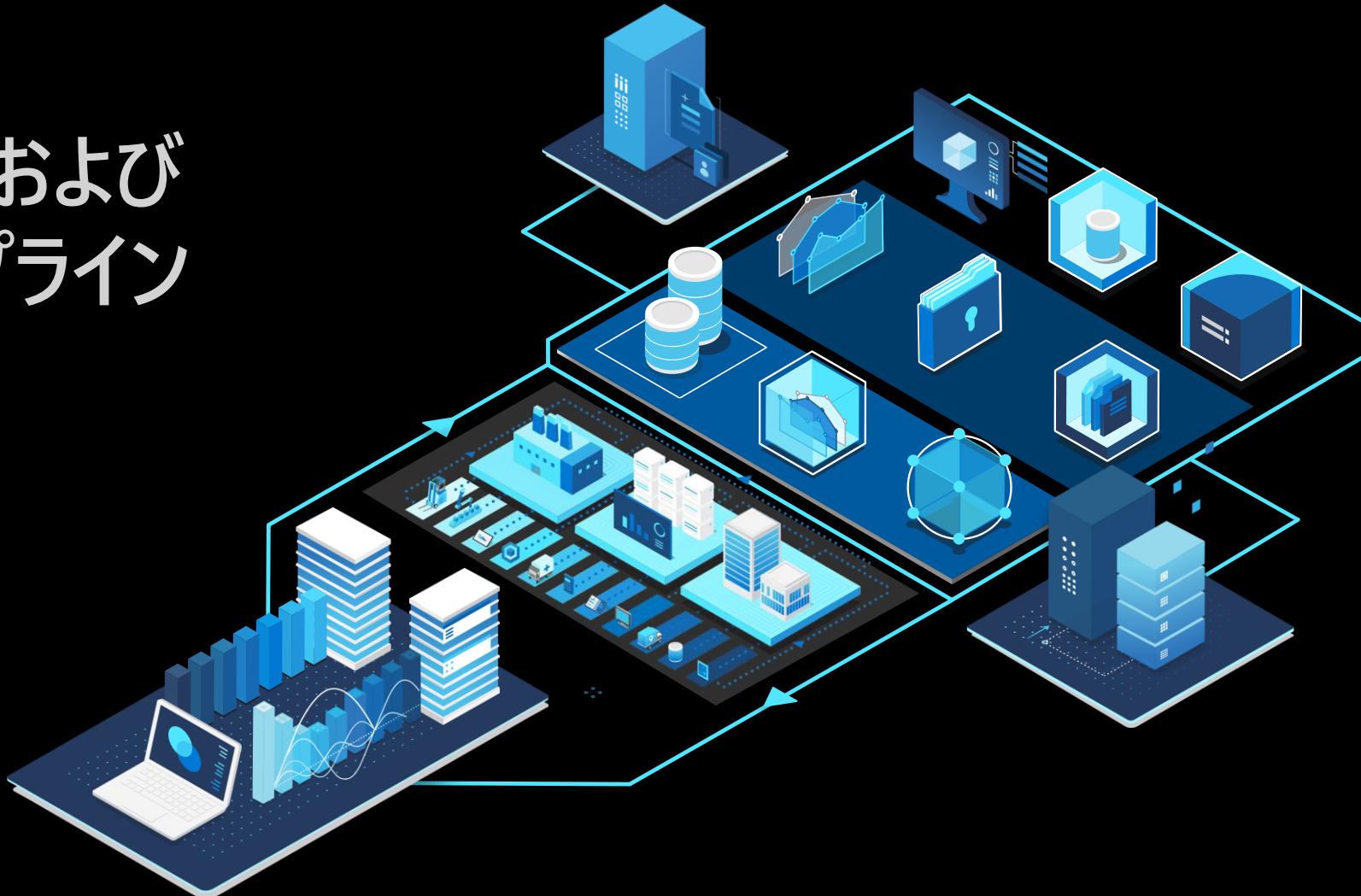


Azure Data Factory および Azure Synapse パイプライン

製品の概要

OpenHack for Lakehouse 向け
ダイジェスト版

2022 年度第 4 四半期



Azure Data Factory

ハイブリッド データ統合、簡素化



使いやすい

- コード不要の ETL/ELT
- 数回のクリックで SSIS を再ホスト
- 組み込みの Git & CI/CD



コスト効率が良い

- 従量制
- フルマネージド、サーバーレス
- オンデマンドでスケーリング



強力

- 組み込みのエンタープライズコネクタ
- 大規模な調整と監視



インテリジェント

- 自律的 ETL
- AI ベースの目的主導のコピー
- 予測パイプライン

すべてのエンタープライズ データに接続

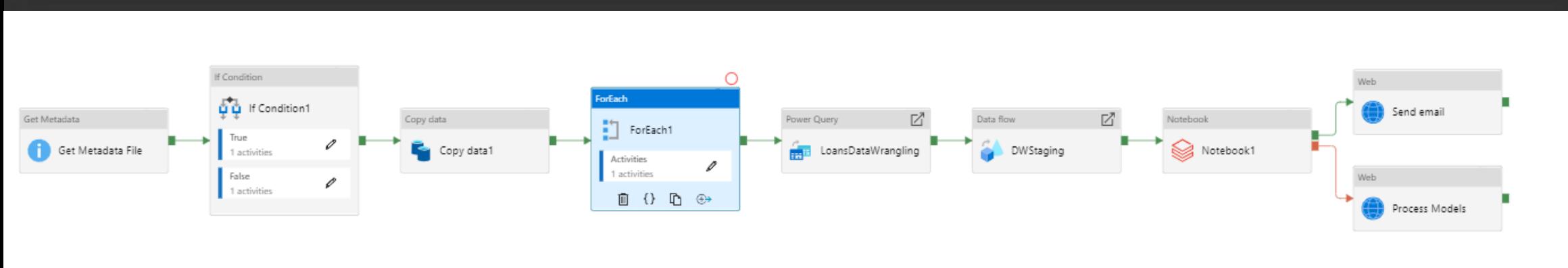


100 以上の組み込みコネクタなど

強力なコード不要のデータパイプラインを構築

ブラウザベースの設計 UI でのサーバーレス オーケストレーション

- データ統合と ETL ワークフローの設計
- ビジネスロジックのサーバーレス実行
- スケジュールまたはイベントベースでの実行のためのトリガーの設定
- ソース管理と CI/CD のための Git リポジトリとの統合
- データファクトリ全体にわたるコラボレーション



コード不要またはコード中心の変換を選択

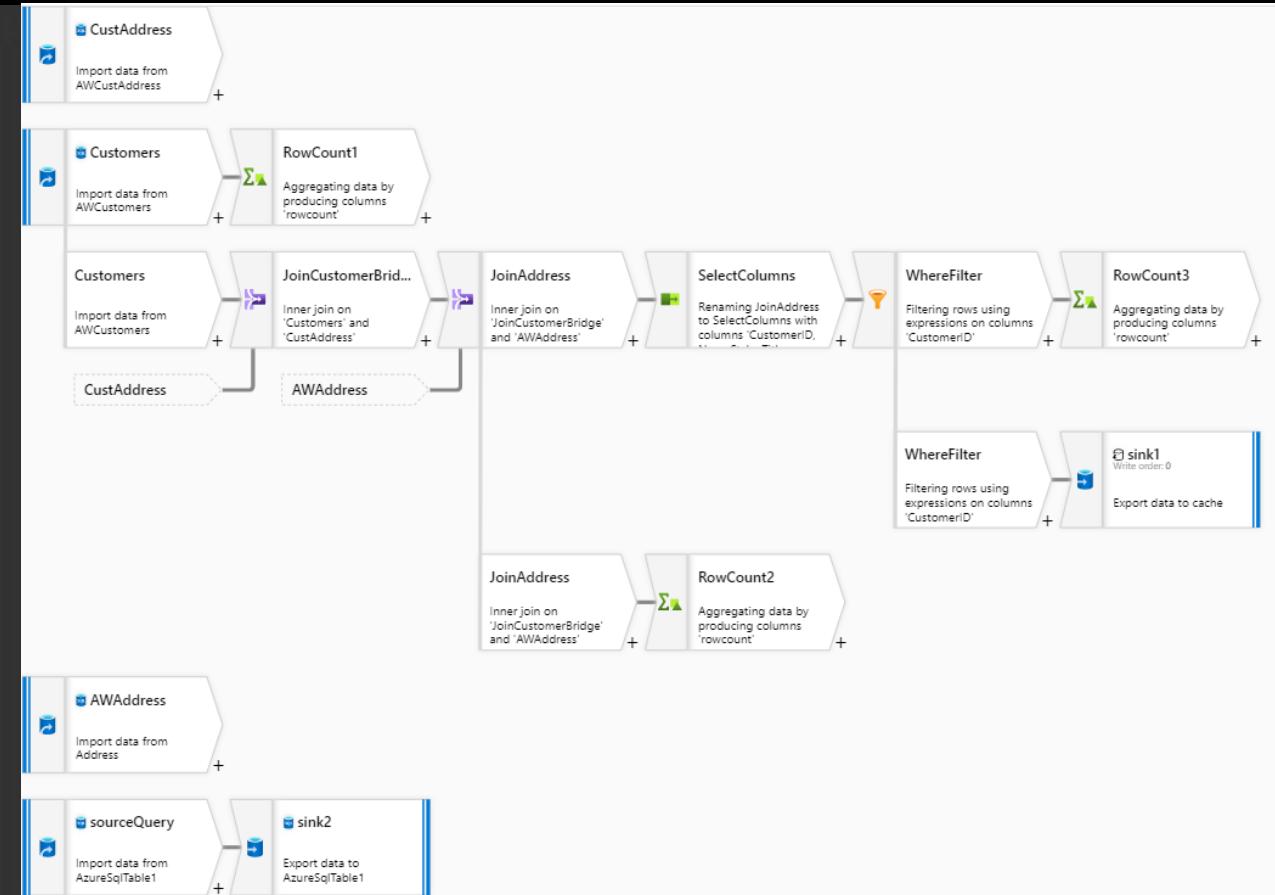
マッピングデータフローによるコード不要の大規模な変換



直感的な環境で ETL と ELT のプロセスを構築して実行

設計、テスト、メンテナンスに費やす時間を減らしてビジネスロジックに集中

- データのクレンジング、変換 (transformation)、集約、変換 (conversion)
 - Spark の実行を通じてクラウド規模に移行
 - 回復力の高い、簡単に構築できるデータフロー

