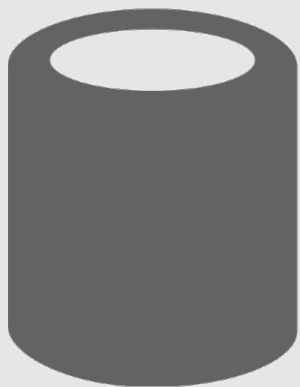


Azure Databricks Workshop

女部田啓太

日本マイクロソフト株式会社

データ分析と分析プラットフォームの変遷



-
- ✓ データ量の増加によるリソースの枯渇
 - ✓ 多様なデータへの対応に難あり

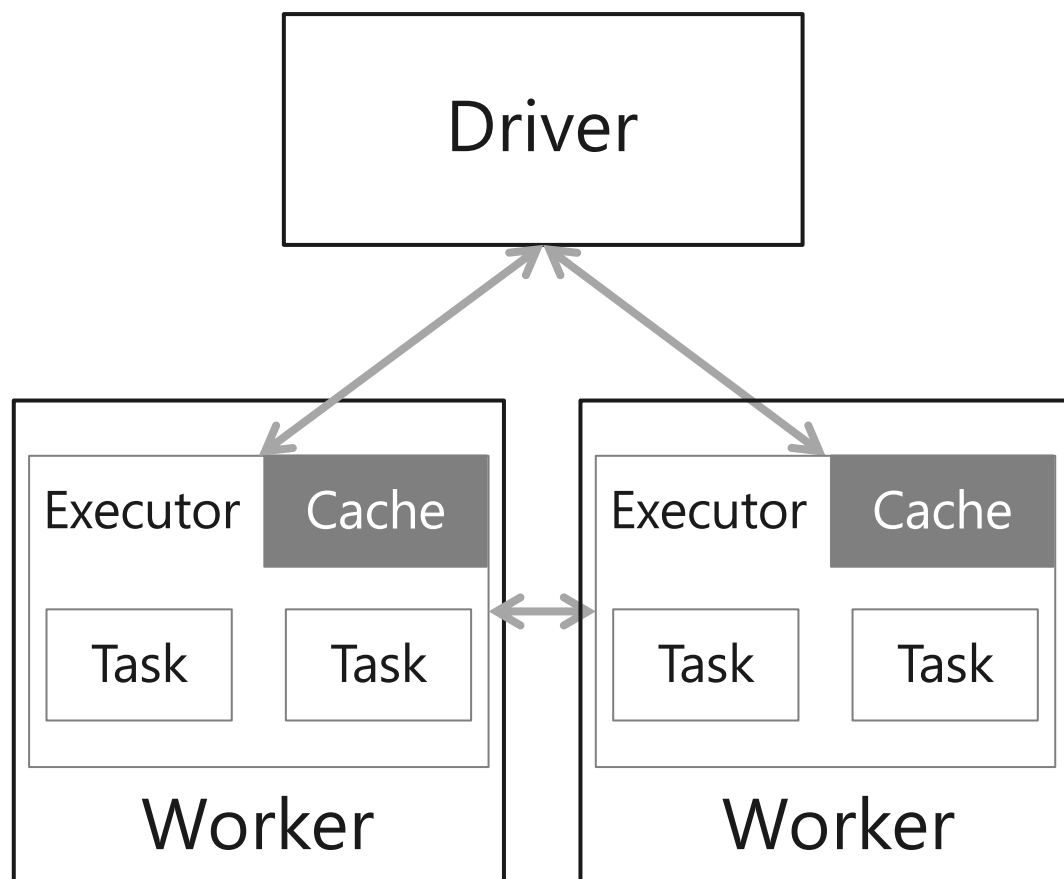
-
- ✓ データ分散処理基盤
 - ✓ レイテンシー大
 - ✓ バッチ処理向き

