュリティ
ロイヤルティ	注文	イベント	SNS	価格	行動ログ

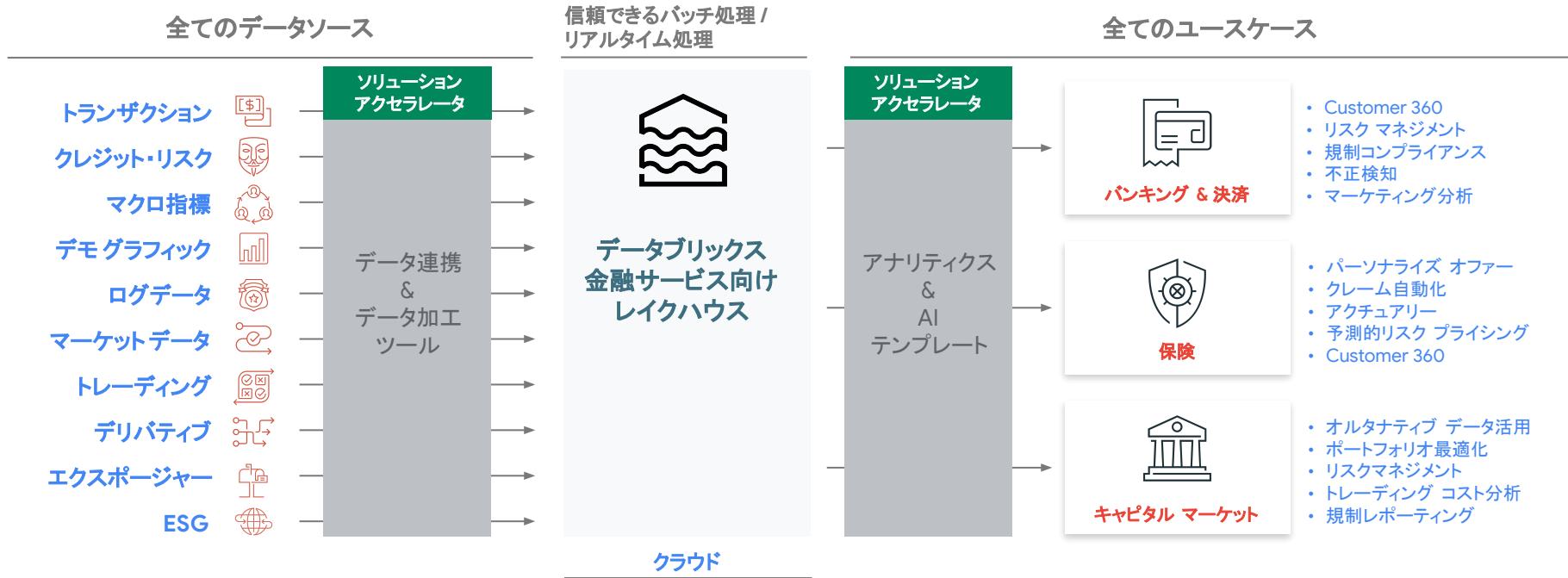
Lakehouse for Retail アーキテクチャ概要

Databricks により良いオペレーションを実現し、次世代リテール ビジネスを可能にします。



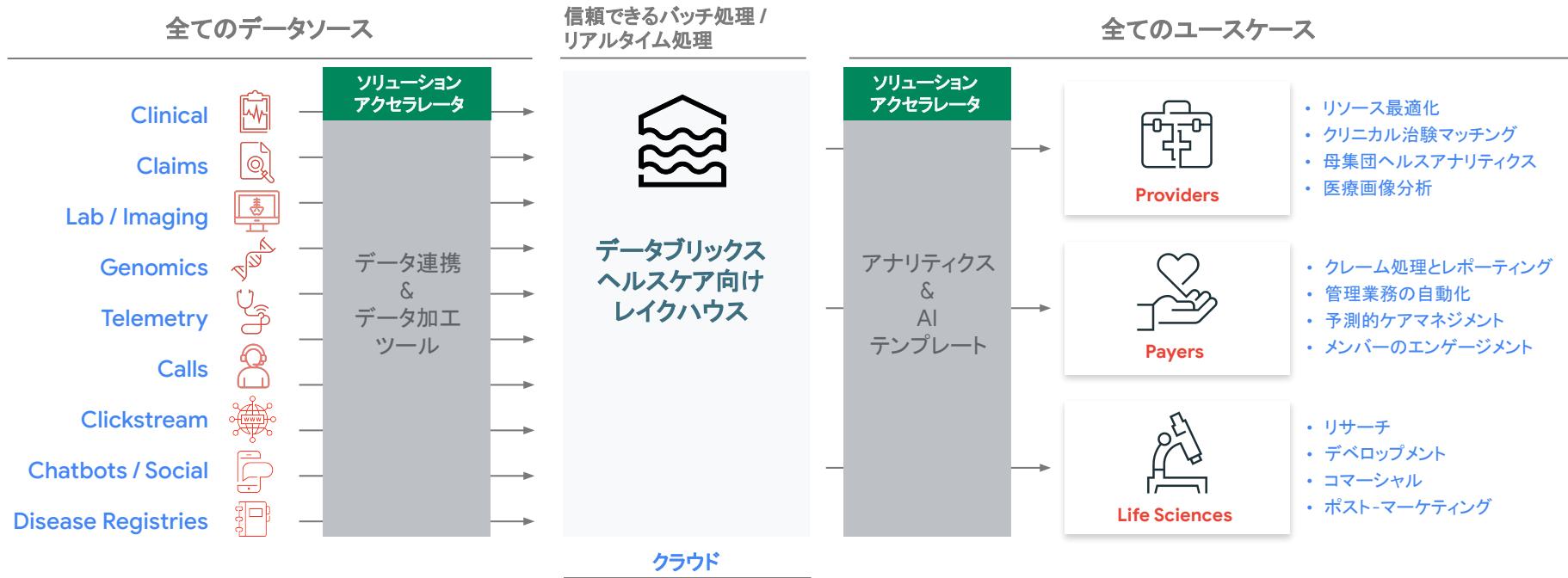
Lakehouse for Finance アーキテクチャ概要

Databricks により良いオペレーションを実現し、次世代金融ビジネスを可能にします。



Lakehouse for Health Care アーキテクチャ概要

Databricksにより良いオペレーションを実現し、次世代ヘルスケア&ライフサイエンス・ビジネスを可能にします。



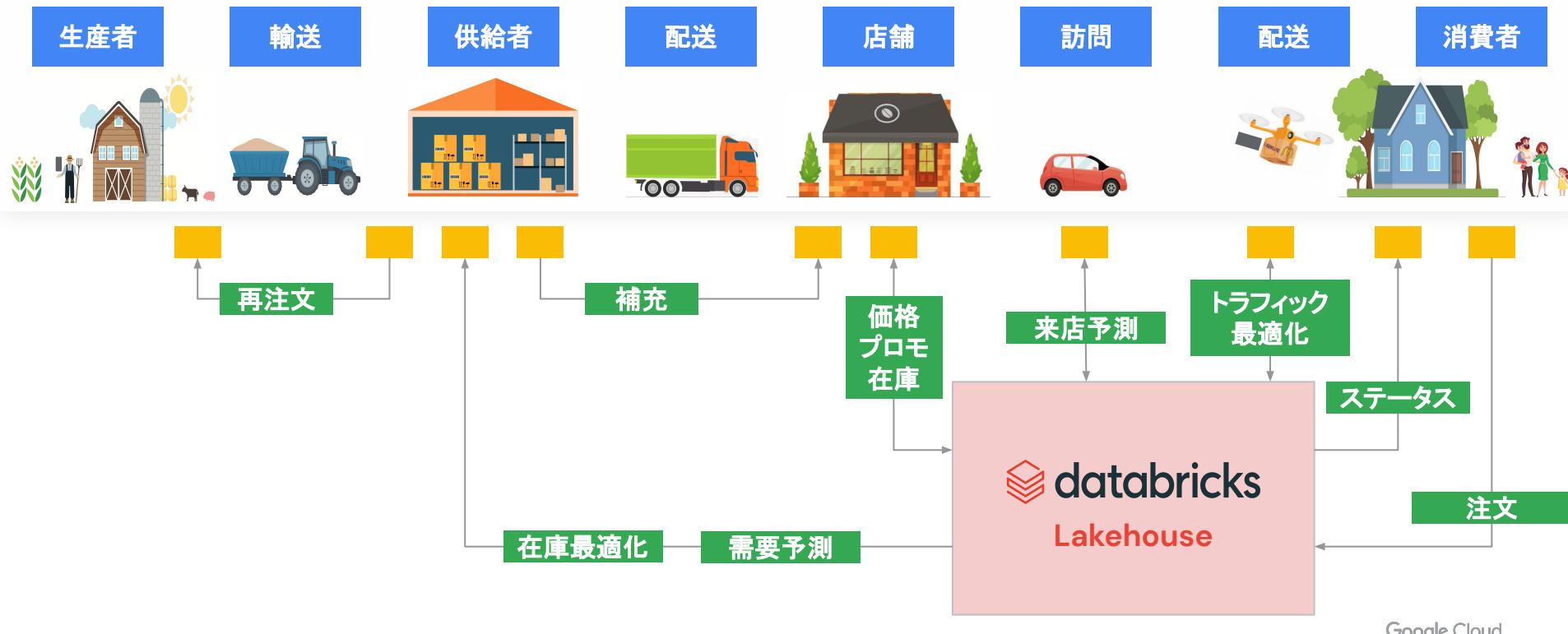


DS の業務、課題、そして解決策

1. DS の業務 + 課題
2. 機械学習 End to End フローでの解決策
3. システム インフラ観点での解決策

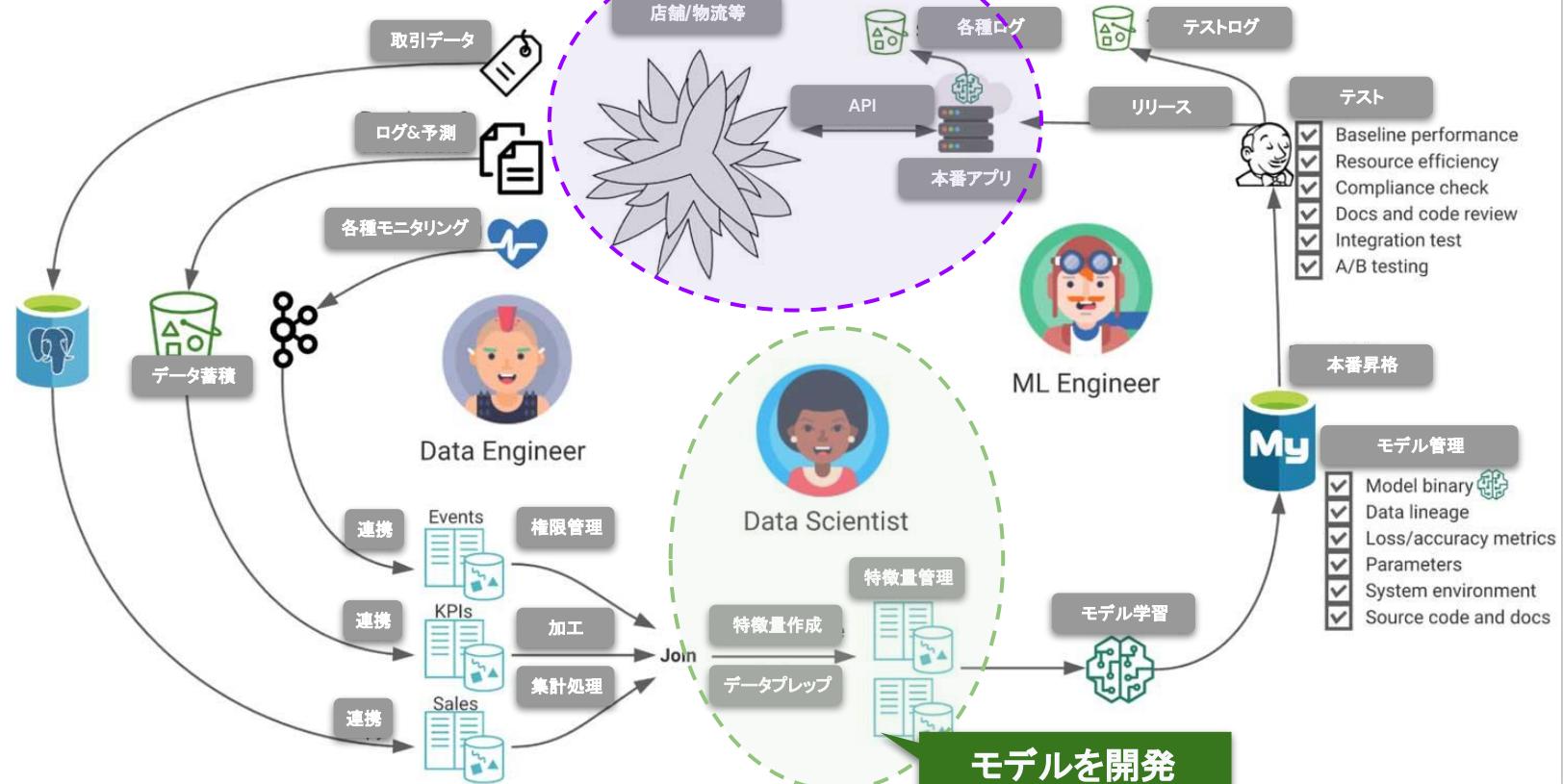
流通業のユースケース例

多くのステークホルダーの活動をリアルタイムで調整。数分の遅れが数百億の損失につながる可能性あり。



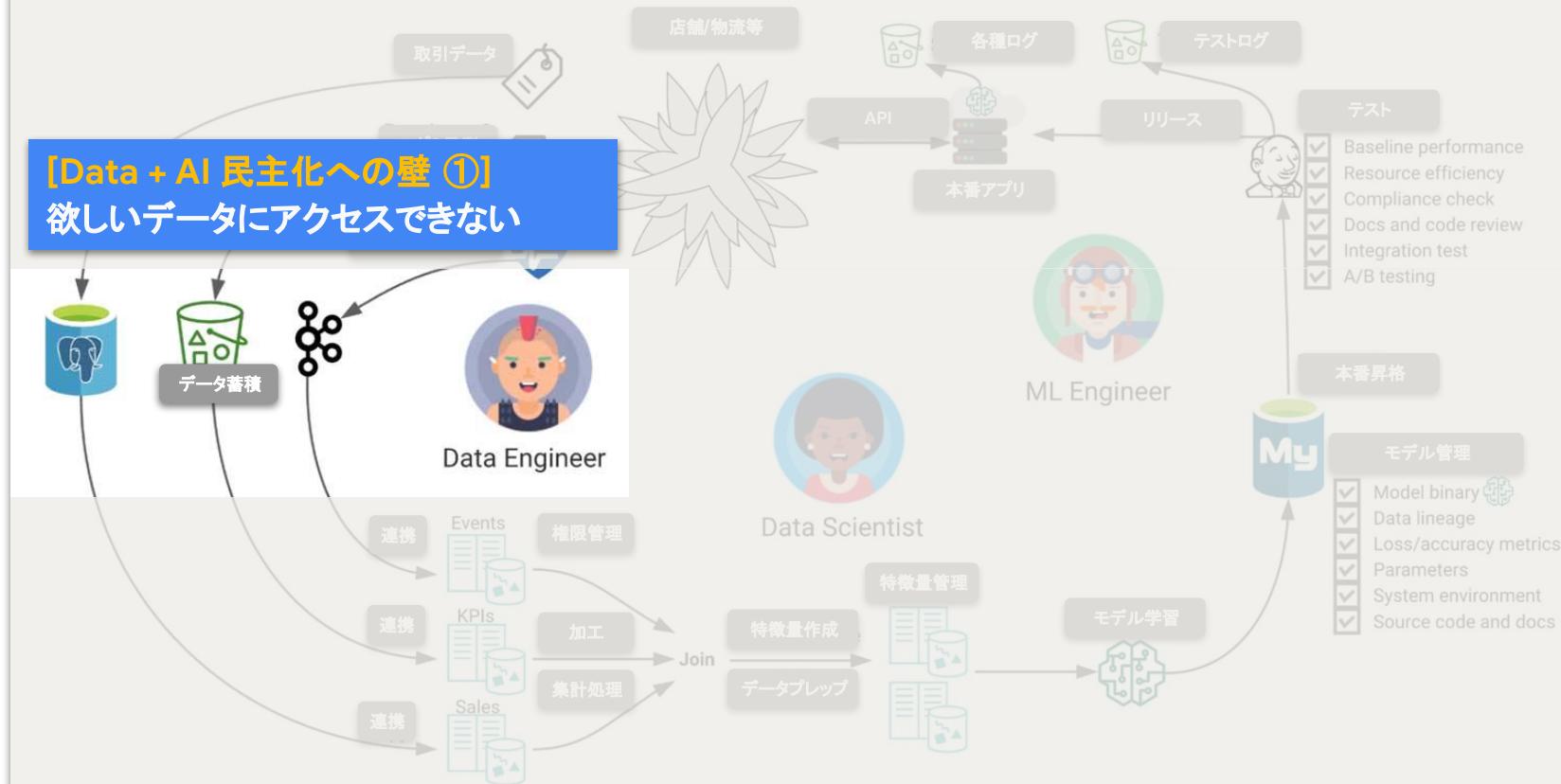
より精緻なフローで見ると…

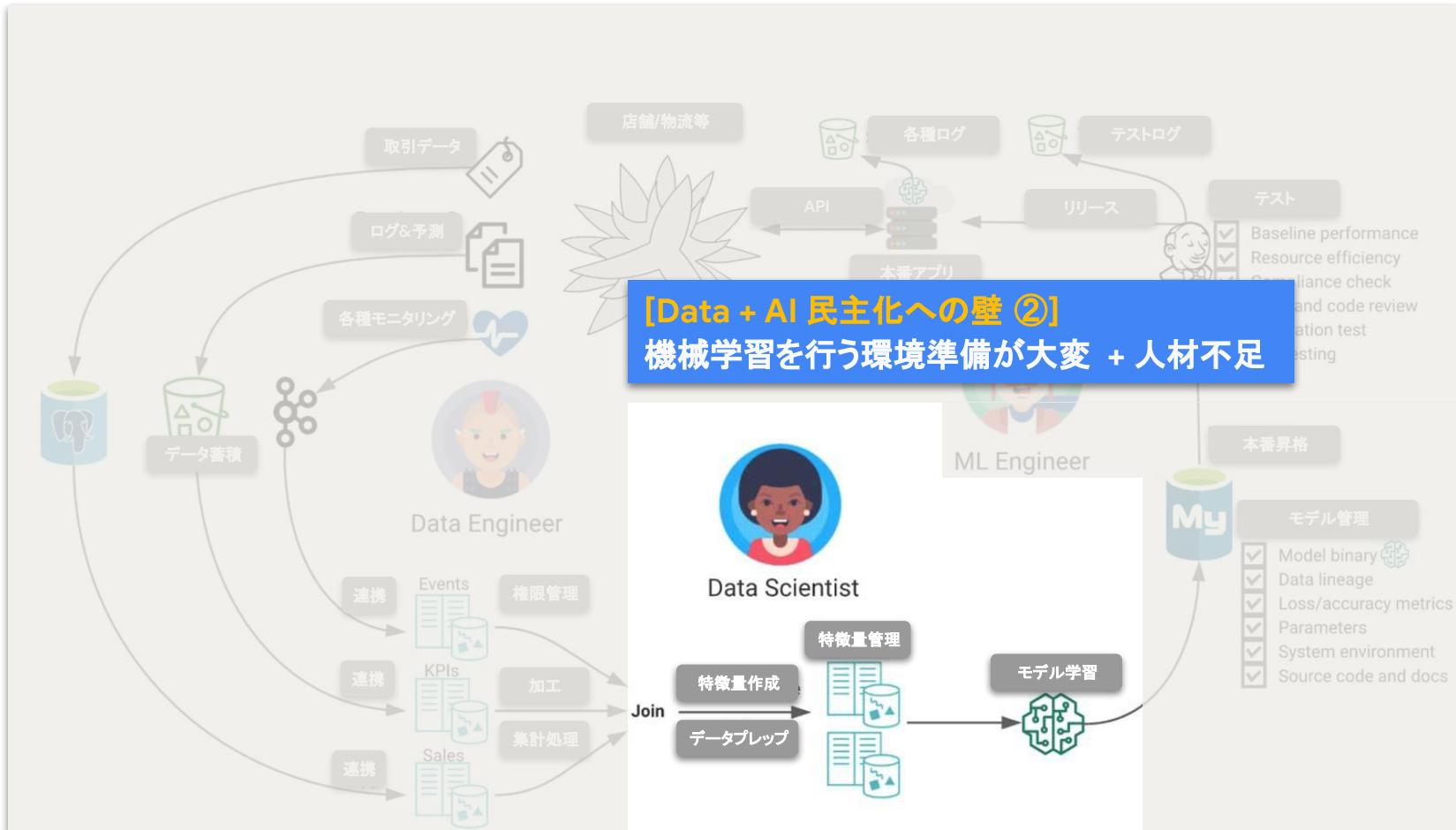
需要予測が稼働



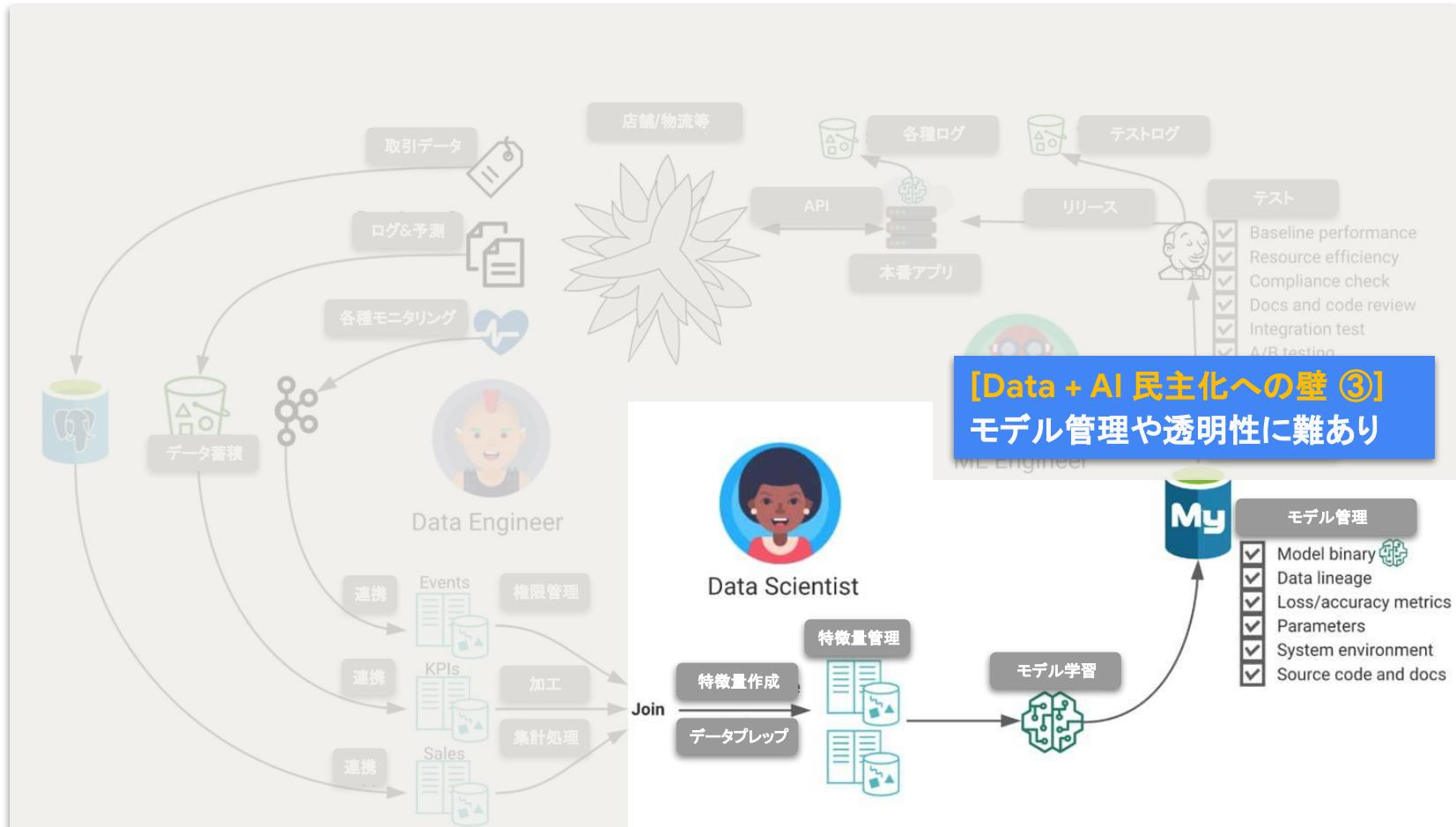
モデルを開発

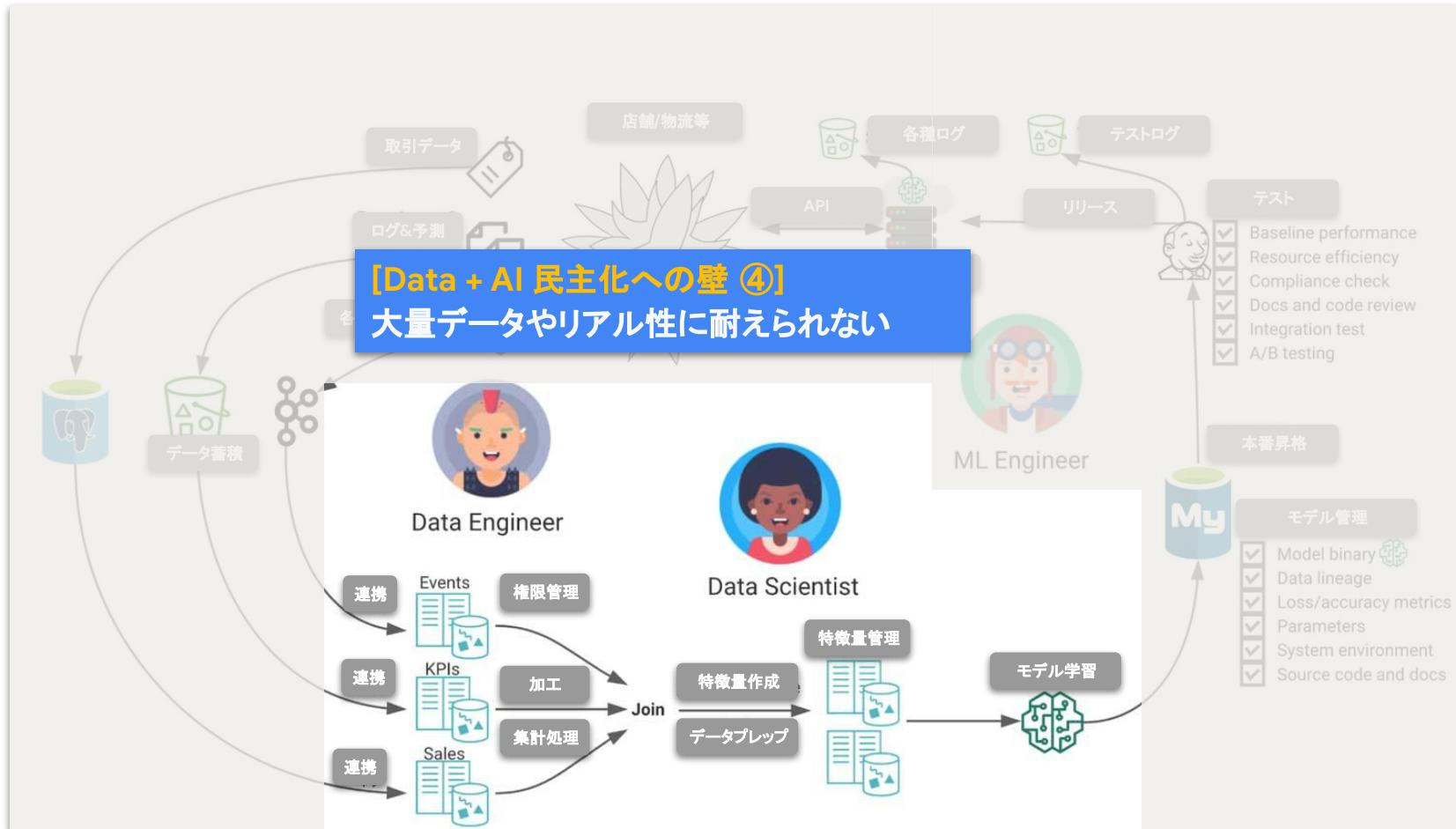
[Data + AI 民主化への壁 ①] 欲しいデータにアクセスできない





[Data + AI 民主化への壁 ②] 機械学習を行う環境準備が大変 + 人材不足







DS の業務、課題、そして解決策

1. DS の業務 + 課題
2. 機械学習 End to End フローでの解決策
3. システム インフラ観点での解決策

End to End フロー一例

これらの環境を数クリックで構築可能です





Machine Learning End to End Demo 概要

実際にRawデータから加工してモデル学習&デプロイまで構築するデモになります。以下のようなパイプラインを想定しております。

```

graph LR
    DS[Cloud Storage] --> RD[raw data (bronze)]
    RD --> FS[Feature Store]
    FS --> EXP[Experiment  
ハイパラメータ探索、モデル学習、評価]
    EXP --> MLR[MLflow Registry]
    MLR --> ME[model endpoint]
    API[RestAPI] --> ME
    EXP --> CM[Create Model]
    CM --> ML[Model Load]
    ML --> INF[推論  
batch / Streaming]
    INF --> DL[DELTA LAKE]
    DL --> DSQL[Databricks SQL  
ダッシュボード]
    ALERT[アラート] --> DSQL
  
```

Data Source

raw data (bronze)

Feature Store