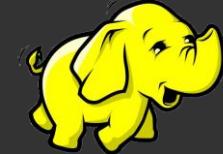


コード不要またはコード中心の変換を選択

コード中心のコンピューティングのための完全に完結した運用化



Databricks、Synapse Spark
ノートブック、Jar、Python



Azure HDInsight
Hive、Pig、Spark、MapReduce、
Streaming



Synapse SQL プール、Azure
SQL DB & SQL Server
ストアド プロシージャ、
スクリプト アクティビティ



機械学習
バッチ実行、更新リソース、
Azure ML Execute Pipeline



Azure Functions
関数呼び出し

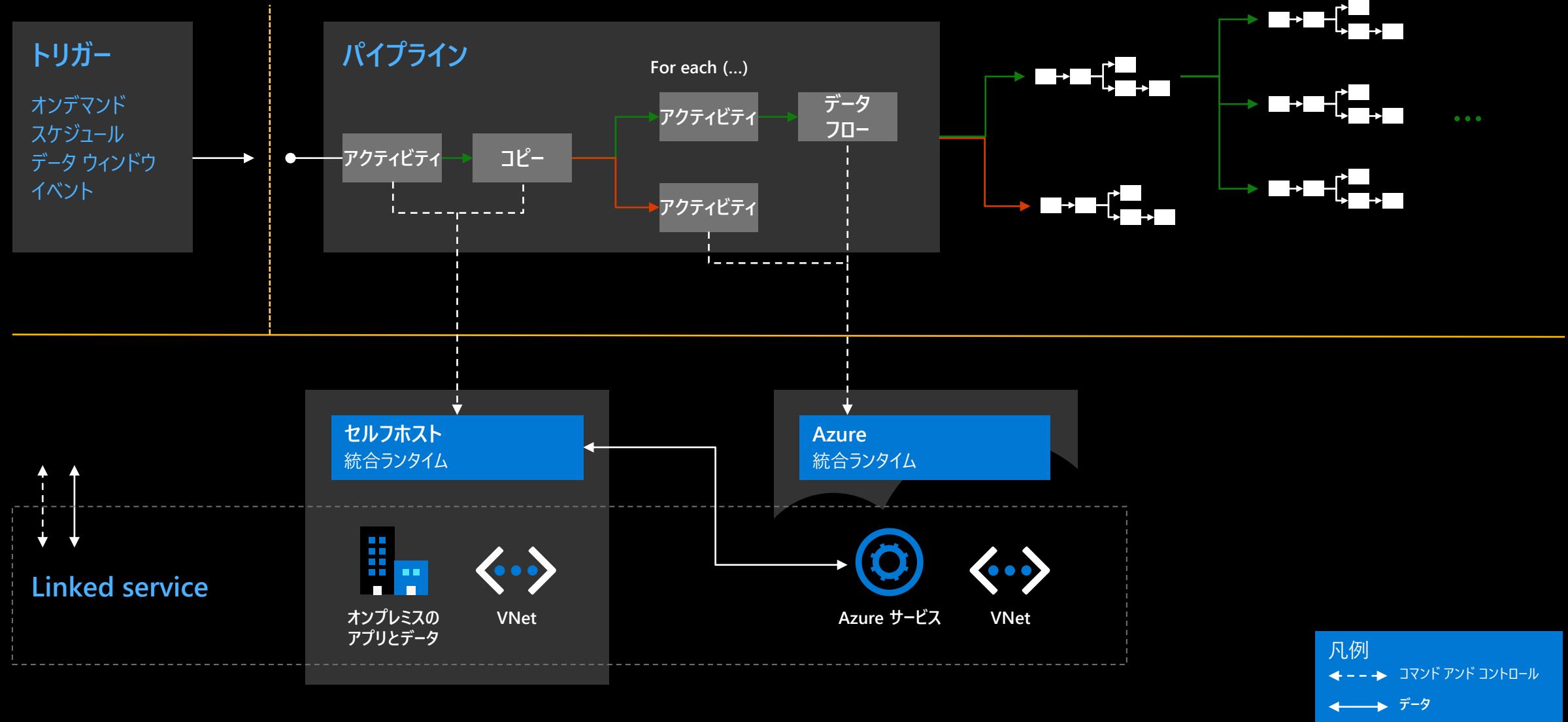


Azure Batch
カスタム実行可能ファイル



Azure Data Lake Analytics
Data Lake Analytics U-SQL

Azure Data Factory の概念



マネージド仮想ネットワーク & マネージドプライベートエンドポイント

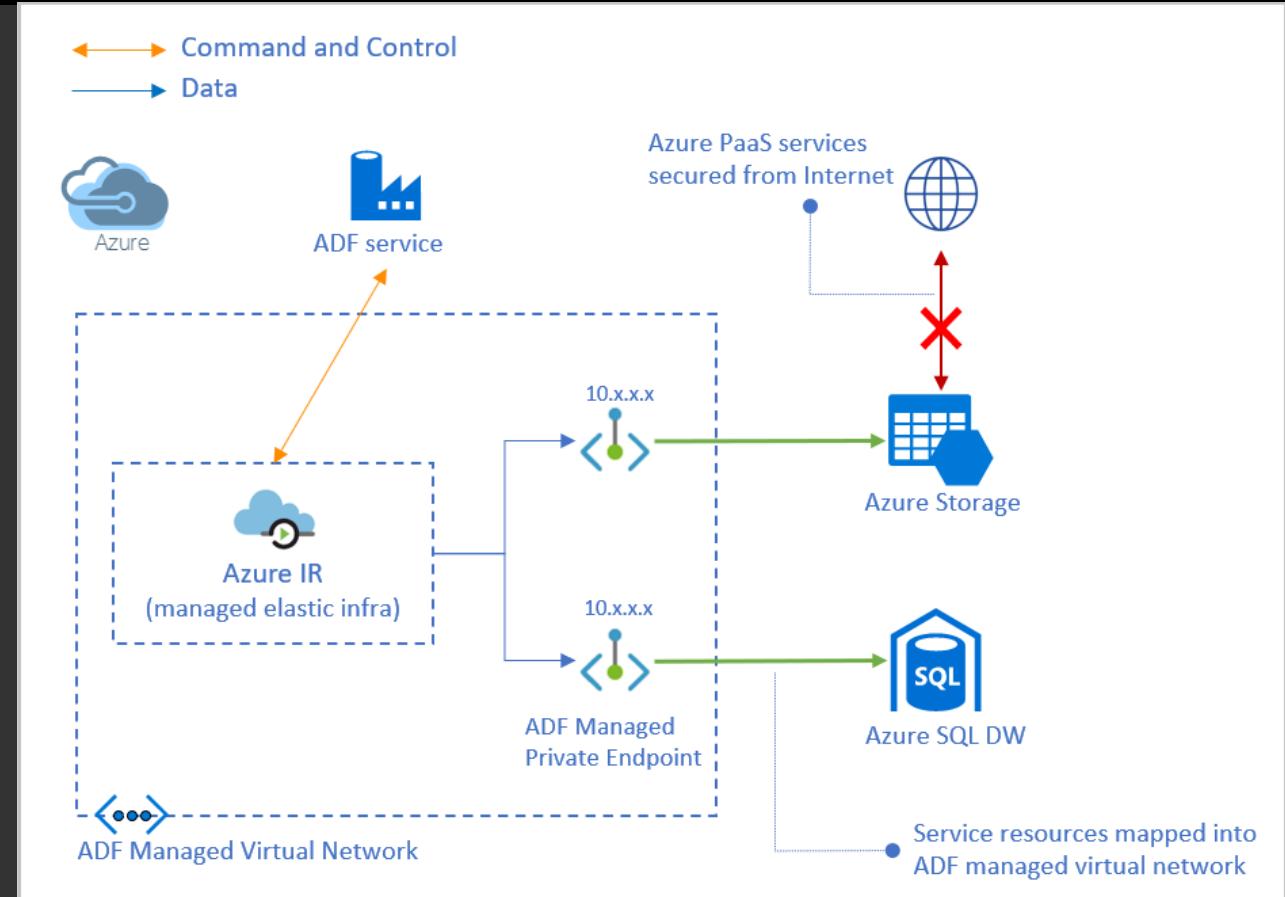
マネージド仮想ネットワーク内の統合ランタイムから安全にデータストアに接続

Data Factory が管理するマネージド仮想ネットワーク (マネージド VNET) に統合ランタイムをプロビジョニングできる

マネージド VNET に作成されるマネージドプライベートエンドポイントを利用してことで統合ランタイムとデータストアの間で Microsoft のバックボーン ネットワークを通じて安全にデータをやりとりできる

Azure Private Link をサポートする多くのデータソースに加えて Azure Databricks もマネージドプライベートエンドポイントに対応

[マネージド仮想ネットワークとマネージドプライベートエンドポイント - Azure Data Factory | Microsoft Learn](#)

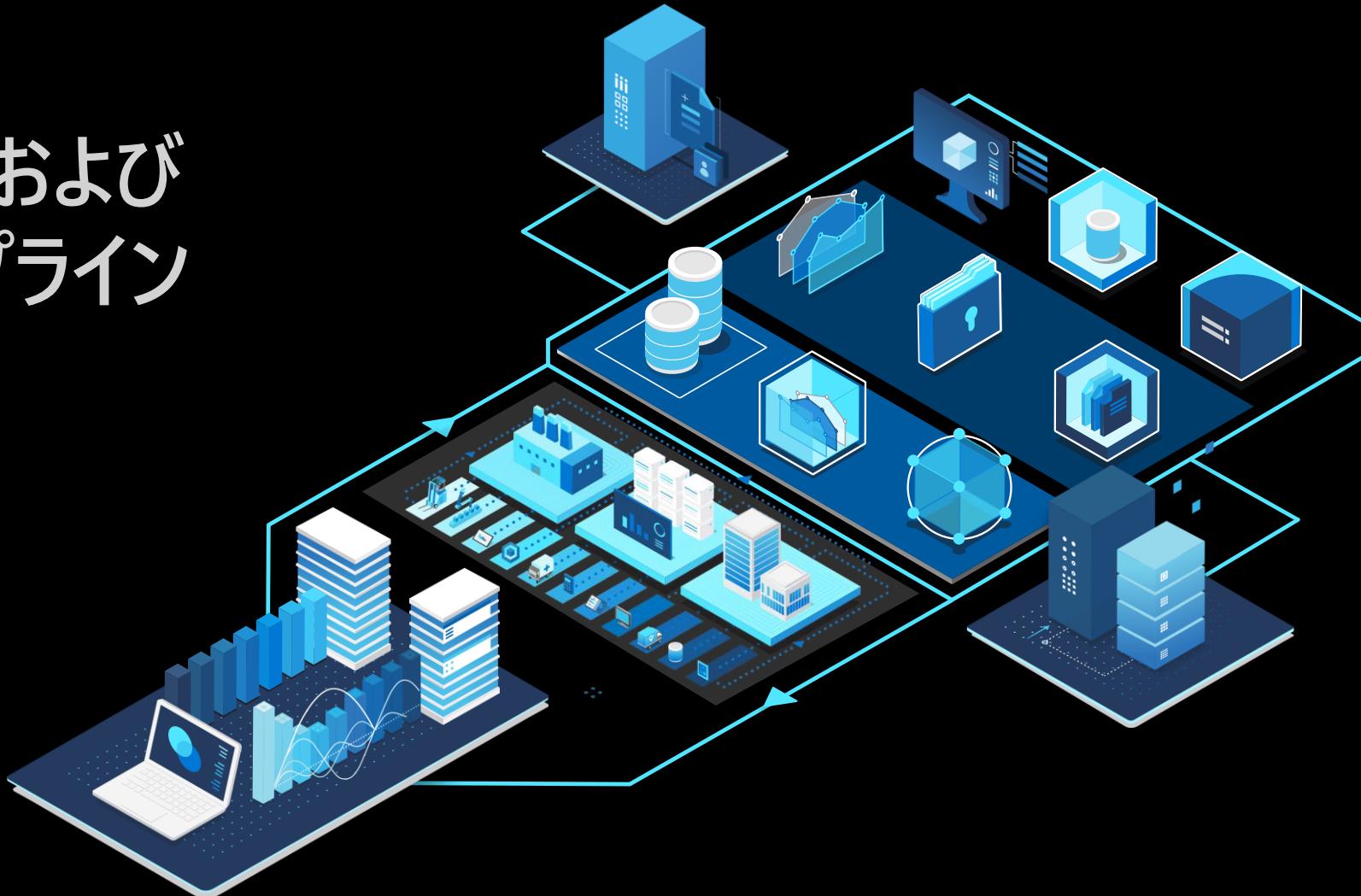


ダイジェスト版のスライドは以上
以降はフル版の製品概要(必要に応じて参照)