-
- ✓ インメモリ処理
 - ✓ 様々な分析ワークロードに対応可能

Apache Spark



Spark Architecture Overview



2009 : UC バークレー AMPLab にて
スタート

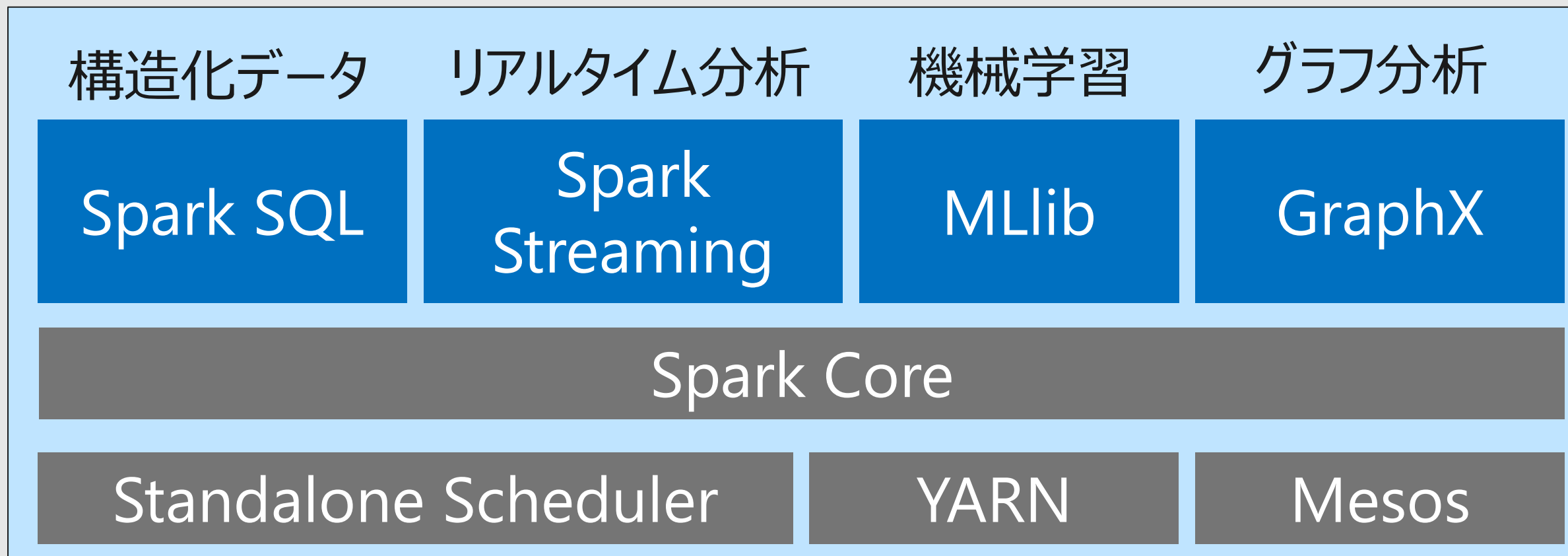
2010 : オープンソース化

2013 : Apache トッププロジェクトへ

データセットのキャッシュにより、
反復処理の多い、機械学習で
効果を発揮

Apache Spark 基本コンポーネント

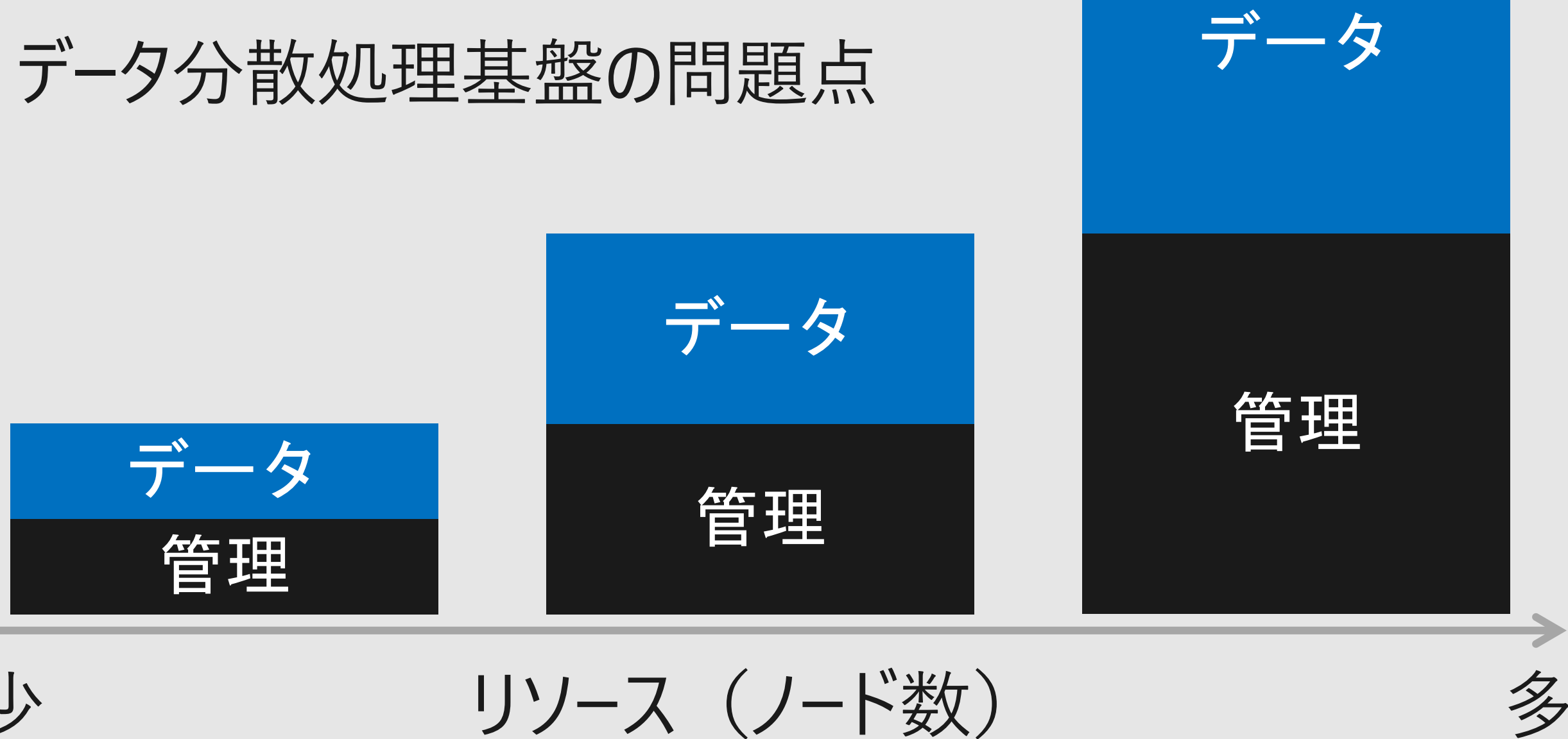
- データ分析に必要な機能を 1 プラットフォームで提供
- Python、R、SQL、Scala、Java が利用可能



