Qlik

ホワイトペーパー

ベストプラクティスガイド

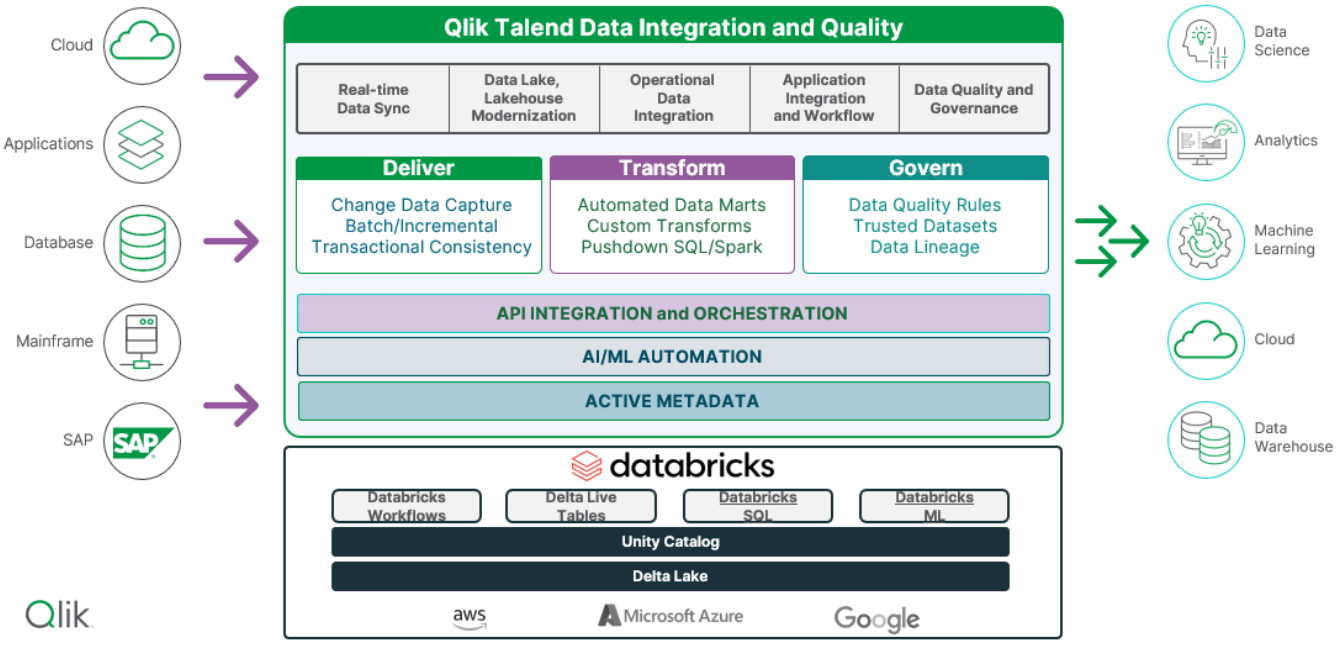
Qlik Talend™ Cloud および Databricks®

目次

はじめに

Qlik® では、Databricks Lakehouse Platform のお客様向けに、強力なリアルタイムのデータパイプライン自動化ソリューションを提供しています。Qlik データ統合ソリューションは、変更データキャプチャ（CDC）テクノロジーを活用し、複数のデータソースから Databricks Lakehouse Platform への継続的なデータストリームを確保することで、機械学習（ML）、人工知能（AI）、そして DataOps の取り組みを加速します。

Qlik データ統合ソリューション



Qlik が提供する 2 つのソリューション、Qlik Replicate および Qlik Cloud Data Integration は、企業による様々なソースやプラットフォーム間でのデータ管理を可能にします。Qlik Replicate は、手作業でのコーディングやスクリプトを必要とせず、オンプレミスおよびクラウドソースからクラウドデータウェアハウスに対して、リアルタイムでデータを複製および更新できるデータ統合ソフトウェアです。Qlik Cloud Data Integration は、データの取り込み、登録、変換、統合など、様々なデータ統合タスクを実行するデータパイプラインの生成機能を提供するクラウドベースのサービスです。両ソリューションは共に、リレーショナルデータベース、ビッグデータプラットフォーム、クラウドストレージ、SaaS アプリケーションなど、幅広いデータソースやターゲットをサポートします。ただし、機能、価格、展開オプションについては、2 つのソリューション間でいくつかの違いがあります。

主な違いは次のとおりです：

- Qlik Replicate は、オンプレミスまたはクラウド環境にインストールできるスタンドアロン製品ですが、Qlik Cloud Data Integration は、Qlik Cloud プラットフォーム上で実行されるフルマネージドサービスです。
- Qlik Replicate は、変更データキャプチャ（CDC）、遅延併合（delayed merge）、スキーマ進化（schema evolution）、競合解決（conflict resolution）など、データレプリケーションのためのより高度な機能を提供します。一方、Qlik Cloud Data Integration は、データクレンジング、エンリッチメント（強化）、ブレンディング、プロファイリングなどのデータ変換と統合に重点を置いています。
- Qlik Replicate は、ソースの数やレプリケートされるデータ量に基づいたソース単位の価格モデルとなっております。また、Qlik Cloud Data Integration は、各パイプラインの実行に使用されるクレジット数に基づいた利用量ベースの価格モデルとなっています。

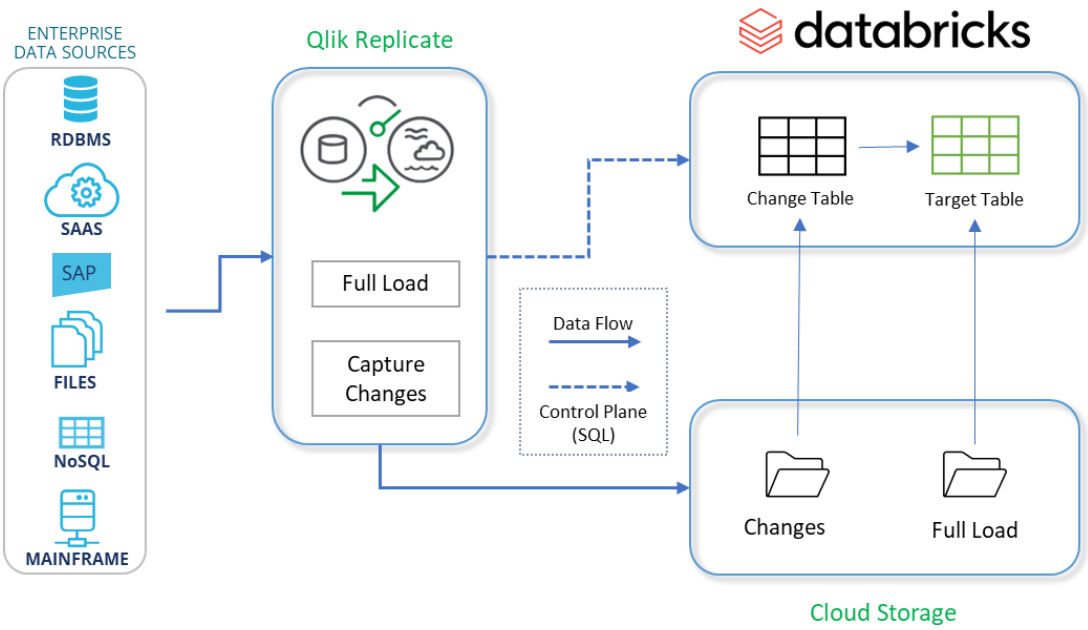
要約すれば、Qlik Replicate および Qlik Cloud Data Integration は、企業のユースケースやデータアーキテクチャに応じて共に使用することも、また別々に使用することもできる補完的なソリューションと言えるでしょう。どちらのソリューションも、最新の分析ニーズに対応できる、高速で信頼性が高く、スケーラブルなデータ統合機能を提供することを目的としています。

Qlik Replicate

Qlik Replicate® は、複数のデータ ソース（Oracle、Microsoft SQL Server、SAP、メインフレームなど）から Databricks Lakehouse Platform への CDC データの移動を継続的に自動化します。これによりお客様は、手作業でデータを抽出し、API / スクリプト経由でそれを転送し、スライス、ステージング、インポートするといった負荷の高い作業を回避することができます。