Azure Data Factory および Azure Synapse パイプライン

製品の概要

2022 年度第 4 四半期



マイクロソフトのデータ統合

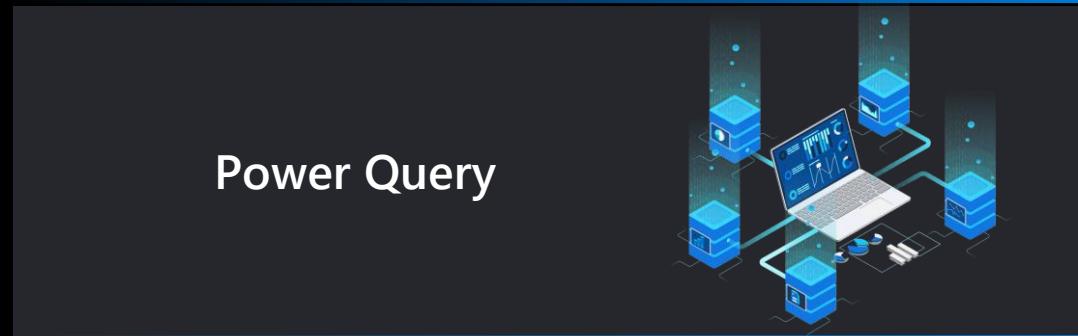
製品ポートフォリオ

プロフェッショナル データ統合

市民データ統合



Azure Data Factory と
SQL Server
Integration Services



Power Query

フル マネージドのサーバーレス データ統合サービス
100 を超える組み込みのコネクタを追加コストなしで使用してデータ ソースを
視覚的に統合し、ETL プロセスをコード不要で簡単に構築

強力でスマートなデータ準備
お客様がより早く結果を明らかにできるように、多数のマイクロソフト人気製品に
Power Query をシームレスに統合して使いやすさを向上



Azure Synapse Analytics



制限のない分析

データ統合、エンタープライズ データ ウェアハウス、
ビッグ データ分析を 1 つのエクスペリエンスに統合する
制限のない分析サービス

お客様の動向

<https://aka.ms/adf/customerstories>



groupm accenture



HEALTH • HYGIENE • HOME



Azure Data Factory

地理的な拠点

-
- 1. 米国西部
 - 2. 米国西部 2
 - 3. 米国西部 3
 - 4. 米国東部
 - 5. 米国東部 2
 - 6. 米国中部
 - 7. 米国中北部
 - 8. 米国中南部
 - 9. 西ヨーロッパ
 - 10. 北ヨーロッパ
 - 11. 東日本
 - 12. ブラジル南部
 - 13. ノルウェー東部
 - 14. オーストラリア東部
 - 15. 中国東部 2
 - 16. 中国北部
 - 17. 東アジア
 - 18. 東南アジア
 - 19. スイス北部
 - 20. 南アフリカ北部
 - 21. フランス中部
 - 22. 韓国中部
 - 23. アラブ首長国連邦北部
 - 24. ドイツ西部
 - 25. インド中部
 - 26. インド南部
 - 27. JIO インド西部
 - 28. 英国西部
 - 29. 英国南部
 - 30. 米国中西部
 - 31. カナダ東部
 - 32. カナダ中部
 - 33. 西日本
 - 34. ナショナル クラウド × 5

34 以上

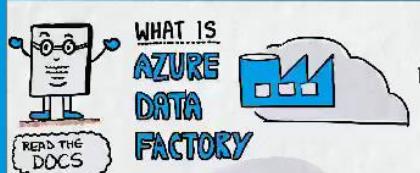
世界中の
Azure リージョン

12

言語

ナショナル クラウド

米国政府、
中国 & ドイツ



A CLOUD-BASED DATA INTEGRATION SERVICE THAT ORCHESTRATES DATA MOVEMENT & TRANSFORMATION BETWEEN DIVERSE DATA SOURCES & CLOUD COMPUTE RESOURCES AT SCALE



7 THINGS TO KNOW ABOUT AZURE DATA FACTORY

1 ENTERPRISE READY

Data integration at cloud scale!

2 ENTERPRISE DATA READY

90+ connectors! It just works.

3 CODE-FREE TRANSFORMATION

UI driven mapping data flows

4 RUN CODE ON ANY AZURE COMPUTE

For hands-on data transformations

5 MANY SSIS PACKAGES RUN ON AZURE

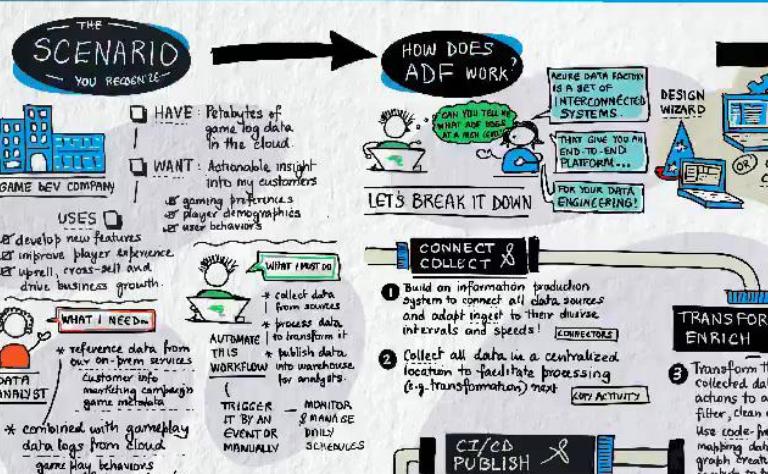
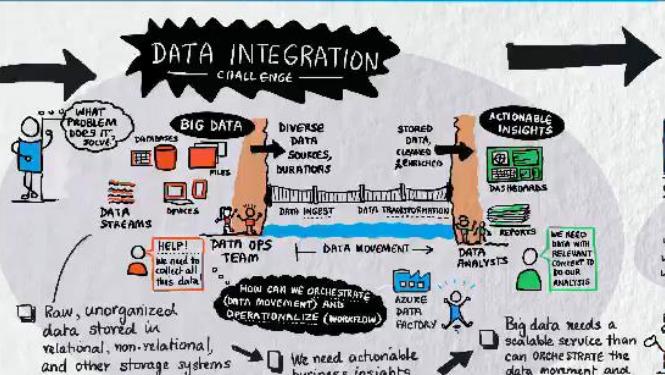
(Batch on-prem in ADF + 3 steps)

6 ADF CAN MAKE DATA OPS SEAMLESS

source control, automated deploy, simple templates

7 SECURE DATA INTEGRATION

Managed virtual networks protect sensitive data, SSO, multi-factor authentication

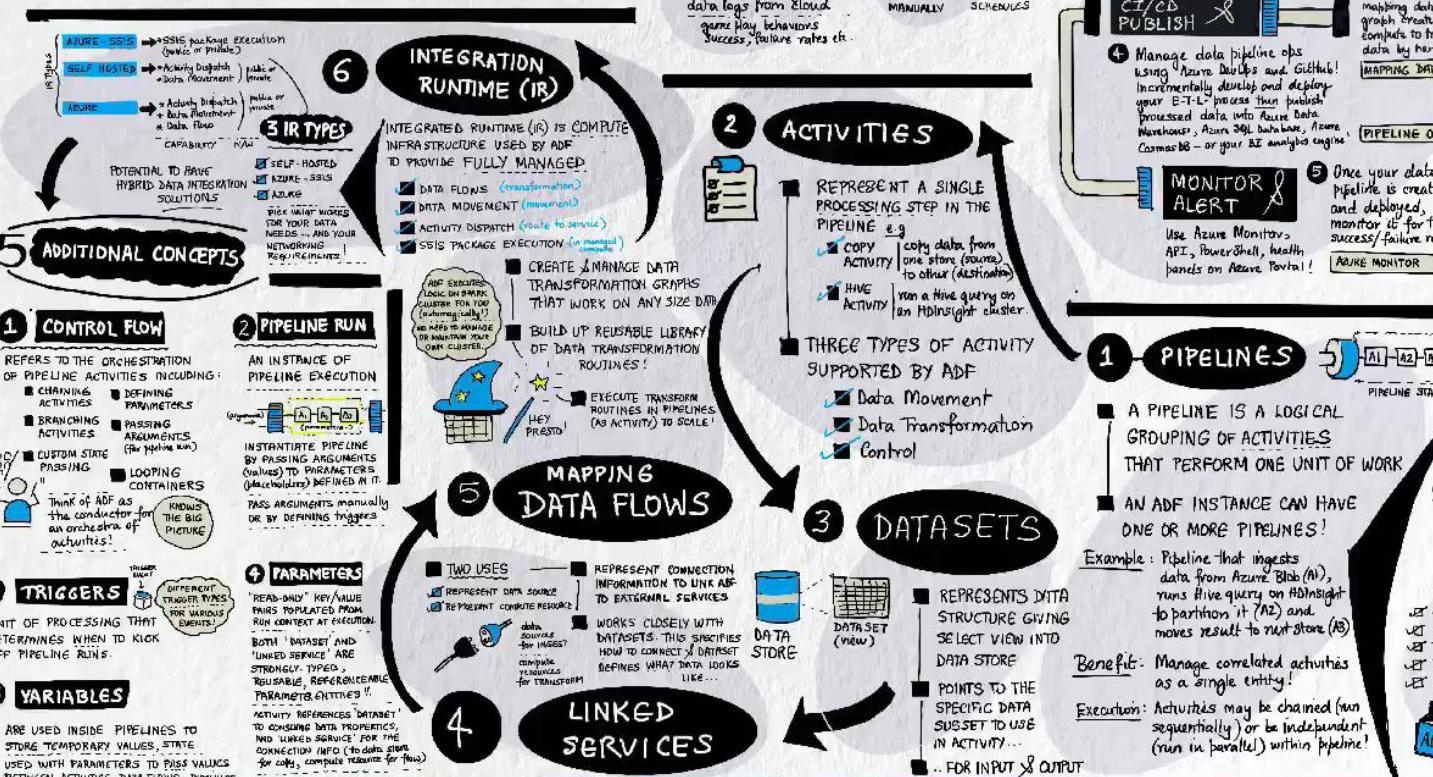


LET'S TALK DATA INTEGRATION PATTERNS

The most common pattern is ETL

- E = EXTRACT = connect to sources
- T = TRANSFORM = process the data for analysis
- L = LOAD = move the data to data warehouse or analytics engines for business insights

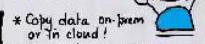
Another pattern is ELT where raw (native) data is itself loaded (stored) before the transformation phase...



AZURE DATA FACTORY HAS CODE-FREE ETL AS A SERVICE!

You always have the option to do hand-coded transformation using Azure compute... but mapping data flows provide a UI-based wizard to simplify your pipeline setup..

5 ASPECTS IT COVERS!



- 1 INGEST DATA
- 2 CONTROL FLOW
- 3 DATA FLOW
- 4 SCHEDULE OPS
- 5 MONITOR OPS

* Copy data on-prem or in cloud!

- * 90+ native connectors
- * Serverless, auto-scaling!
- * Design (via UI) or Create (via SDK) your data pipelines
- * Utilize workflow constructs (loops, parameters, variables...) for efficiency
- * Code-free data transforms (execute automatically in script)
- * Scale out with Azure Integration Runtimes: generate data flows with SDK (code) or designer (UI)
- * Build/Maintain jobs/schedules for your data pipelines
- * Options include: Wall Clock, Event-based, Trickle Windows, Sharding
- * View active executions and pipeline history - details of activity/pipeline
- * Establish alerts for key events or progress notifications



LET'S TALK ABOUT COMPONENTS OF AZURE DATA FACTORY



HERE ARE THE KEY TERMS AND CONCEPTS YOU NEED TO KNOW TO USE ADF!



AND A FEW KEY TERMS

- > CONTROL FLOW
- > PIPELINE RUN
- > ACTIVITIES
- > DATASETS
- > LINKED SERVICES
- > DATA FLOWS
- > INTEGRATION RUNTIMES



BENEFIT: Manage correlated activities as a single entity!

EXECUTION: Activities may be chained (run sequentially) or be independent (run in parallel) within pipeline!