best notebook

AutoML (option)

mlflow

Create Model

Model Load

推論 (batch / Streaming)

アラート

Databricks SQL (ダッシュボード)

データ連携

データ加工

特微量管理

機械学習

モデル管理

各種テスト

デプロイ

本番実行

モデル監視

レポート

モデル改善



Data Science & E...

作成

ワークスペース

リポジトリ

最近使用したアイ...

検索

データ

クラスター

ジョブ

Partner Connect

ヘルプ

設定

field-eng-west
shunichiro.takeshita@...

メニュー オプション

Partner Connect

パートナーコネクトを利用すると、Azure Databricksワークスペースが特定のパートナーソリューションに数分で接続できるようになります。パートナーにアカウントがない場合は、パートナーコネクトがトライアルアカウントの開設をサポートします。

ワークスペースをこちらに掲載されていない別のパートナーソリューションに接続することもできます。詳細を表示

すべてのカテゴリ

データ取り込み



Fivetranの自動化されたデータ統合は、スキーマやAPIの変更に適応し、信頼性の高いデータアクセスを確保し、すぐに使えるスキーマで分析を簡素化します。



RiveryはクラウドネイティブのELT+プラットフォームです。データ取り込み、変換、オーケストレーション、リバースETLを通じて、Databricksのワークフロー全体を加速します。

Partner Connect にて簡単に
Databricks にデータ連携することも
可能です！

データの作成と変換



ProphecyはApache Sparkワークフローを視覚的に構築するローコード製品です。Gitのコードで補完され、メタデータ検索、リネージュ、スケジューリングを含みます。

機械学習

データ連携

データ加工

特微量管理

機械学習

モデル管理

各種テスト

デプロイ

本番実行

モデル監視

レポート

モデル改善

databricks

Machine Learning

Create

Workspace

Repos

Recents

Search

Data

Compute

Jobs

Experiments

Feature Store

Models

Partner Connect

Help

01_create_DeltaLake Python

test-jpn-psa-10.3ML

公開クラウドストレージから学習データを取得します (WASBプロトコル)
sourcePath = 'wasbs://public-data@sajpstORAGE.blob.core.windows.net/customer.csv'

df_customer_csv = spark.read.format('csv').option("header","true").option("inferSchema", "true").load(sourcePath)
display(df_customer_csv)

(3) Spark Jobs

df_customer_csv: pyspark.sql.dataframe.DataFrame = [customer]

Table Data Profile

	customerID	gender	seniorCitizen	partner	dependents	tenure	phoneService	multipleLines
