データブリックスを使ってみよう

ETL,BI,MLなどのあらゆる分析ユースケースを一つのプラットフォームで実現し、組織全体でのデータとAIの効率的な利活用を促進

弥生 隆明 / Paulo Gutierrez
データブリックス・ジャパン株式会社　ソリューションアーキテクト

databricks

アジェンダ

- 会社概要
- AI・機械学習プロジェクトにおける課題
- レイクハウスプラットフォーム

databricks

自己紹介

弥生 隆明 (やよい たかあき)

ソリューションアーキテクト

- 前職はコンサルティングファーム、総合電機メーカーにてデータ分析・Webサービス構築などに従事。インド赴任経験あり。
- 最近驚いたこと：インドカレーのサグカレーが好きなのですが、先日インド人の知り合いから「サグはほうれん草ではなくて草、ほうれん草はパラク」と教えてもらったことです。

@taka-yayoi

@taka_aki

About Me

Paulo Gutierrez （パウロ）
ソリューションアーキテクト

Spark, Flink, Arrow, Elasticsearch, MongoDB など

カメラ、２輪車、コーヒー、自作キーボード、アウトドア

@tokyodataguy
@juanpaulo

databricks

アジェンダ

- 会社概要
- AI・機械学習プロジェクトにおける課題
- レイクハウスプラットフォーム

databricks

databricks

データ&AIカンパニー

レイクハウス

データ、分析、AIに関わるワークロードを統合する
シンプルなプラットフォーム

顧客数

5000+

世界中で利用

オリジナルの開発者が設立

APACHE Spark™ DELTA LAKE mlflow™

米国本社: Databricks Inc.
設立: 2013年
所在地: アメリカ カリフォルニア州 サンフランシスコ市
社員数: 2,000 以上
拠点数: 11ヶ国 16拠点
沿革: カリフォルニア大学バークレー校から発祥
代表者: アリ・ゴディシ (共同設立者 & CEO)

日本法人: データブリックス・ジャパン株式会社
設立: 2019年
所在地: 東京都港区六本木1-4-5アークヒルズサウスタワー16階
代表者: 竹内　賢佑

主要な指標

5,000 以上の顧客

450 以上のパートナー

ARR(年間定額収益)は約430億円
(2021年1月)

シリーズGの資金調達により約3兆円市場価値
(2021年2月)

事業戦略

AI, ビックデータに特化

クラウド上のみで提供
(Microsoft Azure, AWS, GCP, Alibaba Cloud)

積極的なオープンソース化

Apache Spark　mlflow　DELTA LAKE　re:dash

市場における評価

ガートナー社 マジック・クアドラント・レポート
「2021年 データサイエンス & 機械学習 部門」
"リーダー" 企業に指名 (3年連続)