ADF TOOLKIT



SKETCHTHEDOCS · dev | @SKETCHTHEDOCS

Azure Data Factory

ハイブリッド データ統合、簡素化



使いやすい

- コード不要の ETL/ELT
- 数回のクリックで SSIS を再ホスト
- 組み込みの Git & CI/CD



コスト効率が良い

- 従量制
- フルマネージド、サーバーレス
- オンデマンドでスケーリング



強力

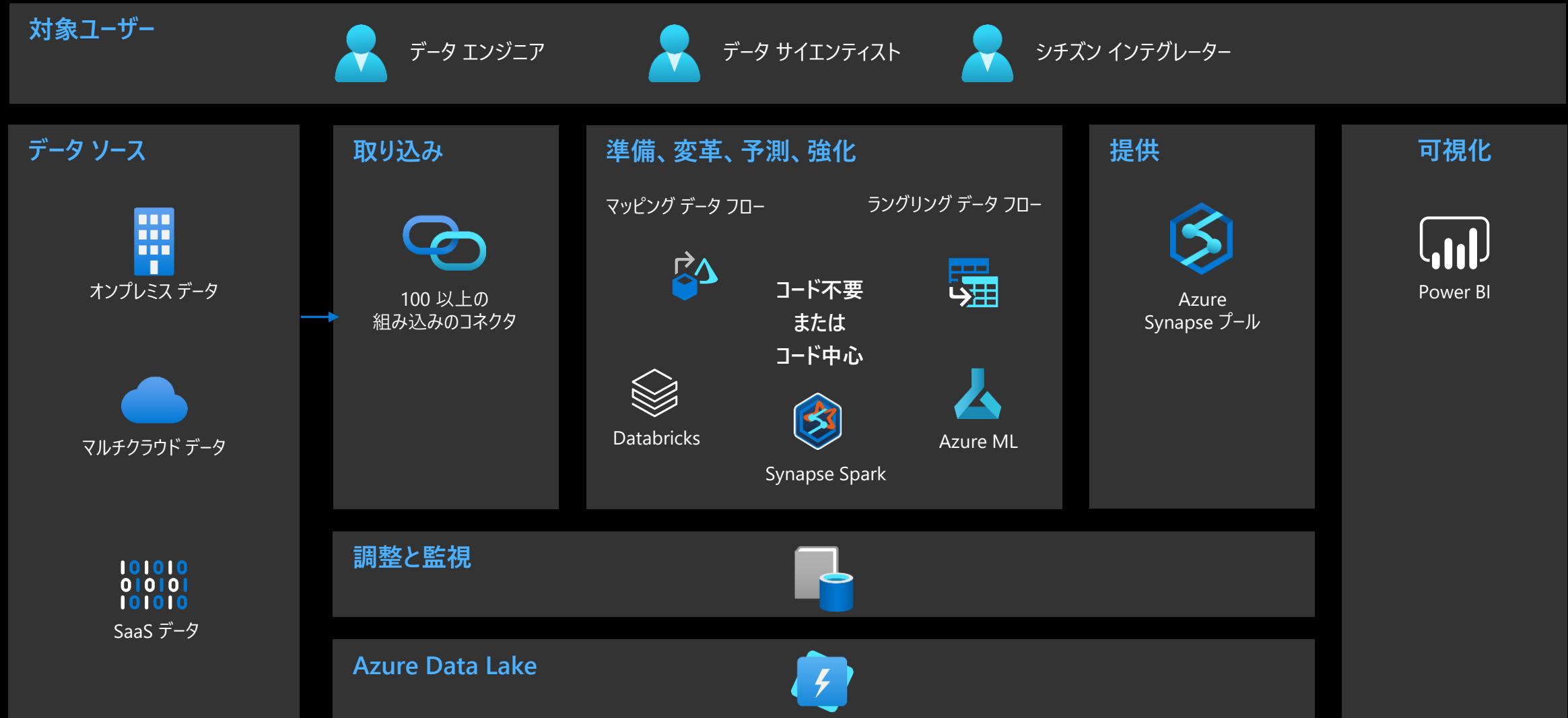
- 組み込みのエンタープライズコネクタ
- 大規模な調整と監視



インテリジェント

- 自律的 ETL
- AI ベースの目的主導のコピー
- 予測パイプライン

分析を最大限に活用



すべてのエンタープライズ データに接続

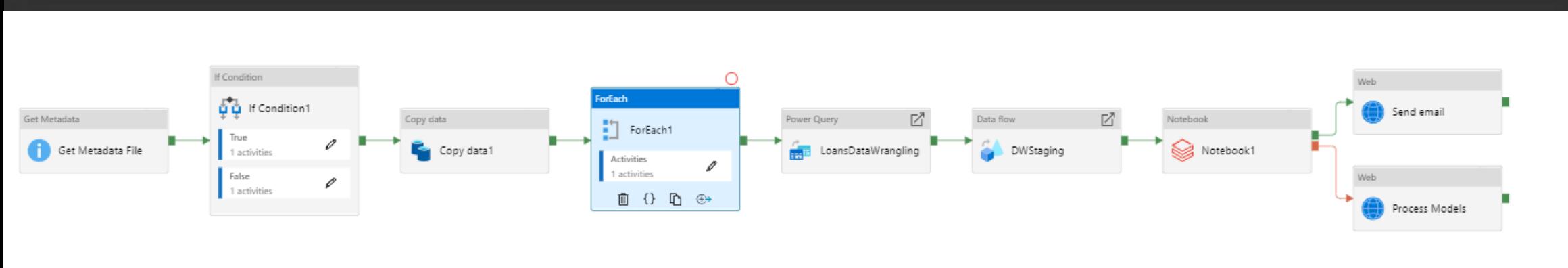


100 以上の組み込みコネクタなど

強力なコード不要のデータパイプラインを構築

ブラウザベースの設計 UI でのサーバーレス オーケストレーション

- データ統合と ETL ワークフローの設計
- ビジネス ロジックのサーバーレス実行
- スケジュールまたはイベントベースでの実行のためのトリガーの設定
- ソース管理と CI/CD のための Git リポジトリとの統合
- データファクトリ全体にわたるコラボレーション



コード不要またはコード中心の変換を選択

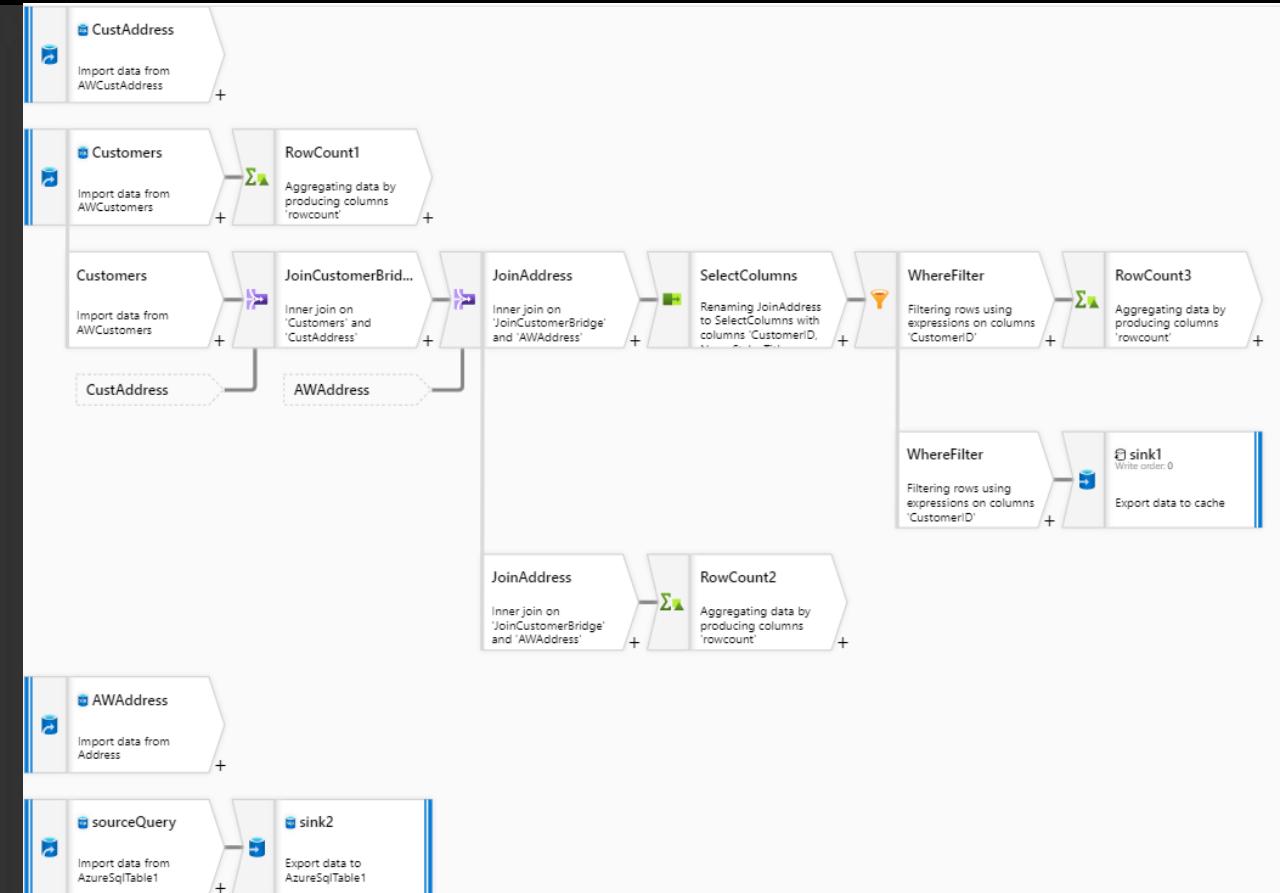
マッピング データ フローによるコード不要の大規模な変換



直感的な環境で ETL と ELT のプロセスを構築して実行

設計、テスト、メンテナンスに費やす時間を減らしてビジネス ロジックに集中

- データのクレンジング、変換 (transformation)、集約、変換 (conversion)
- Spark の実行を通じてクラウド規模に移行
- 回復力の高い、簡単に構築できるデータ フロー