1	7590-VHVEG	Female	0	Yes	No	1	No	No phone service
2	5575-GNVDE	Male	0	No	No	34	Yes	No
3	3668-QPYBK	Male	0	No	No	2	Yes	No
4	7795-CFOCW	Male	0	No	No	45	No	No phone service
5	9237-HQITU	Female	0	No	No	2	Yes	No
6	9305-CDSKC	Female	0	No	No	8	Yes	Yes

Truncated results, showing first 1000 rows.
Click to re-execute with maximum result limits.

Command took 14.09 seconds -- by shunichiro.takeshita@databr

多様なデータ読み込み、パートナーコネクトでGUIベースでデータ連携も(その後、Delta Lake化)

テーブル表形式で直ぐにデータの中身を確認可能、コード書かずにグラフ化も可能

databricks Machine Learning

+ Create

Workspace

Repos

Recents

Search

Data

Compute

Jobs

Experiments

Feature Store

Models

Partner Connect

Help

01_create_DeltaLake Python

test-jpn-psa-10.3ML

PySpark Pandas API を使って前処理を実施

多くの Data Scientist は、pandasの扱いになれており、Spark Dataframeには不慣れです。Spark 3.2より Pandas APIを一部サポートしました。

これにより、Pandasの関数を使いながら、Sparkの分散機能も使うことが可能になります。

Requirement Spark3.2以降の機能なため、**DBR 10.2以上**のクラスター上で実行する必要があります

```
Cmd 10
1 import pyspark.pandas as ps
2
3 # Convert to koalas
4 data = df_bronze.to_pandas_on_spark()
5
6 # OHE
7 data = ps.get_dummies(data,
8                         columns=['gender', 'partner', 'dependents',
9                                   'phoneService', 'multipleLines', 'internetService',
10                                  'onlineSecurity', 'onlineBackup', 'deviceProtection',
11                                  'techSupport', 'streamingTV', 'streamingMovies',
12                                  'contract', 'paperlessBilling', 'paymentMethod'], dtype = 'int64')
13
14 # Convert label to int and rename column
