

alteryx

WHITEPAPER

Databricks + Alteryx: Ein Leitfaden mit Best Practices

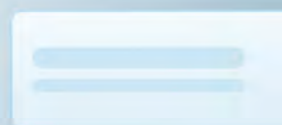
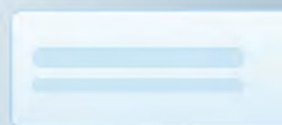
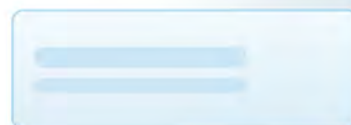
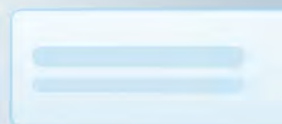
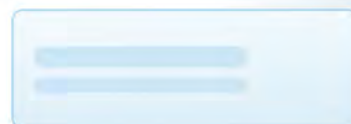
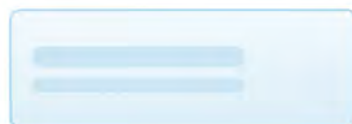


Databricks

alteryx

AI PLATFORM
FOR ENTERPRISE ANALYTICS

Search



Inhalt

03 Einleitung

05 SaaS-Lösung: Databricks + Designer Cloud

Alteryx über Databricks Partner Connect starten

In Alteryx Analytics Cloud eine Verbindung zu Databricks herstellen

Sicherheits- und Governance-Funktionen

Pushdown-Ausführung in Databricks mit Alteryx Designer Cloud

10 Desktop-Lösung: Databricks + Designer Cloud

Alteryx über Databricks Partner Connect starten

In Alteryx Analytics Cloud eine Verbindung zu Databricks herstellen

Sicherheits- und Governance-Funktionen

Pushdown-Ausführung in Databricks mit Alteryx Designer Cloud

17 Fazit und Ressourcen

Einleitung: Demokratisierung von Daten und KI im Lakehouse

Sowohl Alteryx als auch Databricks sind davon überzeugt, dass mehr Menschen mit unterschiedlichen Qualifikationen die beste verfügbare KI-Technologie und -Architektur nutzen können, um mit Daten mehr zu erreichen.

Die Herausforderung besteht darin, dass Business Analysts nicht unbedingt über die spezifischen technischen Fähigkeiten oder Programmierkenntnisse verfügen, um Daten zu nutzen. Deswegen müssen Sie Geschäftsanwender:innen für Ihr Data-Lakehouse-Projekt gewinnen. Und hier kommt Alteryx ins Spiel.

Mit einer intuitiven No-Code-Benutzeroberfläche für Datenvorbereitung, -umwandlung, -analyse, Data Science und KI/ML ermöglicht es die Alteryx AI Platform for Enterprise Analytics Geschäftsanwender:innen abteilungsübergreifend, die Vorteile der Databricks Data Intelligence Platform und Ihrer Databricks Lakehouse-Architektur zu nutzen.

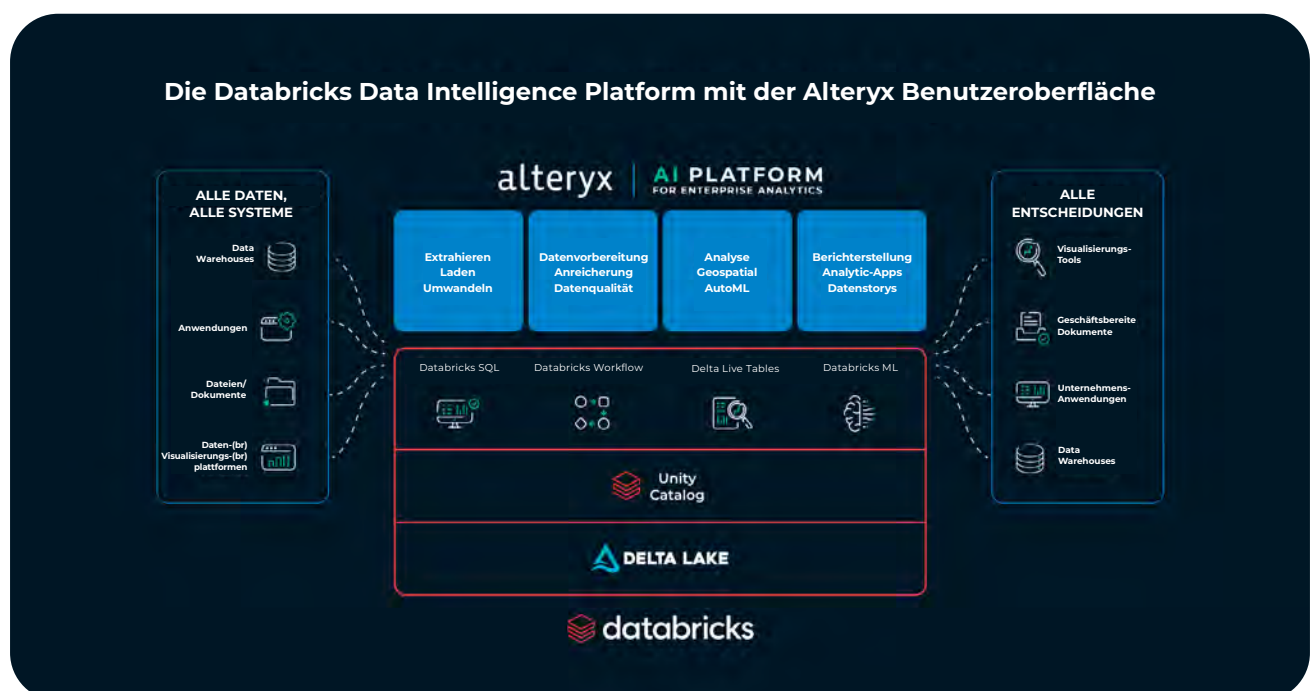
Alteryx macht mit Databricks die gesamte Datenreise von Anfang bis Ende zugänglich. Von der Verbindung zu Daten bis hin zur Ausführung von Transformationen in Databricks – mit der intuitiven Benutzeroberfläche von Alteryx ist dies alles

möglich. Stellen Sie sich Alteryx als die einfach zu bedienende, unternehmensfreundliche Benutzeroberfläche für Databricks vor.

Die Alteryx AI Platform for Enterprise Analytics ist vollständig für die Databricks Spark-, SQL-, Unity Catalog- und Delta Lake-Architektur zertifiziert. Alteryx verfügt über die Interoperabilität mit Databricks, die Sie benötigen, um den Zugriff zum Lakehouse zu demokratisieren und gleichzeitig Ihre Workflows sicher, verwaltet und für die Databricks Engine optimiert zu halten.

Dieses Whitepaper hilft Ihnen, die einzigartigen Fähigkeiten dieser sich hervorragend ergänzenden Plattformen optimal zu nutzen. Wir werden die folgenden Themen behandeln:

- Die Vorteile der Verwendung von Alteryx mit Databricks
- Einrichtung Ihrer Databricks-Verbindung mit Alteryx
- Die Vorteile der Pushdown-Ausführung in Databricks nutzen
- Wie jedes dieser Dinge sowohl für Alteryx Designer Cloud als auch für Alteryx Designer Desktop aussieht, je nachdem, welche Alteryx Lösung Ihren Anforderungen entspricht.



