



Hands-Onワークショップ

90分で触れる データブリックスの機能

データブリックス・ジャパン株式会社
2023年9月14日



本日のアジェンダ

- ワークショップ環境のセットアップ
- データブリックスの概要
- データブリックスの機能ワークショップ・デモ
 - テーブル作成とUnity Catalogへの登録
 - Python/SQLを使用したETL処理の実行
 - SQLでの分析/ダッシュボードでのデータの可視化
 - AutoMLによる機械学習モデルのトレーニングとモデル管理



プロダクトセーフハーバーステートメント

この情報は、データブリックスの一般的な製品の方向性を概説するために提供されるものであり、**情報提供のみを目的**としています。データブリックス のサービスを購入するお客様は、現在利用可能なサービス、特徴、機能のみに依拠して購入を決定してください。将来見通しに関する記述に記載されている未発表の機能または特徴は、データブリックスの裁量で変更される可能性があり、計画通りまたは全く提供されない可能性があります。



はじめに: ワークショップ環境へのログイン

Workshop資料
bit.ly/3ReQNFn



Workshop環境登録
bit.ly/3ZfBCh8

Step 1: ワークスペースへの登録

本日のワークショップで使用する一時的なワークスペースは、サードパーティベンダーによって管理されています。
ワークショップのハンズオン部分のために、この簡単な登録フォームに必要事項を記入してください。

The registration form is titled 'Register Now' and includes fields for 'First Name*' and 'Last Name*'. It also has an 'Email*' field and a checkbox for accepting terms and conditions. To the right, there are four callout boxes with labels:

- 名前(姓)
- 名前(姓)
- Emailアドレス
- 同意にチェック
- Submitをクリック

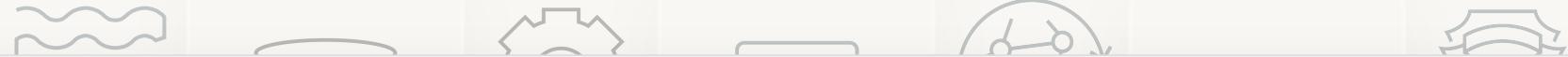
ログイン関係のFAQ

過去のワークショップで多くいただいた質問を記載いたします。

質問	回答
トレーニング環境はいつまで使用できますか？	本日のトレーニング終了(11:30)まで使用可能です。 トレーニング以降は無償のDatabricks Community Editionか 2週間のフリートライアルをご活用ください。 https://databricks.com/jp/try-databricks
本トレーニングのスライドは、どこからダウンロードできますか？	bit.ly/3ReQNFn 上記サイトの「1. 本トレーニングのスライド資料」のリンクにPDFがござります。
ノートブックで利用している多目的クラス(AAll-Purposeクラスタ)やJobsクラスタと、SQLウェアハウスはどう違うのでしょうか？	クエリを処理するためのコンピュートリソースという役割は同じです。一方で、clusterとSQLウェアハウス(cluster)は利用料金が異なります。 https://databricks.com/jp/product/aws-pricing
ハンズオン環境のログイン情報が記載されたメールが届きません。 環境にログインできません。	Cloudlabsへの登録がお済みでない方 こちら を参照頂き、アカウント作成をお願いします。 Cloudlabsへの登録がお済みの方 Cloudlabsからのメールを検索頂くか、On24のQ/A機能より、ご登録済みのemail addressをご連絡ください。弊社側で、ログイン情報を確認いたします。



多くのプラットフォームをつなぎ合わせる必要がある



これら全ては非常に高価で複雑です

Data
Lake

Data
Warehouse

Orchestration

Business
Intelligence

Data Science
& ML

Streaming

Governance

データのサイロは
高価なオペレーションコストを引
き起こします

一貫性が無いポリシーは
データの信頼性を損ないます

バラバラなツールはチーム間の
生産性を悪化させます

データレイクハウスは異なるアプローチを取ります

複数ペルソナをサポートする単一のプラットフォーム



BI & 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 & ML

企業全体におけるすべてのデータアクセスに対する
単一のセキュリティ、ガバナンスモデル

構造化データ、準構造化データ、非構造化データ
すべてを格納、管理する単一のプラットフォーム



クラウドデータレイク
すべての生データ
(ログ、テキスト、音声、動画、画像)