Qlik Replicate は、このアーキテクチャ内で次の機能を実行します：

1. ターゲットのインスタンス化
    a. ソースから変換された適切なデータタイプを使用して、DELTA 形式でターゲットテーブルを作成
    b. ソースから初期 / 完全ロードを実行し、対象データをストレージレイヤーへ送信
    c. ストレージレイヤーからデータをロードするためにSpark SQL を Databricks へ送信し、DELTA 形式でテーブルへと変換
2. 変更内容のキャプチャと適用
    a. ログベースの CDC を使用してソースから変更内容をキャプチャ
    b. ターゲットとなる DELTA テーブルに対して変換内容（挿入 / 更新 / 削除）を適用（中間ステージングとしてストレージレイヤーを使用）



Qlik Cloud Data Integration

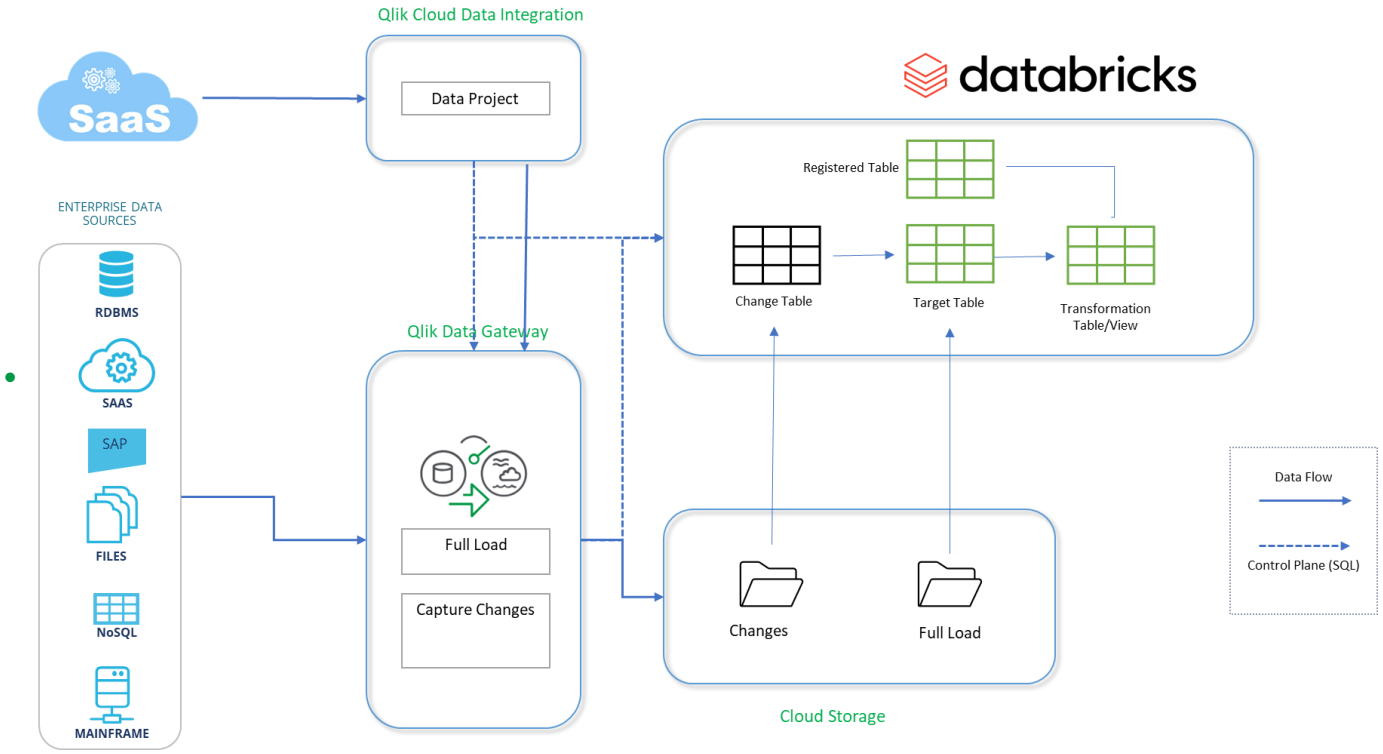
- Qlik Cloud Data Integration は、iPaaS（Integration Platform as a Service）として、データアーキテクチャおよび分析要件に対応する様々なデータ統合タスクを実行するためのデータパイプラインを生成します。オンボーディングデータにより、増分データの移動を自動化することができます。ファイアウォールの内側にあるオンプレミスのデータソースに対しては、安全な接続によってリアルタイムなログベースの変更データキャプチャ機能を活用することができ、また SaaS データソースに対しては、完全ロード機能を使用することができます。一旦データをオンボードしたら、目的に合ったアウトプットを生成するために変換機能を適用したり、データマートのファクトやディメンションなどのパターンを自動化することができます。また、外部ビューおよびライブビューは、データを利用するために生成されます。さらに、Qlik Cloud Data Integration は、完全なタイプ 2 の履歴データストア（HDS）も生成します。

- Qlik Cloud Data Integration から出力された精緻なデータは、次のような様々な目的に使用することができます：

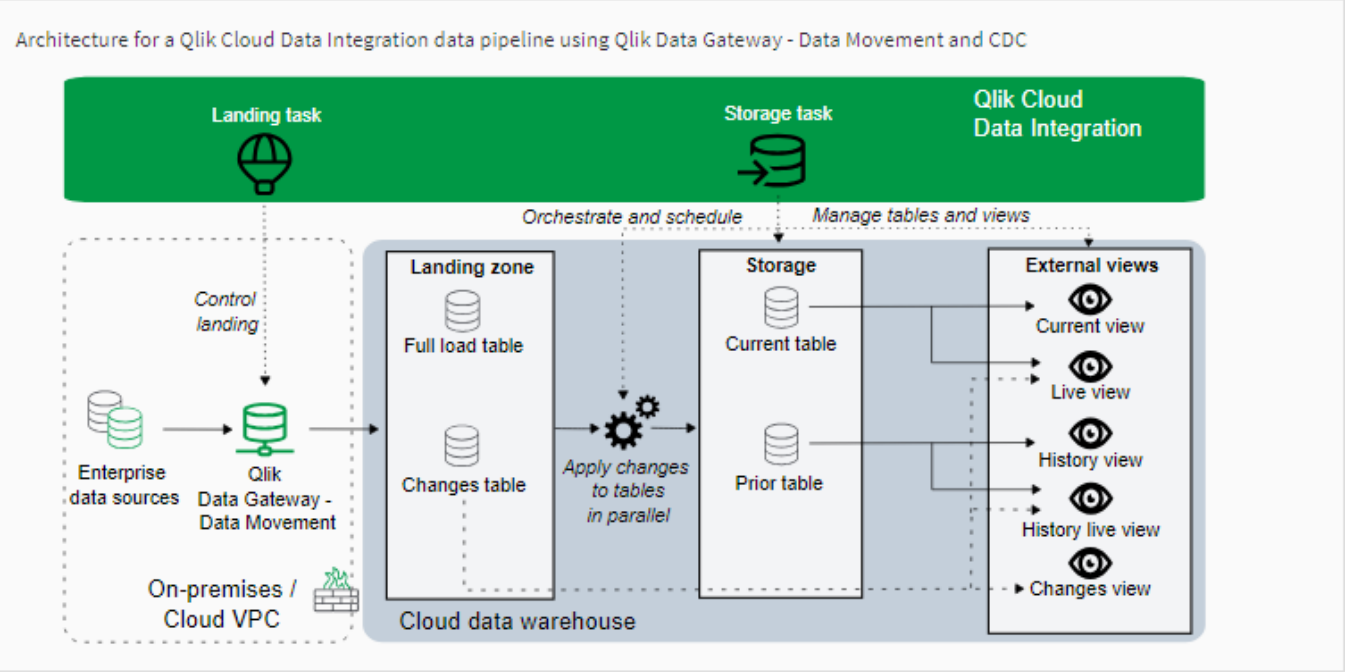
  - リレーショナルデータベース、SAP、メインフレーム、SaaS アプリケーションを含むすべてのエンタープライズソースからのリアルタイムなデータ移動
  - ELT を使用したデータ変換（抽出 / ロード / 変換）
  - Databricks Lakehouse 内での分析用データマートの自動作成
  - 機械学習やその他の取り組みへのサポートに向けたデータリポジトリのモダナイゼーション