15 data['churnString'] = data['churnString'].map({'Yes': 1, 'No': 0})
16 data = data.astype({'churnString': 'int32'})
```

実は、従来の python pandas の知識があれば Spark という並列分散処理機能が利用可能

databricks

Machine Learning ▾

- + Create
- Workspace
- Repos
- Recents
- Search
- Data
- Compute
- Jobs
- Experiments
- Feature Store
- Models
- Partner Connect
- Help

01_create_DeltaLake Python

test-jpn-psa-10.3ML

In [1]:

```
1 df_passengers = pd.read_csv("titanic_full.csv")
2 df_passengers
3 # bamboolib live code export
4 df_passengers['Survived'] = df_passengers['Survived'].astype(bool, errors='raise')
5 df_passengers = df_passengers.rename(columns={'PassengerId': 'PId'})
6 df_passengers
```

Show static HTML History Export Live Code Export

Search transformations or Create plot or Explore DataFrame

Age

Age Rename

Column transformations

Summary Valid Uniques Missing values

	i PId	b Survived	i Pclass	o Name
0	1	False	3	Braund, Mr. Owen Gridley
1	2	True	1	Cumings, Mrs. J. Hatley
2	3	True	3	Heikkinen, Miss. Laina
3	4	True	1	Futrelle, Mrs. Jacob
4	5	False	3	Allen, Mr. William Henry
5	6	False	3	Moran, Mr. James
6	7	False	1	McCarthy, Mr. Timothy J.
7	8	False	3	Palsson, Master. Gosta
8	9	True	3	Johnson, Mrs. O. A.
9	10	True	2	Nasser, Mrs. Nicanor

891 rows x 12 columns

Google Cloud

Databricks Machine Learning Create Workspace Repos Recents Search Data Compute Jobs Experiments Feature Store Models Partner Connect Help

Data Sources: dbfs:/tmp/e2e_psa_jp/e2e_demo/bronze

Description Edit

これらの特徴は、外部ストレージの customer.csv から派生したものです。カテゴリ列に対してダミー変数を作成し、その名前をクリーンアップし、顧客が解約したかどうかの布尔型フラグを追加しました。集計は行っていません。