Vorteile von Alteryx + Databricks

Sicherer Zugriff auf Databricks-Daten über eine No-Code-Benutzeroberfläche

Alteryx bietet mehr als 80 nativ integrierte Datenquellen, einschließlich Databricks. Stellen Sie eine Verbindung zu Databricks her, um auf die von Ihnen benötigten Datasets zuzugreifen, diese Daten umzuwandeln und Erkenntnisse zu gewinnen – und das alles über die GUI-basierte No-Code-Benutzeroberfläche von Alteryx.

Mit Unterstützung für Databricks Unity Catalog können Geschäftsteams mit Data Engineers und Data Scientists an vertrauenswürdigen Daten und KI-Assets zusammenarbeiten und dabei KI-gestütztes Monitoring und Beobachtbarkeit nutzen.

Daten transformieren mit No-Code Pushdown in Databricks

Alteryx bietet eine No-Code-Schnittstelle für die Pushdown-Ausführung in Databricks. Wenn Sie einen Designer Cloud-Workflow mit Pushdown in Databricks ausführen, verwenden Sie die leistungsstarke Databricks Engine, um Ihre Daten zu verarbeiten, ohne zwischen Plattformen wechseln zu müssen. Ihre Anwender:innen brauchen Alteryx nicht zu verlassen und Ihre Daten verlassen das Lakehouse nicht.

Fundiertere Erkenntnisse fördern mit Advanced Analytics

Alteryx ist weit mehr als nur ein Vorbereitungs- und Zusammenführung-Tool. Es ermöglicht es Geschäftsanwender:innen, hochwertige Data-Science-Anwendungsfälle in Angriff zu nehmen. Nutzen Sie nicht nur demografische Standarddaten, sondern auch Verhaltens- und Einzelhandelsinformationen, um einen detaillierten Überblick zu erhalten und Ihre Kundinnen und Kunden wirklich zu verstehen. Bereichern Sie Ihre Analysen mit Karten, Adresslösungen, Fahrzeitfunktionen und einem tieferen Verständnis Ihrer Kundschaft und Standorte. Nutzen Sie die Geofunktionen in Designer und [Alteryx Location Intelligence](#), um mit Databricks-Daten standortbasierte Entscheidungen zu treffen.

Machen Sie Ihre Daten KI-fähig

Es ist eine Herausforderung, Daten für KI/ML-Workloads produktionsbereit zu machen. Das Sammeln, Bereinigen und Vorbereiten von Daten dauert ewig. Und nach all dem müssen Sie immer noch wissen, wie das Modell erstellt wird. Die Alteryx AI Plattform ändert das. Benutzer:innen ohne technischen Hintergrund können Datasets in Designer Cloud mithilfe der No-Code-Benutzeroberfläche vorbereiten und untersuchen sowie Modelle in Alteryx Machine Learning erstellen – und dabei Daten von Databricks nutzen.

Erlangen Sie Erkenntnisse und bessere Antworten, indem Sie Daten für komplexe Analysemethoden nutzen, von der Sentiment-Analyse unstrukturierter Daten bis hin zur Erstellung komplexer R-basierter Modelle mit geringen bis gar keinen Programmierkenntnissen. Aktivieren Sie die Intelligence-Ebene mit Advanced Analytics, wie [Alteryx Machine Learning](#) und [Alteryx Auto Insights](#), um umfassendere Erkenntnisse zu erhalten. Databricks-Benutzer:innen können auch die Vorteile von [Alteryx AiDIN](#) nutzen, der KI-Engine, die die Leistung von generativer KI und Machine Learning (ML) in die Alteryx Plattform einbringt.

Die Gesamtbetriebskosten reduzieren mit einem optimierten Stack

Die Lakehouse-Architektur von Databricks vereint Data Warehouse und Data Lake in einer Plattform. Die Alteryx AI Plattform optimiert dies noch weiter, indem sie Beschäftigten eine leicht zugängliche Benutzeroberfläche bietet, über die sie das Lakehouse nutzen können. Dieser Prozess reduziert die Anzahl der Tools und Stacks, die Sie möglicherweise verwenden, um unterschiedliche Fähigkeiten zu ermöglichen.

SaaS-Lösung: Databricks + Alteryx Designer Cloud

Alteryx bietet Ihnen Optionen, wie und wo Sie mit Ihren Daten arbeiten möchten, abhängig von Ihren Geschäftsanforderungen und Ihrem Data Stack. Alteryx Analytics Cloud (AAC), zu der auch Alteryx Designer Cloud gehört, ist eine browserbasierte, native SaaS-Plattform, mit der Benutzer:innen Databricks verwenden können.

Starten Sie Alteryx Analytics Cloud über Databricks Partner Connect

Wenn Sie noch kein/e Benutzer:in von Alteryx sind, haben Sie die Möglichkeit, Alteryx Analytics Cloud ganz einfach direkt über Ihr Databricks-Konto zu starten. Ihre Databricks-Verbindung ist dabei bereits konfiguriert. Dies ist für Databricks auf AWS verfügbar.

- Wählen Sie in Databricks Partner Connect die Alteryx Designer Cloud-Kachel aus.
- Melden Sie sich auf der Registrierungsseite für eine Testversion an.
- Sobald Sie in AAC angekommen sind, haben Sie Zugriff auf die verschiedenen verfügbaren Anwendungen und Ihre Databricks-Verbindung ist einsatzbereit.
- Von dort aus können Sie auf Daten aus Databricks zugreifen und diese in Ihren Workflows verwenden.

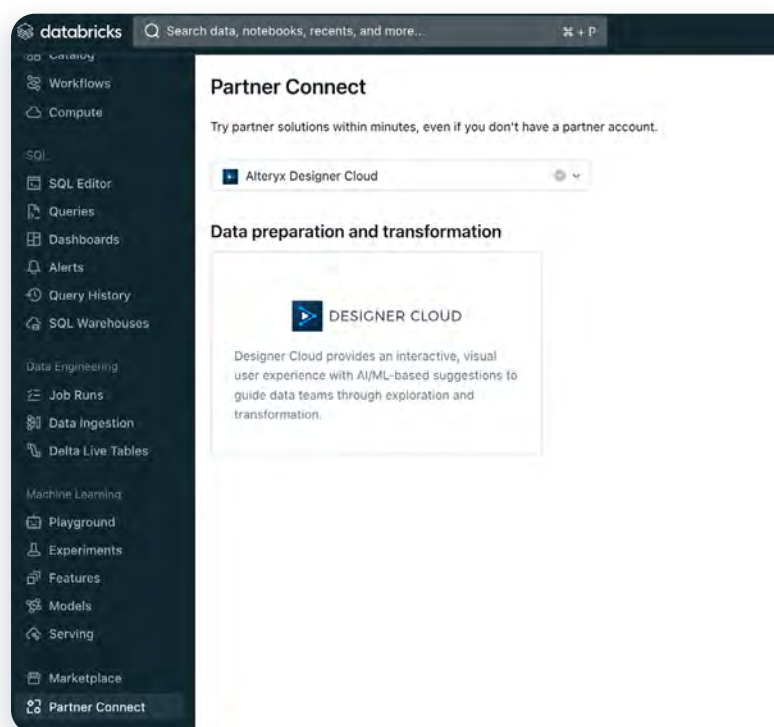
Herstellung einer Verbindung zu Databricks auf AWS oder Azure

Der/die AAC-Administrator:in kann einem AAC-Arbeitsbereich einen oder mehrere Databricks-Arbeitsbereiche hinzufügen, indem er/sie die Databricks-Service-URL und ein persönliches Zugriffstoken (PAT) angibt.