データレイクハウスは異なるアプローチを取ります

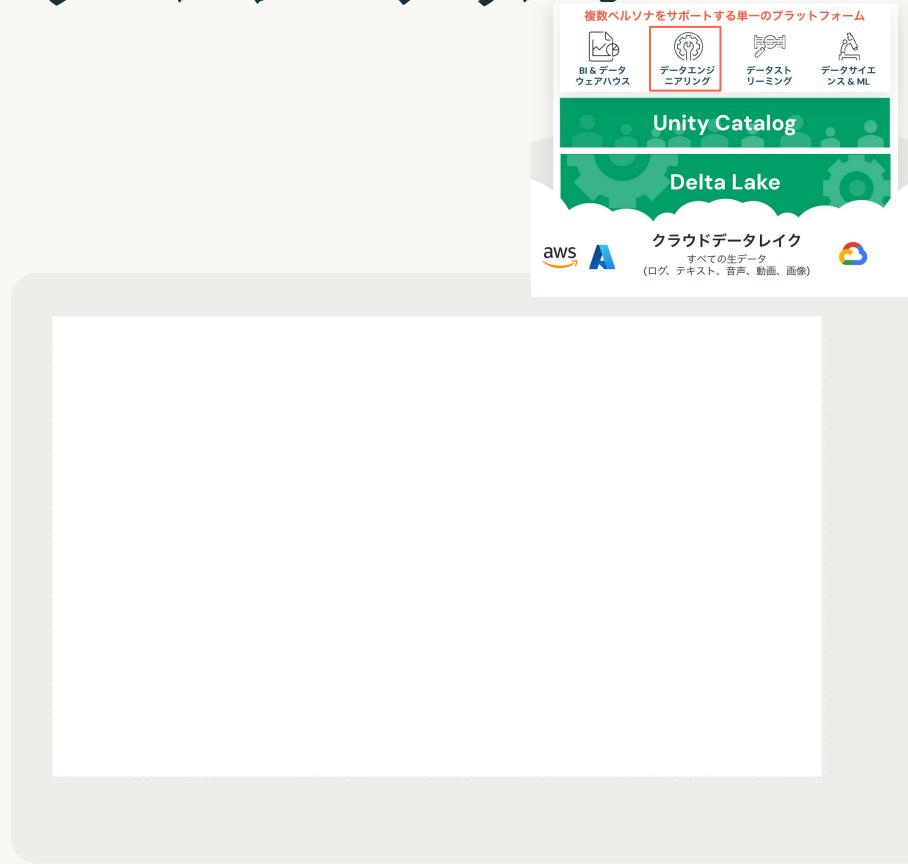


クラウドデータレイク
すべての生データ
(ログ、テキスト、音声、動画、画像)



Databricksにおけるデータエンジニアリング

- Databricksワークフローによるデータオーケストレーション
- Delta Live Tablesによる完全なデータパイプラインの管理
- Delta Lakeによるキュレーテッドデータレイクアプローチを通じてデータエンジニアリングをシンプルに



DatabricksにおけるSQLワークフロー

- Delta LakeにおけるBI、SQLワークフローの優れた性能、同時実行性
- 分析に適したネイティブSQLインターフェース
- Delta Lakeの最新データに直接クエリーすることによるBIツールのサポート
- サーバレスSQLで迅速な立ち上げ（日本では今後ご提供予定）



DatabricksにおけるML & データサイエンス

機械学習

- ・ モデルレジストリ、再現性、本格運用への投入
- ・ 再現性確保にDelta Lakeを活用
- ・ シチズンデータサイエンティストのための AutoML

データサイエンス

- ・ インタラクティブ分析向けコラボレーティブ ノートブック、ダッシュボード
- ・ Python、Java、R、Scalaのネイティブサポート
- ・ Delta Lakeデータのネイティブサポート