チャレンジャー	リーダー
	databricks
ニッチプレイヤー	ビジョナリー

あらゆる産業にサービスを提供しています

アジェンダ

- 会社概要
- AI・機械学習プロジェクトにおける課題
- レイクハウスプラットフォーム

databricks

未来はすぐそこに来ています。しかし、すべての人が享受できている訳ではありません

83% のCEOは戦略的にAIを優先すべきと述べています

MIT Sloan Management Review

$3.9T
約425兆円

が2022年までにAIによってもたらされるビジネス価値となります

Gartner

85% のビッグデータプロジェクトが失敗しています

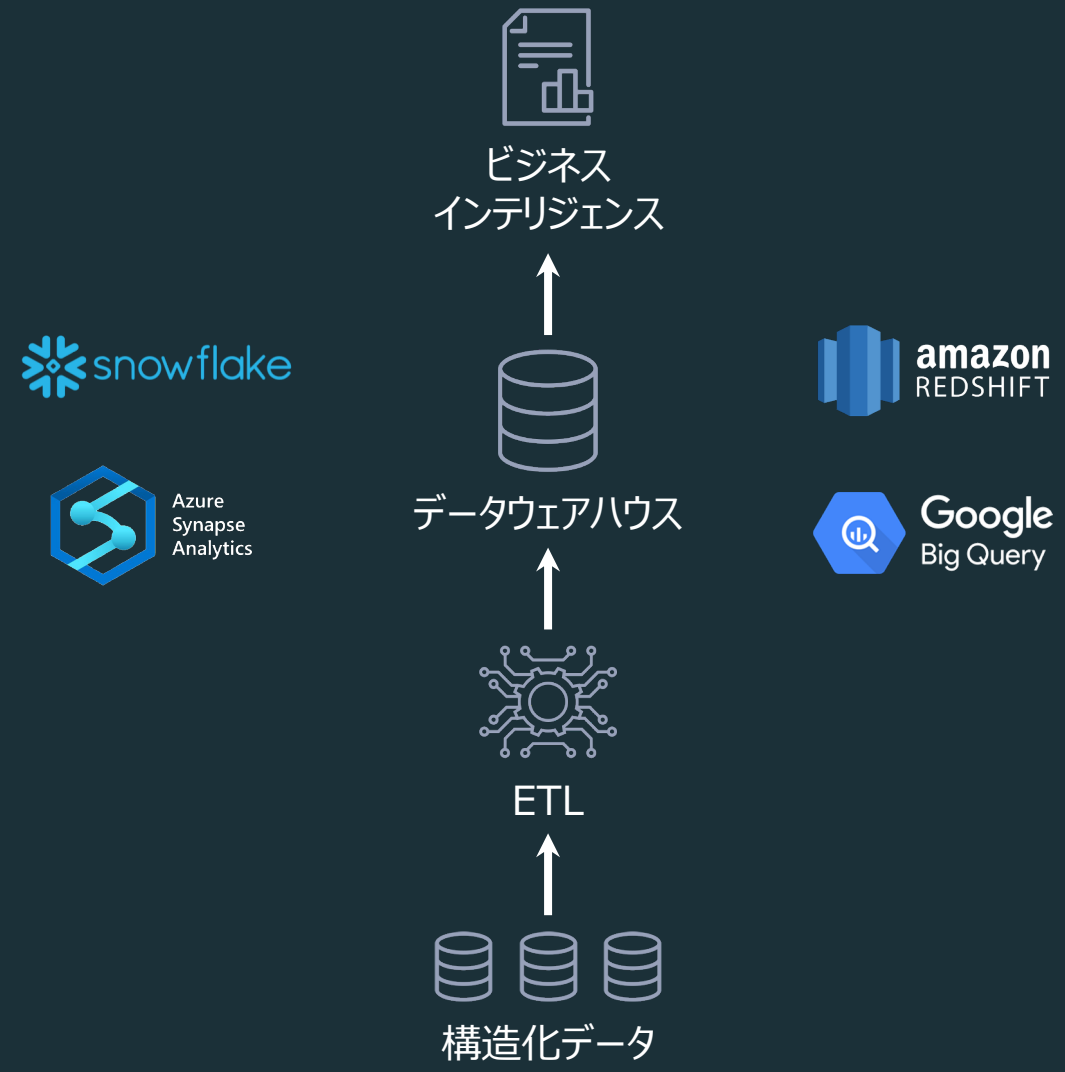
Gartner

87% のデータサイエンスプロジェクトが本格運用に到達していません

VB

データプラットフォームの進化の過程

なぜ、AI・機械学習プロジェクトが暗礁に乗り上げてしまうのか？



データウェアハウス 1990年代～

Pros

- 偉大なるビジネスインテリジェンス(BI)アプリケーション

Cons

- 限定的な機械学習(ML)のサポート
- SQLインタフェースのみを備えたプロプライエタリなシステム

データレイク 2010年代～

Pros

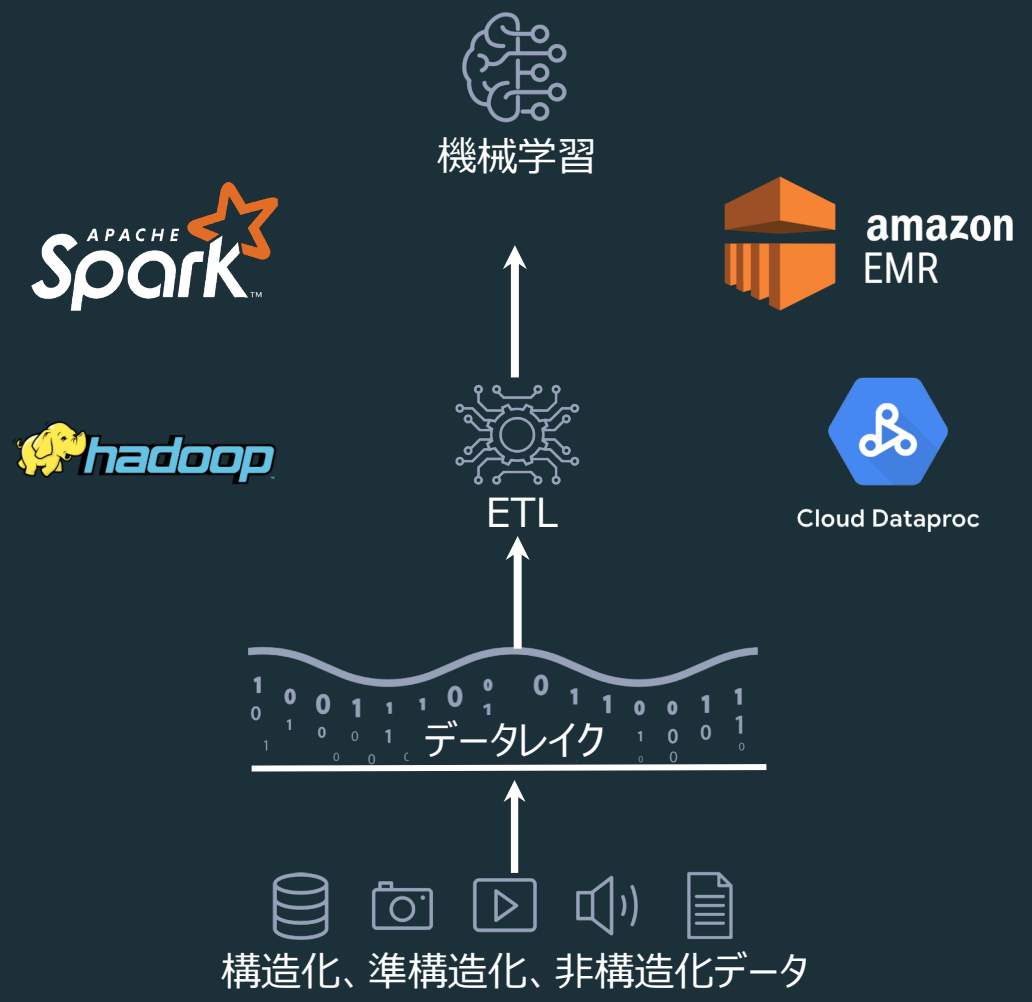
MLをサポート

オープンなフォーマット、
巨大なエコシステム

Cons

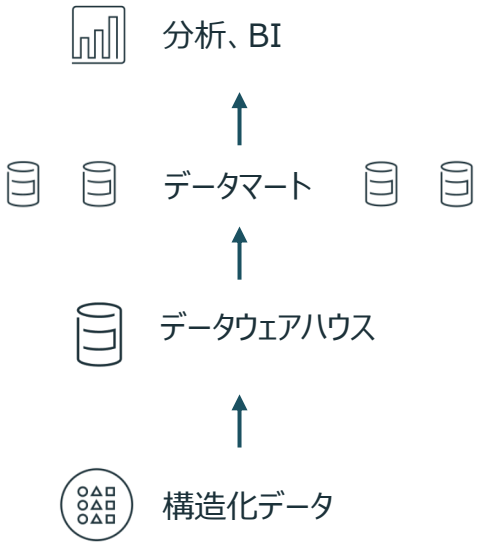
貧弱なBIサポート

複雑化したデータ品質問題

