Open for Innovation
KNIME

KNIME Databricks Integration Benutzer Leitfaden

KNIME AG, Zürich, Schweiz

Version 5.7 (letzte Aktualisierung auf)

Inhaltsverzeichnis

Überblick

KNIME Analytics Platform, ab Version 4.1, enthält eine Reihe von Knoten zur Unterstützung DatabricksTM .

Die KNIME Databricks Integration ist auf der KNIME Hubraum .

> Neben dem standardmäßig bezahlten Service bietet Databricks auch eine Kostenlose Ausgabefür Test- und Bildungszwecke.

Erstellen eines Databricks-Clusters

Für eine detaillierte Anleitung zur Erstellung eines Databricks-Clusters folgen Sie bitte der Tutorial von Databricks bereitgestellt. Während der Cluster-Erstellung könnten die folgenden Features wichtig sein:

- Autoscaling: Die Aktivierung dieser Funktion ermöglicht es Databricks, Arbeiter dynamisch zu lokalisieren für den Cluster abhängig vom aktuellen Lastbedarf
- Autoterminierung: Geben Sie eine Inaktivitätsperiode an, nach der der Cluster endet automatisch. Alternativ können Sie die Option aktivieren Cluster auf Kontext beenden zerstören im Dialog Konfigurationsdialog Databricks Environment node erstellen, um zu beenden der Cluster, wenn der Spark Kontext zerstört wird, z.B. wenn der Destroy Spark Kontext Knoten wird ausgeführt. Für weitere Informationen über die Cluster auf Kontext zerstören beenden Checkbox oder der Destroy Spark Context-Knoten, bitte überprüfen Sie die Abschnitt.

<a href="#page7" style="color: #ff6

Alle Benutzer, die einen Spark-Kontext erstellen möchten, benötigen die "Can Manage"-Berechtigung, um in der Lage zu sein, KNIME Job Jars hochladen. Für weitere Details siehe Databricks Dokumentation.

Geteilte Cluster werden nicht unterstützt. Um Spark zu verwenden, müssen Sie eine persönliche Compute erstellen Cluster.

Bei der Erstellung eines Clusters von Version 15 oder neuer, die funker.databricks.driver.dbfsLibraryInstallationAllowed = true muss eingestellt werden aktiviert. Für weitere Details klicken Hier.

> Die Autoskalierung und Auto-Terminierung Funktionen, zusammen mit anderen Funktionen während Cluster-Erstellung könnte nicht in der freien Databricks-Community verfügbar sein Ausgabe.

Kontakt zu Databricks

Die Databricks Workspace Connector-Knoten ermöglicht die Verbindung zu allen unterstützten Databricks Services. Der Nod unterstützt entweder die Authentifizierung per Personal Access Token (PAT) oder Microsoft Entra ID. Um Microsoft Entra ID zu verwenden, verbinden Sie den Knoten mit einem Microsoft Authenticator Node. Andere Authentifizierungsmethoden wie OAuth für Benutzer (OAuth U2M) und OAuth für Dienstleitungen (OAuth M2M) werden über den Secret Store unterstützt KNIME Business Hub und KNIME Community Hub.

Verbinden Sie mit einem Databricks-Cluster

In diesem Abschnitt wird beschrieben, wie Sie den Databricks-Umweltknoten erstellen konfigurieren, um eine Verbindung herzustellen zu einem Databricks-Cluster innerhalb der KNIME Analytics Platform.

Bitte stellen Sie vor der Verbindung mit einem Cluster sicher, dass der Cluster bereits in Databricks für weitere Informationen über einen Databricks-Cluster erstellen.

Nach der Erstellung des Clusters öffnen Sie den Konfigurationsdialog der Databricks-Umgebung erstellen Knoten. Bei der Konfiguration müssen Sie folgende Informationen bereitstellen:

1. ANHANG Die vollständige Databricks-Bereitstellungs-URL: Die URL ist jedem Databricks zugeordnet Bereitstellung. Zum Beispiel, wenn Sie Databricks auf AWS verwenden und sich anmelden http://1234-5678-Abcd.cloud.databricks.com/ , das ist Ihre Databricks URL. Die URL ist nur erforderlich, wenn der Knoten nicht mit dem Databricks Workspace Connector-Knoten verbunden ist.