レイクハウスガバナンスのための Unity Catalog

すべてのデータ資産のガバナンス、管理

- ウェアハウス、テーブル、カラム
- データレイク、ファイル
- 機械学習モデル
- ダッシュボード、ノートブック

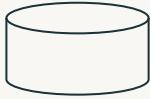
機能

- データリネージ
- 属性ベースのアクセス制御
- セキュリティポリシー
- テーブル、カラムレベルのタグ
- 監査
- データ共有

The screenshot shows the Databricks Unity Catalog interface. On the left, there's a sidebar with navigation icons and a search bar. The main area displays a table named 'city_data'. The table has a 'Description' section with a note about raw events from across the LOC platform. It also includes sections for 'Recommendations', 'Owners' (with icons for pii, cost, and inventory), 'Frequent users', 'ABAC Policies' (with checkboxes for pii, cost, and inventory), and 'Top Queries' (listing 'Champions stats 1H 2021', 'Gameplay analysis Q2 2021', 'Gameplay hours LOC semi-finals', and 'Purchase prediction model 2021'). Below these is a 'Statistics' section showing 'Format: Delta', 'Updated 5 hours ago', 'Created 1 year ago', and 'Size: 500 GB'. To the right of the table, there's a 'Schema' section with columns for ProductID (integer), Status (string), PricingTier (float), DistributionTier (string), and AccountInfo (string). There are also tabs for 'Add description', 'Add Tag', and 'Edit'. A 'Lineage' section shows a flowchart starting from 'intbdcall' leading to 'Firehose', which then branches to 'disp_rec' and 'inventory'. At the bottom, there's a 'Privileges' section with a table for 'User / Group' (listing 'bi_team' and 'dev') and a 'Permissions' table (listing 'Select, Modify, Manage' for bi_team and 'Select' for dev). The last section is 'Data Sample' with a table showing sample data for ProductID, Status, PricingTier, DistributionTier, and AccountInfo.



ワークショップ完了後の目標



クラウドストレージ



bronze
未加工データ



silver
エンハンス+クレンジング済
データ



gold
BI向け集計データ



BI ダッシュボード

借入情報
(生データ)
json

借入情報
(生データ)

借入情報
(履歴データ)
csv

借入情報
(履歴データ)

顧客情報
delta

顧客情報

mlflow

Unity Catalog

ワークショップ完了後の目標

[1] クラウドストレージのデータを日次で処理



クラウドストレージ



bronze
未加工データ



silver
エンハンス+クレンジング済
データ



gold
BI向け集計データ

借入情報
(生データ)

json

借入情報
(生データ)

借入情報
(履歴データ)

csv

借入情報
(履歴データ)

借入情報
(顧客情報付与)

顧客情報

delta

顧客情報

Unity Catalog

[3]国別の借入状況を
ダッシュボードで確認可能



BI ダッシュボード

mlflow

[4]国別の借入状況を予測(デモ)

ワークショップでの想定業務フロー

利用データのサンプル
アウトプットの要望

データ処理のプロトタイプ実
装・アウトプット検証

実装したデータ処理を
自動化・高度化

アウトプットの
活用

ノートブック

コーディング
Python/SQLの動作
コラボレーション機能

1_SingleNotebook

Unity Catalog

権限管理
カラム/依存関係の確認

ワークフロー

2_Workflow
Delta Live Table

DBSQL
BI
ダッシュボード

3_SQL

AI/
ML
Mlflow / AutoML
FeatureStore

読込フォルダ



ソースコード読み込み



個人用の環境設定

設定用ノートブックに個人用の設定値を入力します
(入力完了後本ノートブックは実行不要です)

The screenshot shows a Databricks workspace interface. On the left, the sidebar displays a 'ワークスペース' section with a 'workshop-New' notebook selected. Below it is a '並べ替え: 名前' dropdown menu. Under the '新規' section, there are several notebooks listed: '0_DataPreparation', '1_SingleNotebook', '2_Workflow', '3_SQL_Dashboard', and 'setting', which is currently highlighted.

The main area contains two code cells:

Cmd 1

```
1 USE_CATALOG="workshop"
2 USE_DATABASE="workshop_your_name_here" # この値を設定
3 spark.conf.set("demo_setting.use_catalog", USE_CATALOG)
4 spark.conf.set("demo_setting.use_database", USE_DATABASE)
```

A red hexagonal callout labeled '1 文字列を設定ください' points to the second line of code where 'workshop_your_name_here' is highlighted.