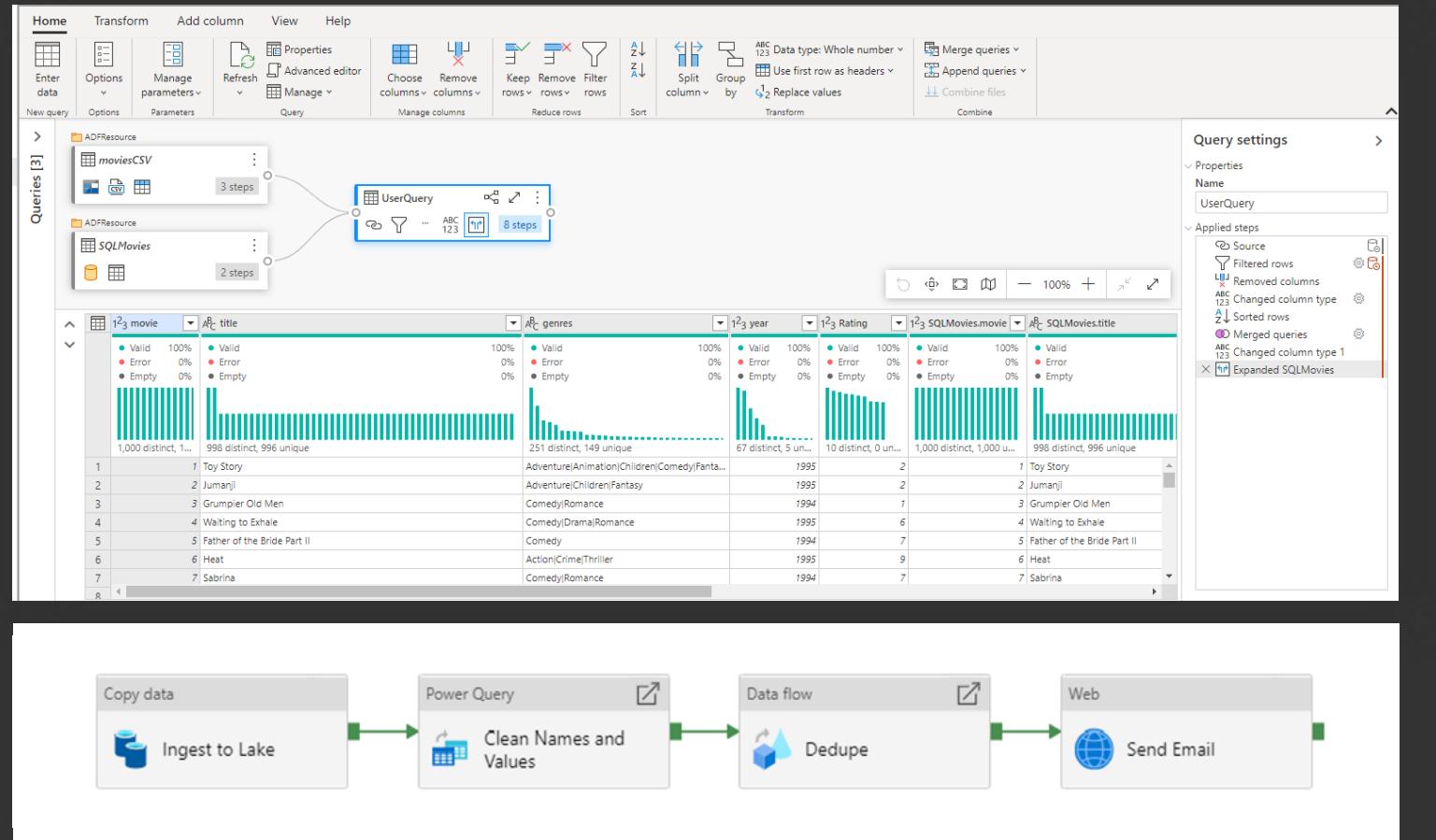
ADF での Power Query: 大規模なデータ ランキング

Spark を実行してコード不要の
アジャイルなデータ ランキングと
準備を実行

使い慣れた Microsoft Excel に
似た UI で素早く簡単に開始

テーブルの結合、列の追加、
行の削減など、さまざまな
ランディング機能から選択

ワンストップショットでランディング
データ フローの作成、デバッグ、
スケジュール、監視を実行



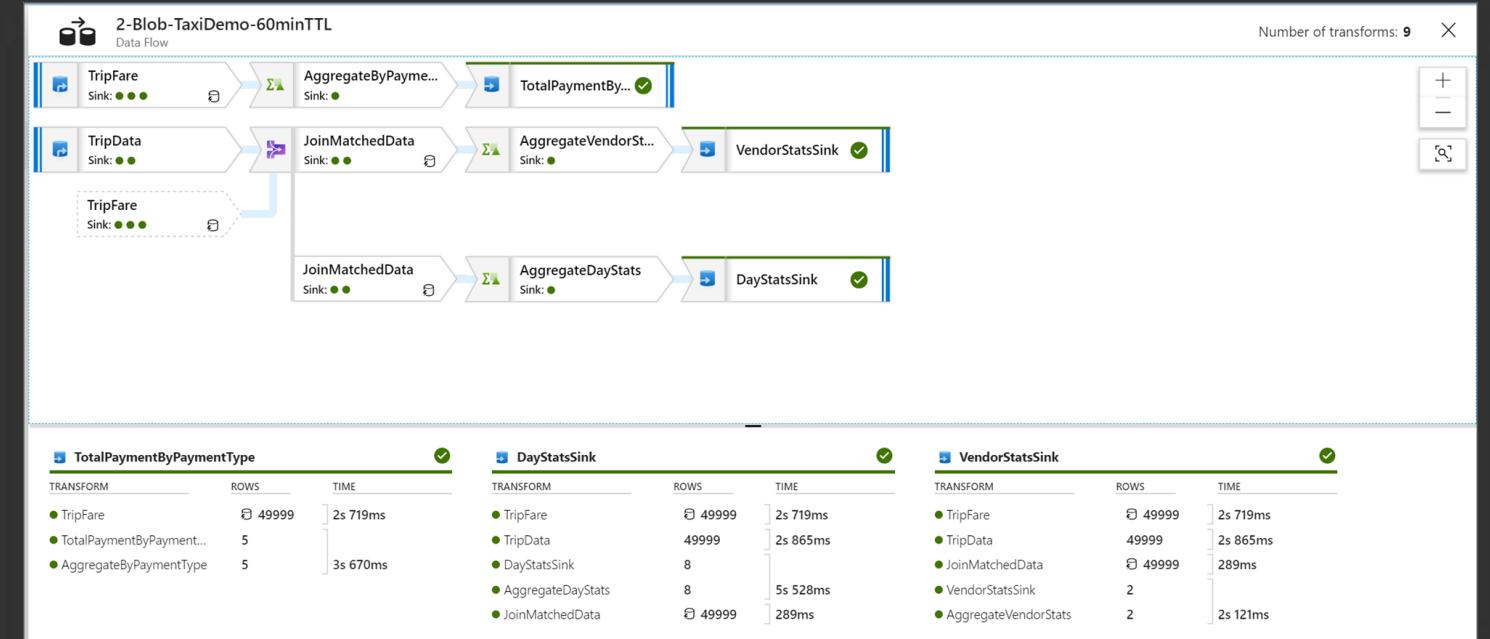
監視: クラス最高の管理

既成のテンプレートと UI ベースの
ビューでパイプラインとアクティビティの
実行を監視して、展開、テスト、
メンテナンスのコストを削減

表現力のある言語によりクエリを実行
親子パイプライン間の運用系統を参照

診断ログ、メトリックとアラート、
イベントを表示するために
Azure Monitor を統合

パイプラインとアクティビティの再記述

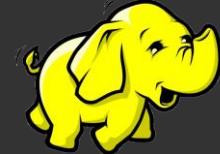


コード不要またはコード中心の変換を選択

コード中心のコンピューティングのための完全に完結した運用化



Databricks、Synapse Spark
ノートブック、Jar、Python



Azure HDInsight
Hive、Pig、Spark、MapReduce、
Streaming



Synapse SQL プール、Azure
SQL DB & SQL Server
ストアド プロシージャ、
スクリプト アクティビティ



機械学習
バッチ実行、更新リソース、
Azure ML Execute Pipeline



Azure Functions
関数呼び出し

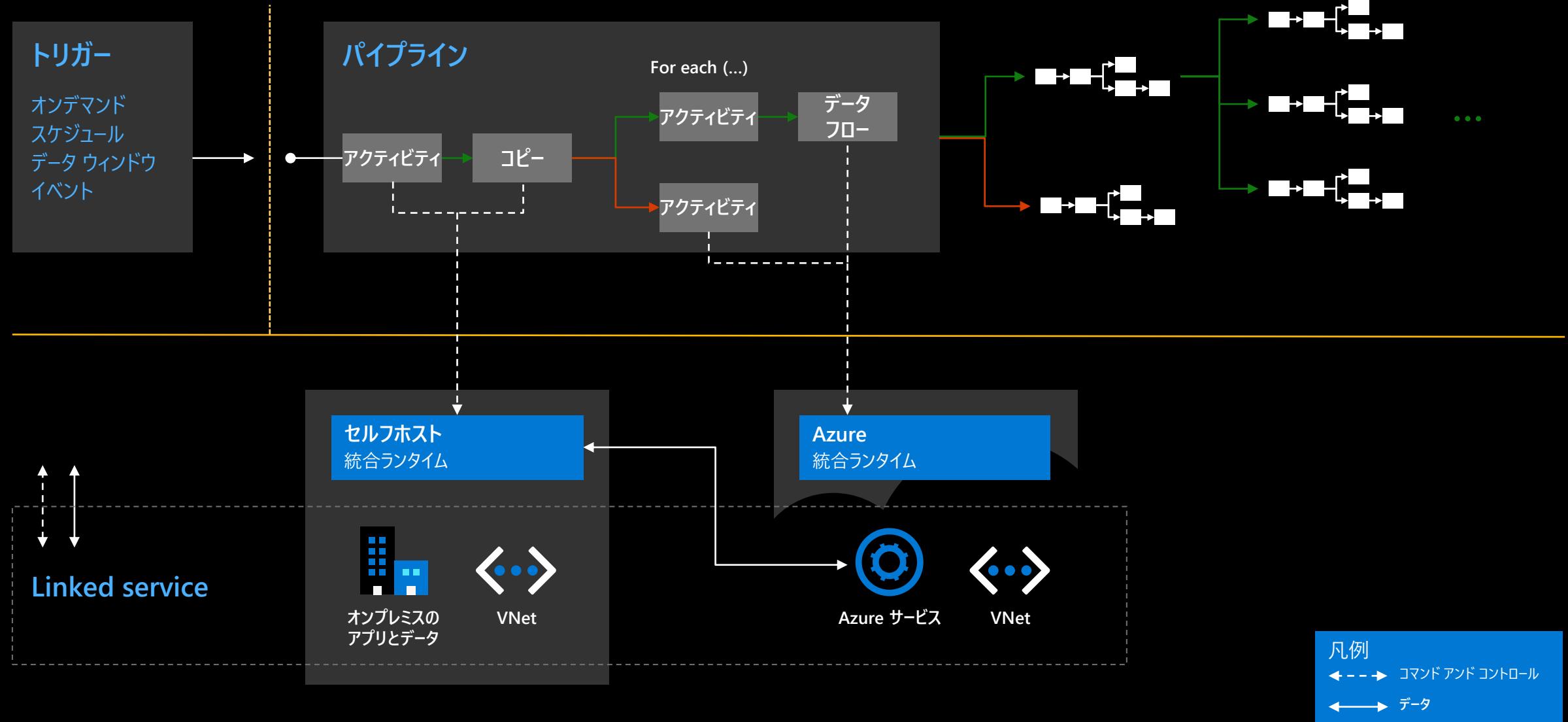


Azure Batch
カスタム実行可能ファイル



Azure Data Lake Analytics
Data Lake Analytics U-SQL

Azure Data Factory の概念



マネージド仮想ネットワーク & マネージドプライベートエンドポイント

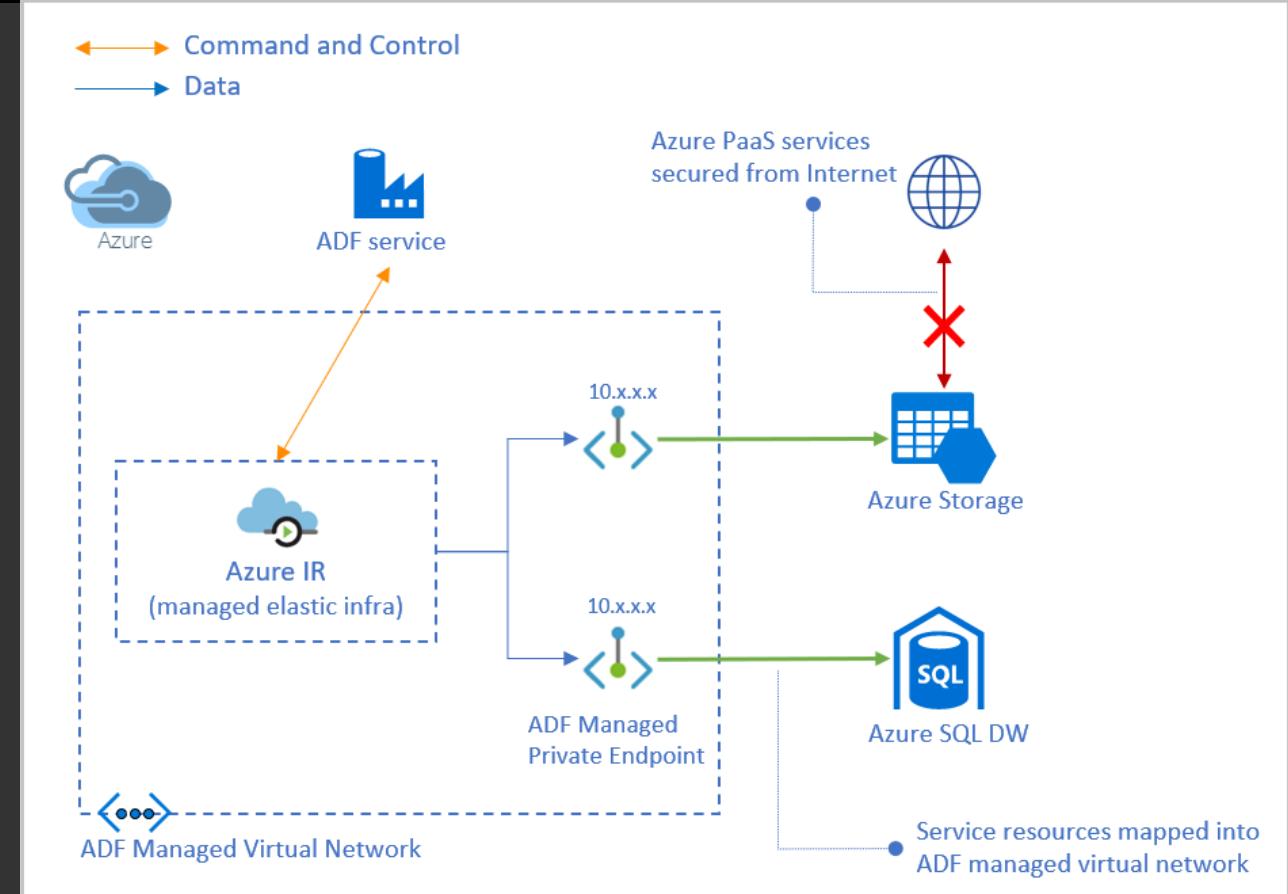
マネージド仮想ネットワーク内の統合ランタイムから安全にデータストアに接続

Data Factory が管理するマネージド仮想ネットワーク (マネージド VNET) に統合ランタイムをプロビジョニングできる