データ分散処理基盤の問題点



データ分散処理基盤の問題点



管理負荷の増加が分析の大きな障壁

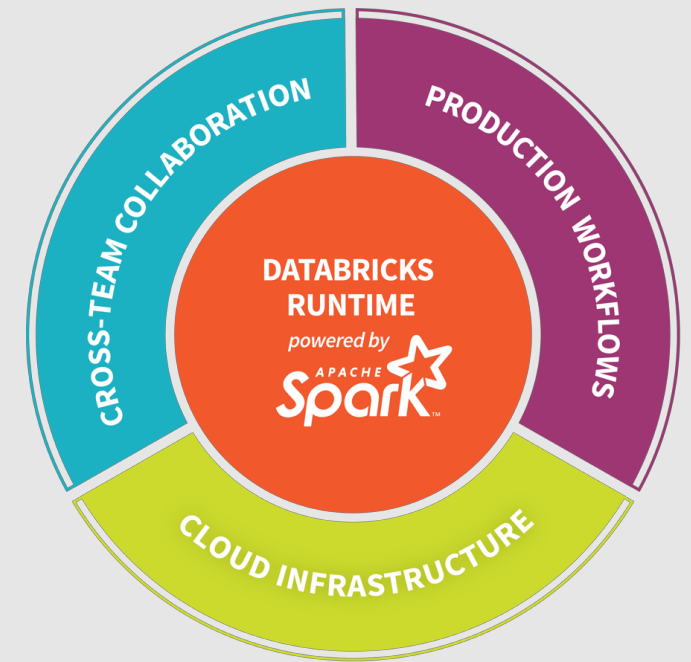
分析に“集中できる”データ分散処理基盤



必要最低限の管理で分析に集中

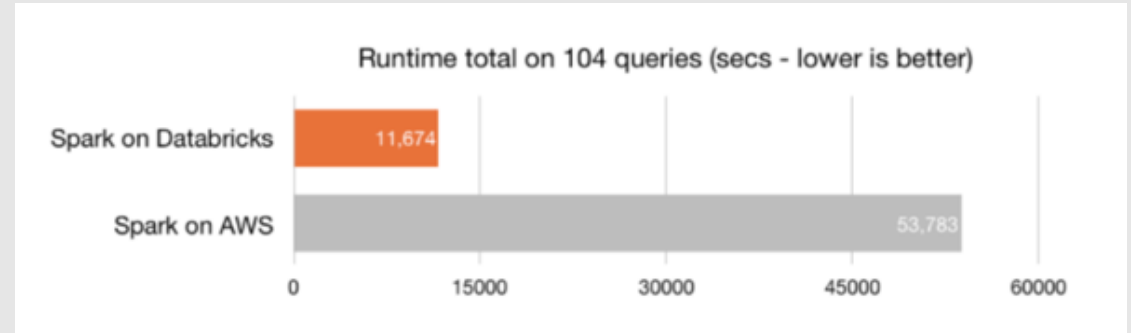
Databricks

- UC バークレー AMP Lab メンバーの Apache Spark 開発チームによって 2013 年に設立
- OSS Apache Spark の最大コントリビューター (75%)
- Spark をベースとした “Unified Analytics Platform” を提供
 - Databricks Runtime
 - Databricks Delta
 - Databricks Collaborative Notebooks