Voraussetzungen

1. Konfigurieren Sie für Databricks auf AWS Ihre Basisspeichenumgebung auf S3 und deaktivieren Sie ADS, bevor Sie Databricks-Arbeitsbereiche einrichten. Weitere Informationen finden Sie unter [AWS als privater Datenspeicher](#).

2. Konfigurieren Sie für Azure Databricks Ihre Basisspeichenumgebung für ADLS und deaktivieren Sie ADS, bevor Sie Databricks-Arbeitsbereiche einrichten. Weitere Informationen finden Sie unter [ADLS als privater Datenspeicher](#).
3. Sie haben einen Databricks E2-Arbeitsbereich.
4. Rufen Sie ein persönliches Zugriffstoken (PAT) von Databricks ab. [Erfahren Sie](#), wie Sie Ihren PAT generieren.
5. Bitten Sie eine/n Administrator:in, den Databricks-Arbeitsbereich in AAC zu konfigurieren. [Erfahren Sie](#), wie Sie einen Databricks-Arbeitsbereich konfigurieren.
6. Verbinden Sie AAC mit Ihrem Databricks-Arbeitsbereich. [Erfahren Sie](#), wie Sie Ihren Databricks-Arbeitsbereich verbinden.



Erstellen Ihrer Databricks-Verbindung in AAC

Wenn Sie Databricks Partner Connect nicht verwenden oder Ihre Verbindung manuell einrichten müssen, finden Sie in der [Dokumentation hier](#) weitere Informationen zum Einrichten Ihrer Databricks-Verbindung.

1. Wählen Sie auf der AAC-Landingpage in der globalen Kopfzeile „Verbindungen“ aus.
2. Wählen Sie oben auf der Seite die Option „Verbindung erstellen“ aus.
3. Suchen Sie nach Databricks und wählen Sie dann die Databricks-Verbindung aus.
4. Geben Sie einen eindeutigen Verbindungsnamen und eine optionale Verbindungsbeschreibung ein.
5. Wählen Sie unter „Arbeitsbereichsname“ einen verfügbaren Databricks-Arbeitsbereich aus.
6. Das persönliche Zugriffstoken für den ausgewählten Arbeitsbereich wird automatisch ausgefüllt.
7. (Optional) Geben Sie Optionen für die Verbindungszeichenfolge ein. Einzelheiten zu dieser Option finden Sie weiter unten auf dieser Seite.
8. Wählen Sie „Verbindung testen“, um zu prüfen, ob Ihr Arbeitsbereich eine korrekte Verbindung herstellt.
9. Wählen Sie „Erstellen“.

Sie können jetzt über die [Datenseite](#) in AAC auf Ihre Databricks-Tabellen zugreifen.

Create Connection

Databricks
Import and publish
[Learn more](#)

CONNECTION INFORMATION

Connection Name
Databricks Connection

Connection Description (optional)
For demos

Workspace Name
Select an option

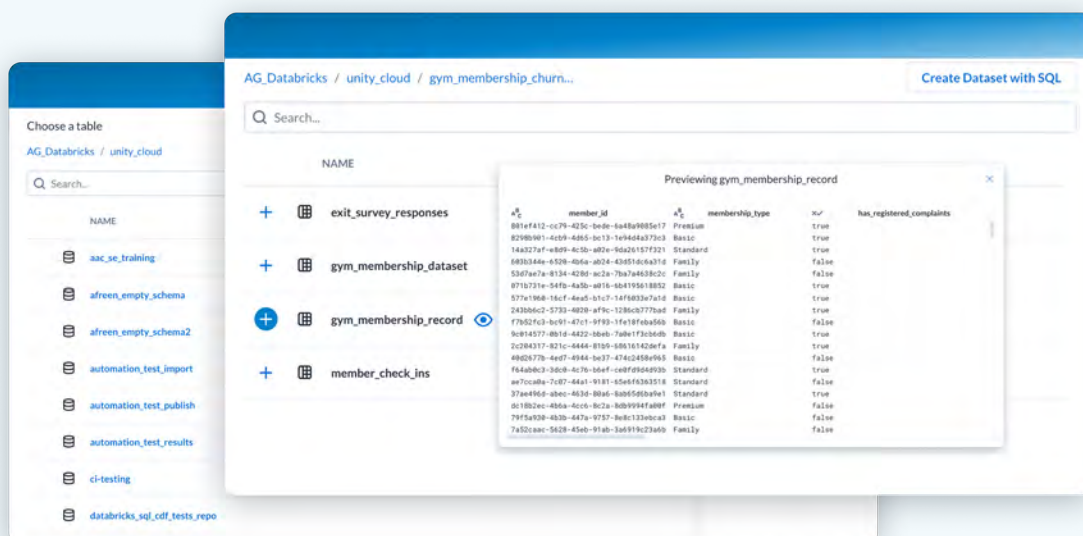
You can create only 1 connection for each Databricks workspace.

Personal Access Token
.....

This is your personal access token generated in Databricks.

Back Cancel Create

Sobald die Databricks-Verbindung festgelegt ist, wird die [AAC-Datenimportfunktion](#) verwendet, um Tabellen im Databricks Unity Catalog sowie im [Apache Hive-Metastore anzuzeigen und zu untersuchen](#). Wenn Sie die Tabelle auswählen, werden Metadaten über die Tabelle und eine kleine Stichprobe angezeigt.



Nutzung der Databricks Engine für die Pushdown-Ausführung

Sobald Sie Zugriff auf die benötigten Databricks-Daten haben, können Endbenutzer:innen mit der Vorbereitung, Umwandlung und Analyse der Daten in Designer Cloud beginnen. Mit [Designer Cloud](#) können Sie Databricks als Ausführungs-Engine nutzen, ohne dass Programmierung oder zusätzliche Schritte erforderlich sind. Ein Geschäftsanwender kann in Designer Cloud einen Workflow mit No-Code-Modulen erstellen und dann den gesamten Workflow mit Databricks ausführen.

Pushdown-Verarbeitung in Databricks Serverless SQL Warehouse

In Situationen, in denen sich alle Workflow-Eingaben und -Ausgaben in Databricks befinden, generiert Designer Cloud die Umwandlungslogik in natives Databricks SQL, das in Databricks weitergegeben und ausgeführt wird.

Dies ist die empfohlene Ausführungsmethode, da keine Daten aus Databricks ausgehen und Databricks SQL Warehouse und/oder Spark Cluster Compute für die Workflow-Ausführung verwendet werden. Je nach Ihren Anforderungen bietet Alteryx flexible Optionen.