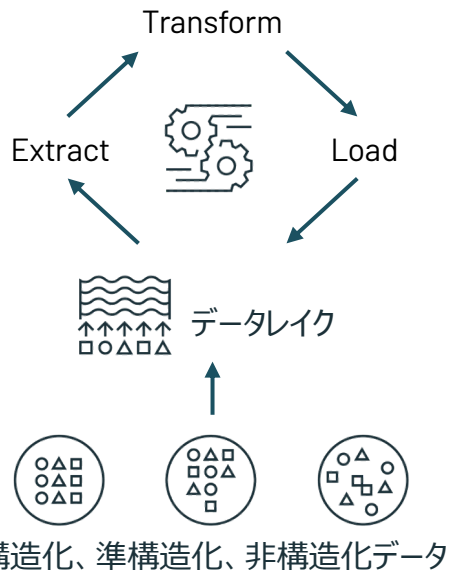


そして現在 多くの企業がデータウェアハウス・データレイクと格闘しています

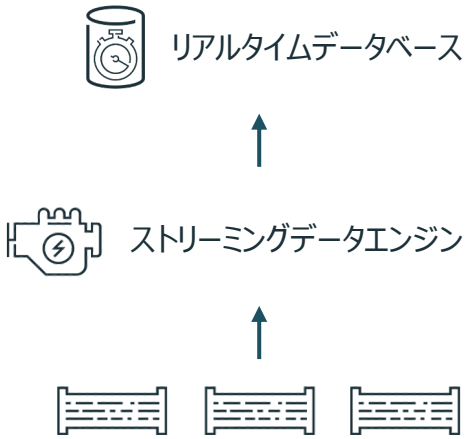
データウェアハウス



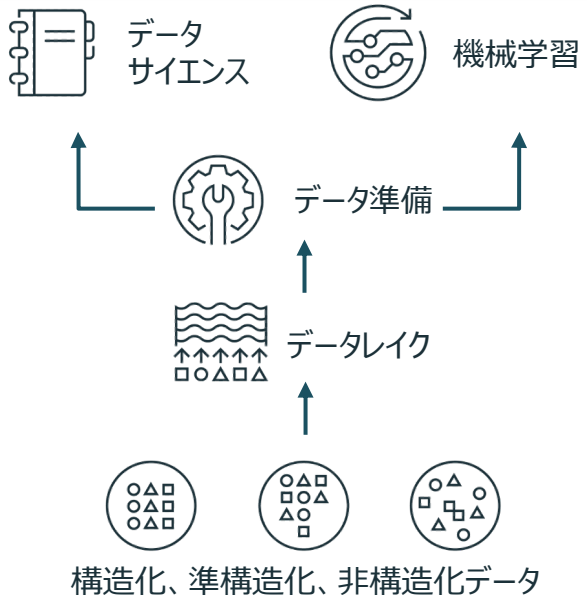
データエンジニアリング



ストリーミング



データサイエンス & 機械学習



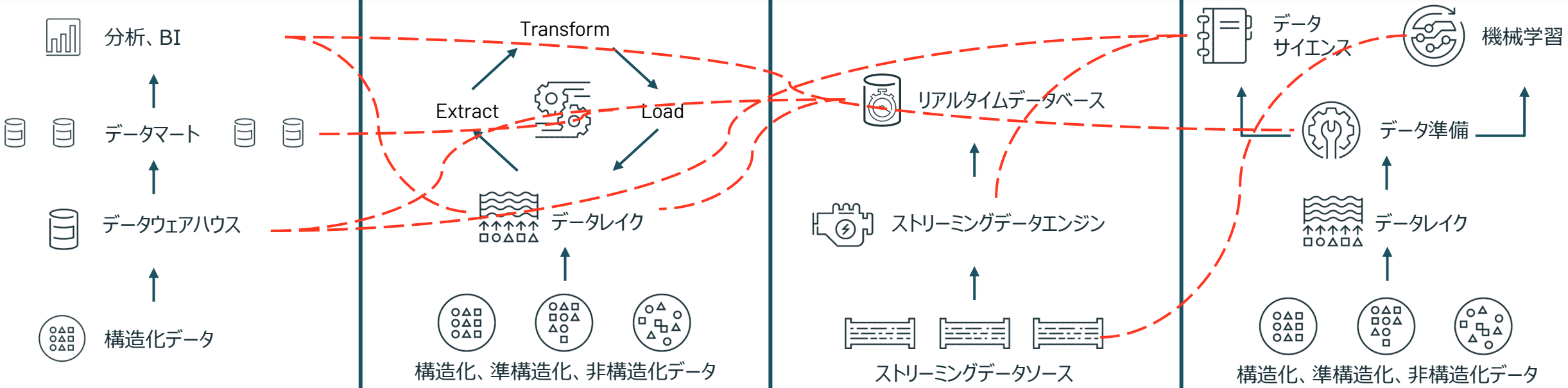
サイロ化はデータアーキテクチャをより複雑にしています

そして現在 多くの企業がデータウェアハウス・データレイクと格闘しています

データウェアハウス		データエンジニアリング		ストリーミング		データサイエンス & 機械学習	
Amazon Redshift	Teradata	Hadoop	Apache Airflow	Apache Kafka	Apache Spark	Jupyter	Amazon SageMaker
Azure Synapse	Google BigQuery	Amazon EMR	Apache Spark	Apache Flink	Amazon Kinesis	Azure ML Studio	MatLAB
Snowflake	IBM Db2	Google Dataproc	Cloudera	Azure Stream Analytics	Google Dataflow	Domino Data Labs	SAS
SAP	Oracle Autonomous Data Warehouse			Tibco Spotfire	Confluent	TensorFlow	PyTorch