このアーキテクチャにおいて、Qlik Cloud Data Integration は次の機能を実行します：

- ターゲットのインスタンス化
    - ソーステーブルから変換された適切なデータタイプを使用して、DELTA 形式でターゲットテーブルを作成
    - ソースから初期 / 完全ロードを実行し、データをストレージレイヤーへ送信
        - SaaS アプリケーション – 直接接続
        - リレーショナルデータベース – データゲートウェイを使用
    - ストレージレイヤーからデータをロードするためにSpark SQL を Databricks へ送信し、DELTA 形式でテーブルへと変換
- 変更内容のキャプチャと適用
    - ログベースの CDC を使用してソースから変更内容をキャプチャ
        - SaaS アプリケーション – 直接接続
        - リレーショナル データベース – Qlik Data Gateway を使用
    - ターゲットとなる DELTA テーブルに対して変換内容（挿入 / 更新 / 削除）を適用（中間ステージングとしてストレージレイヤーを使用）
- SparkSQL を Databricks へ送信するための変換を実行



- このアーキテクチャを別の視点で捉える方法としては、“特化した機能を持つタスク”という概念を通して考える形態があります。



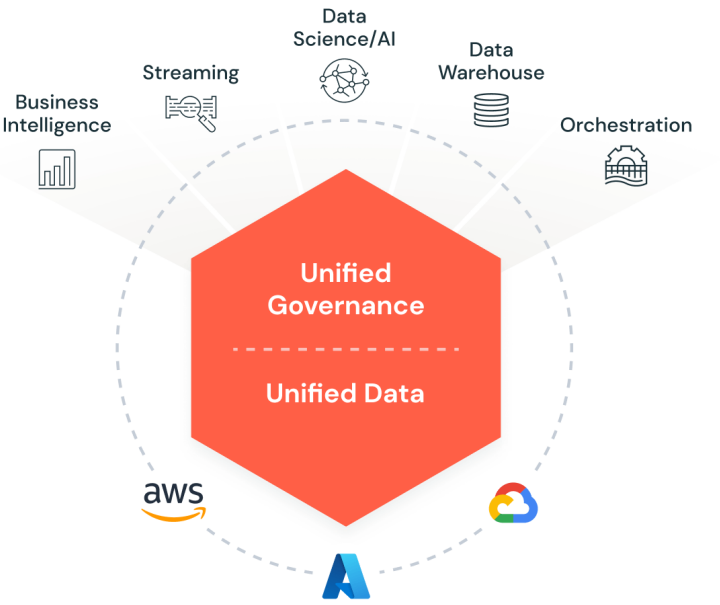
1. ランディングタスク - 様々なソースから指定されたランディングゾーンへのデータのシームレスな転送を監視します。図中のイラストは、Qlik Data Gateway - Data Movement を利用し、変更データキャプチャ（CDC）を介してデータソースへアクセスすることで、データが常に最新の状態に保たれていることを示しています。さらに、Qlik Cloud Data Integration のソース接続機能を使用することで完全ロードを実行できるため、スケジュール化された定期的なリロード対応が可能となります。
2. ストレージタスク - テーブルおよび外部ビュー両方の作成と管理を含め、ストレージテーブルへのデータの適用を監視します。この極めて重要なタスクは、Qlik Cloud Data Integration 環境内でデータの整合性とアクセシビリティを維持する上で中核的な役割を果たします。ストレージタスクは、データアプリケーションのタイミングを制御するだけでなく、ストレージインフラストラクチャに対する情報のシームレスな統合を保証し、Qlik Cloud Data Integration プラットフォームの全体的な効率や機能を強化します。
3. 変換タスク - データパイプライン内で、再利用可能なルールベースのデータ変換機能を生成することができます。これらの変換機能は、シームレスにデータオンボーディングプロセスに組み込んだり、再利用可能な変換データタスクとして設定したりすることができます。その柔軟性については、行レベルの変換を実行やカスタム SQL を使用したデータセットにまで及び、これらはテーブルとして実体化することも、またダイナミックビューとして臨機応変に変換内容を適用することも可能です。

4. データマートタスク - データのオンボードが正常に完了すると、ストレージタスクまたは変換タスクから取得した情報を使用してデータマートを生成することができます。ビジネス要件に合わせて、複数のデータマートを作成することが可能です。理想的には、これらのデータマートは、営業部門など組織の特定の部門内またはユニット内で、分析目的で収集された集約データのリポジトリとして機能するか、また、ML プロセスで使用される機能として公開される必要があります。

クイック比較		
	Qlik Replicate	**Qlik Cloud Data Integration**
導入	クライアント管理（オンプレミスまたはクラウド上で実行される VM）	iPaaS（サービスとしての統合プラットフォーム）
価格設定	ソース（コア / 容量 / ボリューム） ターゲット数	消費ベース
サポート対象ソース/ターゲット	最新のリストを取得する場合には、サポート対象のソースエンドポイントおよび サポート対象のターゲットエンドポイントを確認してください。	SaaS アプリケーションおよびデータベース – 最新のリストを取得する場合には、（データの統合 | Qlik Cloud Help）およびターゲットプラットフォームへの接続を参照してください。
レプリケーショントポロジー	- 1 対 1 - 1 対多 - 多対 1 - 双方向	- 1 対 1
導入形態	- オンプレミス - クラウド上の Compute Engine（AWS / GCP / Azure） - Linux または Windows	- SaaS - Linux ボックス（オンプレミスまたはクラウド環境）内に展開された Data Gateway
ユースケース	- データの移行 - データの複製 - データのストリーミング - データの変換（メモリ内、行内）	- データの移行 - データの複製 - データの変換（メモリ内、行内） - データの変換（SQL を使用した ELT） - データマートオートメーション（SQL を使用）

Databricks Data Intelligence Platform

Databricks Data Intelligence Platform は、データレイクとデータウェアハウスの要素を組み合わせることで、データウェアハウスの信頼性、強力なガバナンス、パフォーマンスと、データレイクのオープン性、柔軟性、機械学習サポートを実現します。このアプローチによって、これまでデータエンジニアリング、分析、BI、データサイエンス、機械学習を分離させ、また複雑化させてきたデータサイロを排除することで、最新のデータスタックをシンプル化することができます。柔軟性を最大限に高めるために、本プラットフォームはオープンソースおよびオープンスタンダードに基づいて構築されています。また、データ管理、セキュリティ、ガバナンスに対する共通のアプローチにより、組織はより効率的に事業を運営できるだけでなく、より迅速なイノベーションが可能となります。



SQL Warehouses vs. 一般的なコンピューティングクラスター

Qlik ソリューションは、Databricks SQL Warehouses および Compute Cluster の両方をサポートしていますが、これらはクラウド環境でデータを処理する 2 つの異なる方法です。 どちらの場合も、Qlik ソリューションは SparkSQL コマンドを送信してデータを処理し、他のサポート機能（例えば、notebooks in Scala など）には依存しません。Databricks SQL Warehouses と一般的なコンピューティングクラスターのどちらを選択するかは、各プロジェクト特有の要件や目標に応じて決定されます。その際、考慮すべき要素としては、次のようなものがあります：

- データ量
- CDC の頻度：SQL Warehouses は初回の起動が速く、一方、クラスターは最初のコマンドへの応答が遅くなることがあります。
- 拡張性および弾力性のレベル：SQL Warehouses は、同時利用ユーザーやクエリの要求に合わせて自動的にスケールアップまたはスケールダウンすることができます。一般的なコンピューティングクラスターも拡張が可能ですが、さらに手作業での対応や調整が必要になります。

- セキュリティおよびガバナンスのレベル：SQL Warehouses は、暗号化、認証、承認、監査、コンプライアンスなどの組み込み型のセキュリティ機能を提供します。一般的なコンピューティングクラスターでもセキュリティ対策を実装できますが、より多くの構成や管理が必要になります。

要約すると、Databricks SQL Warehouses および一般的なコンピューティングクラスターは、どちらもクラウド環境でのデータ処理に向けた強力で信頼性の高いソリューションと言えます。ただし、それぞれ異なる強みや弱みがあるため、どちらかを選択する前に慎重に検討する必要があります。

Qlik ソリューション向けの一般的なコンピューティングクラスターに関する推奨事項

免責事項

以下の推奨事項は、参考および情報提供のみを目的としており、特定の要件に関して Qlik やそのパートナーが実施したプロジェクトおよび POC に基づいた内容となっています。ネットワークトポロジー、レイテンシー、テーブル構造、更新頻度、ドライバーのバージョンなどのいくつかの要因により、特定のユースケースで必要となる設定に影響がおよぶ可能性があります。 お客様には、必要なスコープの設定と慎重な精査を行い、独自の構成を決定するよう推奨しています。