マネージド VNET に作成されるマネージドプライベートエンドポイントを利用してことで統合ランタイムとデータストアの間で Microsoft のバックボーン ネットワークを通じて安全にデータをやりとりできる

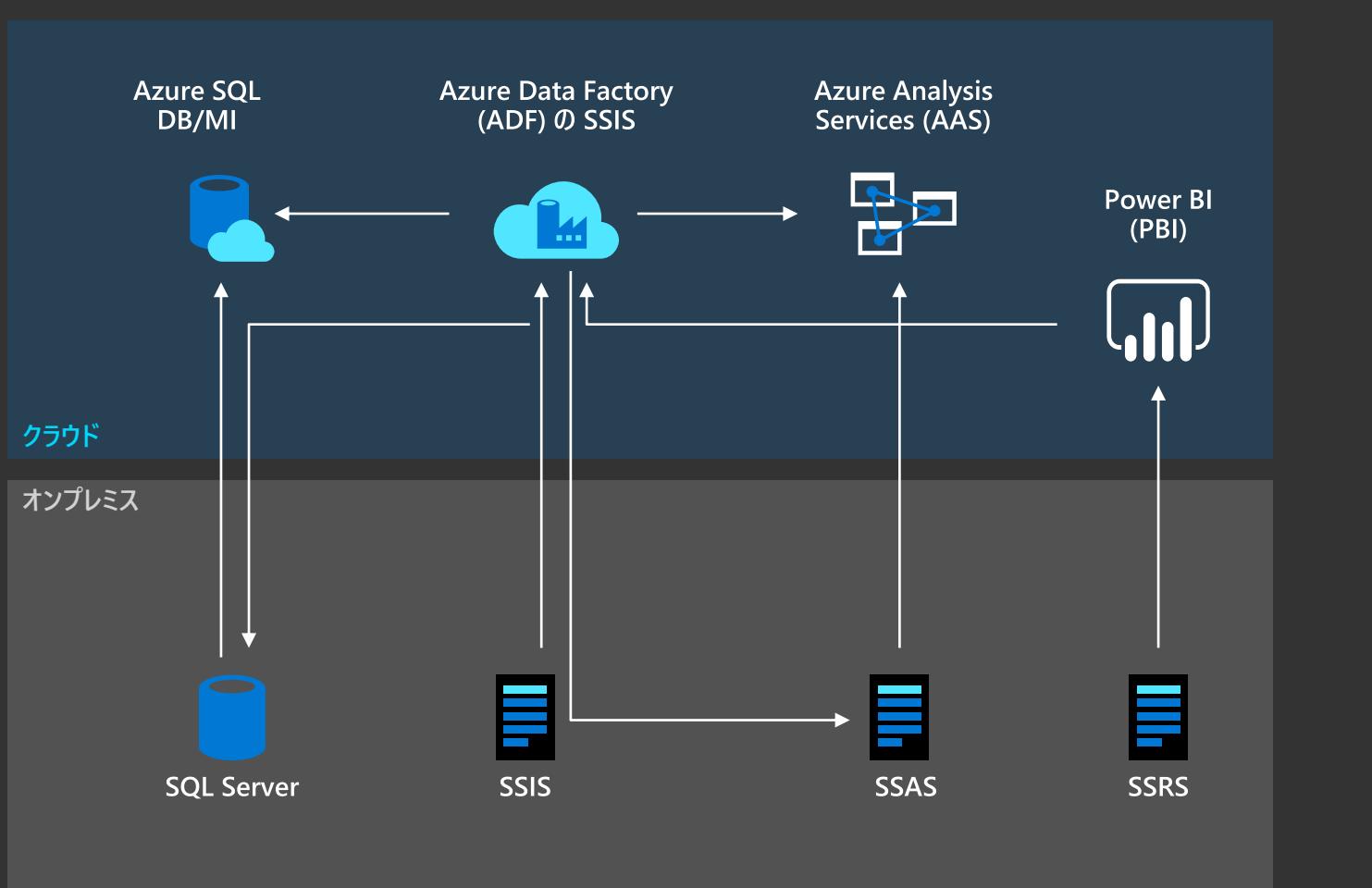
Azure Private Link をサポートする多くのデータソースに加えて Azure Databricks もマネージドプライベートエンドポイントに対応

[マネージド仮想ネットワークとマネージドプライベートエンドポイント - Azure Data Factory | Microsoft Learn](#)



SSIS を正常に再ホスト

1. オンプレミスの SSIS
インフラストラクチャの維持に
必要な運用コストを削減
2. Azure 対応 SQL Server Data Tools により、再ホスト プロセスを
合理化
3. Azure Monitor と Azure DevOps により、Data Factory で
ETL を運用化



SSIS の最新化によるメリット

- より大きな可能性 - 最大 88%* のコスト削減を実現
SQL Server と SSIS を同時に最新化して、TCO を大幅に削減
- 完全な互換性のある唯一のソリューション
すべてのワークロード、カスタムコンポーネント、サードパーティコンポーネントをクラウドに移行
- 使いやすい
[Azure 対応 SQL Server Data Tools](#) を使用して、ハウツー ドキュメントを参照
- クラウドのメリットを実感
自分のデータに集中して、後は専用のデータ統合機能に任せる
- Azure Synapse で制限のない洞察を引き出す
専用のオンデマンド リソースまたはプロビジョニング済みのリソースを使用して、条件に合わせて大規模にデータをクエリ

Azure Data Factory
MDW のためのクラウドネイティブな ELT PaaS

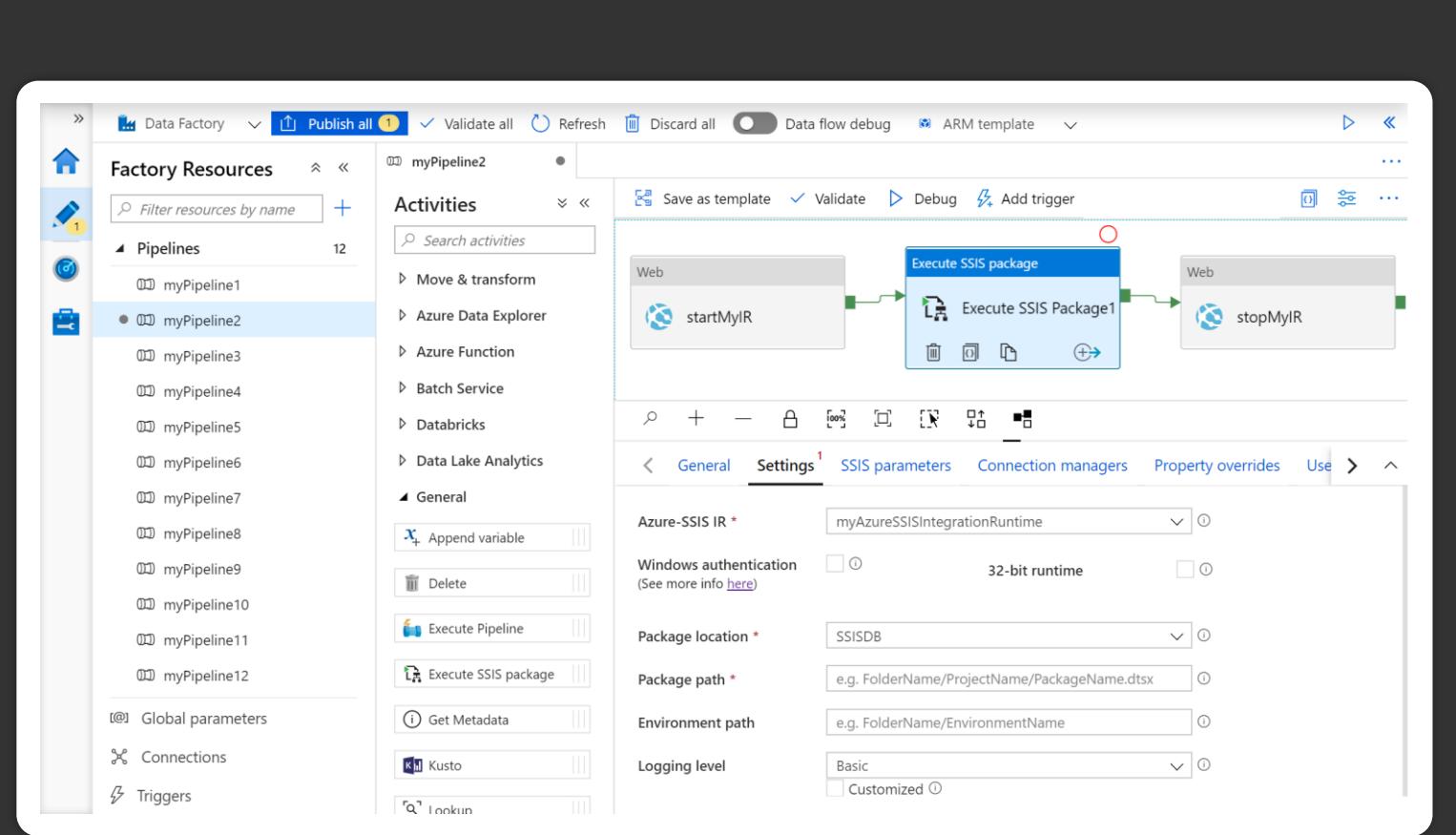
Azure Data Factory の SSIS
SQL Server の最新化のための ETL PaaS

SSIS
オンプレミス/IaaS の SQL Server を管理するための ETL

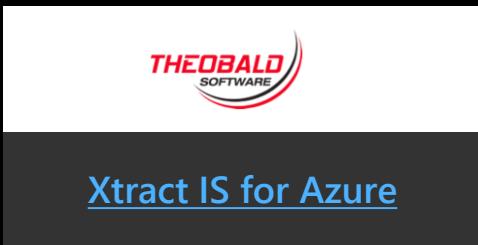


Data Factory で SSIS パッケージを強化

1. 膨大な量の生データを精緻化して、実用的なビジネス上の洞察を獲得
2. 複雑な ETL/ELT プロセスを開発して、データサイロの統合を支援
3. ネイティブな Data Factory 機能を備えたファーストクラスの SSIS アクティビティとして SSIS パッケージを補完