断絶したシステムとプロプライエタリなデータフォーマットはシステムの統合を妨げています

サイロ化はデータアーキテクチャをより複雑にしています



そして現在 多くの企業がデータウェアハウス・データレイクと格闘しています

データウェアハウス	データエンジニアリング	ストリーミング	データサイエンス & 機械学習

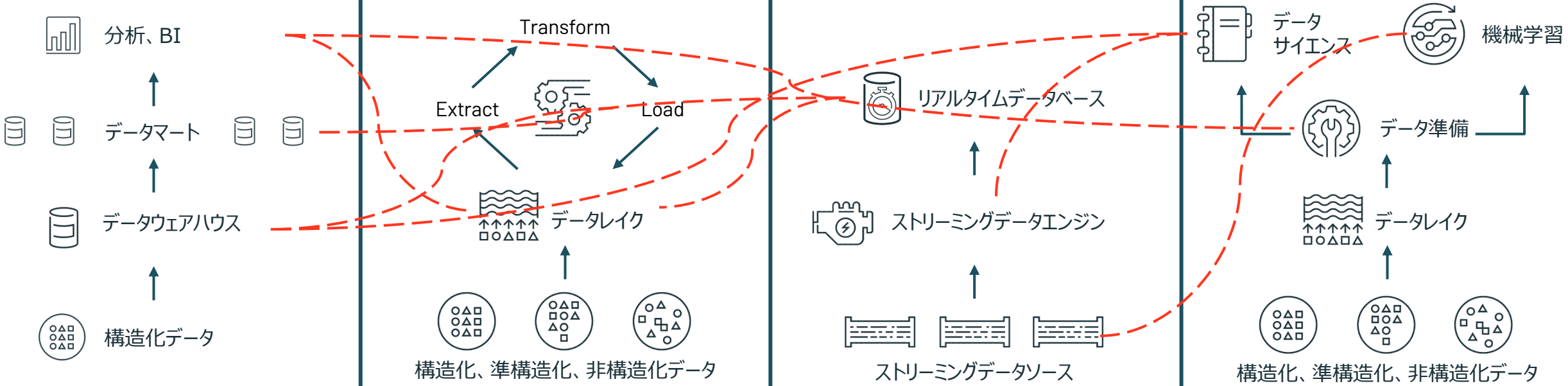
サイロ化したチームの生産性は低下します

データアナリスト | データエンジニア | データエンジニア | データサイエンティスト

断絶したシステムとプロプライエタリなデータフォーマットはシステムの統合を妨げています

Amazon Redshift	Teradata	Hadoop	Apache Airflow	Apache Kafka	Apache Spark	Jupyter	Amazon SageMaker
Azure Synapse	Google BigQuery	Amazon EMR	Apache Spark	Apache Flink	Amazon Kinesis	Azure ML Studio	MatLAB
Snowflake	IBM Db2	Google Dataproc	Cloudera	Azure Stream Analytics	Google Dataflow	Domino Data Labs	SAS
SAP	Oracle Autonomous Data Warehouse			Tibco Spotfire	Confluent	TensorFlow	PyTorch