1. **Databricks ランタイム**

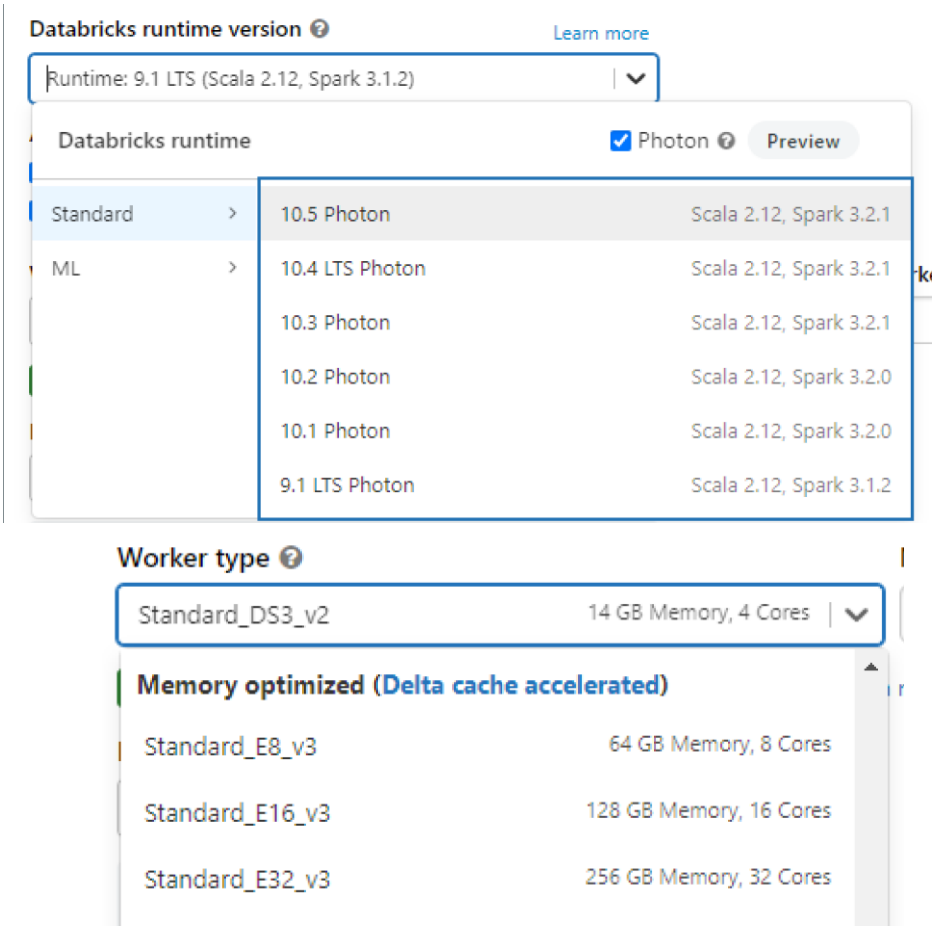
クラスターを構成する際には、必ず Qlik Replicate および Qlik Cloud Data Integration (http://help.qlik.com) でサポート対象の Databricks Runtime をご確認ください。

2. **Photon をサポートする Databricks のランタイムバージョン**

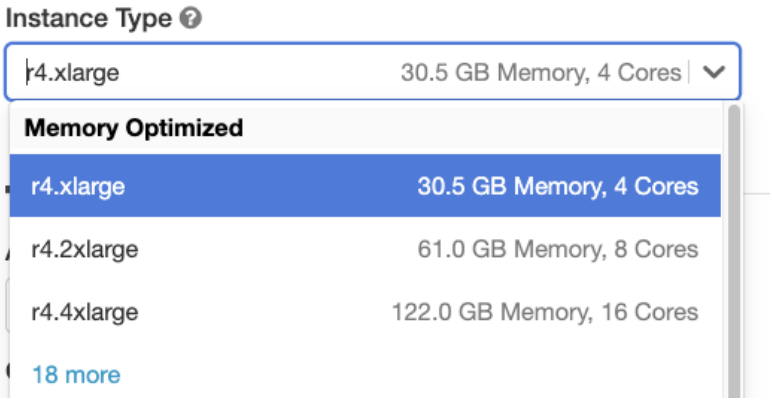
クラスターを構成する際には、汎用クラスターをサポートする **[Databricks runtime version（Databricks のランタイムバージョン）]** の選択で"Photon"を選びます。Photon は Databricks 上のネイティブなベクトル化クエリーエンジンであり、Apache Spark と直接互換性を持つように設計されています。Photon は高性能ランタイムの一部であり、既存の SQL および DataFrame API の呼び出しを高速に実行し、ワークロードあたりのトータルコストを削減します。Photon の詳細については、こちらのドキュメント https://docs.databricks.com/runtime/photon.html をご覧ください。

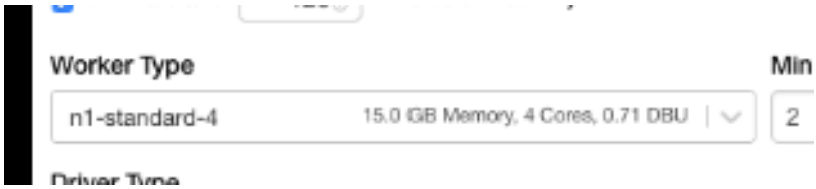
3. **"Memory optimized – Delta cache accelerated（メモリの最適化 – Delta キャッシュの高速化）"を選択します。**

クラスターを構成する際には、必ず **[Worker type（ワーカータイプ）]** で"Memory optimized – Delta cache accelerated（メモリの最適化 – Delta キャッシュの高速化）"を選択してください。



(*) 上記のリストは Azure Databricks を前提とした内容となっており、AWS または GCP を使用している場合には内容が変更される可能性があります。



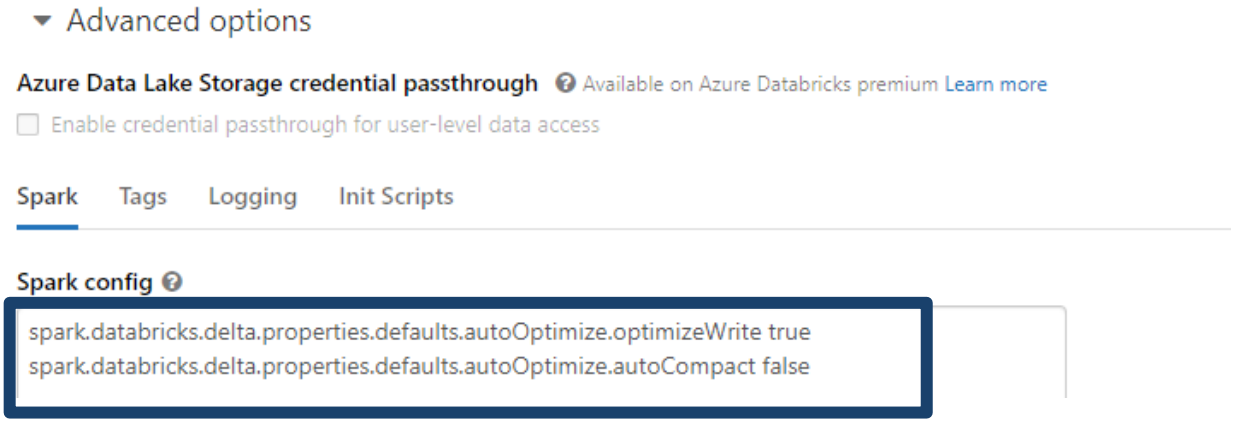


4. Auto-Optimize（自動最適化）オプションの構成

クラスターに構成を追加して、optimizeWrite を有効化し、また autoCompact を無効化します。autoCompact の無効化は、リアルタイムの CDC 更新によって（レイテンシーの拡大につながる可能性がある）シリアルコンパクションがトリガーされることを防ぐために不可欠となります。この対応を行う際には、クラスターの **[Advanced options（詳細オプション）]** に関する **Spark** セクションに以下の行を追加します。

```
smile.databricks.delta.properties.defaults.autoOptimize.optimizeWrite true
spark.databricks.delta.properties.defaults.autoOptimize.autoCompact false
```

クラスターの構成に関する詳細については、https://docs.databricks.com/clusters/configure.html をご覧ください。



5. 定期的なテーブルの最適化

Delta Lake 内のテーブル最適化に向け、notebook をスケジュールことが重要となります。これにより、取得されたデータに対するクエリの速度が向上します。テーブルを最適化するための notebook のサンプルについては、ドキュメント https://docs.microsoft.com/en-us/azure/databricks/delta/optimizations/file-mgmt をご覧ください。