Empfohlenes Warehouse: Databricks SQL Serverless

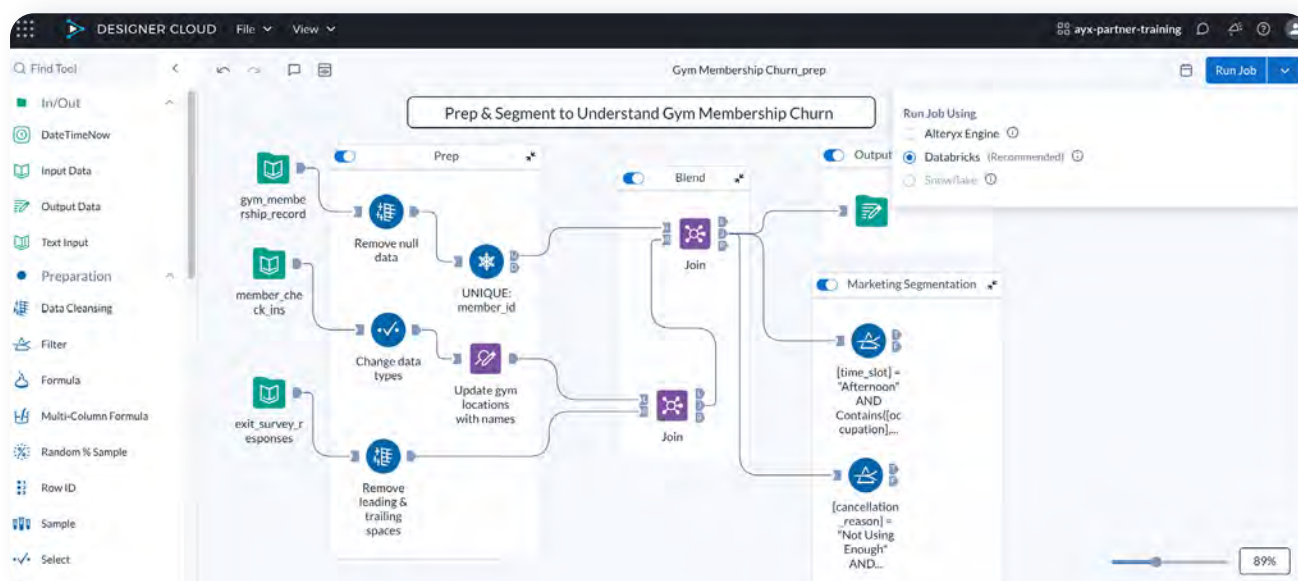
Alteryx unterstützt zwar verschiedene Databricks-Compute-Engines (Databricks SQL Warehouse, verfügbar im Serverless- oder Classic-Compute-Modus, und Databricks Spark), um Flexibilität zu bieten und unterschiedlichen Kundenanforderungen

gerecht zu werden. Es wird jedoch generell empfohlen, Databricks Serverless SQL Warehouse mit Alteryx zu verwenden, da dies mehrere Vorteile bietet:

- Databricks Serverless SQL-Warehouse wurde mit sofortigem Zugriff und elastischem Computing konzipiert, ohne Verzögerung bei der Ausführung von Abfragen und mit dynamischer Skalierung entsprechend Ihrer Workloads.
- Durch die dynamische Skalierung von Rechenressourcen, die durch Machine-Learning-Algorithmen optimiert werden, werden niedrigere Infrastrukturkosten erreicht, ohne dass ein manuelles Herunter- und Hochfahren von Clustern erforderlich ist.
- Kein Verwaltungsaufwand, da Databricks Serverless SQL Warehouse ein vollständig verwalteter Service ist, der Kapazitätsmanagement, Leistungsabstimmung, Monitoring und Support von Cluster-Ressourcen überflüssig macht.
- Leistungs- und Kostenoptimierungen mit Predictive I/O und Ergebnis-Caching.

Databricks Spark-Verarbeitung

Jede AAC-unterstützte Eingabe für jede unterstützte Ausgabe kann mit Databricks Spark als Nicht-Pushdown-Engine-Option verarbeitet werden. Databricks Spark ist eine gute Wahl für die Verarbeitung großer Datensätze. Für jeden Auftrag wird vorübergehend ein Spark Auftrags-Cluster erstellt.



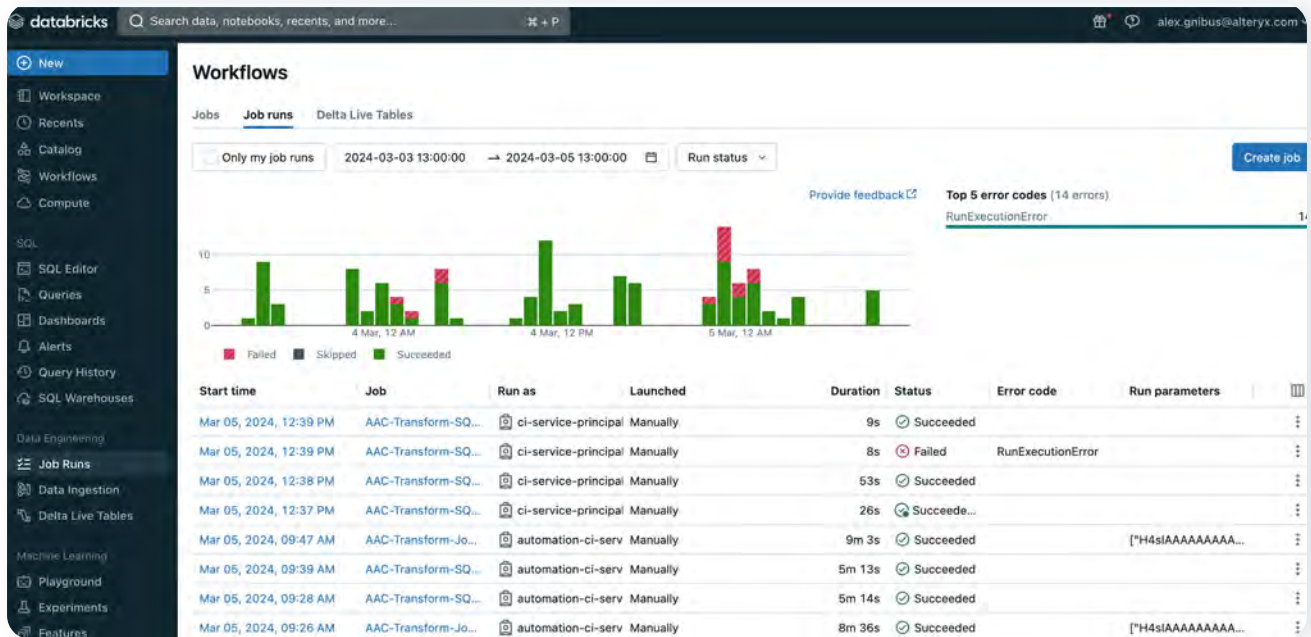
Monitoring von Designer Cloud-Aufträgen, die in Databricks ausgeführt werden

Im Abschnitt „Ausführungsphasen“ der Detailseite für AAC-Aufträge wird „Databricks SQL“ oder „Spark (Databricks)“ für die Transformationsumgebung für Designer Cloud-Aufträge angezeigt, die in Databricks ausgeführt werden. Auftragsstatus, Laufzeitdauer und Ausgabedaten werden ebenfalls auf dieser Seite angezeigt.