サイロ化はデータアーキテクチャをより複雑にしています

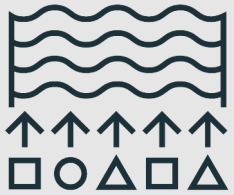


アジェンダ

- 会社概要
- AI・機械学習プロジェクトにおける課題
- レイクハウスプラットフォーム

databricks

データ
レイク



レイクハウス

データ、分析、AIに関わるすべての活動を
統合するプラットフォーム

データ
ウェアハウス


databricks
レイクハウスプラットフォーム
シンプル
オープン
コラボレーティブ
データエンジニアリング
BI & SQL
アナリティクス
リアルタイムデータ
アプリケーション
データサイエンス &
機械学習
データマネジメント & ガバナンス
DELTA LAKE
オープンなデータレイク
構造化データ
準構造化データ
非構造化データ
ストリーミング
aws
Microsoft
Azure
Google Cloud

Databricksの提供する機能

全データ及び機械学習ライフサイクルをサポートする レイクハウスプラットフォーム

1 高品質なデータにアクセス可能

APACHE Spark™

DELTA LAKE

re'dash

さまざまなデータソースから集約された
高品質なデータセットにアクセス可能

2 データサイエンスチームの生産性向上

Databricks
データサイエンス
ワークスペース
(MLランタイム上で稼働)