6. 自動スケーリング

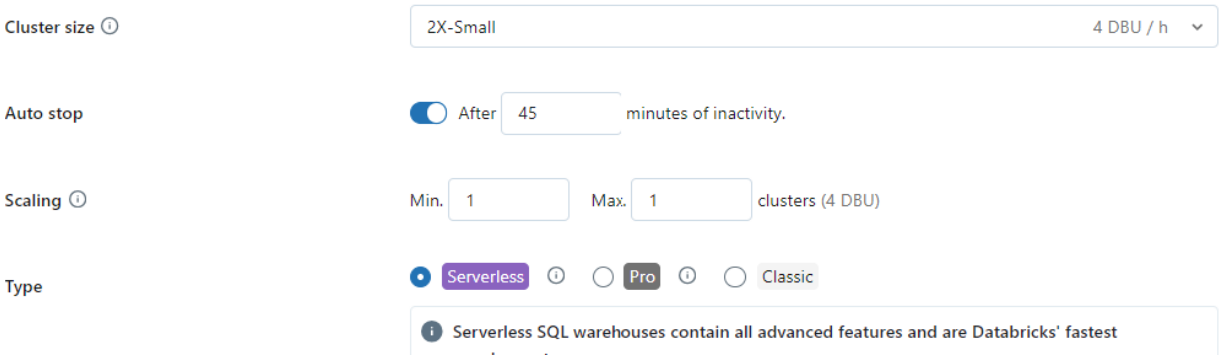
CDC がもたらすワークロード量の変動性を考慮し、ワークロードおよびタスクのテスト結果に基づいて構成内容を見直しモニタリングしながら、使用状況に応じて構成内容を調整することを推奨しています。クラスターのパフォーマンス監視方法については、Databricks のドキュメント (https://docs.databricks.com/clusters/clusters-manage.html#monitor-performance) をご覧ください。

Qlik ソリューション向けの SQL Warehouses に関する推奨事項

免責事項

以下の推奨事項は、参考および情報提供のみを目的としており、特定の要件に関して Qlik とそのパートナーが実施したプロジェクトと POC に基づいた内容となっています。ネットワークトポロジー、レイテンシー、テーブル構造、更新頻度、ドライバーのバージョンなどのいくつかの要因により、特定のユースケースで必要となる設定に影響がおよぶ可能性があります。 お客様には、必要なスコープの設定と慎重な精査を行い、独自の構成を決定するよう推奨しています。

SQL Warehouses の場合は、（クラスターと比較して）ウェアハウスレベルで構成するオプションがはるかに少なくなっています。利用可能な構成を以下に示します。



1. ウェアハウスのタイプ

本ドキュメントの執筆時点では、3 つのウェアハウスタイプがあります。SQL Warehouses についての一般的な説明については、ドキュメント What are SQL Warehouses?（SQL Warehouses とは?）をご覧ください

	Photon Engine	Predictive IO	Intelligent Workload Management
Serverless	X	X	X
Pro	X	X	
Classic	X		

パフォーマンスと同時実行性の観点から、サーバーレスのウェアハウスを使用してタスクの全体的なパフォーマンスを向上させるという対応が一般的に推奨されています。一部の環境やアカウントでは、このオプションが利用できない場合があります。このような状況では、Pro ウェアハウスが推奨されます。

2. **スケーリング**

   タスクパラメーター"**Maximum number of tables to load in parallel（並行してロードするテーブルの最大数）"**（Replicate の場合）、または "**Maximum number of database connections（データベース接続の最大数）"（Qlik Cloud Data Integration の場合）** に基づいてこちらのパラメーターを増加させます。一般的な経験則では、1 つのウェアハウスクラスターで 2 ～ 3 のテーブルまたは接続を並行して処理します。

3. **クラスターのサイズ**

   このパラメーターは、処理される対象のデータよって内容が大きく異なります。テーブルの数、各テーブルのカラム数とそのタイプ、更新頻度など、いくつかのパラメーターが全体のパフォーマンスに影響を与える可能性があります。

   一般的には、コストとパフォーマンスのバランスが良いサイズ（例えば "Medium"）から始め、ベースラインと比較しながらこのパラメーターを調整（増減）させ、テストするという形態が推奨されています。

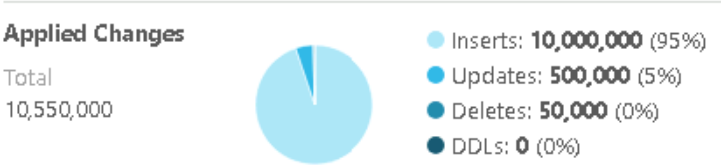
SQL Warehouses と一般的なコンピューティングクラスターのパフォーマンス比較

免責事項

このテストはラボ環境で実施されたものであり、実際の環境を反映したものではありません。実施結果については、ソースのタイプ、トポロジ、ボリューム、レコードサイズ、ソースデータベースの調整、その他の変数によって異なる場合があります。この実施結果は正式な推奨事項ではなく、お客様のテストの参考情報として利用されるべきものです。

環境:

- ソース
  - VM 上で実行される PostgreSQL
  - 7 つのカラムと 3,600 万件のレコードを格納した 1 つのテーブル
  - 主キー（1 カラム）
  - 以下のプロファイルを使用した CDC スクリプトのテスト



- ターゲット（Azure 上の Databricks）– ADLS gen2 でのステージング

SQL Warehouse		クラスター	
Type	Pro	Summary	
Cluster size	Medium	1-4 Workers	64-256 GB Memory 8-32 Cores
Auto stop	After *120* minutes of inactivity	1 Driver	14 GB Memory, 4 Cores
Scaling	Cluster count: Active *2* Min *1* Max *4*	Runtime	10.4.x-scala2.12
Channel	Current (v 2023.40)	Photon / Standard_E8d_v4 / Standard_DS3_v2 / 5–18 DBU/h	
Spot instance policy	Cost optimized		

- Windows 上の Qlik Replicate 2023 年 11 月（2023.11.0.149）
- プロセス
  - 2 つのタスク（同じソースで異なるターゲットエンドポイント）
  - タスク 1 の完全ロードを開始し、完了を待ちます
  - タスク 2 の完全ロードを開始し、完了を待ちます
  - 変更をシミュレートする SQL スクリプトを開始します両タスクが並行して実行され、Databricks 上の異なるスキーマに配信されます

実行結果：

Qlik Enterprise Manager を使用してすべての統計を収集したところ、SQL Warehouses とクラスターの使用によるパフォーマンスの違いは特に見られませんでした。コストの観点から見ると、通常、SQL Warehouses の方がコスト効率に優れており、クラスターよりも優先使用される選択肢となる可能性が示唆されます。

Task Name	Server Name	Server Utilization			Full Load		Change Processing		
		Avg Disk Usa...	Avg Memory	Avg Task CP...	Max Load D...	Avg Target T...	Total Applied...	Avg Apply T...	Avg Apply L...
BP Cluster	CLEAS Repli...	50.53 MB	590.48 MB	0.7	00:15:52.509	88,989	30,545,514	92,558	10
BP Warehouse	CLEAS Repli...	66.56 MB	700.27 MB	0.7	00:16:12.018	88,999	28,640,297	108,994	20

Qlik Cloud Data Integration に関する推奨事項

1. タスク内テーブルの選択

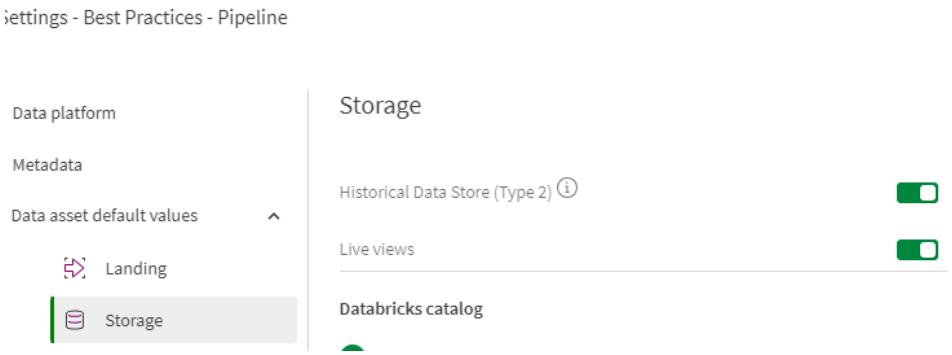
Databricks では、独自のタスクに対して多くの処理を実行する大きなテーブルまたは幅の広い（カラム数が多い）テーブルについては、分離することを推奨しています。このアプローチを使用することで、パフォーマンスの監視や特定のタスクへのウェアハウスの割り当てが容易になります。