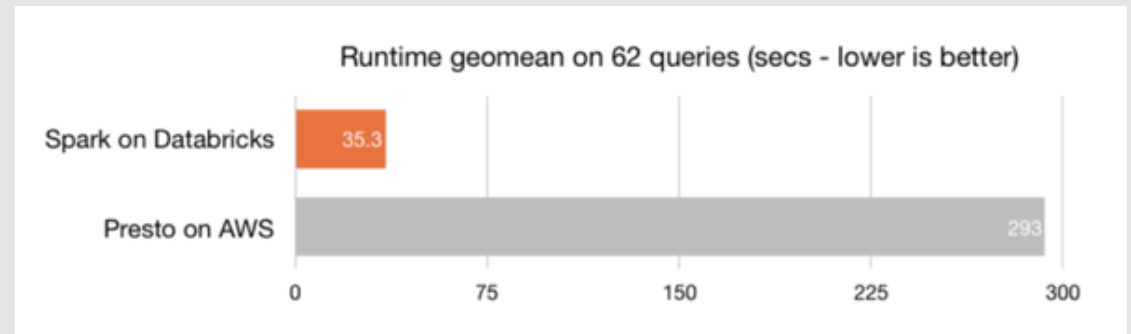


Databricks パフォーマンス - TPC-DS(DBR3.0)

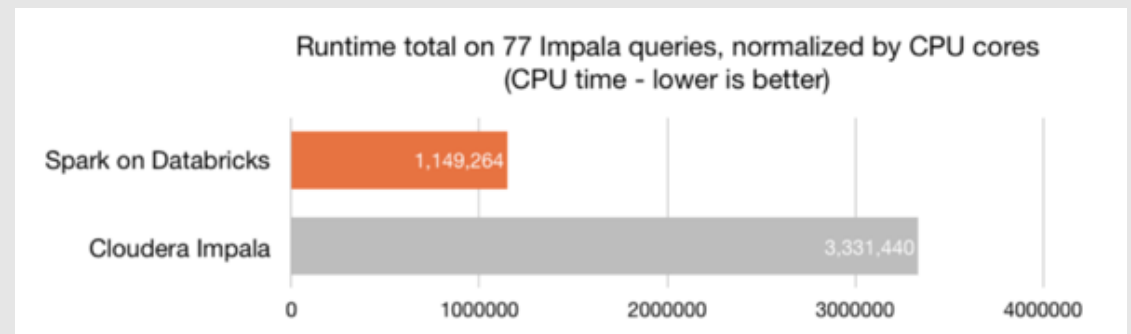
OSS Spark の **5** 倍



Presto の **8** 倍



Impala の **3** 倍



Azure Databricks

Azure の 1st パーティの PaaS サービスとしてご提供



Azure Databricks

Spark SQL

Spark
Streaming

MLlib

GraphX

Spark Core API

R

SQL

Python

Scala

Java

マネージドサービス

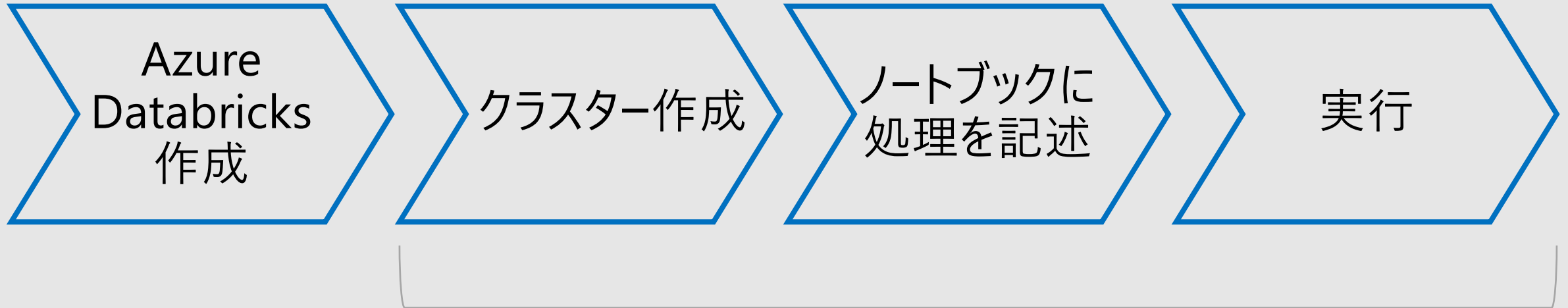
エンタープライズ
セキュリティ

データサービスとの連携

Azure Databricks 基本操作

Azure Databricks 利用の流れ


Azure Portal



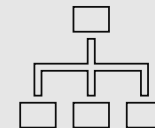
Azure Databricks





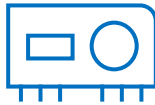









Azure Databricks 基本機能一覧

	Workspace	<ul style="list-style-type: none">✓ フォルダ構造✓ ノートブックやライブラリを保存
	Cluster	<ul style="list-style-type: none">✓ クラスターの作成、編集、クローンなどの管理✓ オートスケール機能や自動停止機能を設定
	Notebook	<ul style="list-style-type: none">✓ クラスターに対して実行する処理の記述✓ 複数人でコラボレーション機能
	Job	<ul style="list-style-type: none">✓ ノートブックの処理をスケジュール実行
	Table	<ul style="list-style-type: none">✓ データを永続化し、テーブルとして格納
	Library	<ul style="list-style-type: none">✓ ライブラリをクラスターにインストール