特徴量リスト

Feature	Data Type	Models	Endpoints	Jobs	Notebooks
churn	INTEGER	-	-	-	-
contract_Month-to-month	LONG	-	-	-	-
contract_Oneyear	LONG	-	-	-	-
contract_Twoyear	LONG	-	-	-	-
customerID	STRING	-	-	-	-
dependents_No	LONG	-	-	-	-

加工済みデータを特徴量ストアとして管理、データチーム内で共有・再利用可能

特徴量がどのモデル・エンドポイント・ジョブ・ノートブックで利用されたかのリネージ



[Experiments](#) > Configure AutoML experiment

Configure AutoML experiment

1 Configure 2 Train 3 Evaluate

AutoML Experiment Configuration

* Compute: test-jpn-psa-10ML

* ML problem type: Classification

* Dataset: customer_churn_demo.churn_features

* Prediction target: churn

* Experiment name: churn_churn_features-2022_02

Advanced Config

Start AutoML

Schema:

Column name	Data type
customerID	string
seniorCitizen	double
tenure	double
monthlyCharges	double
totalCharges	string
churn	int
gender_Female	bigint
gender_Male	bigint
partner_No	bigint
partner_Yes	bigint
dependents_No	bigint
dependents_Yes	bigint
phoneService_No	bigint
phoneService_Yes	bigint
multipleLines_No	bigint

特徵量

予測対象

AutoML 実行

Notebook プログラムが自動生成される、専門家によりコード改善

21-11-20-15:16-Prophet-89564fbc9006ba858485d5b607fb16 Python

Detached Load Data Train Prophet model Aggregate data by ... Display the results Save the model The predicted ... Plot the forecast w... Plot the forecast co... Plot the forecast with change points and trend

```
Cond 28
1 from prophet.plot import plot_plotly, plot_components_plotly
2
3 # Choose a random ID from "id_list" for this run
4 id = set[forecast_pd.index[0].tolist()][0]
5 model = prophet_model._model_.impl._python_model.model[id]
6 forecast_pd = forecast_pd[forecast_pd['ds'].dt.year > 2018]
7 forecast_pd = forecast_pd[forecast_pd['ds'].dt.year < 2021]
8 fig
9
Out[28]:
```

lw lm Sm lv ss

2000

1500

1000

500

0

Jan 2018 Jul 2018 Jan 2019 Jul 2019 Jan 2020 Jul 2020 Jan 2021

ds

データ連携

データ加工

特微量管理

機械学習

モデル管理

各種テスト

デプロイ

本番実行

モデル監視

レポート

モデル改善

databricks

Machine Learning ▾

Create

Workspace

Repos

Recents

Search

Data

Compute

Jobs

Experiments

Feature Store

Models

Partner Connect

Help

View notebook for best model View data exploration notebook

Refresh Compare Delete Download CSV test_accuracy_score All time

Columns Only show differences

metrics.rmse < 1 and params.model = "tree"

Search Filter Clear

Showing 44 match

アルゴリズム

生成された Notebook

モデル精度

	Start Time	Duration	Run Name	User	Source	Version	Models	Metrics
<input type="checkbox"/>	3 minutes ago	7.0s	xgboost	shun...	Notebook:	-	sklearn	test_accuracy: 0.823
<input type="checkbox"/>	4 minutes ago	6.6s	random_for...	shun...	Notebook:	-	sklearn	test_accuracy: 0.823
<input type="checkbox"/>	4 minutes ago	9.1s	lightgbm	shun...	Notebook:	-	sklearn	test_accuracy: 0.817
<input type="checkbox"/>	3 minutes ago	6.2s	lightgbm	shun...	Notebook:	-	sklearn	test_accuracy: 0.812
<input type="checkbox"/>	4 minutes ago	11.2s	xgboost	shun...	Notebook:	-	sklearn	test_accuracy: 0.812
<input type="checkbox"/>	4 minutes ago	5.9s	lightgbm	shun...	Notebook:	-	sklearn	test_accuracy: 0.801
<input type="checkbox"/>	4 minutes ago	6.4s	logistic_reg...	shun...	Notebook:	-	sklearn	test_accuracy: 0.801
<input type="checkbox"/>	4 minutes ago	6.4s	decision_tr...	shun...	Notebook:	-	sklearn	test_accuracy: 0.801
<input type="checkbox"/>	3 minutes ago	6.8s	random_for...	shun...	Notebook:	-	sklearn	test_accuracy: 0.796
<input type="checkbox"/>	4 minutes ago	7.5s	xgboost	shun...	Notebook:	-	sklearn	test_accuracy: 0.796

Google Cloud

機械学習の実験結果は MLflow にて一元管理される

データ連携

データ加工

特徴量管理

機械学習

モデル管理

各種テスト

デプロイ

本番実行

モデル監視

レポート

モデル改善

Databricks Machine Learning ページ

Date: 2022-02-20 21:31:53 Source: Notebook: XGBoost User: shunichiro.takeshita@databricks.com

Duration: 7.0s Status: FINISHED Lifecycle Stage: active

Description Edit
Customer Churn モデル

Parameters (72)

モデルごとの管理内容

- ・関連 Notebook
- ・利用データバージョン
- ・パラメータ群
- ・メトリックス

Name	
classifier	XGBClassifier(base_score=None, booster=None, colsample_bylevel=None, colsample_bynode=None, colsample_bytree=0.3866165988537786, enable_categorical=False, gamma=None, gpu_id=None, importance_type=None, int...
classifier__base_score	None
classifier__booster	None
classifier__colsample_bylevel	None
classifier__colsample_bynode	None

Experiments

Feature Store

Models

Partner Connect

Help

Google Cloud

モデル開発環境の定義

Full Path:dbfs:/databricks/mlflow-tracking/4016443028876321/15b088e... Register Model

MLflow Model

The code snippets below demonstrate how to make predictions using the logged model. You can also register it to the model registry to version control and deploy as a REST endpoint for real time serving.

モデルインプット定義
モデルアウトプット定義

Model schema
Input and output schema for your model. Learn more

Name	Type
Inputs (46)	
customerID	string
seniorCitizen	double
tenure	double

Make Predictions
Predict on a Spark DataFrame:

```
import mlflow
logged_model = 'runs:/15b088e9243c442d917028c1faa38382/model'

# Load model as a Spark UDF. Override result_type if the model does not return double values.
loaded_model = mlflow.pyfunc.spark_udf(spark, model_uri=logged_model, result_type='double')

# Predict on a Spark DataFrame.
columns = list(df.columns)
df.withColumn('predictions', loaded_model(*columns))
```

databricks

Machine Learning ▾

- + Create
- Workspace
- Repos
- Recents
- Search
- Data
- Compute
- Jobs
- Experiments**
- Feature Store
- Models
- Partner Connect
- Help

▼ Artifacts

model

- MLmodel
- conda.yaml
- input_example.json
- model.pkl
- requirements.txt
- test_confusion_matrix.png
- test_precision_recall_curve.png
- test_roc_curve.png
- training_confusion_matrix.png
- training_precision_recall_curve.png
- training_roc_curve.png**
- val_confusion_matrix.png
- val_precision_recall_curve.png
- val_roc_curve.png

Full Path:dbfs:/databricks/mlflow-tracking/4016443028876321/15b088e... ↗

Size: 13.39KB

ROC curve

True Positive Rate (Positive label: 1)

False Positive Rate

Pipeline (AUC = 0.88)

特徴量の重要度や ROC 曲線などの評価指標も自動で保存される

データ連携

データ加工

特徴量管理

機械学習

モデル管理

各種テスト

デプロイ

本番実行

モデル監視

レポート

モデル改善



Machine Learning

Create

Workspace

Repos

Recents

Search

Data

Compute

Jobs

Experiments

Feature Store

Models

Partner Connect

Help

Registered Models > customer-churn > Version 1

Version 1

Use model for inference

Registered At: 2022-02-20 21:53:36

Creator: shunichiro.takeshita@databricks.com Follow Status: Following

Last Modified: 2022-02-20 21:57:31

Source Run: xgboost

Stage: None

Description

Edit

AutoMLで顧客離反モデルを作成
ベストモデルを本番環境へ登録

Pending Requests

Request

Request by

Actions

Transition to → Production

shunichiro.takeshita@databricks.com

Approve

Reject

C

承認ワークフロー

Tags

Schema

Request transition to → Staging

Request transition to → Production

Request transition to → Archived

Transition to → Staging

Transition to → Production

Transition to → Archived

Google Cloud

06_batch推論 Python

test-jpn-psa-10.3ML

モデルのロード

mlflow には pyfunc というライブラリがあり、モデルをロードしてUDF化してくれる関数も用意してあります。

```
1 import mlflow.pyfunc
2
3 model_name = f"{prefix}_churn_model" # ご自分のmodel nameに変更ください
4 model_version = 'staging' # model_version = 'production' ## <= このようにproduction/stagingも指定可能
5
6 # Load model from registry
7 loaded_model = mlflow.pyfunc.spark_udf(spark, model_uri=f"models:{model_name}/{model_version}")
```

Command took 1.80 seconds -- by a user at 2022/2/13 12:44:34 on unknown cluster

Spark DataFrameによる推論

Spark DataFrameの場合、バッチとストリーミングを両方とも扱える点と、SparkAPIを使った分散処理が出来るため大量のデータを非常に高速に処理実行することができます。