The screenshot shows the Databricks Analytics Cloud interface. The top navigation bar includes 'Overview', 'Library', 'Connections', 'Data', 'Schedules', and 'Jobs'. The main content area displays details for Job 395952, which finished on Thursday at 1:39 PM. The 'Overview' tab is selected. On the left, under 'Output data', it says 'Preview not available.' Below that, 'Execution stages' are listed: 'Transform' (Completed Last Thursday at 1:39 PM, started at 1:30 PM, ran for 9 min) and 'Publish' (Completed Last Thursday at 1:39 PM, started at 1:39 PM, ran for <1 sec). The environment for both is 'Spark (Databricks)'. On the right, the 'Job summary' shows Job ID 395952 and status 'Completed'. The 'Execution summary' shows Job type 'Manual', User 'Alex Gnibus', Start time 'February 29th 2024, 1:30 pm', Finish time 'February 29th 2024, 1:39 pm', Last update 'February 29th 2024, 1:39 pm', Duration '9 minutes', and Environment 'Spark (Databricks)'. The 'Optimization summary' shows 'Optimization' as 'Enabled'.

Zusätzliche Databricks-spezifische Auftragsinformationen werden auf der Seite „Auftragsausführungen“ der Databricks-Benutzeroberfläche angezeigt. Workflow-Aufträge, die mit Databricks SQL ausgeführt wurden, enthalten „SQL“ im Auftragsnamen. Aufträge ohne „SQL“ im Namen wurden mit Databricks Spark ausgeführt. Wenn Sie den Link „Auftragsname“ auswählen, wird die Seite mit den Auftragsausführungsdetails angezeigt, die Abfrageinformationen für SQL-Aufträge und Links zu Protokollen für Spark-Aufträge enthält.

The screenshot shows the Databricks Jobs page for Job 313394937277062-run-71757970261727. The left sidebar contains navigation links like 'Workspace', 'Recent', 'Catalog', 'Workflows', 'Compute', 'SQL Editor', 'Overview', 'Dashboards', 'Alerts', 'Query History', 'SQL Workflows', 'Data Engineering', 'Job Runs', 'Data Ingestion', 'Delta Live Tables', 'Machine Learning', 'Playground', 'Experiments', 'Features', 'Models', 'Serving', 'Marketplace', and 'Partner Connect'. The main content area shows 'Stages for All Jobs' with a table of execution details. The table has columns: Stage ID, Description, Submitted, Duration, Tasks: Succeeded/Total, Input, Output, Shuffle Read, and Shuffle Write. The table shows 18 completed stages and 13 skipped stages. The first stage (ID 30) is 'com.databricks.jobserver.transformer.spark.Transform... collect at Checkpoint scale 576' with a duration of 1 s and 1 task succeeded. The last stage (ID 11) is 'com.databricks.jobserver.transformer.spark.Transform... count at joinTransform scale 210' with a duration of 14 s and 200 tasks succeeded.



Workflows orchestrieren mit Databricks

Die Orchestrierung in AAC, bekannt als [Plans](#), ist eine Reihe von Funktionalitäten, die die geplante Ausführung von Aufgabensequenzen in der Alteryx Analytics Cloud unterstützen. Diese Aufgaben können externe Prozesse, Datenumwandlungsaufträge, HTTP-Anforderungen und mehr sein.

Es sind die folgenden Aufgabentypen verfügbar:

Workflow-Aufgabe	Ermöglicht die Ausführung eines Ihrer Designer Cloud-Workflows als Teil eines Plans, den Sie in der Alteryx Analytics Cloud erstellen. Diese Aufgabe könnte beispielsweise verwendet werden, um einen Designer Cloud-Workflow zu planen, der einen Auftrag in Databricks ausführt.
Desktop-Workflow-Aufgabe	Orchestrieren Sie die Ausführung Ihrer Alteryx Designer Desktop-Workflows, die in Alteryx Analytics Cloud hochgeladen wurden.
ML-Prognose-Aufgabe	Erstellen Sie mit Ihrem Alteryx Machine Learning-Projekt Prognosen zu neuen Daten.
ML-Upload-Aufgabe	Laden Sie neue Trainingsdaten in Ihr Machine-Learning-Projekt hoch.
HTTP-Aufgabe	Eine Anforderung, die im Rahmen eines Planlaufs an einen Dritt-Server übermittelt wird.
Slack-Aufgabe	Senden Sie eine Nachricht mit Informationen über den Planlauf an einen bestimmten Slack-Kanal.

Desktop-Lösung: Databricks + Alteryx Designer

Sehen wir uns nun Alteryx Designer für Desktop an. Designer ist das Spitzenprodukt von Alteryx mit einer robusten Analyseplattform, die mehr als 300 Analysemodule bietet. Viele Alteryx Benutzer:innen bevorzugen die Verwendung von Designer für Desktop mit Databricks oder in Kombination mit AAC.

Herstellen einer Verbindung mit Databricks in Alteryx Designer

Es gibt verschiedene Verbindungen, die Sie mit Databricks herstellen können, abhängig von der Datenmenge, mit der Sie arbeiten, den Transformationen, die Sie durchführen müssen, und der Verarbeitungs-Engine, die Sie bevorzugen.

Alteryx Designer ermöglicht den sicheren, kontrollierten Zugriff auf Ihre Databricks-Plattform über OAuth-Authentifizierung oder Standard-Personal Access Token (PAT) und Databricks Unity Catalog-Unterstützung.

Databricks Unity Catalog

Databricks Unity Catalog wird mit Alteryx Designer 23.2 und höher unterstützt.

Beachten Sie beim Konfigurieren einer Databricks Unity Catalog-Verbindung Folgendes:

- Databricks Unity Catalog wird nur mit Alteryx Data Connection Manager ([DCM](#)) unterstützt.
- Databricks Unity Catalog wird nur mit DSN-losen Verbindungen unterstützt.
- Das Schreiben in Databricks Unity Catalog wird nur unterstützt mit den [In-DB-Tools](#).
- Wir unterstützen die In-DB-Zusammenführung für Databricks Unity Catalog. Hier geht's zum [In-DB-Daten-schreiben-Tool](#).

The screenshot shows the 'Connection Manager' window in Alteryx Designer. The left sidebar has three tabs: 'Data Sources', 'Credentials', and 'Alteryx Links'. The main area is titled 'Databricks Demo AWS' and contains the following fields:

- Data Source Name ***: Databricks Demo AWS
- Technology**: Apache Spark ODBC DSN-less with Simba Databricks Unity Catalog
- Host ***: [redacted].cloud.databricks.com
- Driver ***: Simba Spark ODBC Driver
- Port**: 443
- Http Path ***: /sql/1.0/warehouses/[redacted]

Below these fields is a 'Connections' section with a '+ Connect Credential' link. It includes:

- Authentication Method ***: A dropdown menu set to 'Username and password'.
- Test Connection**: A button with a play icon.
- Cancel**: A button.
- Save and Link**: A blue button.
- USERNAME AND PASSWORD ***: A section with a 'Credential' dropdown set to 'DBX Demo (c49e3d38)'.
- ☒ **Allow connection for SDK**

In der [vollständigen Dokumentation](#) finden Sie eine schrittweise Anleitung dazu.

Wählen Sie bei Verwendung von Alteryx Designer 23.2 den Apache Spark ODBC DSN-IOS mit Simba Databricks Unity Catalog-Treiber aus.

Standard-Lesen/Schreiben: Ein Standard-Workflow verwendet die In-Memory Alteryx AMP Engine, um Transformationen mit einem der über 300 verfügbaren Analysemodule auszuführen. Lesen Sie Daten ein und schreiben Sie sie zurück in Databricks.