Cmd 2

```
1 %sql
2 use catalog "${demo_setting.use_catalog}";
3 create database if not exists ${demo_setting.use_database};
4 use database ${demo_setting.use_database};
```



ノートブックの基本動作

“ワークスペース”
からファイルを探索

The screenshot shows a Databricks notebook interface. On the left, a sidebar titled "ワークスペース" lists various notebooks and clusters. A red hexagon labeled "1" highlights the "新規" (New) button and the "ワークスペース" (Workspace) button. The main area displays a notebook titled "1_ノートブックの基本". The top bar includes a dropdown for "Python", a "接続" (Connection) dropdown set to "接続", and a "スケジュール" (Schedule) button. A red box highlights the "接続" button. Below the top bar, there are two buttons: "Run All Above" and "Run All Below". The notebook content shows a cell with the following Python code:

```
1 print("Pythonを実行中です!")
```

The output of the cell is:

Pythonを実行中です!

コマンド所要時間: 0.04秒 -- 実行者:a user 実行日:2023/9/12 11:47:26 実行場所:unknown compute

Below the cell, a section titled "複数の言語を利用" (Use multiple languages) provides information about magic commands for other languages like %python and %sql.

On the right side of the interface, there is a vertical sidebar with icons for different tools and a red hexagon labeled "3" highlighting the "UI上の ▶ ボタンか Shift + Enter でセルを実行" (Run cell via UI button or Shift + Enter) text.

このボタンからクラスタ
(計算リソース)を選択し接続

2

UI上の ▶ ボタンか
Shift + Enter
でセルを実行

Unity Catalogでのデータ確認

The screenshot shows the Databricks interface with the following highlights:

1. ワークスペースメニュー
2. カタログメニュー
3. 依存関係タブ
4. リネージュグラフを表示ボタン

画面内容:

- カタログ: workshop > workshop_masataka_yokota > cleaned_new_txs
- 所有者: odl_instructor_1063322@datarickslabs.com
- タグ: タグを追加
- LiveStream of new transactions, cleaned and compliant
- 依存関係: サンプルデータ、詳細、権限、履歴、依存関係、洞察
- リネージュグラフを表示
- テーブル名: workshop.workshop_masataka_yokota.new_txs
- ノートブック: Upstream
- ワークフロー: Downstream
- バイブレイン: Downstream
- ダッシュボード: 依存関係データは過去90日間のデータから保存されます
- ノード: historical_txs, new_loan_balances_by_country, new_txs, raw_historical_loans, raw_txs, ref_accounting_treatment, total_loan_balances, workshop_yukiteru_koide_2_here, workshop_yukiteru_koide_here, workshop_yyl



Workflowによるパイプライン構築(1)

ジョブ作成画面への遷移

ワークフロー

ジョブ ジョブ実行 Delta Live Tables

名前

タグ

作成者

トリガー

最近のラン

ジョブを作成

名前	作成者	トリガー	最近のラン
job	odl_user_106342e	---	---
job_workshop_yokota	odl_instructor_10e	---	---
sample_job_masataka	odl_instructor_10e	---	---

ジョブに最初のタスクを作成

ワークフロー > ジョブ > 新規

job_sample_my

ジョブ名入力

workshop_workflow_<your_name>

task_bronze

ノートブック

ワークスペース

2_WorkFlow/0_workflow_loan_bronze

Lab-Cluster-1063322

タスク名*

task_bronze

種類*

ノートブック

ソース*

ワークスペース

パス*

2_WorkFlow/0_workflow_loan_bronze

クラスター*

Lab-Cluster-1063322

UI | JSON

タスク名	task_bronze
種類	ノートブック
ソース	ワークスペース
パス	2_WorkFlow/0_workflow_loan_bronze
クラスター	割り当てられたクラスタ

Workflowによるパイプライン構築(2)

タスクの追加

ワークフロー > ジョブ > [フィードバックを提供]

sample_job_masataka ☆

ジョブの実行 タスク

The screenshot shows the Databricks Workflow UI. At the top, there is a card for a task named 'bronze' with the ID '....Workflow/0_workflow_loan_bronze' and the cluster 'Lab-Cluster-1063322'. Below this, a blue button labeled '+ タスクを追加' is highlighted with a red box and the number '6'. A modal window is open, showing the configuration for a new task. The 'タスク名' field is set to 'bronze'. The '種類' dropdown is set to 'ノートブック'. The 'ソース' dropdown is set to '...322@databrickslabs.com/workshop/2_Workflow/0_workflow_loan_bronze'. The 'クラスター' dropdown is set to 'Lab-Cluster-1063322'. The 'UI' tab is selected at the bottom right of the modal. At the bottom of the main screen, there is a 'タスクを保存' button highlighted with a red box and the number '8'.