ただしPandasにも対応しているためどちらも利用することができます。

Schedule Share

Notebook をジョブ化するだけ、しかもデータ量に応じて自動伸縮するクラスターで稼働

databricks

07_create_alert_dashboard Python

Schedule Share

test-jpn-psa-10.3ML

CustomerChurn Dashboard + Add tag

Share Schedule Refresh

クエリーを保存し、add dashboard を選択します。新規の場合はダッシュボードを作成します。

解約予測顧客リスト - CustomerList

prediction	customerID	Name	Age	gender	Country	techSupport	tenure	monthlyCharges	email	internetService	paymentMethod	AccountCreation	seniorCitizen
1.00	8779-QRDMV	duis non	63	Male	NY	No	1.00	39.65	incidunt.official@ex.com	DSL	Electronic check	2021-09-06	1.00
1.00	4929-XIHVW	in sint	28	Male	UT	No	2.00	95.50	aliqua.deserunt@laborum.com	Fiber optic	Credit card (automatic)	2021-10-19	1.00
0.00	7760-DYPDY	do ex	53	Female	NC	No	2.00	80.65	cillum.iure@est.com	Fiber optic			
0.00	3071-VYPO	deserunt aliquip	65	Male	UT	No	3.00	89.85	occaecat.excepteur@culpa.com	Fiber optic			
1.00	7273-TEFQD	pariatur elit	66	Male	DE	No	3.00	41.15	fugiat.exercitation@sed.com	DSL			

1 2 3 4 5 ... 8 >

⌚ 19 minutes ago

解約予測顧客地図
解約リスク顧客の地域になります。

4.00
3.00
2.00
1.00

⌚ 19 minutes ago

Confusion Matrix
予測モデルの精度をチェックします

		0	1	Totals
prediction	0	1		
churn	6,147	27	6,174	
0	1,638	31	1,669	
1	5,509	26	5,535	
Totals	6,985	58	7,043	

⌚ 19 minutes ago

ダッシュボードはSubscription設定をすることで、定期的にメール配信することができます。

機械学習モデルの精度を経過観察するダッシュボード

Customer-360

チャネル別の観点

Viz-001 - C360-005

601.9 K
(3.708)

ATM利用者

⌚ a minute ago

- ATM利用者は微増
- 前年同月比3.7%増加

ATM - C360-006

ym

atm_20 atm_40 atm_60 atm_80

⌚ a minute ago

- 店舗利用者は微減
- 前年同月比2.2%減少

Viz-002 - C360-005

312.3 K
(-2.228)

店舗利用者

⌚ a minute ago

- 店舗利用は横ばい

branch_20 branch_40 branch_60 branch_80

ym

⌚ a minute ago

- Web利用者は微増
- 前年同月比3.7%増加

Viz-003 - C360-005

136.7 K
(3.708)

Web利用者

⌚ a minute ago

web_20 web_40 web_60 web_80

⌚ a minute ago

- Web利用者は微増
- 前年同月比3.7%増加

Viz-001 - C360-004

ATM APP WEB
CON STORE CALL

Viz-002 - C360-004

Stage: 3 43.5%
3456 of 7947

databricks

Customer-360 + Add tag

Share Schedule Refresh ...

商品別の観点 および 顧客セグメント別の観点

年齢別の商品利用傾向

- 年代別の利用商品の全体傾向
- 低年齢層は預金から
- 中年齢層から投資系を利用
- 高年齢層は預金割合金額が大

Viz-001 - C360-001

a minute ago

預金から投資商品へ

Viz-001 - C360-008

a minute ago

お客様の声

コールログのトピック

Viz-001 - C360-010

お客様の声

Viz-001 - C360-011

category	percentage
4_やや不満	29.2%
2_満足	41.7%
1_非常に満足	16.7%
3_普通	8.33%
5_非常に不満	4.17%

Table - C360-012

warn	evaluation	doc
✓	非常に満足	親身になって対応してくれます。これまで普通預金口座のみの利用で、メインバンクとしては活用していませんでしたが、定期預金を勧められて新たに口座開設をすることになりました。他行と比較してどのようなメリットがあるのか尋ねたところ、スタッフさんが一つ一つ丁寧に説明してくれたので、大きな信頼感を持って手続きを進めることができました。お得なキャンペーンや優遇金利についても熟知されていて、自信を持って説明する姿がとても印象的でした。ネットバンキングにも対応している旨の説

データ連携

データ加工

特徴量管理

機械学習

モデル管理

各種テスト

デプロイ

本番実行

モデル監視

レポート

モデル改善

Customer-360

Share Schedule Refresh :

顧客別/商品別/チャネル別タッチポイント/LTV/レコメンド/アップセル率/解約率

セグメント
商品利用
チャネル利用
今後の予測
キャンペーン状況

		id	image	segment	cust_cnt	deposit	invest	loan	atm	store	web	salary_est	ltv_pred	upsell_pred	churn_pred	camp1	camp2	camp3
1		Age-30-Female		123	✓	High	200	50	10	✓	✗	✓						
2		Age-30-Female		1,234	✓	✓	✗	✓	✗	✗	✗	Mid	150	30	15	✓	✓	✓
3		Age-30-Female		2,345	✓	✓	✗	✓	✓	✓	✗	Mid	100	25	25	✗	✗	✓
4		Age-30-Female		3,456	✓	✗	✗	✓	✗	✗	✗	Low	50	10	30	✗	✗	✓
5		Age-30-Male		123	✓	High	200	50	10	✓	✗	✓						
6		Age-30-Male		1,234	✓	✓	✗	✓	✗	✗	✗	Mid	150	30	15	✓	✓	✓
7		Age-30-Male		2,345	✓	✓	✗	✓	✓	✓	✗	Mid	100	25	25	✗	✗	✓
8		Age-30-Male		3,456	✓	✗	✗	✓	✗	✗	✗	Low	50	10	30	✗	✗	✓

⌚ a minute ago

データ連携

データ加工

特徴量管理

機械学習

モデル管理

各種テスト

デプロイ

本番実行

モデル監視

レポート

モデル改善



Machine Learning

Create

Workspace

Repos

Recents

Search

Data

Compute

Jobs

Experiments

Feature Store

Models

Partner Connect

Help

Registered Models

Permissions

Share and serve machine learning models. [Learn more](#)

Create Model

churn

X

元に戻り、再学習！

Name	Latest Version	Staging	Production	Last Modified	Tags	Serving
customer-churn	Version 1	—	Version 1	2022-02-20 22:17:30	—	—
hhar_churn	Version 1	—	—	2021-09-19 14:23:58	—	—
brheid_churn	Version 1	Version 1	—	2021-03-09 03:44:44	—	—