   Schlussfolgerung

   Die URL sieht anders aus, je nachdem, ob sie auf AWS eingesetzt wird oder Azure.

   In der kostenlosen Databricks Community Edition ist die Bereitstellungs-URL https://community.cloud.databricks.com/ .

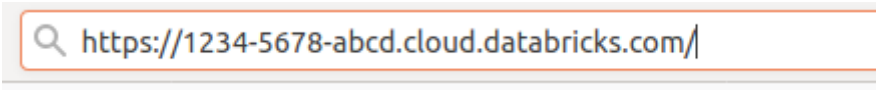

Abbildung 1. Databricks-Bereitstellungs-URL auf AWS

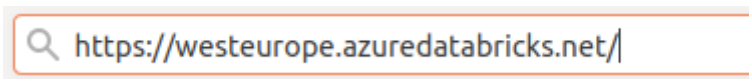

Abbildung 2. Databricks Bereitstellung URL auf Azure

2. Die Cluster-ID: Cluster-ID ist die einzigartige ID für einen Cluster in Databricks. Um den Cluster zu erhalten ID, klicken Sie auf die Cluster Tab im linken Bereich und wählen Sie dann einen Clusternamen aus. Sie finden

die Cluster-ID in der URL dieser Seite /#/settings/clusters/

id>/Konfiguration .

Die URL in der kostenlosen Databricks-Community-Edition ist ähnlich wie bei der

...

)

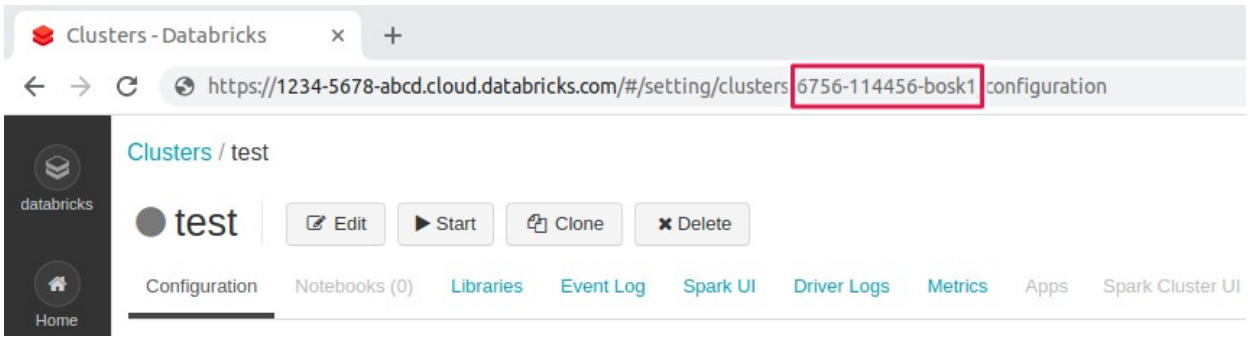

Abbildung 3. Cluster-ID auf AWS

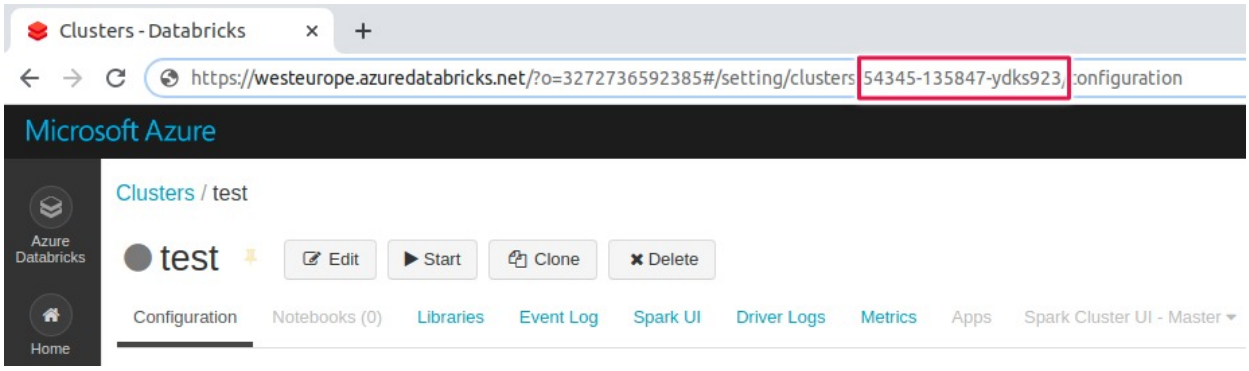

Abbildung 4. Cluster ID auf Azure

3. Workspace ID: Workspace ID ist die einzigartige ID für einen Databricks Workspace. Nur

   verfügbar für Databricks auf Azure, oder wenn Sie die kostenlose Databricks Community Edition verwenden. Für

   Databricks auf AWS, lassen Sie einfach das Feld leer.