Cluster

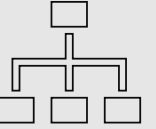


- クラスター構成を選択

クラスターモード	Databricks Runtime	Virtual Machine
<div> Standard</div> <div>OR</div> <div> High Concurrency</div>	<div>Databricks Runtime Version</div> <div></div> <div>Spark Version 2.X.X</div>	<div></div> <div></div> <div></div>

構成はクラスター作成後でも変更可能

Autoscaling と Auto Termination

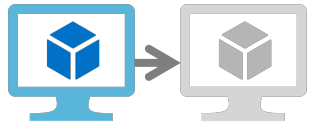


- ムダな稼働を減らして、コストの最適化が可能



Autoscaling

選択したノード数内でリソース負荷状況を基に、自動的にスケールアップ / ダウン



Auto
Termination

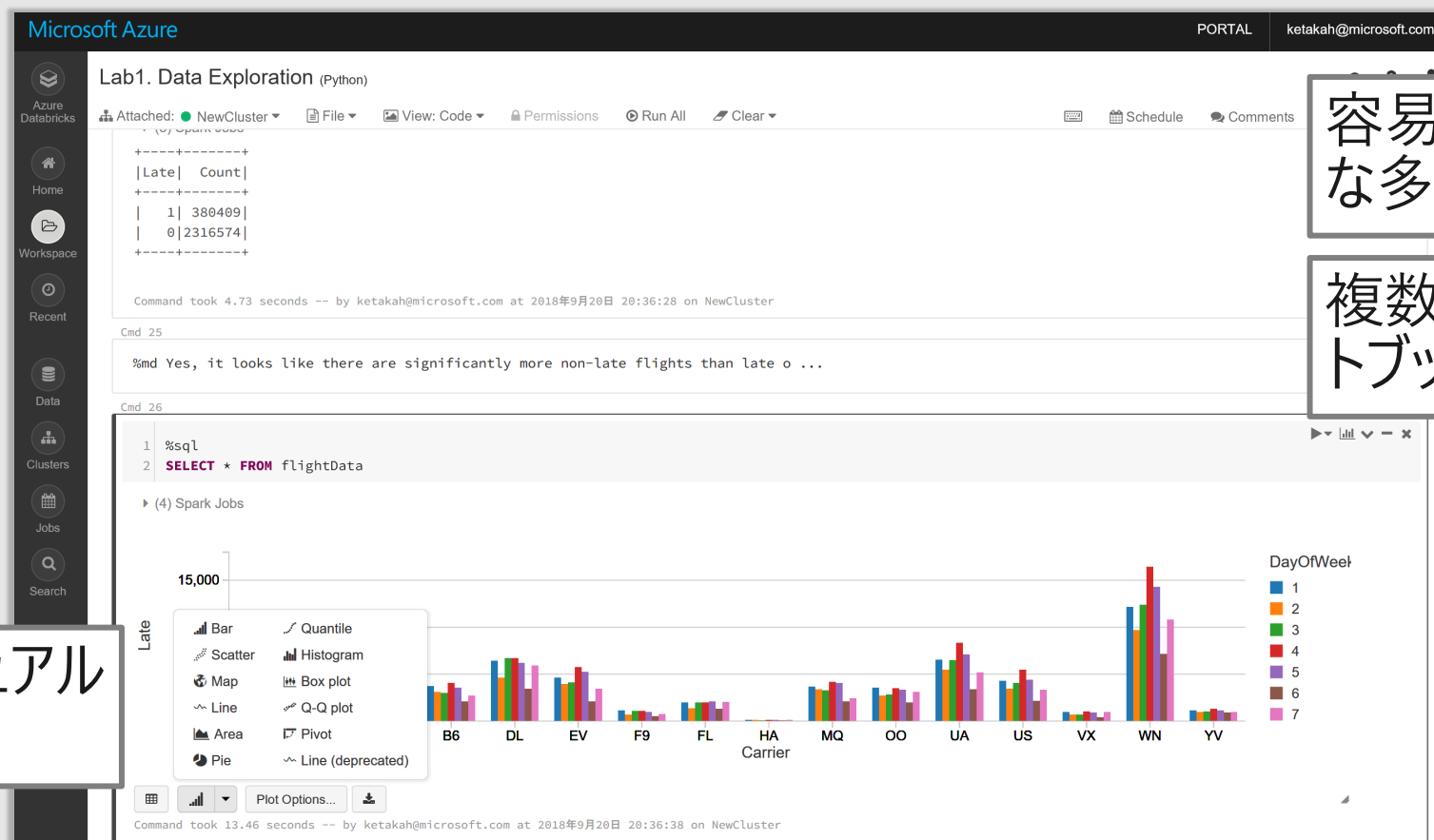
指定した時間の間、クラスターに対して処理が実行されない場合、クラスターを自動的に終了