2. タスク内での変換

Lakehouse に取り込まれるトランザクション性の高いデータのスループットを最適化することが目的の場合には、タスクレベルでの変換を最小限に抑えることが推奨されます。このアプローチでは、データをそのまま Lakehouse に取り込み、利用可能なすべてのデータエンジニアリング機能を活用して変換を行います。このプロセスは、技術的には ETL（抽出-変換-ロード）から ELT（抽出-ロード-変換）への変換と呼ばれます。

3. ストレージ / 変換における履歴データストア（タイプ 2）

Qlik Cloud Data Integration は、デフォルトでタイプ 2 SCD（Slowly Changing Dimension）コンセプトに基づき、現在および履歴データを長期間にわたり保持および管理する履歴データストア（タイプ 2）アセットを作成します。削除内容も含めたレコードの全バージョンが保持され、各レコードがアクティブだった期間を示す日付が付与されます。この情報が不要な場合は、プロジェクト設定のストレージまたは変換セクションで当機能を無効にすることができます。これにより、ワークスペースのリソース（ストレージスペースやコンピューティング）が節約されます。



4. 変換タスクのマテリアライゼーション（具体化）

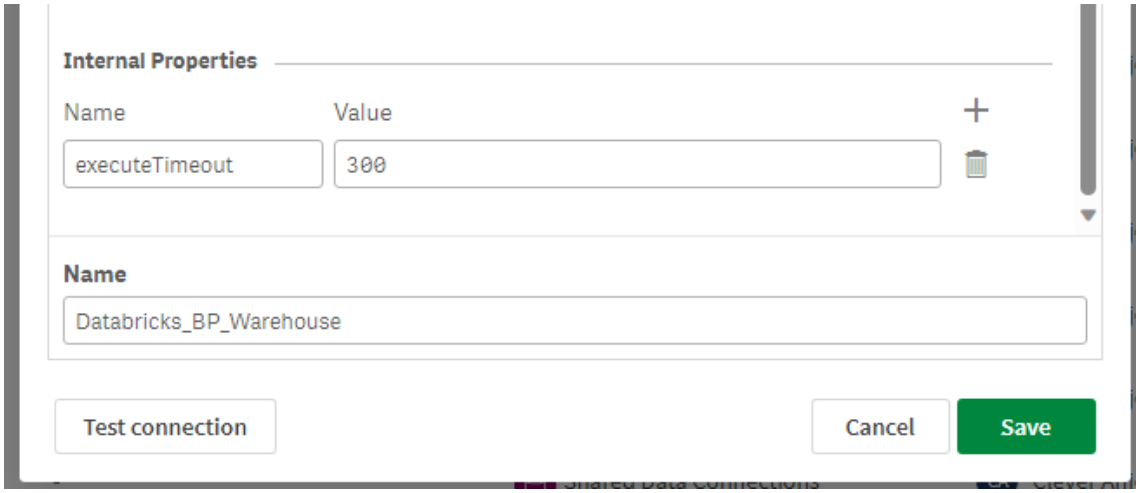
変換タスクから生成されたデータは、次の 2 つの方法で公開することができます：

- ビュー - このエンティティに対するすべてのクエリは、ストレージゾーンの元のテーブルに対して実行されます。
- テーブル - ストレージゾーンから得られたデータに基づき、スケジュールされた方法でテーブルへデータが入力されます。

どちらを選択するかは、変換済みデータへのアクセス頻度によって決定されます。変換頻度が低いデータはビューとして公開される可能性が高く、変換頻度が高いデータはテーブルとして保持する必要があります。両方を使用する必要がある場合には、2 つ（またはそれ以上）の変換タスク（1 つはビューに基づき、もう 1 つはテーブルに基づく）を使用するというオプションもあります。

5. 接続レベルでのタイムアウト

Qlik Cloud Data Integration の最適な動作を保証するためには、**executeTimeout** と呼ばれる内部プロパティを 300 より大きい値で構成することが重要となります。この構成により、Qlik Cloud Data Integration システムは障害を登録する前に少なくとも 5 分間の待機期間を維持するようになります。この結果、仮にウェアハウスが停止状態であった場合でも、その初期化に向けた十分な時間が確保されます。これは、ウェアハウスの起動時間が長くなるシナリオで特に重要となります。



Qlik Replicate に関する推奨事項

1. **タスク内テーブルの選択**

   Databricks では、独自のタスクに対して多くの処理を実行する大きなテーブルまたは幅の広い（カラム数が多い）テーブルについては、分離することを推奨しています。このアプローチを使用することで、パフォーマンスの監視や特定のタスクへのクラスターの割り当てが容易になります。

2. **タスク内での変換**

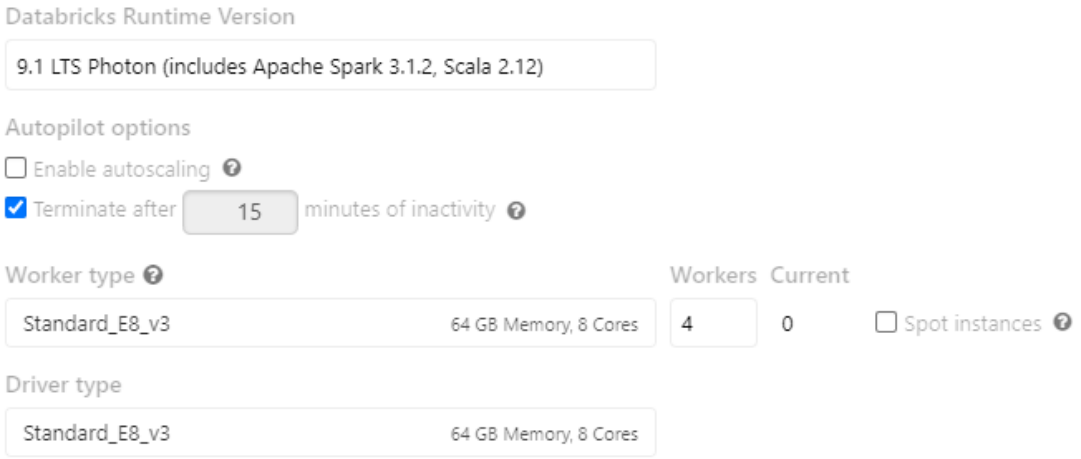
   Lakehouse に取り込まれるトランザクション性の高いデータのスループットを最適化することが目的の場合には、タスクレベルでの変換を最小限に抑えることが推奨されます。このアプローチでは、データをそのまま Lakehouse に取り込み、利用可能なすべてのデータエンジニアリング機能を活用して変換を行います。このプロセスは、技術的には ETL（抽出-変換-ロード）から ELT（抽出-ロード-変換）への変換と呼ばれます。

3. **ファイルサイズの設定**

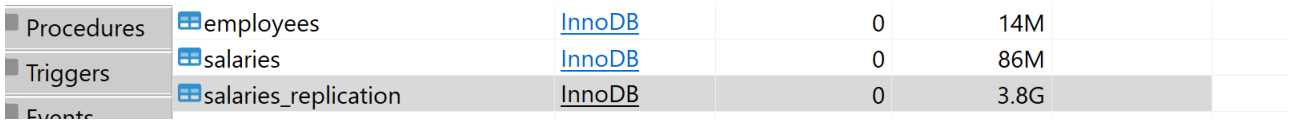
   接続レベルについては、データスループットを向上させる Qlik Replicate パラメーターがあります。これは **[Maximum file size（MB ：最大ファイルサイズ）]** と呼ばれるもので、接続の **Advanced（詳細）**設定のタブ内にあります。

   デフォルト値は 100MB で、このパラメーターはテーブルにロードされる前にステージング領域にアップロードされるファイルのサイズを示します。100M のレコード（ソース上の約 3.8Gb のデータ）を持つテーブルに対し、このパラメーターを変更した場合の影響を次に示します。このパラメーターに関して"ゴールデンルール"は存在しませんが、通常、ファイルサイズが大きくなるほどデータ転送のパフォーマンスが向上します。このため、初回の完全ロード時には非常に重要となります。