In-DB (Pushdown-Ausführung in Databricks): Alteryx Designer verfügt über eine Untergruppe von In-DB-Modulen, die die Abfrageausführung in Databricks verschieben, sodass abgefragte Daten nicht mehr in den Systemspeicher geladen werden müssen. In-DB bietet eine einfache Möglichkeit, große Datasets zu analysieren, ohne Daten einschränken oder Stichproben durchführen zu müssen.

Bulk Loading: Alteryx Designer unterstützt auch eine Databricks Delta Lake Bulk-Verbindung durch die Verwendung von Cloud-Speicher wie S3 oder Azure Data Lake Storage.

So konfigurieren Sie die einzelnen Verbindungstypen

In der [vollständigen Dokumentation](#) finden Sie eine schrittweise Anleitung dazu.

Support	Standard-Workflow-Verarbeitung
Nicht datenbankinterne Verbindung	• Installieren und konfigurieren Sie den Apache Spark-ODBC-Treiber. • Wenn Sie Simba Spark ODBC Driver DSN konfigurieren, geben Sie SparkThriftServer (Apache Spark 1.1 und höher) als Servertyp an. • In Alteryx können Sie beim Einrichten der Databricks-Verbindung im Verbindungsmanager Ihre Datenquelle entweder mit Apache Spark ODBC DSN oder DSN-IOS konfigurieren.
Datenbankinterne Verbindung	• Installieren und konfigurieren Sie den Apache Spark-ODBC-Treiber. • Wenn Sie Simba Spark ODBC Driver DSN konfigurieren, geben Sie SparkThriftServer (Apache Spark 1.1 und höher) als Servertyp an.
Lesen	• Richten Sie die Databricks-Verbindung ein. Beim Einrichten der Databricks-Verbindung im Verbindungsmanager können Sie wählen, ob Sie Ihre Datenquelle entweder mit Apache Spark ODBC DSN oder DSN-IOS konfigurieren möchten.
Schreiben	• Richten Sie die Databricks-Verbindung ein. Beim Einrichten der Databricks-Verbindung im Verbindungsmanager können Sie wählen, ob Sie Ihre Datenquelle entweder mit Apache Spark ODBC DSN oder DSN-IOS konfigurieren möchten.
Delta Lake-Verbindung	• Das Bulkloading in Delta Lake wird über Cloud-Speicher wie S3 oder ADLS unterstützt. • Richten Sie die Verbindung für Delta Lake auf AWS oder Delta Lake auf Azure ein, indem Sie Details Ihres AWS S3-Buckets bzw. Azure ADLS-Containers angeben.

Databricks Azure OAuth-Authentifizierung

Richten Sie die OAuth-Authentifizierung für Ihre Databricks-Verbindung zum Lesen und Schreiben ein.

- Platzieren Sie in Designer das Eingabe-Tool in den Arbeitsbereich (Canvas).
- Markieren Sie das Kontrollkästchen Data Connection Manager (DCM) verwenden.
- Wählen Sie „Verbindung einrichten“.
- Gehen Sie im Fenster „Datenverbindungen“ zu „Datenquellen“ und wählen Sie „Databricks ODBC“ aus.
- Wählen Sie + Datenquelle hinzufügen aus.

Beim Hinzufügen einer Datenquelle können Sie die Verbindung [mit oder ohne DSN auswählen](#). Befolgen Sie die Schritte in den Abschnitten „Verbindung mit DSN“ und „Verbindung ohne DSN“ in der Dokumentation.

Die OAuth-Authentifizierung für Bulkloading mit Databricks ist derzeit für Azure Data Lake Storage verfügbar.

Staging & Bulkloading mit Databricks Delta Lake Bulk Connection

Es gibt zwei Methoden zum Laden von Daten:

Einfügungen: Die grundlegende ODBC-Methode zum Schreiben von Daten besteht darin, jeweils einen einzelnen Datensatz einzufügen. Die Module „Datenausgabe“ und „In-DB verbinden (Schreiben)“ unterstützen Einfügungen. Beschränken Sie die Verwendung dieser Methode auf Fälle, in denen Sie mit kleinen Datasets arbeiten oder aus anderen Gründen nicht in der Lage sind, Bulkloads durchzuführen. Um Einfügungen zum Laden von Daten in Databricks zu verwenden, wählen Sie in der Konfiguration

des Datenausgabe-Tools aus dem Menü Datenverbindungen > Datenquellen die Option „ODBC“ aus.

Bulkloading: Bei der Arbeit mit großen Datasets empfiehlt es sich, die Databricks Delta Lake Bulk Connection zu verwenden. Mit dieser Methode werden Datenstapel aus Dateien im Cloud-Speicher (S3 oder ADLS) oder durch Kopieren von Datendateien von einem lokalen Rechner zu einem internen (d. h. Databricks) Cloud-Speicherort geladen, bevor die Daten mit dem Befehl COPY in Tabellen geladen werden.

Best Practices für In-DB-Pushdown in Databricks

Um Ihnen zu helfen, die Databricks Engine zu nutzen und Datenbewegungen zu minimieren, führt Alteryx Pushdown mit No-Code Tools in Designer durch. Alteryx unterstützt die folgenden Engines:

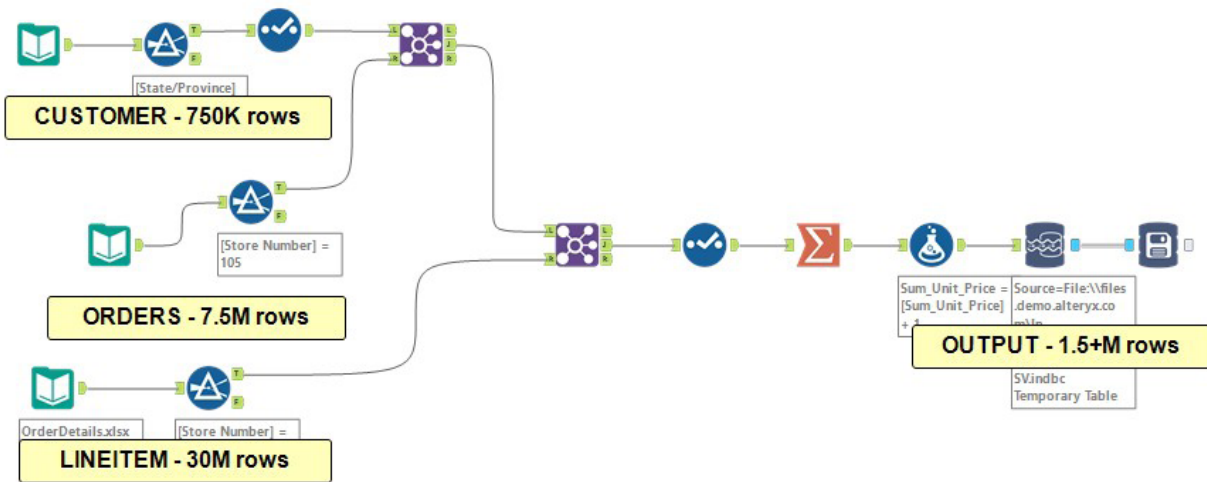
- Databricks Spark
- Databricks SQL Warehouse

Benutzer:innen können die Databricks-Verbindung in Alteryx so konfigurieren, dass eine bestimmte Compute Engine verwendet wird, indem sie den HTTP-Pfad in der Konfiguration angeben.