クラスターの用途に合わせて設定

Notebook



- Jupyter Notebook のコーディングのし易さと Apache Zeppelin のビジュアライゼーション

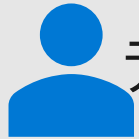


容易にコーディング可能な多機能エディター

複数言語を同一のノートブックで記述可能

結果を複数のビジュアルで表示可能

コラボレーション機能



データサイエンティスト



ビジネスアナリスト



データエンジニア

変更履歴

Comments

Revision history

Git: Not linked

September 22, 22:57 PM JST

● Kenta Takahashi

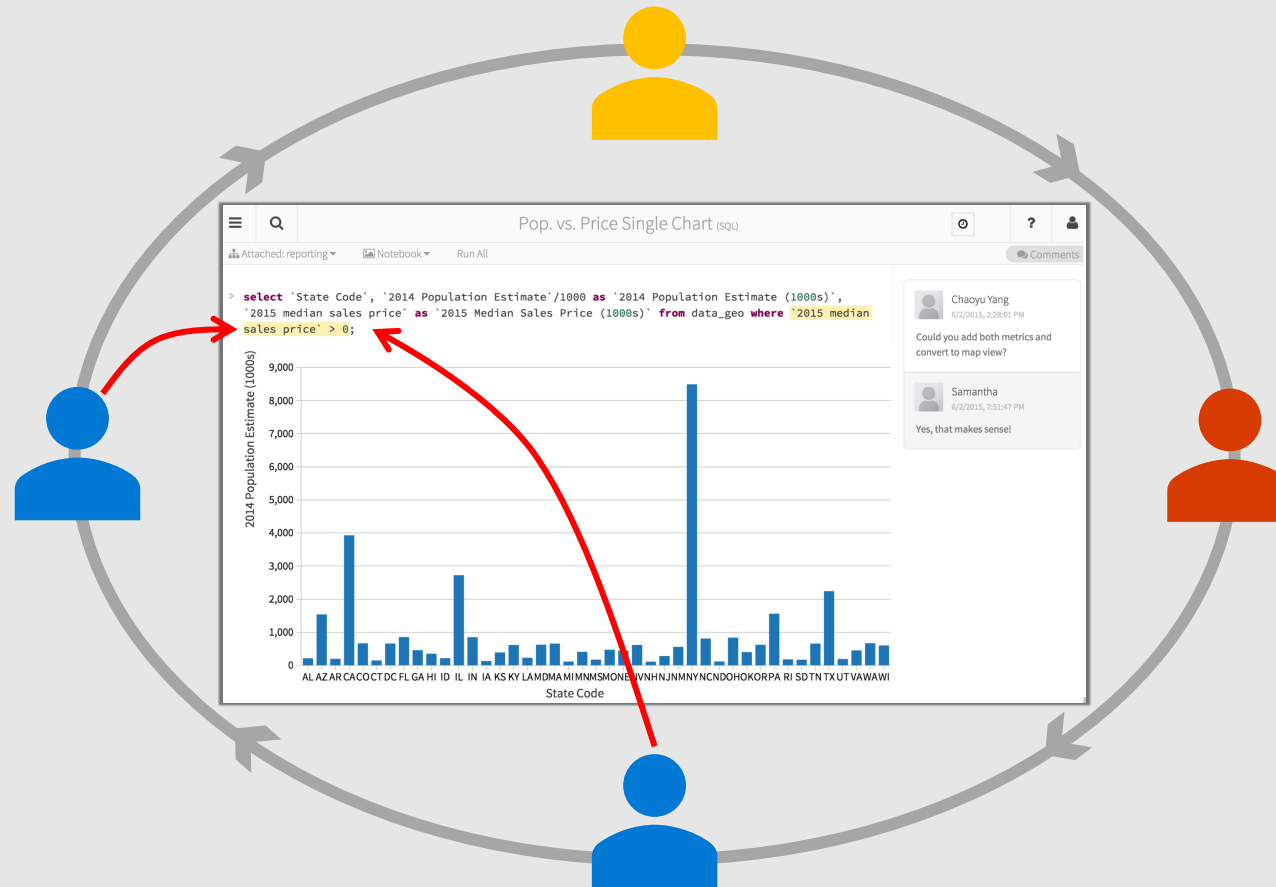
All changes saved [Save now](#)

September 22, 22:54 PM JST

● Kenta Takahashi

September 20, 12:19 PM JST

● Kenta Takahashi



コード単位でコメント (コミュニケーション)

Comments

Chaoyu Yang
6/2/2015, 2:28:01 PM
Could you add both metrics and convert to map view?

Samantha
6/2/2015, 7:51:47 PM
Yes, that makes sense!