   Die Workspace-ID finden Sie auch in der Bereitstellungs-URL. Die Zufallszahl nach

   = ist beispielsweise die Workspace-ID, http:///?

   = 327273659238 5 .

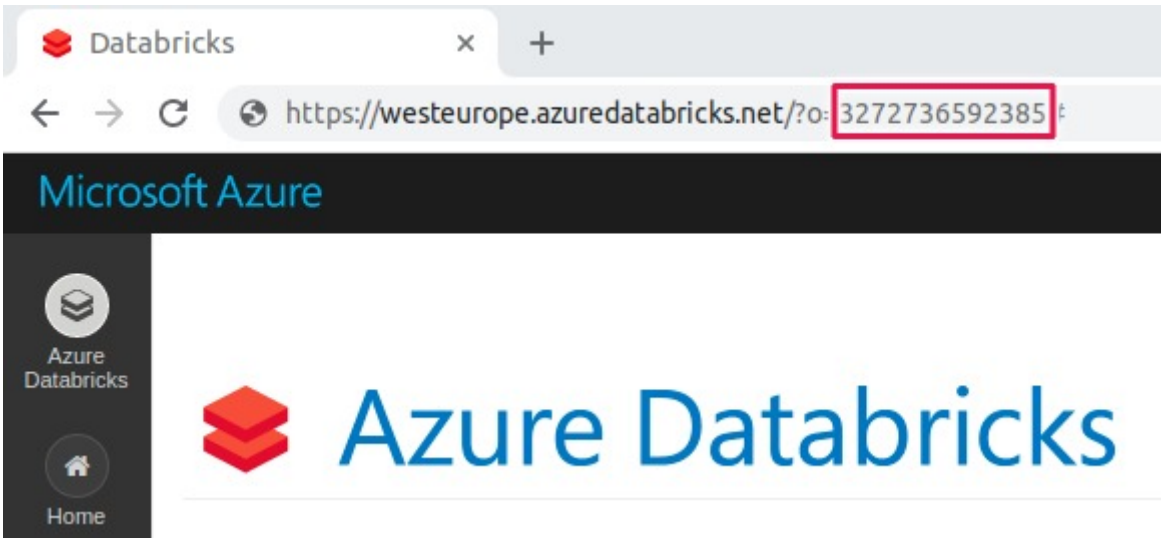

Abbildung 5. Workspace-ID auf Azure

Für weitere Informationen zu den Databricks-URLs und -IDs überprüfen Sie bitte die Dokumentation von Databricks .

L 347 vom 20.12.2013, S. 1). Authentication: Token wird als Authentifizierungsmethode in Databricks. Um einen Zugriffstoken zu generieren:

a. Klicken Sie in Ihrem Databricks-Workspace auf das Benutzerprofil-Symbol oben rechts Ecke und wählen Benutzereinstellungen

B. Navigieren Sie zum Zugang zu Token Register

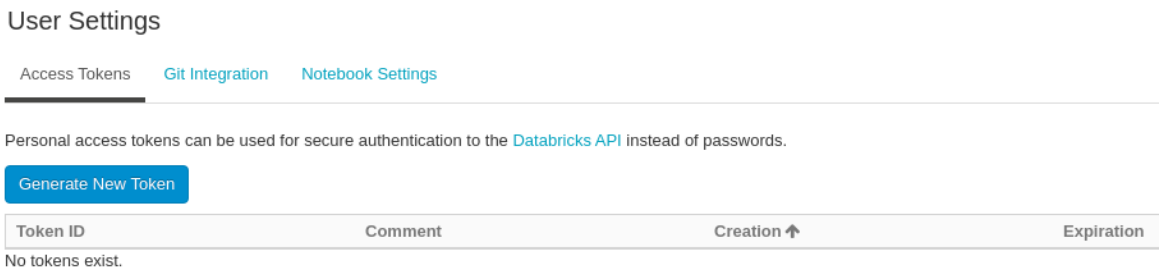

Abbildung 6. Die Registerkarte Access Tokens

c. Klicken Sie auf Neue Token generieren , und optional die Beschreibung und den Token eingeben Leben. Am Ende klicken Sie auf die Generieren Knopf.