1つのプラットフォームにて多種多様な
ツール・言語・フレームワークを
利用可能

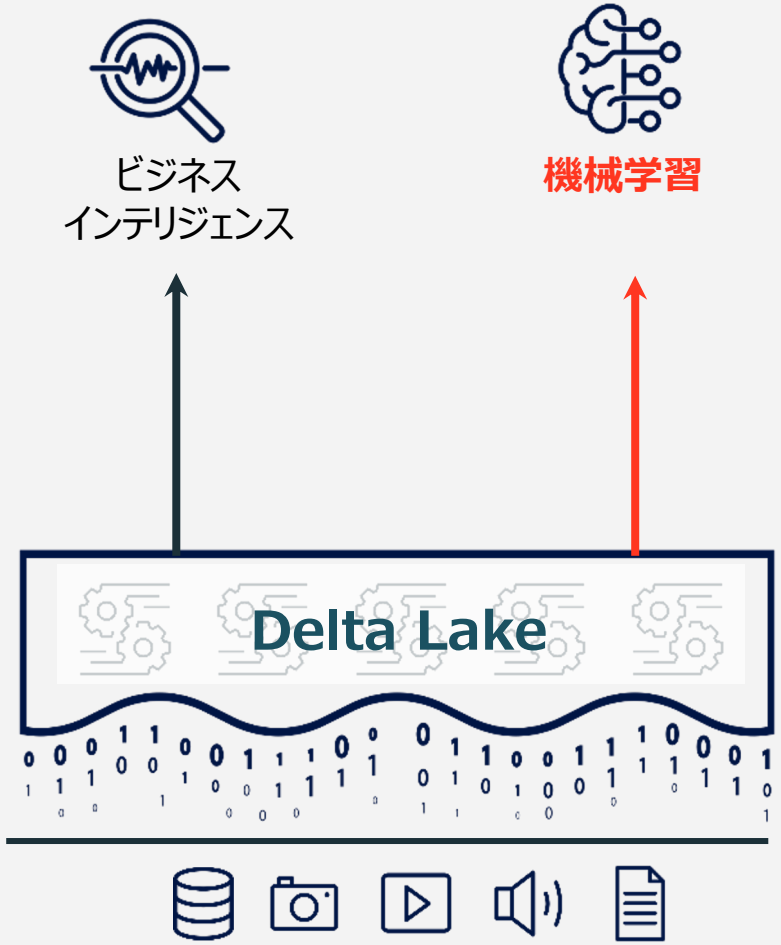
3 標準化された機械学習ライフサイクル

mlflow™

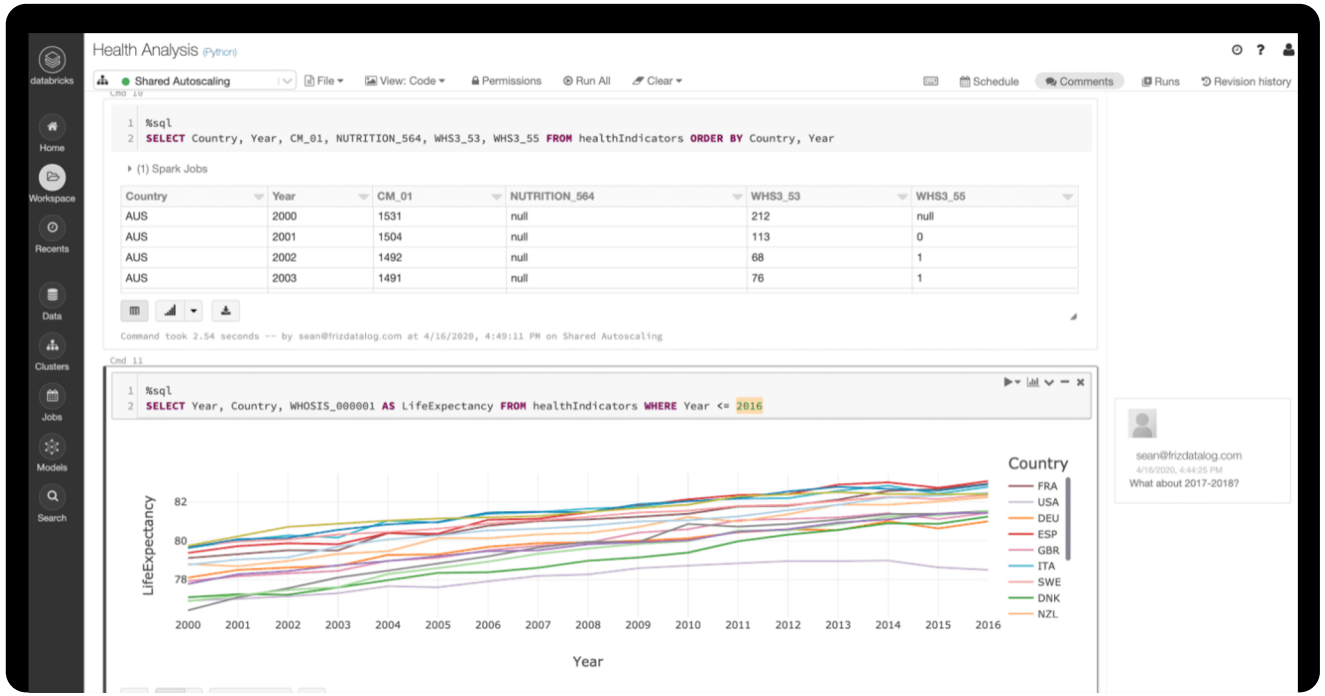
機械学習モデルをステージングから本番環境へ
シームレス・セキュアに連携可能

databricks