Die Leistungsvorteile der Verwendung von In-DB-Tools können erheblich sein. Sehen Sie sich beispielsweise den Leistungsunterschied zwischen den folgenden Standard- und In-DB-Workflows an, die beide das gleiche Ergebnis liefern:

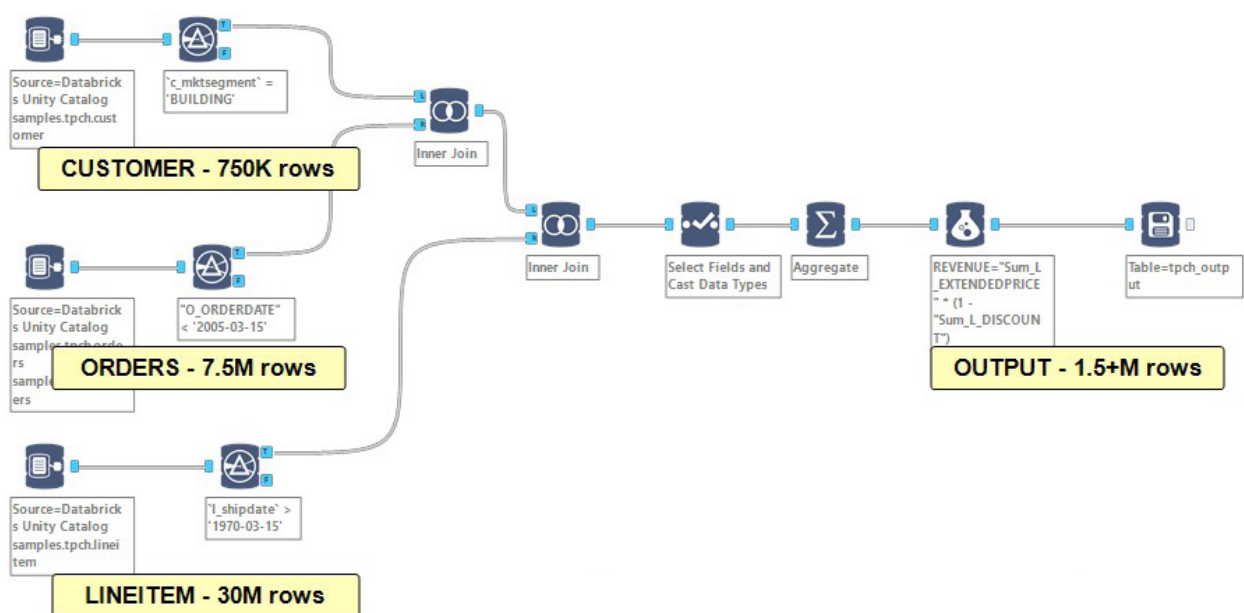
Standard-Workflow

DURCHSCHNITTliche LAUFZEIT: 8 MINUTEN



In-DB-Workflow

DURCHSCHNITTliche LAUFZEIT: 10 SEKUNDEN



Einige Tipps, wie Sie das In-DB-Pushdown optimal nutzen:

- Verwenden Sie In-Database-Tools so früh wie möglich im Datentransformationsprozess. Versuchen Sie, die Vorbereitungs- und Zusammenführungs-, Filter- und Transformationsprozesse früher in Ihren Workflow zu verschieben.
- Wenn Ihr Workflow Designer Standard-Tools erfordert, sollten Sie das Dataset aufbereiten, filtern und vorbereiten, sodass die Datentransferabfrage an Alteryx auf ein Mindestmaß reduziert wird.
- Sobald das Dataset in der Datenbank vorbereitet ist, verwenden Sie Alteryx für Prognosen, Geodaten und Datenanreicherung.
- Im [Überblick zu In-Database](#) finden Sie weitere Informationen dazu.

Anhand dieser Tabelle können Sie erkennen, wo ein datenbankinternes Tool anstelle eines Standardtools verwendet werden kann.

In-DB-Tool	Toolbeschreibung	Standard-Tool analog	Anmerkungen
In-DB durchsuchen 	Sie können sich Ihre Daten an einer beliebigen Stelle in einem In-DB-Datenstrom ansehen. Bei jedem Suchvorgang in In-DB wird eine Datenbankabfrage ausgelöst (x Datensätze), die die Leistung des Gesamt-Workflows beeinträchtigen kann.	Durchsuchen 	Verwenden Sie das In-DB-durchsuchen-Tool großzügig, um zu sehen, was sich innerhalb des Workflows abspielt. Entfernen Sie alle Durchsuchen-Module, bevor Sie den Workflow in den Produktiveinsatz übernehmen.
In-DB verbinden 	Stellen Sie mithilfe von Databricks SQL oder dem interaktiven Abfrage-Generator eine Datenbankverbindung für einen In-DB-Datenstrom her.	Eingabedaten 	Schreiben Sie SQL-Anweisungen, nutzen Sie den interaktiven Abfrage-Generator und geben Sie Pre- und Post-SQL-Anweisungen an.
Eingehender Datenstrom 	Bringen Sie Daten aus einem Standard-Workflow in einen In-DB-Workflow.	Ausgabedaten 	Schreiben Sie eine temporäre oder permanente Tabelle in Databricks.
Ausgehender Datenstrom 	Leiten Sie Daten aus einem In-DB-Datenstrom in einen Standard-Workflow, mit der Option, die Datensätze zu sortieren.	Eingabedaten 	Benutzerfreundliches Tool. Sie haben die Möglichkeit, die Datensätze vor dem Ausgeben zu sortieren.
Dynamische In-DB-Eingabe 	Übernehmen Sie einen In-DB-Verbindungsnamen und Abfragefelder aus einem Standarddatenstrom und geben Sie sie in einen In-DB-Datenstrom ein.	Keines	Im Artikel Dynamic Select with In-DB Building Blocks in der Alteryx Community finden Sie ein Beispiel für einen Anwendungsfall.