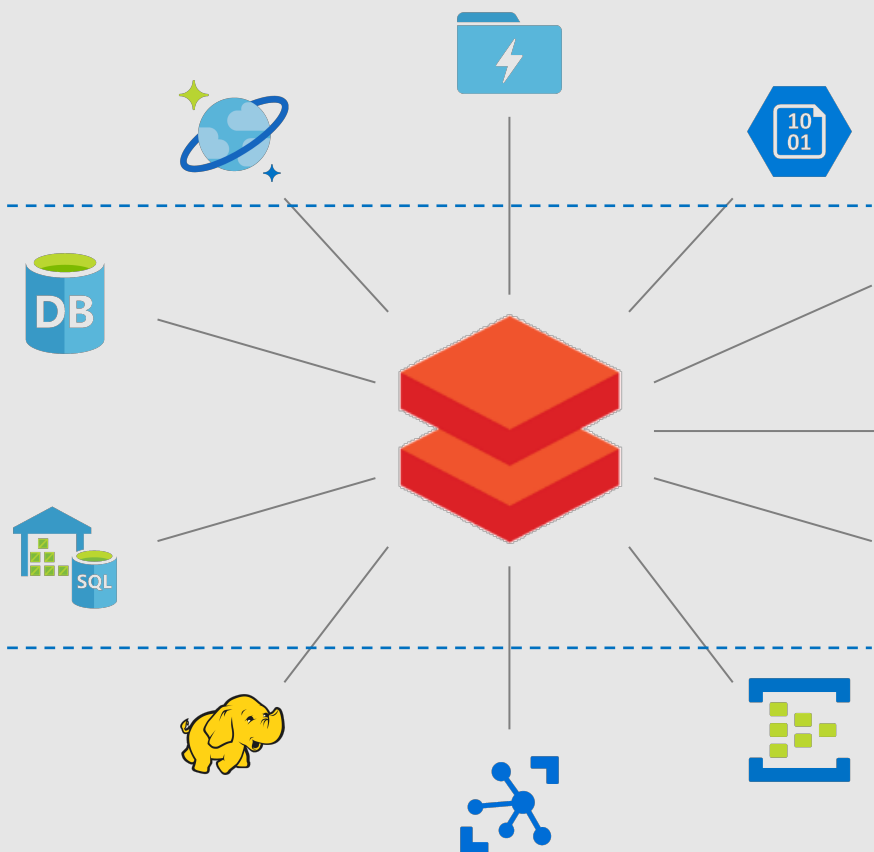
Azure データサービス統合

多様な Azure データサービスとの接続機能を提供

非構造化データ

リレーショナルデータ

ストリーミングデータ



可視化



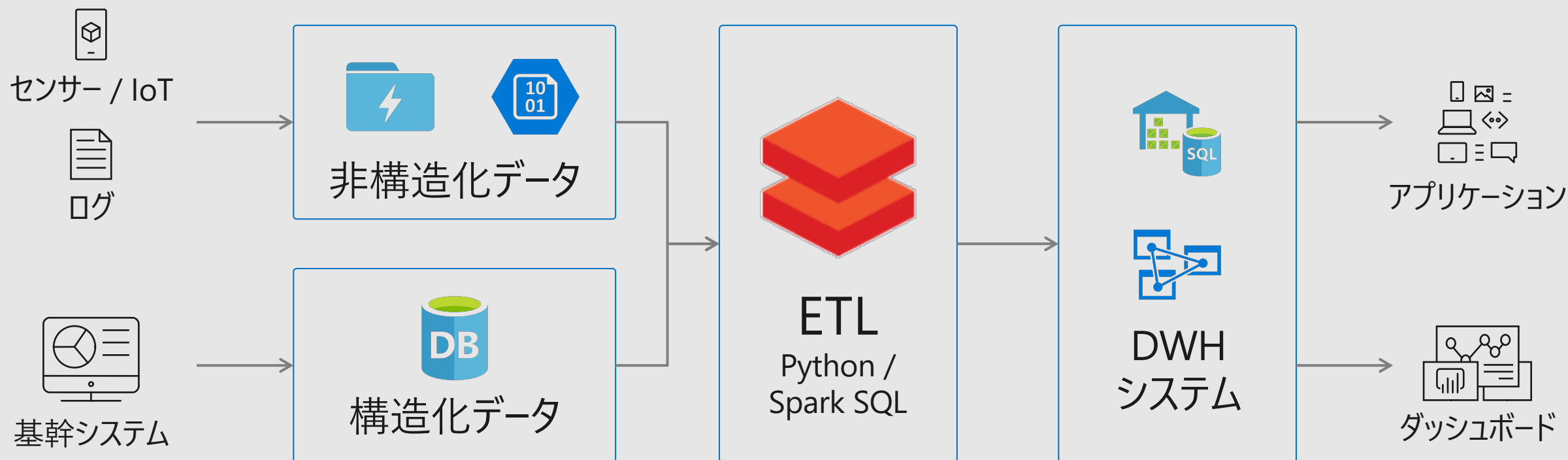
マシンラーニング



データ連携

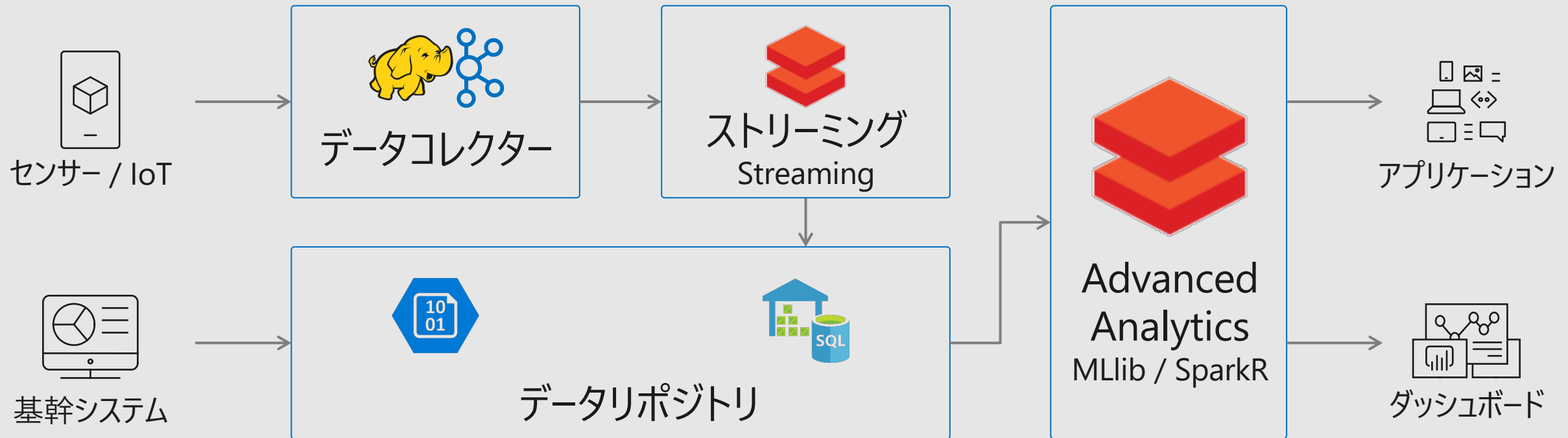
Azure Databricks Use Case – ETL

- 大量データを収集し、加工・集計して DWH システムへロード



Azure Databricks Use Case – Advanced Analytics

- 蓄積された大量データを活用して高度な分析を実施



Machine Learning on Azure Databricks

Advanced Analytics



Azure Infrastructure



- クラスター上の大容量のデータセットに対して、並列分散でマシンラーニングが可能
 - 並列対応された ML アルゴリズムの提供
 - Azure Databricks 上にプリビルド
 - パラメータチューニングの高速化
- 3rd パーティライブラリを利用可能
 - H2O Sparkling Water
 - sciKit-learn
 - XGBoost

Azure Machine Learning service との連携

Advanced Analytics



Azure Databricks



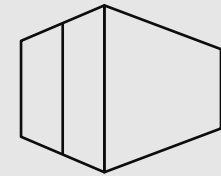