タスク名	task_silver
パス	2_workflow/1_workflow_loan_silver
依存先	task_bronze

タスク名	task_gold
パス	2_workflow/2_workflow_loan_gold
依存先	task_silver

上記の条件以外はbronzeと同様

SQLクエリ実行/可視化

The screenshot illustrates the Databricks SQL interface with various numbered callouts:

- 1**: SQLエディタ (SQL Editor) ボタン
- 2**: 可視化 (Visualization) パネル
- 3**: 実行 (Execute) ボタン
- 4**: クエリ結果 (Results) テーブル
- 5**: バー視覚化 (Bar Visualization) パネル
- 6**: 保存 (Save) ボタン

クエリ結果 (4) の一部:

#	acc_fv_change	_taxes
1		883
2		576
3		1028
4		125
5		725

バーグラフ (5) のY軸: SUM(count)

クエリ (3):

```
1 | select * from new_txs
```

メッセージ (Bottom Left):

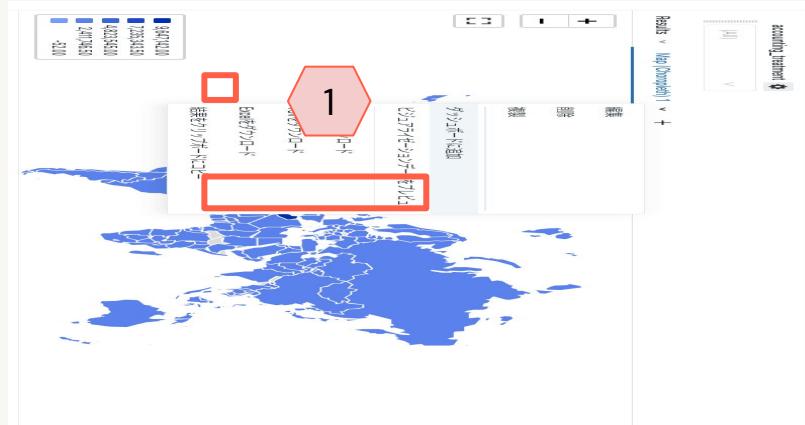
ノートは、エラーを診断するやクエリを定期的に答えることで作ります。間違いがあるため、必ず事前に確認してください。(More)

メッセージ (Bottom Right):

query to something
① 3秒 420ミリ秒 | 1,000行が返されました

ダッシュボードの作成・編集

ダッシュボードへのグラフ追加



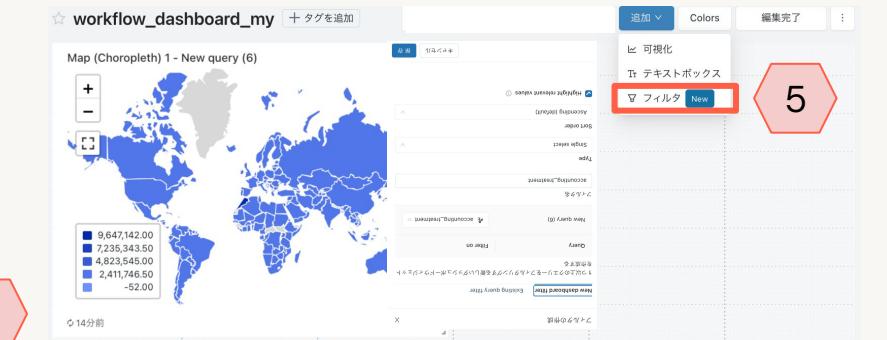
この可視化を追加するダッシュボードを選択:

- workflow_dashboard_my
- 新規ダッシュボードを作成

キャンセル 追加

3

ダッシュボードの編集(フィルタ追加)





DATA+AI WORLD TOUR



@DatabricksJP



bit.ly/DatabricksJP



[databricks.com
/jp/blog](https://databricks.com/jp/blog)



本イベントのX(旧Twitter)ハッシュタグ

#DAIWT