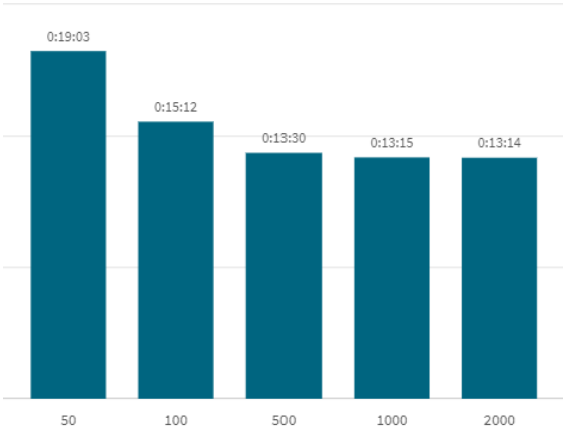
   クラスター構成：



ソーステーブル（Azure RDS MySQL）



最大ファイルサイズ（Mb）	完全ロードの経過時間
50	0:19:03
100	0:15:12
500	0:13:30
1000	0:13:15
2000	0:13:14



上の表に示されるように、ファイルサイズをデフォルト値（100MB）から 500MB に増やすことで非常に良好なパフォーマンスの改善が見られましたが、500MB を超えてさらにファイルサイズ増やした場合には、パフォーマンスへの影響ははるかに小さいものとなっています。

4. バッチチューニング設定

Qlik Replicate のマイクロバッチ変更は、Databricks Delta への配信を最適化し、またタスクに対するバッチチューニング構成は、Databricks に送信されるマイクロバッチのサイズに影響を与えます。

a. **処理モードの変更：** Databricks のターゲットに対しては、バッチ最適化の適用のみがサポートされています。

b. **複数のテーブルに対するバッチ化された変更の同時適用：**このオプションでは、Databricks へデータをアップロードして適用するために並行して動作するスレッドの数を構成します。

デフォルト値は 5 で、最大値は 50 となります。この値を増加させると、特定のバッチ内に CDC を含むテーブルが多数ある場合には、スループットが向上しますが、追加のクラスターリソースが必要になる場合もあります。このモードの制限事項については、Replicates ヘルプガイドでご確認ください。

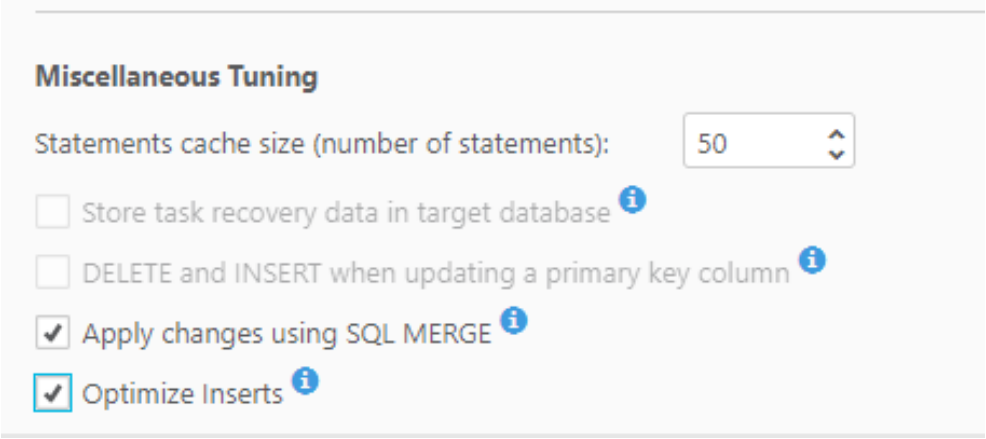
c. バッチ変更に関する適用間隔の設定では、マイクロバッチの時間とサイズを構成します。

d. **Longer than (seconds)：** ここでは、バッチ変更をそれぞれ適用する間に待機する最小時間を指定します。デフォルト値は 1 ですが、これは通常、Databricks Delta の適用プロセスに対しては低すぎる値です。この値を増やすと、ターゲットに変更が適用される頻度は減少しますが、バッチのサイズは増加します。つまり、追加のレイテンシーを伴なう代償としてより大きなバッチが作成できるようになるのです。値 60 から始めて、追加のレイテンシーが許容できる場合には、さらにこの値を増やすことを推奨しています。場合によっては、より大きなバッチを待機させることで、実際のスループットやレイテンシーが改善されることがあります。

e. **But less than (seconds)：** この値には、（タイムアウトを宣言する前に）バッチの変更を適用するたびに待機する最大時間を指定します。言い換えれば、許容可能な最大のレイテンシーということになります。このデフォルト値は 30 です。こちらの値は、**"Longer than (seconds)："**で指定した値に達した後、変更を適用するまでの最大待機時間を決定します。値については 120 に設定することを推奨しています（"Longer Than"の値 60 と組み合わせて、より長い遅延が許容される場合は、この値をさらに高く調整します）。

f. **処理メモリが（MB）を超えたときにバッチを強制適用：**この設定では、バッチ最適化適用モードでの前処理に使用するメモリの最大量を指定します。このデフォルト値は 500 です。最大バッチサイズについては、この値を Qlik Replicate に割り当て可能な最大のメモリ量へ設定します。値 2000 から始めて、Qlik Replicate サーバーに十分なリソースがある場合には、より高い値への調整を検討するよう推奨しています。

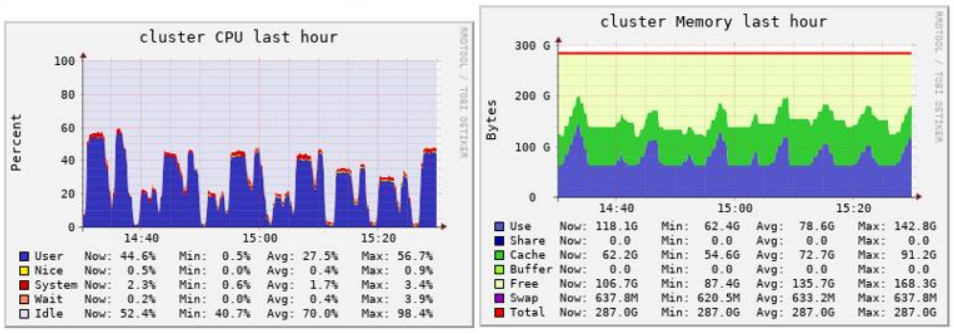
g. **マージを使用した変更内容の適用：**これにより、タスクで SQL MERGE コマンドを使用して、Lakehouse への取り込みを高速化できるようになります。



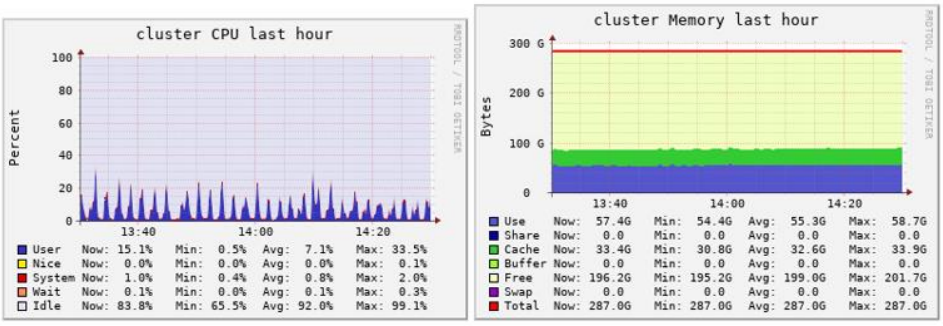
5. 大規模なテーブルのパーティション分割

Databricks は、Delta テーブルをパーティション分割する機能を提供します。適用プロセスのボトルネックとなる可能性がある大規模なテーブルについては、パーティション分割することを推奨しています。現時点で Qlik Replicate は、タスク内でのターゲットパーティショニングの構成をサポートしていません。ターゲットとなる Delta テーブルについては、Qlik Replicate で作成し、適切なパーティションカラムを使用して再作成する必要があります。テーブルがパーティション分割されるよう定義されている場合には、完全ロード時に TRUNCATE を実行するようにタスクを設定することが推奨されています。
パーティショニングは単純な概念ですが、最適なパーティショニングカラムを決定する際には、アプリケーションによってデータがどのように変更されるかを確実に理解する必要があります。カーディナリティー（カラム内のデータ値の一意性）の問題のため、主キーをパーティション分割することは推奨できません。パーティション分割が必要な大規模テーブルは、通常"トランザクション型"となっており、例えば売上データなどがこれに該当します。通常、日付のカラムを選択するか、Replicate 内のターゲットデータセットに YEAR_MONTH カラムを追加することが、適切なパーティション分割の方法となります。以下のグラフは、パーティショニングがクラスターの使用率に、そして最終的にレイテンシーへ与える影響の一例となります。この例では、約 6,800 万のソース行 / 655 GB のデータからなるテーブルで、本番環境の CDC ワークロードが処理されていました。DATE カラムを使用して Delta テーブルをパーティション分割することで、レイテンシーが 73% 削減され、クラスター上のメモリおよび CPU の消費量も大幅に削減されました。