Public Preview
Azure Machine Learning service



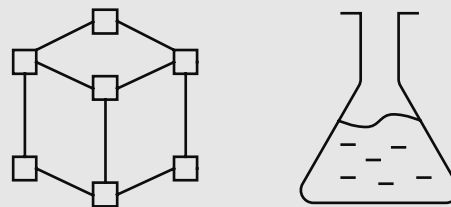
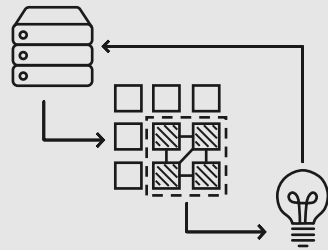
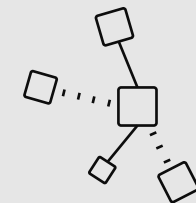
Azure Kubernetes Service / Azure Container Instance



Docker



IoT Edge



- ✓ モデルのトレーニング
- ✓ モデルの評価

- ✓ トレーニング履歴管理
- ✓ モデル管理

ライブラリ AML Service SDK をインポート

ノートブックから、AML service を呼び出す



MICROSOFT CONFIDENTIAL

本資料は情報提供のみを目的としており、本資料に記載されている情報は、本資料作成時点でのマイクロソフトの見解を示したものです。状況等の変化により、内容は変更される場合があります。本資料に表記されている内容（提示されている条件等を含みます）は、貴社との有効な契約を通じて決定されます。それまでは、正式に確定するものではありません。従って、本資料の記載内容とは異なる場合があります。また、本資料に記載されている価格はいずれも、別段の表記がない限り、参考価格となります。貴社の最終的な購入価格は、貴社のリセラー様により決定されます。マイクロソフトは、本資料の情報に対して明示的、黙示的または法的な、いかなる保証も行いません。

© 2019 Microsoft Corporation. All rights reserved.