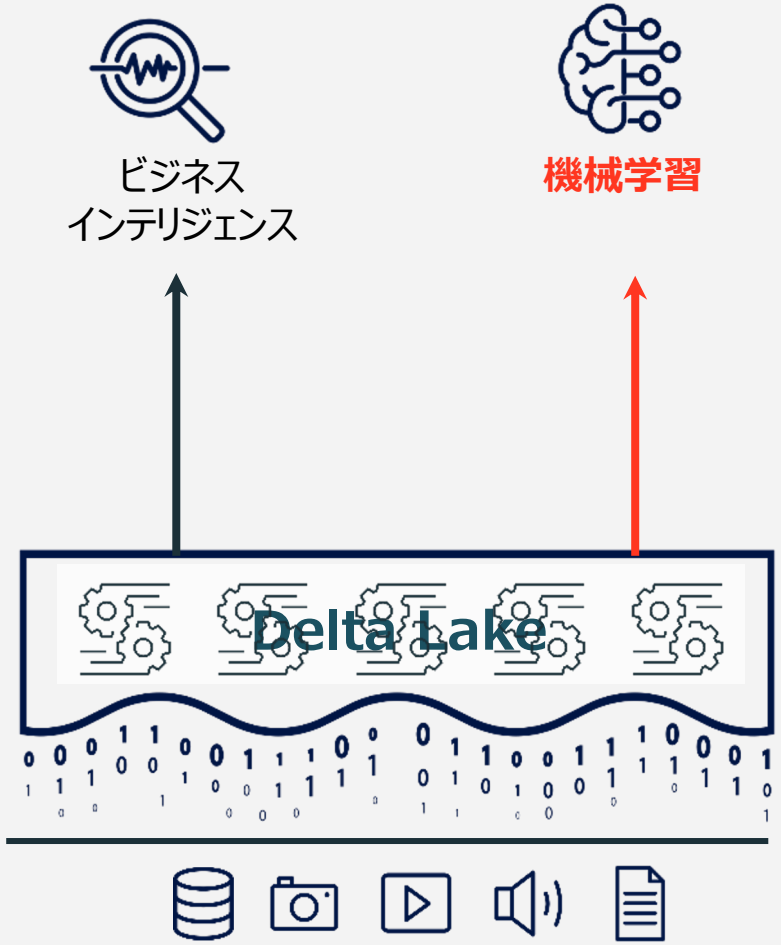
レイクハウスにおけるデータサイエンス



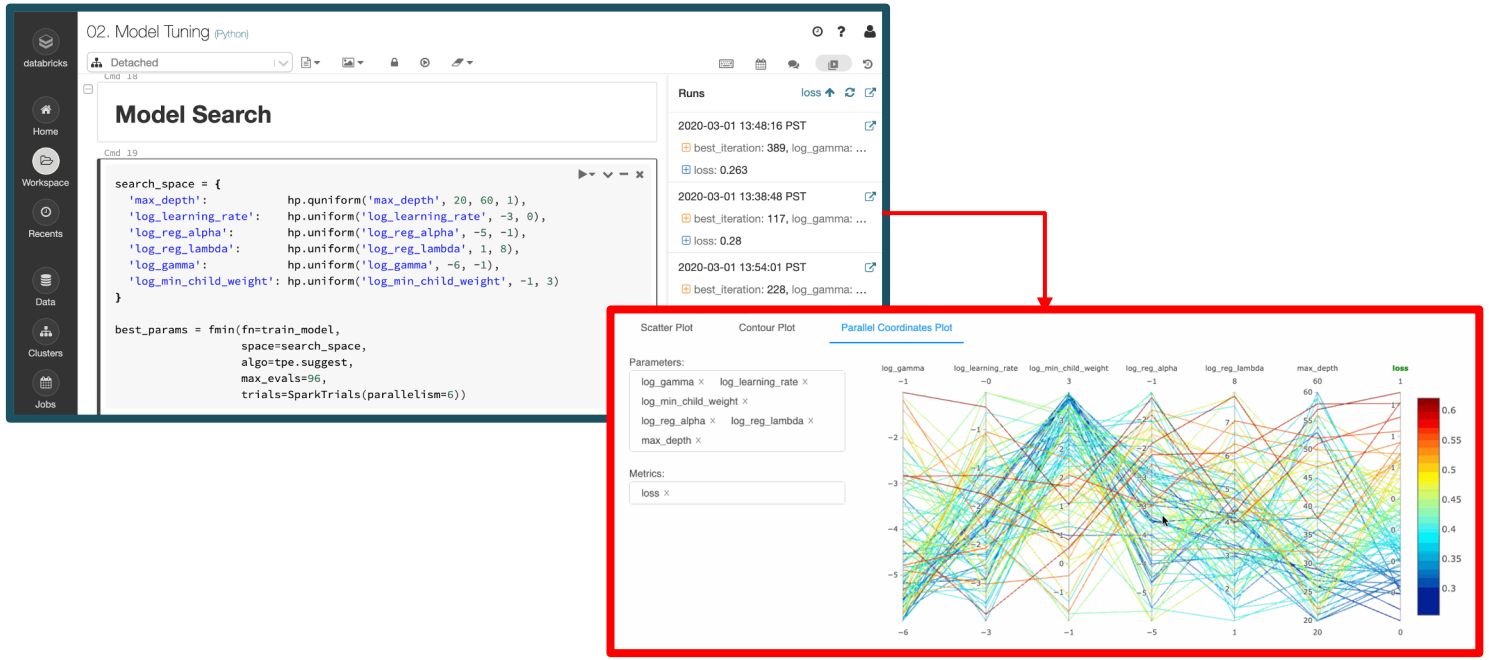
- インタラクティブな分析に適したコラボレーティブノートブックとダッシュボード
- Python、Java、R、Scala、SQLのネイティブサポート
- Delta Lakeデータのネイティブサポート