<

1

>

10 / page

このように Databricks にて 下記フローに対応可能です！

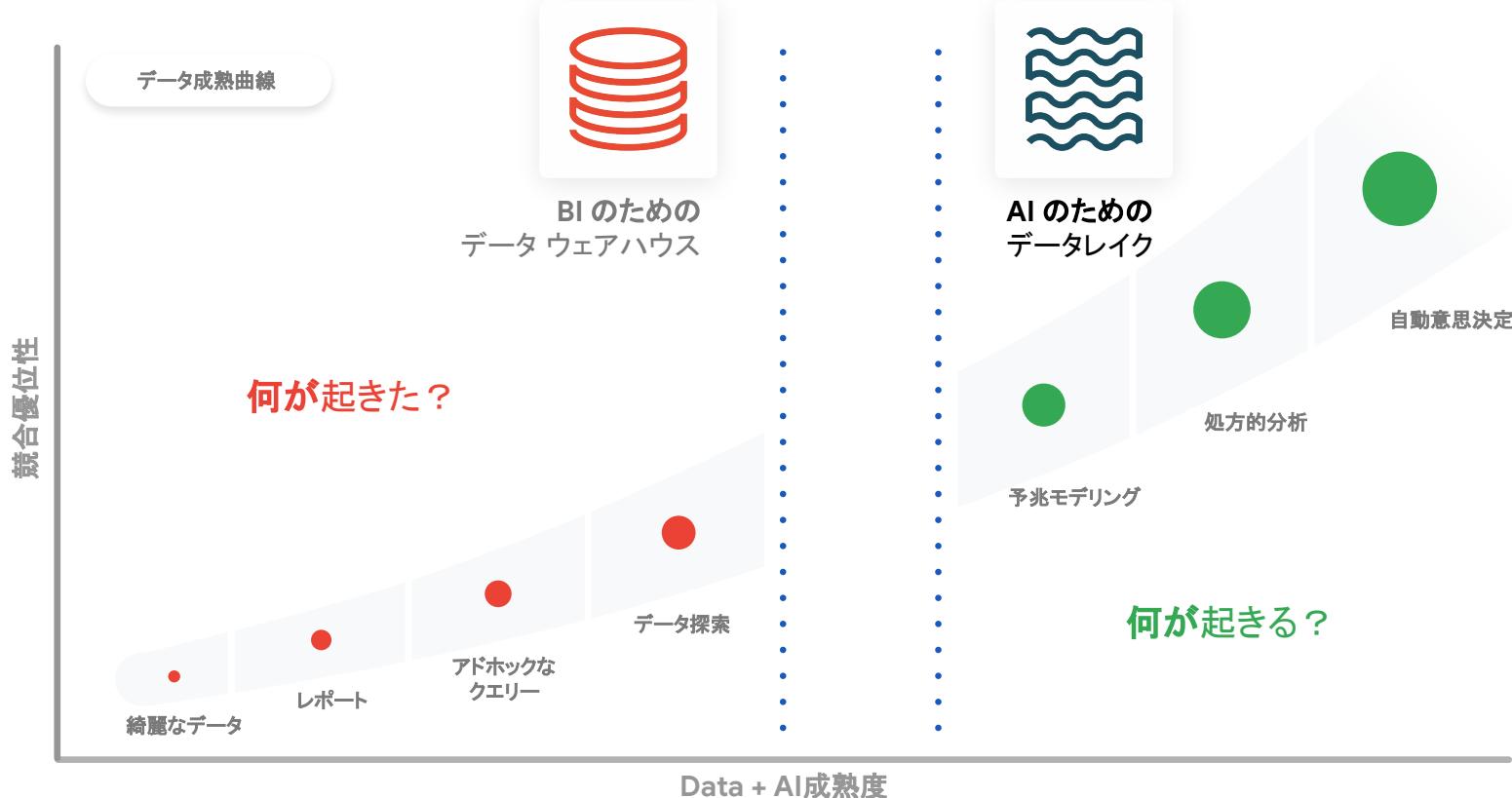




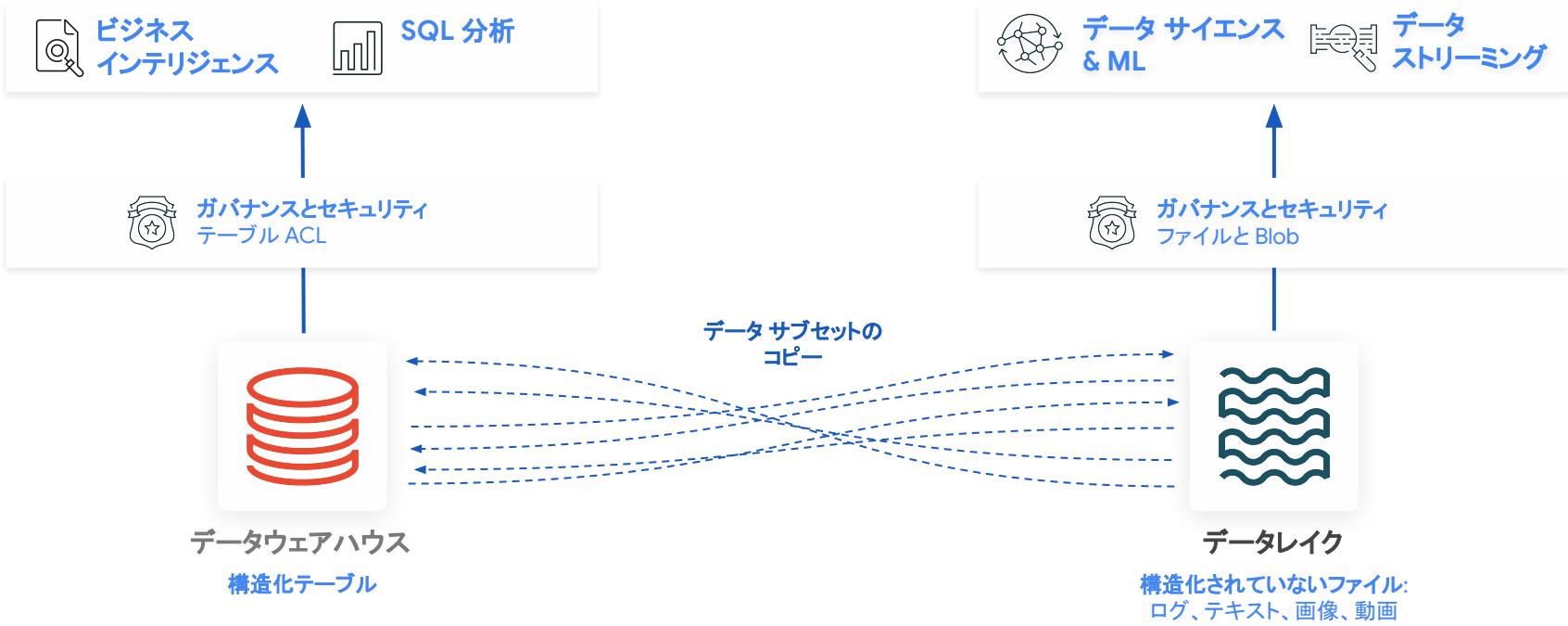
DS の業務、課題、そして解決策

1. DS の業務 + 課題
2. 機械学習 End to End フローでの解決策
3. システム インフラ観点での解決策

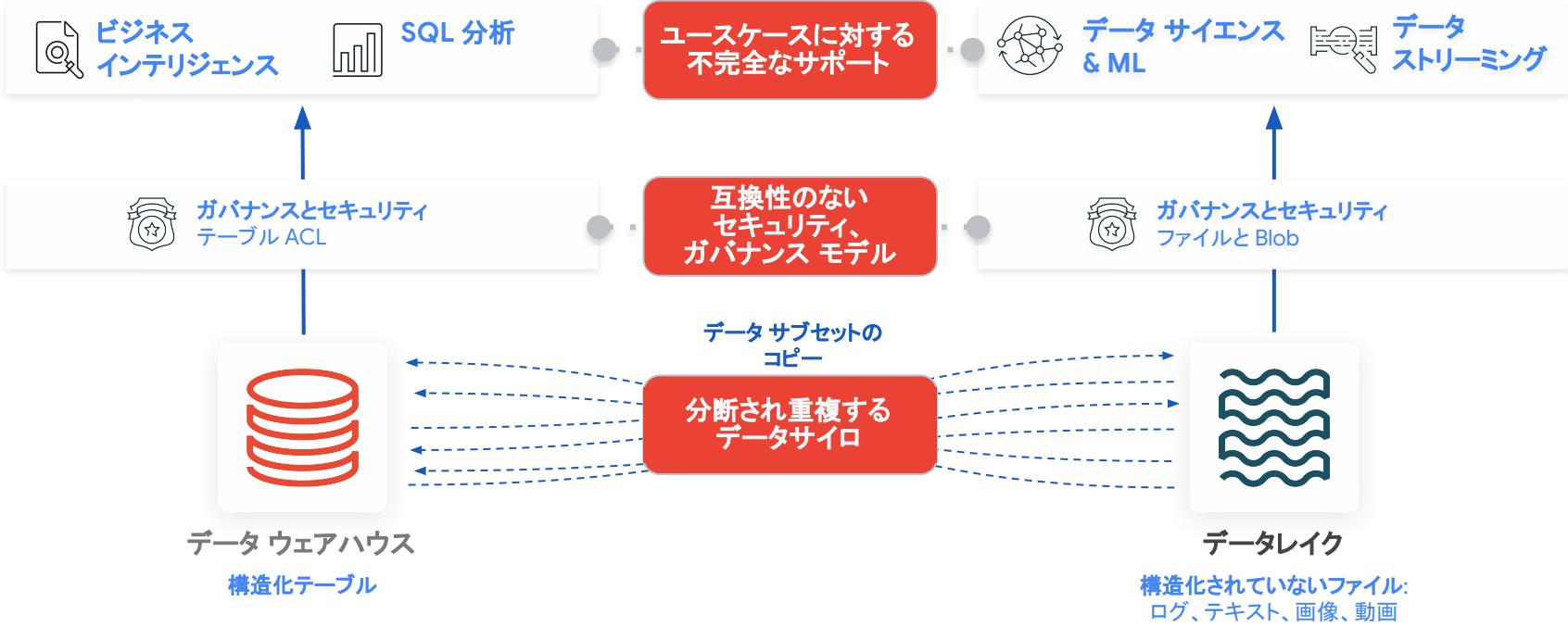
全く異なる、互換性の無い 2つ以上のデータプラットフォームが必要であることを理解しなくてはなりません