マイクロソフトのパートナーによる Data Factory セットアップの最適化



Azure ハイブリッド特典で最大 88%* のコスト削減を実現

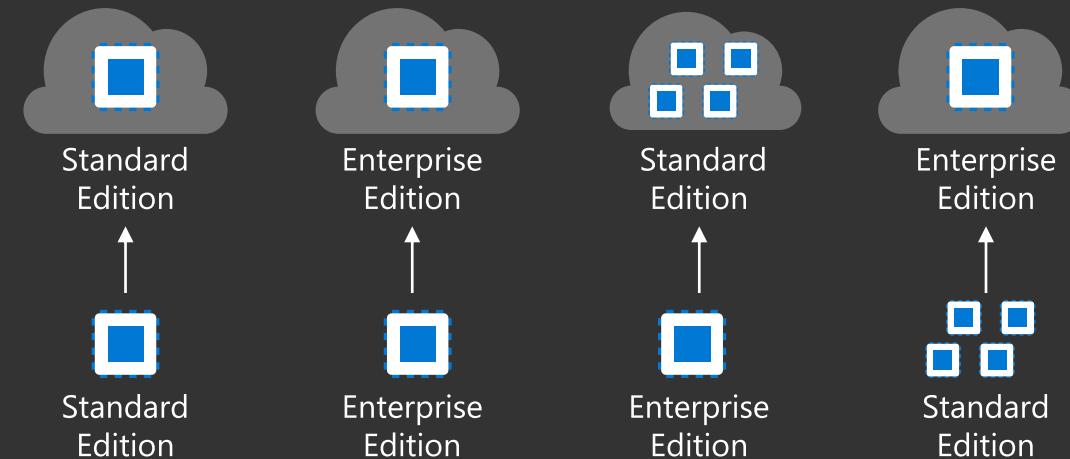
オンプレミスのソフトウェア アシュアランス付き SQL Server ライセンスを Azure へ移行

新しいデータ ファクトリ オファー

2020 年 8 月 1 日以降

- 4 つの SSIS IR Standard Edition 仮想コアごとに 1 つの SQL Server Enterprise Edition コア ライセンス
- 1 つの SSIS IR Enterprise Edition 仮想コアごとに 4 つの SQL Server Standard Edition コア ライセンス

ADF の SSIS/Azure-SSIS Integration Runtime (IR) エディション



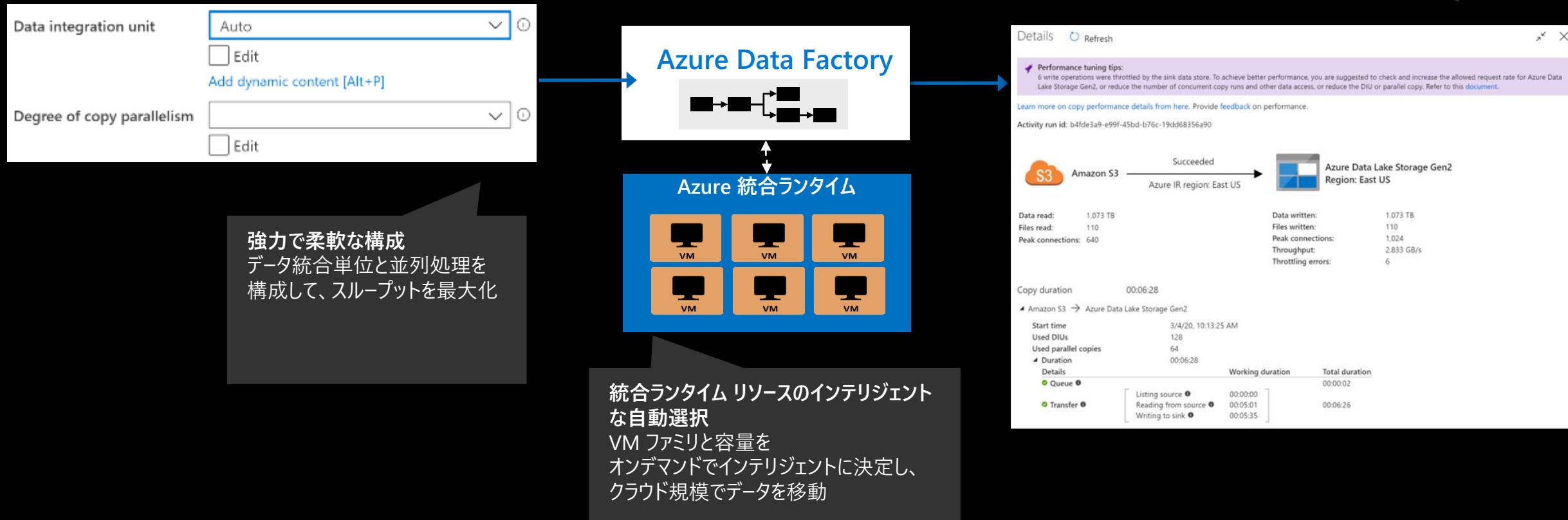
ソフトウェア アシュアランス付き SQL Server エディション

*[Azure Data Factory の価格設定](#)の詳細

Azure Data Factory

グローバル規模のデータの取り込みと移動

インテリジェントな洞察 -
データの取り込みと移動で最高の
パフォーマンスを達成
使用する DIU 数
タイミング内訳
パフォーマンス チューニングのヒント



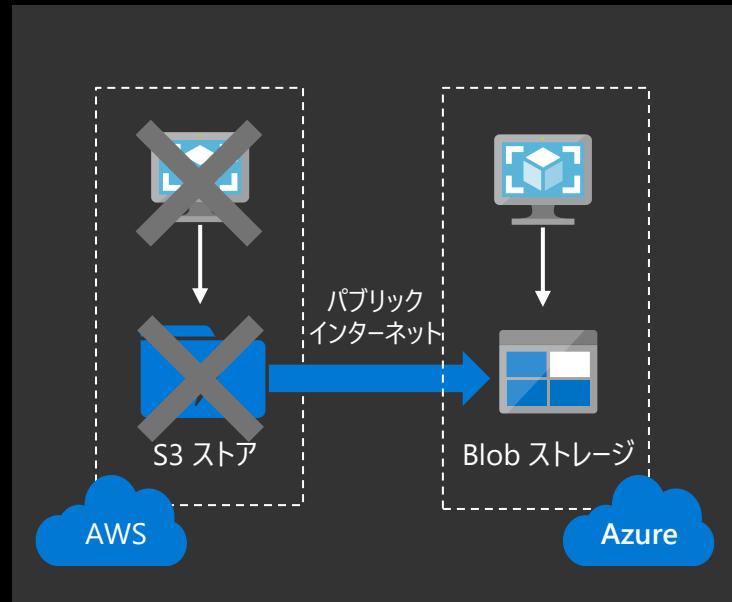
① コピー パフォーマンスの調整

② JIT のコンピューティングのスケーラビリティ

③ 包括的なランタイムの洞察

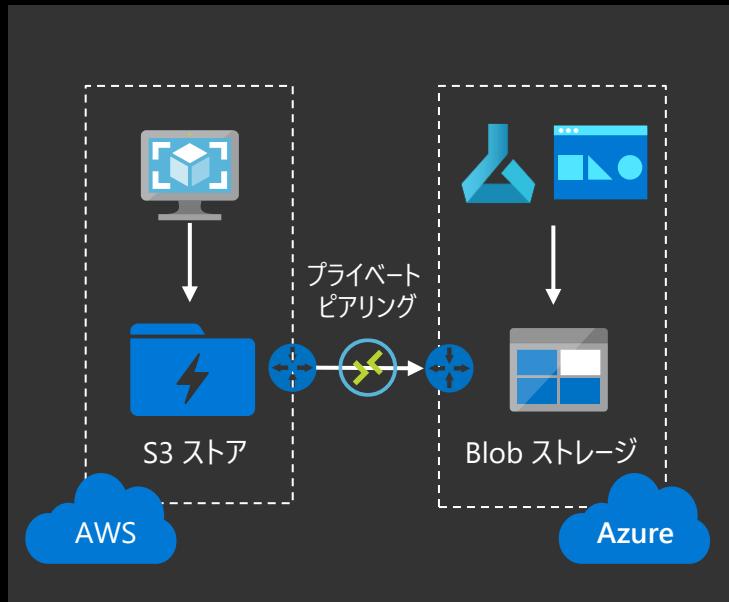
主なシナリオ 1: Data Lake と EDW のデータ移行

- AWS S3 とオンプレミス Hadoop から Azure への PB 規模のビッグ データ ワークロードの移行
- Oracle Exadata、Netezza、Teradata、AWS Redshift から Azure への TB 規模の EDW の移行



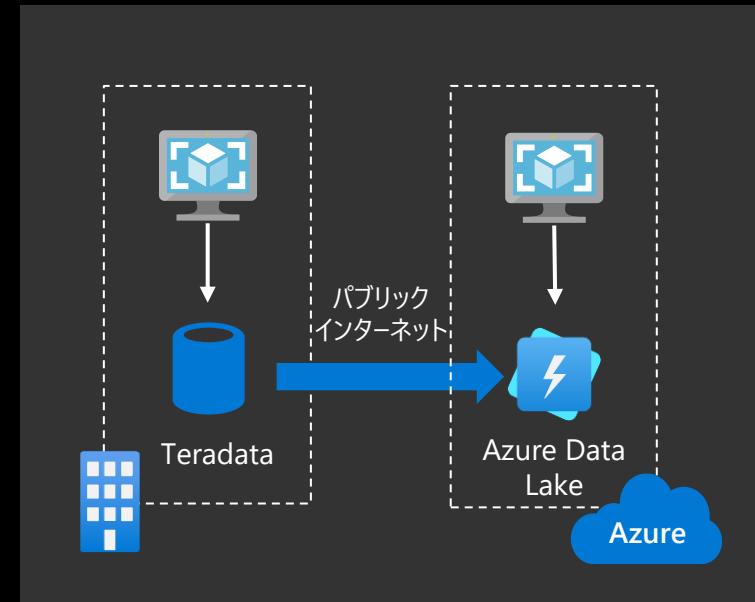
お客様の例: S3 → Blob に 2 PB を
11 日以内に移行

- 初期ロード 1.9 PB、スループット 2.1 GB/s
- 増分ロード 221 TB、スループット 3.6 GB/s



お客様の例: S3 → Blob に 10 Gbps プライベート
リンク経由で同期

- AWS Direct Connect & Azure VNet 間の
プライベート ピアリング
- 初期ロード 1 PB、スループット 787 MB/s
- 増分ロード 1 PB/月



お客様の例: オンプレミスの Netezza から
ADLS に 25 TB を移行

- 夜間の移行時間帯を設計
- DB オーバーヘッドを抑制するために同時接続数を
最大 8 に制限
- 移行は 3 週間で完了

レガシーな ETL ツールからの移行

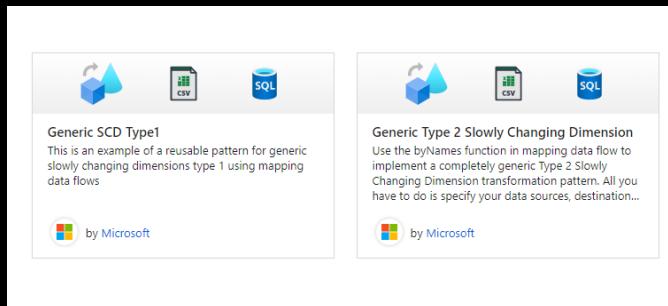
お客様の問題の説明

- オンプレミスの Oracle、Informatica、IBM などのレガシーなデータ ウェアハウスを Azure に最新化
- このお客様は数か月かけて数千の Informatica Power Center ETL を移行