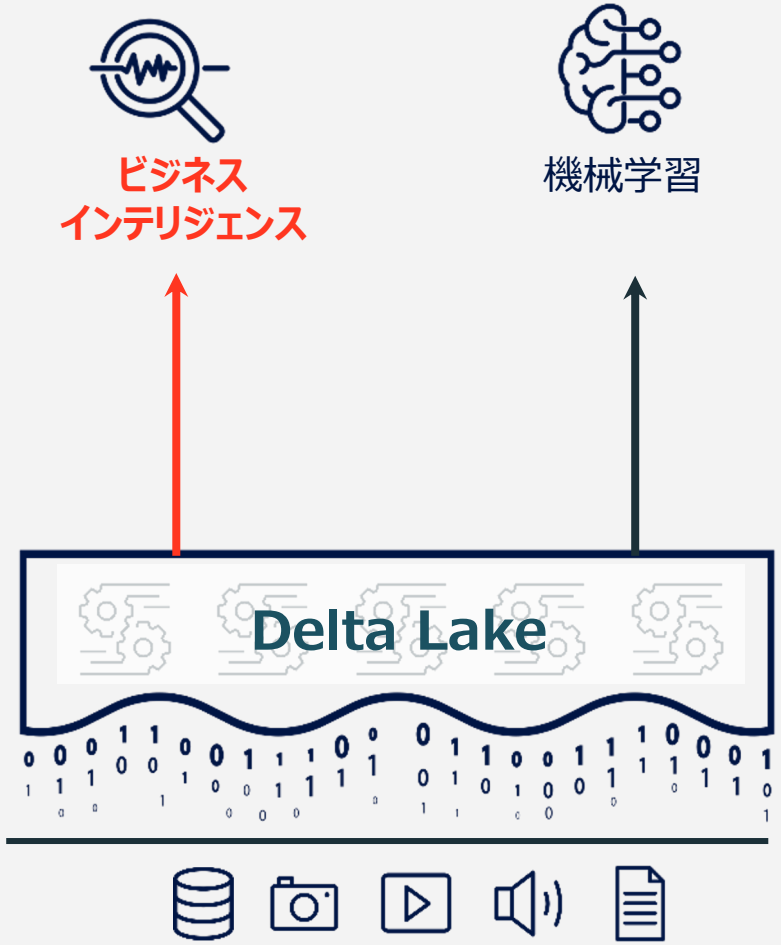
レイクハウスにおける機械学習



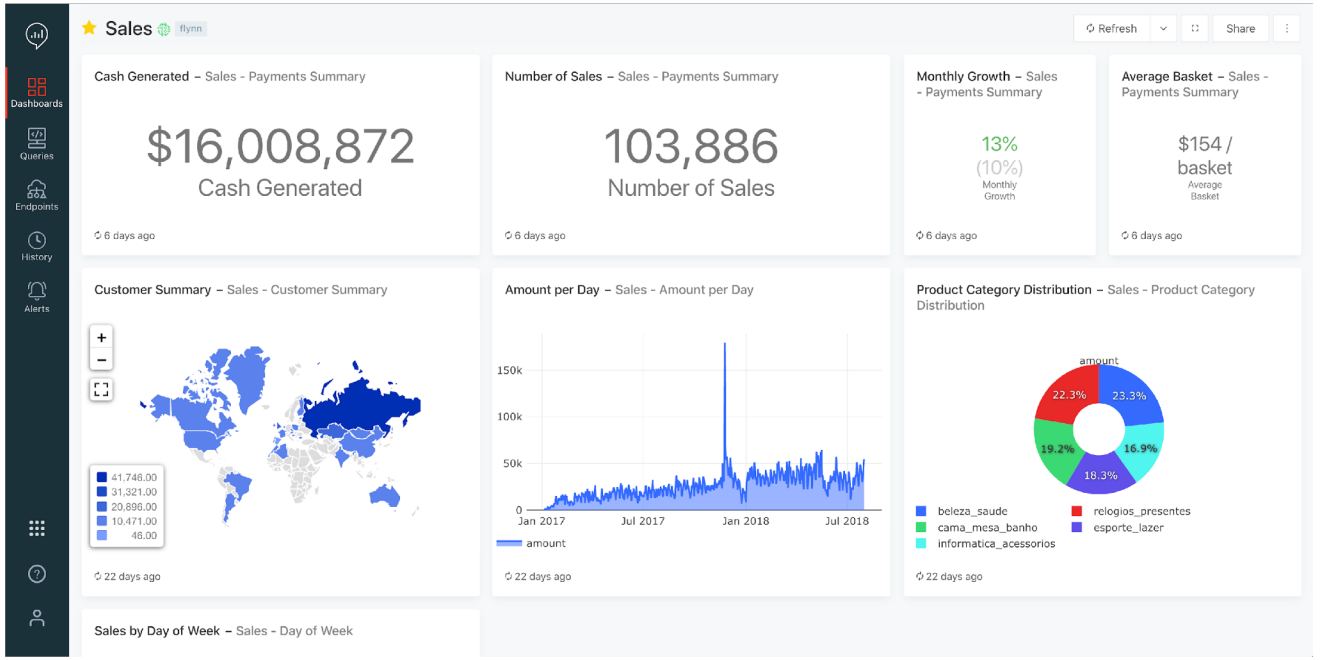
- モデルレジストリ、モデルの再現性、プロダクションへの移行
- Delta Lakeのデータ履歴を活用した再現性の確保



レイクハウスにおけるBI



- Delta Lakeにより実現される高性能なBI、Databricks SQL
- SQLインタフェースのネイティブサポート及びBIツールによるDeltaへの直接接続をサポート



ハンズオンの概要

データブリックスを使ってみよう！

目的：
データブリックスワークスペース上でハンズオンを実施いただき、基本操作を体験していただく

対象：
データ分析に携わる方

アジェンダ：
ワークスペースにおけるノートブックの操作を通じて、データブリックスのメリットを体験いただきます。

1. 探索的データ分析(EDA)、ETL、機械学習モデル構築、モデル配備、BIまでを一つのプラットフォームで実行できます
2. インフラの管理が簡単です
3. データベースのインストールが不要です
4. リモートワークでのコラボレーションが簡単です
5. ノートブックのバージョンが自動で管理されます
6. プログラミング言語を柔軟に切り替えることができます
7. データを簡単に可視化できます
8. 大量データを高速に処理できます

JEDAI
Japan Enduser Group
Databricks Innovation

JEDAIは、データブリックスを最大限ご活用いただくための有益な情報をご提供するとともに、ユーザー同士がつながり、関係を深めることができる場として活動いたします。
2021年は5回の開催を予定しています。ぜひお気軽にご参加ください。

Community Guide

データブリックス・ジャパン株式会社
Senior Customer Success Engineer

徳元 大輔

通信事業者で様々な業務を経験した後にビッグデータ業界に。現在は Databricks Japanでポストセールスの頼れるなんでも屋さんを目指している。趣味は、飲み食べ歩き・キックボクシングと過度なエクササイズ・海外SF小説。好きな映画：ブレードランナー、パルプフィクション。座右の銘：無欲は怠惰の元である。

データブリックス・ジャパン株式会社
TEL.03-6821-1670
E-mail:marketing-jp@databricks.com

プログラム概要

お客様セッション

5・7・9・11・1月開催（予定）

データブリックスをご利用頂いてるお客様企業やデータ＆AIのプロフェッショナルをお招きして、実際現場で苦悩されている点や、さらには普段他では話すことのできないハプニング、ココでしか聞けない開発秘話など、存分にお話いただきます。
セッションの他にも、Q＆Aの時間をたっぷり設け、オンラインの枠を超えた、できる限りインタラクティブなコミュニケーションが取れるようにいたします。

テーマ別セッション

6・8・10・12月開催（予定）

データプロジェクトにおける世界の最新トレンドからAI/機械学習プロジェクトの実運用まで、毎回テーマを設定し、弊社のパートナー様や弊社のデータ&AI プロフェッショナルが、ご説明します。セッションの他にも、Q＆Aの時間をたっぷり設け、オンラインの枠を超えた、できる限りインタラクティブなコミュニケーションが取れるようにいたします。

コミュニティへの登録方法

コミュニティの運営は「Connpass」を利用いたします。
こちらの登録は、Eメールアドレスに加え、各種ソーシャルアカウント（Twitter, Facebook）との連携が可能です。

ご登録はこちら＞　https://jedai.connpass.com/

Thank you