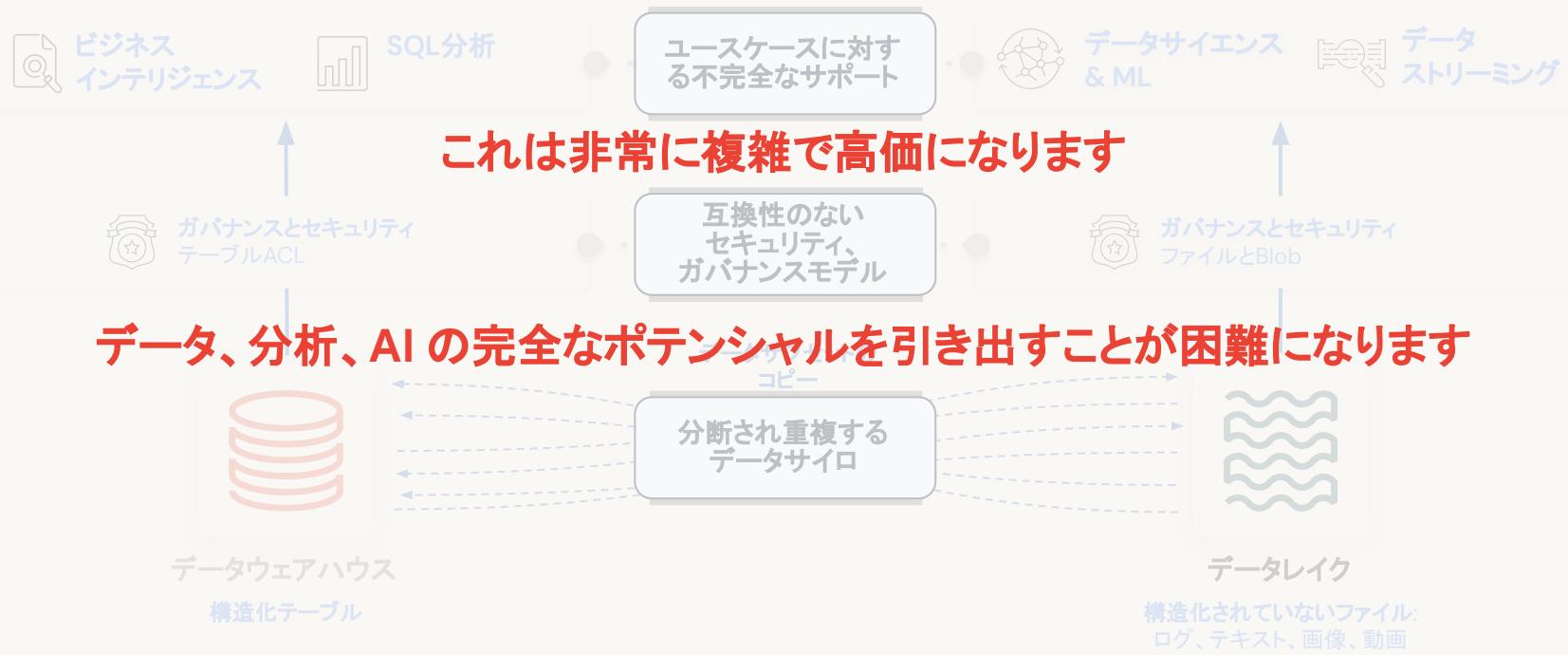
全く異なる、互換性の無い 2つ以上のデータプラットフォームが必要であることを理解しなくてはなりません



全く異なる、互換性の無い 2つ以上のデータプラットフォームが必要であることを理解しなくてはなりません



全く異なる、互換性の無い 2つ以上のデータプラットフォームが必要であることを理解しなくてはなりません

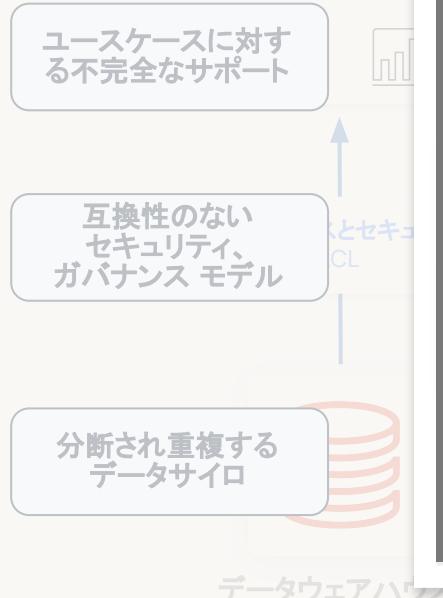


全く異なる、互換性の無い2つ
必要であることを理解した

タプラットフォームが



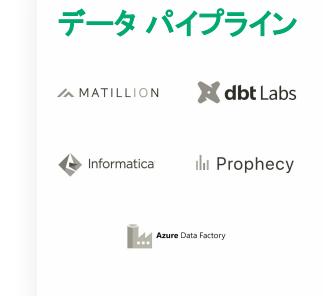
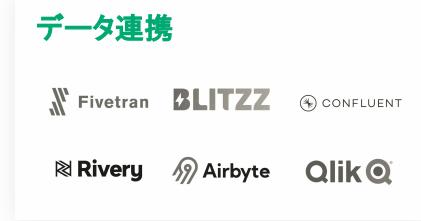
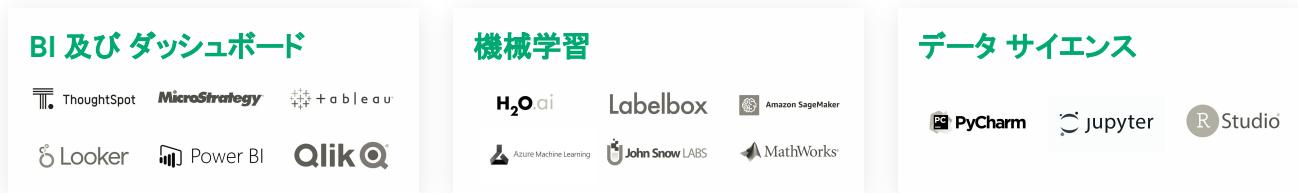
レイクハウス プラットフォーム



構造化されていないファイル:
ログ、テキスト、画像、動画

Google Cloud

Lakehouse は最新のデータスタックにも対応します



Google Cloud

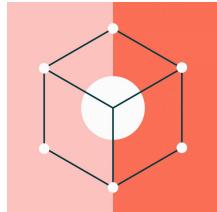


Wrap Up

データサイエンス・機械学習・AIの フル ポテンシャルを引き出すデータブリックス

シンプルな基盤

人材育成



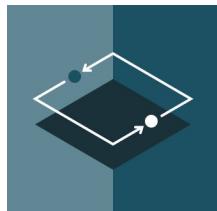
すぐに分析作業をスタート

- クラウドネイティブサービス
- 数クリックで環境準備
- ライブラリがプレインストール



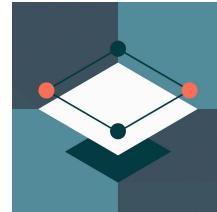
資産をしっかり管理できる

- MLflowでモデル管理
- Delta Lakeで多構造データ管理
- Unity Catalogでデータ辞書化



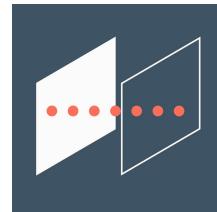
手触り感のある人材育成をご支援

- ソリューション・アクセラレータ
- 業界別の開発済みパッケージ
- 需要予測、不正検知、顧客360、IoT等



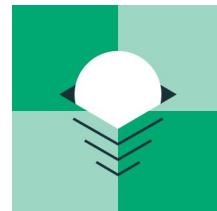
コラボレーションが簡単

- ノートブックで共同作業
- データ、モデル、グラフを共有
- アクセス権限管理が簡単



モデル資産を簡単に本番化

- ペタバイト級へのデータ対応
- リアルタイム処理への対応
- ジョブオーケストレーション



アジャイルにQuick Winをご支援

- Data + AI施策の実現のサポート
- 人材育成、ユースケース選定支援
- レイクハウス、MLOpsのQuickな実現

データサイエンスの仕事が捲ります！！

Next Step

是非下記へお問い合わせください！
marketing-jp@databricks.com



デモ・ハンズオン を見てみたい

- 実環境を使ったデモ や、一時的に触つていただける環境を活用したハンズオンを実施
- デモ・ハンズオン共に、ETL や機械学習といったテーマでご案内

見てみたい！



ご興味のテーマに 関して深堀したい

- 世界のベスト プラクティス にもとづく、具体的な事例をもとに、ワークショップを実施しております！
- セキュリティ・ガバナンス、アーキテクチャー、課金体系等気になる点に関して追加での説明 or 資料共有
- 類似サービスとの比較・相違点等 のご紹介

もっと知りたい！



PoC を 実施してみたい

- 無償で PoC をサポート
- 通常 2 週間の PoC 期間中、Databricks 使用料と技術サポートを無償でご提供
- パブリック クラウドのストレージ、コンピュートコストはお客様ご負担となります (AWS の場合 S3 や EC2 等)

試してみたい！

御社のピックデータ、活用できていますか？

ブリックスちゃんが データエンジニアリングのお悩みを まるっとすぱっと解決！



御社のピックデータ、活用できていますか？

ブリックスちゃんが データサイエンスのお悩みを まるっとすぱっと解決！



<https://dbricks.co/3B1FyEm>



Google Cloud



データブリックス・クイック・スタートガイド

日本初のデータブリックス本を出版しました

- 「データブリックスって聞くけど、一体どういうものなのだろうか」と思われている方、データブリックスを触り始めた方を対象として、データブリックス・ジャパンのエンジニアの有志で本書を執筆しました。本書をご一読いただければ、データブリックスとは何か、データブリックスをどのように使うのかを一通り理解できる内容となっています。
- データとAIを活用して業務を変えたい、機械学習モデルを本格的に運用することを前提としてデータ/AI基盤を構築したいと考えられている方に本書が一助になれば幸いです。



1章 Databricks(データブリックス)とは?

- はじめに
- 背景
- レイクハウスの誕生
- データブリックスとは
- コンセプト
- アーキテクチャ
- 主要機能
- コスト

2章 データブリックスのセットアップ

- データブリックスのセットアップ

3章 データブリックスを使ってみる

- データブリックスのユーザー・グループ
- Databricks クラスター
- Databricks ノートブック
- データブリックスのジョブ

4章 ユースケース別ガイド

- データエンジニアリング
- 機械学習
- BI

5章 ツール連携

- Repos
- Partner Connect

6章 MLOps の実現に向けて

Thank you.