お客様の要件

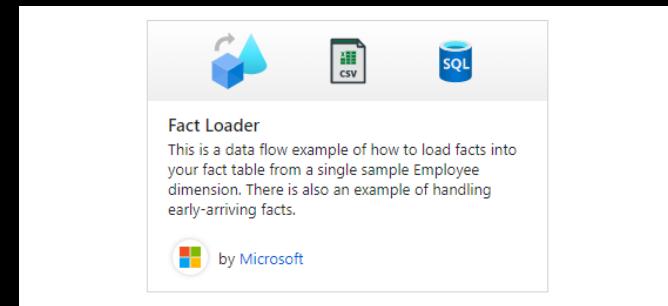
- チームに ETL コードを維持管理するスタッフがいない
- データ エンジニアは INFA ETL に対する既存スキルをデータ フローに簡単に転用できた

- ✓ 1.2 万/週の運用データ フロー パイプラインを実行
- ✓ 2021 年に 5,000 億行を処理



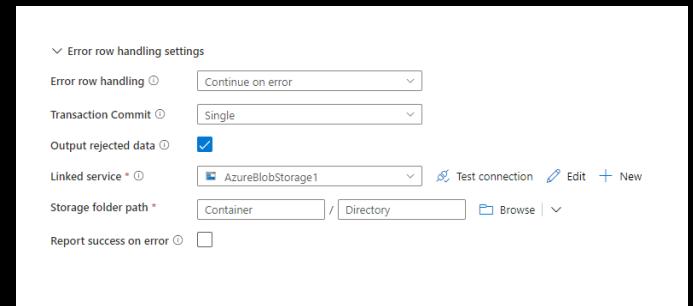
緩やかに変化するディメンション

- 組み込み SCD 機能を利用して、代理キー、キャッシュされた検索、結合を含む、タイプ 1 とタイプ 2 のディメンションを処理



ファクト テーブルの読み込み

- ソース テーブルのプッシュダウンクエリを広く利用し、派生列と集計のカスタム計算により、日々変更されたレコードを検出



エラー行処理

- エラーからの回復力がある連続実行 ETL ジョブを実現する SQL DB と Synapse DW プールのエラー行処理の設計とリリースを推進

ADLA/USQL と ADLS Gen1 から Gen2 への移行とデータフロー

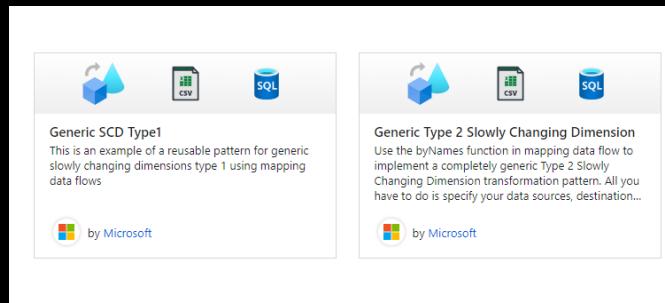
お客様の問題の説明

- ADLA と ADLS Gen1 に基づいて複雑なビッグ データ分析と ETL プロセスを構築
- 最小限の社内データ エンジニアリング リソースで 6 か月以内に 分析プラットフォームを再プラットフォーム化する必要があった

お客様の要件

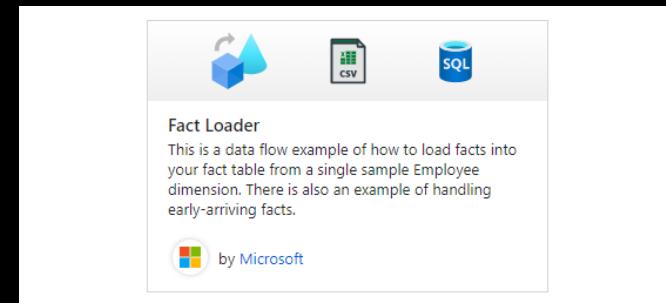
- Data Lake 内の全データを処理し、Snowflake と AAS に 読み込むデータを準備
- ADLS Gen2 に移行するために、ビッグ データ ETL を迅速に 開発して展開するデータフローを選択

- ✓ 2.6 万/週の運用データフロー パイプラインを実行
- ✓ 2021 年に 15 兆行を処理



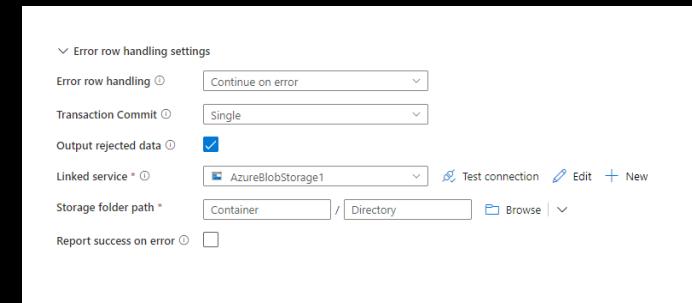
パターンマッチング

- パターンマッチングとパラメーター化を 多用し、汎用的なデータフロー パターン で複雑な ETL ルールを構築



幅広いフロー

- 1 フローあたり 200 件以上の変換を行う データフローで複雑なロジック ルールを 構築し、幅広いフロー処理に対応した 更新 UI 設計を推進



メタデータ検証ルール

- Data Lake では毎日数千のさまざまな 形状のファイルを受信しており、データ フローを利用してメタデータを検証

Azure Data Factory で成功するための行程

Azure Data Factory に関して学ぶべきことは非常にたくさんあるが、心配は無用である。
Data Factory の学習行程をすぐに開始できるように、行程が理解しやすくまとめられている。
毎週、Azure Learn のモジュールで学んでスキルを高め、ステップバイステップのトレーニングで学習し、
実践演習で自分のスキルを試すことができる。

Microsoft Certified:
Azure Data Engineer
Associate 認定の取得

行程をすぐに開始する - <https://aka.ms/adf/azurelearn>



1週目 基本の学習

Azure Data Factory、およびクラウドで大規模なデータ取り込みソリューションを作成できるようにするコアコンポーネントについて学習する

2週目 ペタバイト規模の データ取り込みの体験

Azure Data Factory を使用して、さまざまなデータストア間でデータを取り込むさまざまな方法を確認および学習する

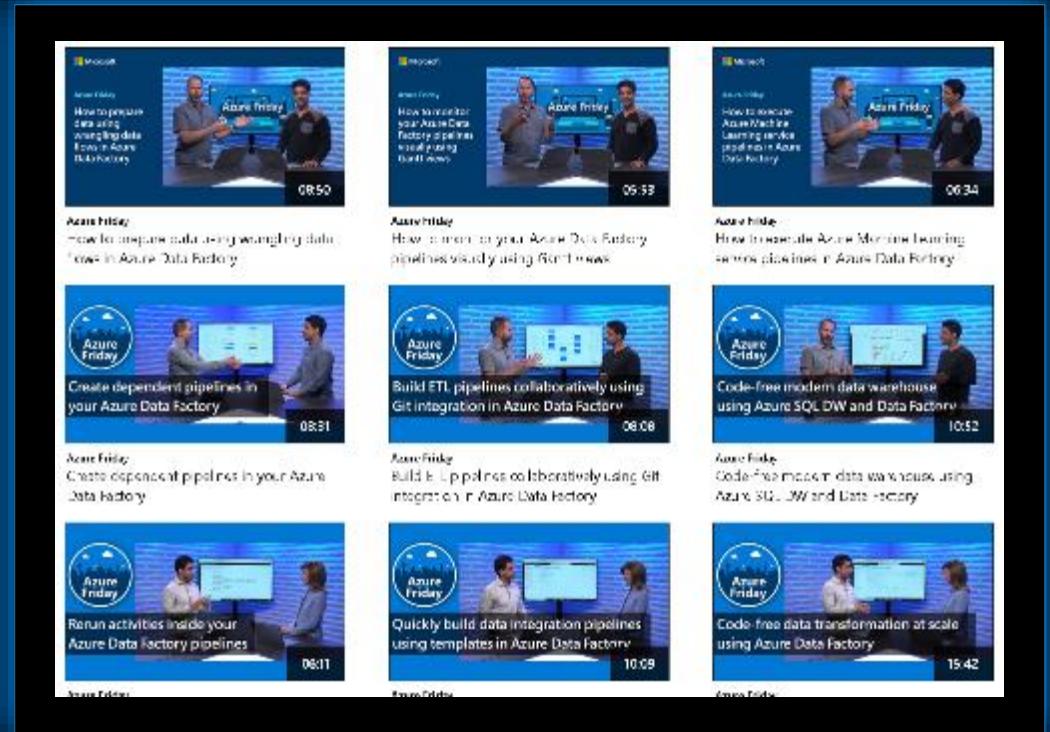
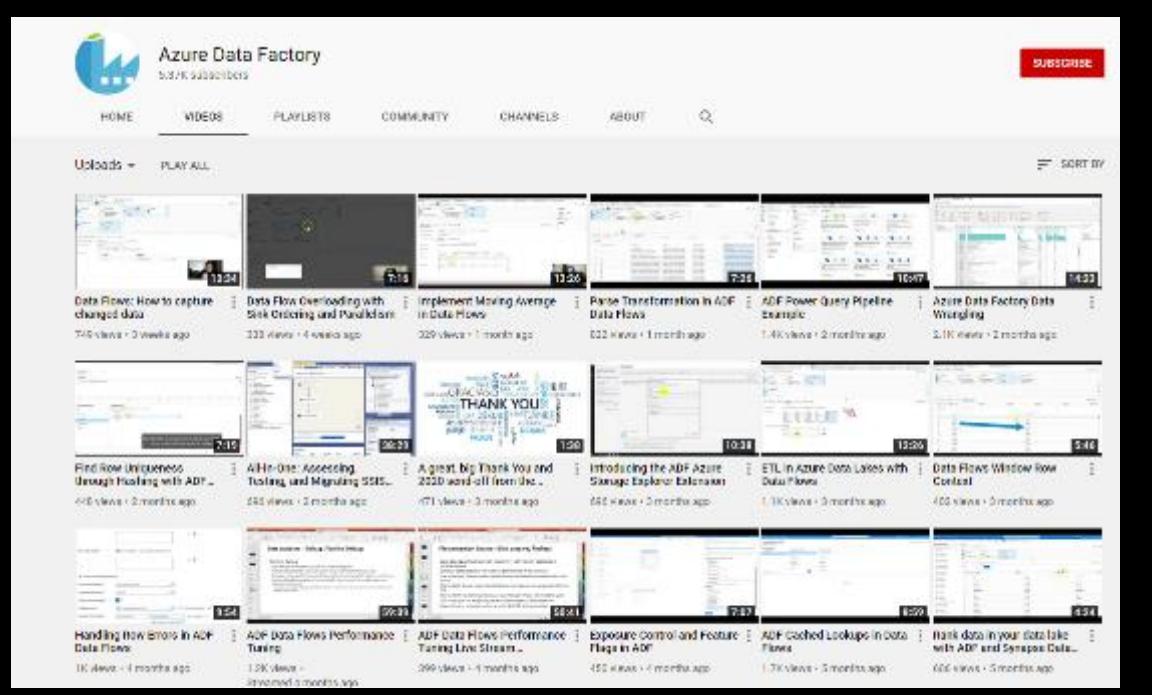
3週目 コード不要の大規模な 変換の学習

Azure Data Factory で、コードを使用せずに一般的なデータ変換とクレンジングを実行する方法について学習する

4週目 データの移動と変換を調整する 方法の学習

Azure Data Factory で他の Azure データサービスを使用して大規模なデータ移動を調整する方法について学習する

Azure Data Factory のビデオ



ADF YouTube チャンネル:
<https://aka.ms/adfvideos>

Azure Friday ビデオ:
<https://channel9.msdn.com/Tags/azure-data-factory>