In-DB-Tool	Toolbeschreibung	Standard-Tool analog	Anmerkungen
Dynamische In-DB-Ausgabe 	Gibt Informationen zu einem In-DB-Workflow an einen Standard-Workflow aus.	Keines	Gibt Metadaten, wie Abfrage und Verbindungsname, an den Workflow aus. Sie können keine Informationen zur Verbindungszeichenfolge ausgeben, es sei denn, Sie wählen die Option „Kennwortentschlüsselung zulassen“ unter „In-DB-Verbindungen verwalten“ aus. Erzielt die schnellsten Ergebnisse bei der Filterung von Dimensionsfiltern, bei denen die Mikropartitionen geclustert sind. Weitere Informationen finden Sie in In-DB Query and the Dynamic Output tool.
In-DB-Filter 	Filtern Sie In-DB-Datensätze mit einem einfachen Filter oder einem benutzerdefinierten Ausdruck unter Verwendung von Databricks SQL.	Filtern 	In-DB- und Standard-Module haben eine ähnliche Konfigurationsschnittstelle.
In-DB-Formel 	Erstellen oder aktualisieren Sie Felder in einem In-DB-Datenstrom mit einem Ausdruck unter Verwendung von Databricks SQL.	Formel 	Einige Formeln im Standardformel-Tool können in der Datenbank in andere Module übersetzt werden. Das Hinzufügen eines Feldes mit einem konstanten Wert kann zum Beispiel in benutzerdefiniertem SQL im In-DB-Verbinden-Tool platziert werden.
In-DB verknüpfen 	Kombinieren Sie zwei In-DB-Datenströme basierend auf gemeinsamen Feldern, indem Sie eine INNERE VERKNÜPFUNG (Inner Join) oder ÄUßERE VERKNÜPFUNG (Outer Join) durchführen.	Verknüpfen  Mehrere verknüpfen 	Verhält sich eher wie eine Datenbankverknüpfung als eine Alteryx Verknüpfung oder ein Mehrere-verknüpfen-Tool. Kann so konfiguriert werden, dass sowohl innere Verknüpfungen als auch Links-, Rechts- und vollständige äußere Verknüpfungen ausgeführt werden. Verwenden Sie dies wie eine SQL-verknüpfen-Klausel. Bei Dimensionen, die Cluster-Schlüssel sind, ist die Leistung besser. Für eine optimale Leistung sollten Sie diese Verknüpfung als materialisierte Ansicht in die Datenbank verschieben.
In-DB-Makroeingabe 	Erstellen Sie eine In-DB-Eingabeverbindung für ein Makro und füllen Sie sie mit Platzhalterwerten.	Makroeingabe 	Das In-DB- und das Standard-Tool haben eine ähnliche Konfigurationsschnittstelle.
In-DB-Makroausgabe 	Erstellen Sie eine In-DB-Ausgabeverbindung für ein Makro.	Makroausgabe 	Das In-DB- und das Standard-Tool haben eine ähnliche Konfigurationsschnittstelle.
In-DB-Stichprobe 	Begrenzen Sie den In-DB-Datenstrom auf eine bestimmte Anzahl von Datensätzen.	Stichprobe 	Fügt eine LIMIT „N“-Klausel hinzu. Verwenden Sie dies beim Prototyping für große Datensets, um die Verarbeitungszeit zu verkürzen. Fügen Sie „Sortieren nach einer beliebigen Dimension, die geclustert wird“ hinzu, um die Partitionsbereinigung zu verbessern.

In-DB-Tool	Toolbeschreibung	Standard-Tool analog	Anmerkungen
In-DB auswählen 	Wählen Sie Felder in einem In-DB-Datenstrom aus, heben Sie die Auswahl auf, ordnen Sie die Felder neu an und benennen Sie sie um.	Auswählen 	Ermöglicht das Kürzen von Feldern und deren Umbenennung. Der Zweck ist etwas anders als beim Auswählen-Modul. Benutzer:innen, die die Fähigkeit des In-Memory-Auswählen-Tools replizieren möchten, Datentypen zu ändern, erstellen ein Formel-Tool mit einer CAST- oder TO_<datatype>-Anweisung.
In-DB zusammenfassen 	Fassen Sie In-DB Daten zusammen, indem Sie diese gruppieren, summieren, zählen und mehr. Die Ausgabe enthält das Ergebnis der Berechnungen.	Zusammenfassen 	Das In-DB- und das Standard-Tool haben eine ähnliche Konfigurationsschnittstelle.
In-DB Zusammenführung 	Kombinieren Sie zwei oder mehr In-DB-Datenströme mit ähnlichen Strukturen basierend auf den Feldnamen oder Positionen. In der Ausgabe enthält jede Spalte die Daten aus den einzelnen Eingaben.	Zusammenführung 	Das In-DB- und das Standard-Tool haben eine ähnliche Konfigurationsschnittstelle.
In-DB-Daten schreiben 	Benutzen Sie einen In-DB-Datenstrom, um eine Tabelle direkt in der Datenbank zu erstellen oder zu aktualisieren.	Ausgabedaten 	Stärker fokussiert als das Ausgabedaten-Tool, da es darauf abzielt, In-DB-Daten zu übernehmen und in eine Tabelle in derselben Datenbank zu schreiben.

Orchestrierung mit Alteryx Server

Alteryx Server wird zum Verwalten, Freigeben, Zusammenarbeiten und Planen von Workflows verwendet. Die Analyseanwendungen von Alteryx Server können genutzt werden, um benutzerfreundliche, browserbasierte Anwendungen zur weiteren Demokratisierung von in Designer erstellten Analyse-Workflows zu entwickeln.

Hinweise zur Ausführung eines Workflows auf Alteryx Server, der eine Verbindung zu Databricks nutzt:

- Damit veröffentlichte Workflows erfolgreich ausgeführt werden können, ist die Installation und Konfiguration der Databricks-Treiber auf Alteryx Server und allen Server- Worker-Knoten erforderlich.
- Es wird empfohlen, beim Speichern und/oder Planen eines Workflows entweder DSN-lose oder In-DB-Verbindungsdateien zu verwenden. In-DB-Verbindungen können nicht von Alteryx Server Admin erstellt und freigegeben werden, jedoch können In-DB-Verbindungsdateien über einen sicheren Netzwerkstandort freigegeben werden.
- Stellen Sie bei der Verwendung von DSN-basierten Verbindungen sicher, dass derselbe DSN auch auf Alteryx Server und Worker-Knoten eingerichtet ist.

Fazit

Wenn Sie die Verbindung zwischen Daten und Werten herstellen möchten, müssen Sie die Verbindung zwischen Ihren technischen und fachlichen Mitarbeiter:innen herstellen. Die Interoperabilität zwischen Databricks und der Alteryx AI Platform erleichtert die Zusammenarbeit zwischen verschiedenen Rollen und Präferenzen und stellt gleichzeitig die richtige Governance und den richtigen Zugriff sicher.

Gemeinsam ermöglichen es Alteryx und Databricks Unternehmen, eine unternehmensweite Kultur der Datenkompetenz aufzubauen und Analysen in großem Umfang durchzuführen. Wenn Sie bereit sind, mehr Ergebnisse für Ihre Daten zu erzielen und KI zu nutzen, werfen Sie einen Blick auf die folgenden Ressourcen.

Ressourcen und nächste Schritte

[Machen Sie eine Tour der Alteryx AI Platform for Enterprise Analytics](#)

[Probieren Sie das Databricks + Alteryx Starter-Kit aus](#)

[Sehen Sie sich das Webinar an: „Loving the Lakehouse for Data Transformation“.](#)

[Beginnen Sie eine kostenfreie Testversion](#)



Über Alteryx

Alteryx ermöglicht mit der AI Platform for Enterprise Analytics umsetzbare Erkenntnisse. Mit Alteryx können Unternehmen intelligentere und schnellere Entscheidungen treffen – mit einer sicheren Plattform, die in lokalen, hybriden und Cloud-Umgebungen eingesetzt werden kann. Mehr als 8.000 Kundinnen und Kunden weltweit verlassen sich auf Alteryx, um Analysen zu automatisieren und dadurch die Umsatzleistung zu verbessern, Kosten zu managen und Risiken in ihrem gesamten Unternehmen zu mindern.

Weitere Informationen erhalten Sie unter www.alteryx.com/de.

Alteryx ist eine eingetragene Marke von Alteryx, Inc. Alle anderen Produkt- und Markennamen sind möglicherweise Warenzeichen bzw. eingetragene Warenzeichen ihrer jeweiligen Eigentümer:innen.