クラスターの使用率 – パーティション分割なし



クラスターの使用率 – パーティション分割



付録 I - Qlik レプリケート接続の作成

	*Databricks 環境の最新要件やサポート対象の構成については、help.qlik.com に掲載されたドキュメントで確認することを強く推奨しています。*

1. Databricks 認証情報および接続に関する詳細情報の取得

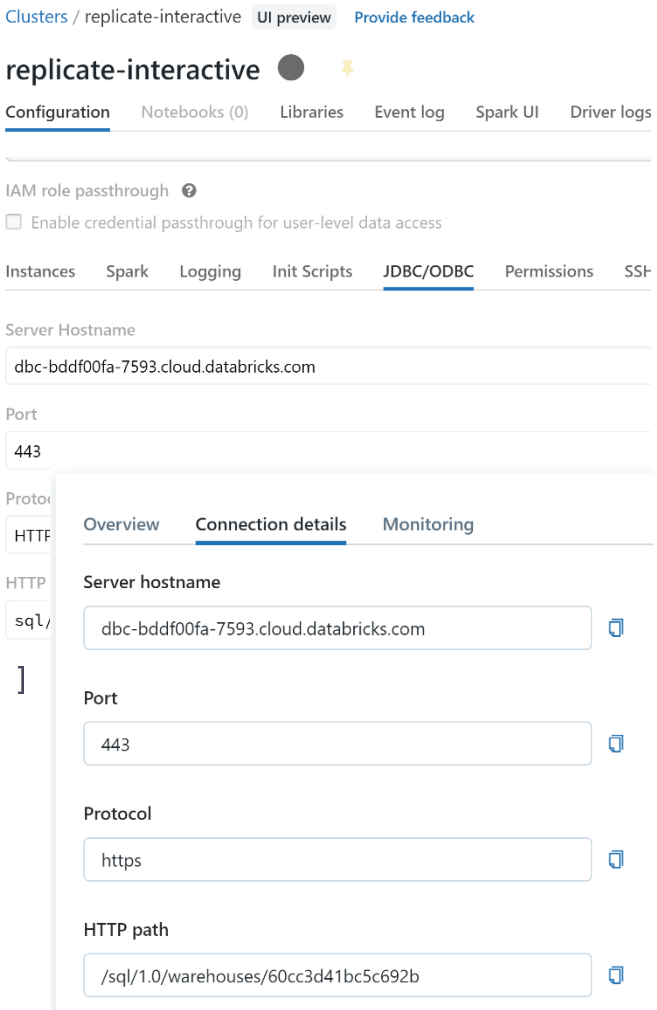
Qlik Replicate で Databricks インスタンスへのデータ接続を作成する際には、以下が必要となります。

1. Server Hostname（サーバーのホスト名）
2. Port（ポート）
3. HTTP Path（HTTP パス）
4. Token（トークン）

- Clusters（クラスター）

  1、2、3 に関する情報については、Databricks コンソールからクラスター構成または SQL エンドポイント構成に移動して取得することができます。

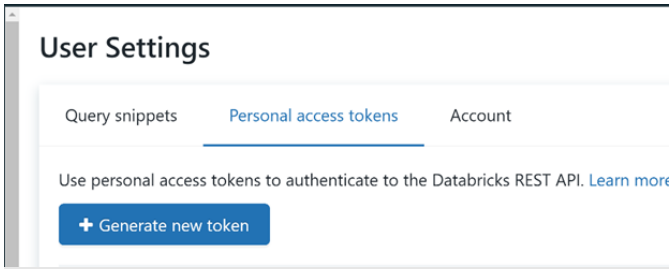
  [Advanced Options（詳細オプション）] セクションにある **[JDBC / ODBC]** タブに表示されています。

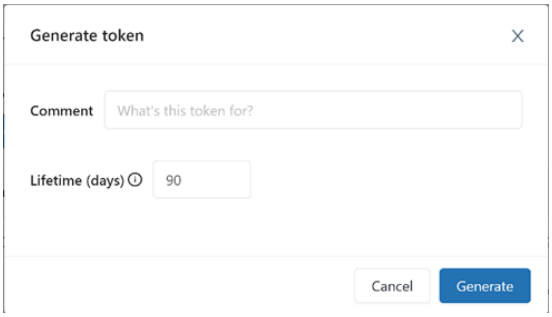


- Databricks SQL Warehouse

  1、2、3 に関する情報については、Databricks コンソールの **[SQL Warehouse]** セクションにある [Connection details（接続の詳細）] タブから取得できます。

アクセストークンを取得する際には、Databricks コンソールの [**User Settings（ユーザー設定）**] セクションへ移動し、[**Generate new token（新しいトークンの生成）**] ボタンをクリックする必要があります。



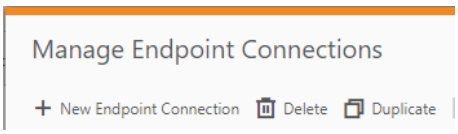


	このダイアログを閉じてしまうと再度トークンを取得できないため、生成されたトークンについては安全な場所に保管することが重要になります。

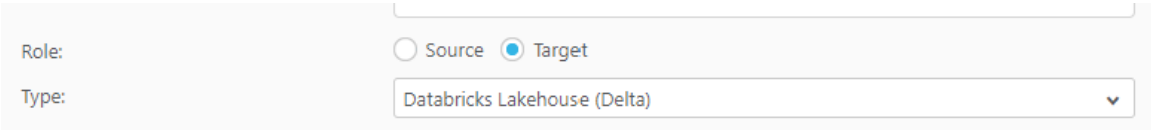
2. レプリケート接続の作成

Databricks への接続を構成する際には、次の手順を実行します。

1) help.qlik.com ドキュメントの"Prerequisites（前提条件）"および"Limitations and considerations（制限事項と考慮事項）"を注意深く確認してください。
2) Simba Spark ODBC ドライバーを Qlik Replicate サーバーマシンへインストールします。必要なバージョンについてはドキュメントで確認します
3) Qlik Replicate Console で、[**Manage Endpoint Connections（エンドポイント接続の管理）**] セクションへ移動します



4) **[New Endpoint Connection（新しいエンドポイントへの接続）]** をクリックします
5) 接続に対して分かりやすい名前と説明を付けます。
6) **[Role（ロール）]** については **[Target（ターゲット）]** を選択し、**[Type（タイプ）]** は [Databricks Lakehouse (Delta)] を選択します。



7) **[Databricks ODBC Access]** セクションでは、*[Obtaining Databricks credentials and connection details （Databricks 資格情報と接続の詳細の取得）]* セクションで収集した情報が必要になります。
    1. **[Host（ホスト）]** - Databricks ワークスペースのホスト名

2. **[Port（ポート）]** - ワークスペースへのアクセスに使用するポート
3. **[Token（トークン）]** - ワークスペースにアクセスするための個人トークン
4. **[HTTP Path（HTTP パス）]** - クラスターへのパス
5. **[Use Unity Catalog（ユニティカタログを使用）]** をオンにします（ワークスペースの UC が有効になっている場合）
6. **[Database（データベース）]** - Databricks ターゲットデータベース（スキーマ）の名前。既存のスキーマを参照できます
7. テーブルを **Managed（管理対象）** または **Unmanaged（管理対象外）** のどちらで作成するのかを選択します。

8) [**Staging（ステージング）**] セクションでは、ステージングリポジトリの**Storage type（ストレージタイプ）**と認証情報の種類を指定する必要があります。

Qlik は、複雑なデータ環境を実用的なインサイトへと変換し、戦略的なビジネス成果を促進します。全世界で 40,000 社を超えるお客様にサービスを提供する Qlik のポートフォリオでは、高度なエンタープライズグレードの AI / ML や広範なデータ品質を最大限に活用しています。また、優れたデータ統合およびデータ統制、多様なデータソースに対応する包括的なソリューションを提供しています。Qlik の直感的でリアルタイムの分析は、隠れたパターンの発見や複雑なビジネス課題の解決、新たなビジネスチャンスの獲得を支援します。当社の AI / ML ツールは実用的かつスケーラブルであり、より適切な意思決定をより迅速に導きます。Qlik は顧客の戦略的パートナーとして、プラットフォームに依存しないテクノロジーと専門知識で、顧客の競争力を高めます。

Qlik.com