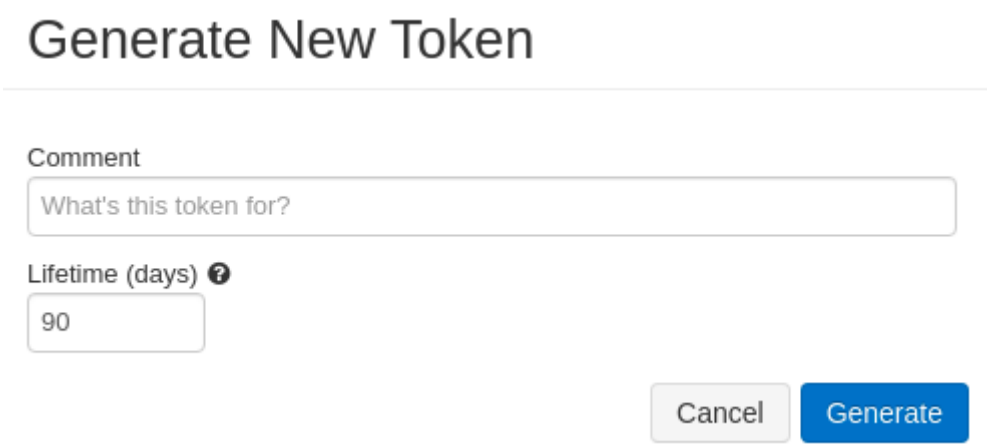

Abbildung 7. Neue Zeichen generieren

d. Speichern Sie das generierte Token an einem sicheren Ort.

Für weitere Informationen zu Databricks Zugriff auf Token, bitte überprüfen Sie die Datenbrände Dokumentation .

Access Token ist leider nicht in der kostenlosen Databricks Community verfügbar Ausgabe. Bitte verwenden Sie den Benutzernamen und die Passwortoption als Alternative.

Nach dem Ausfüllen aller notwendigen Informationen im Databricks-Umweltknoten erstellen Konfigurationsdialog, führen Sie den Knoten aus. Bei Bedarf wird der Cluster automatisch gestartet. Warte. bis der Cluster fertig wird. Dies kann einige Minuten dauern, bis die gewünschte Wolke Ressourcen werden zugewiesen und alle Dienste werden gestartet.

Der Knoten verfügt über drei Ausgänge:

- Red Port: JDBC-Verbindung, die die Verbindung zu KNIME-Datenbankknoten ermöglicht
- Blue Port: DBFS-Verbindung, die die Verbindung zu Remote-Datei-Handling-Knoten als sowie Funkknoten
- Grauer Port: Spark Kontext, der die Verbindung zu allen Spark-Knoten ermöglicht.

Die File Handling-Knoten sind unter IO im Knoten-Repository von KNIME Analyseplattform.

Diese drei Ausgangsports ermöglichen es Ihnen, eine Vielzahl von Aufgaben auf Databrick-Clustern über KNIME Analytics Plattform, z.B. die Verbindung zu einer Databricks-Datenbank und die Durchführung Datenbankmanipulation über KNIME DB-Knoten oder Ausführen von Spark-Jobs über KNIME Spark-Knoten, während Sie den gesamten Berechnungsprozess in den Databricks-Cluster einschieben.

Erweiterte

Um erweiterte Optionen zu konfigurieren, navigieren Sie auf die Erweiterte Tab in der Databricks erstellen Umweltknoten. Beispielsweise können die folgenden Einstellungen nützlich sein:

- Spark Kontext erstellen und Spark Kontextport aktivieren Checkbox wird standardmäßig aktiviert Betrieb KNIME Spark Jobs auf Databricks. Wenn Ihr Cluster jedoch mit Zugang zu den Produkten Steuerung , bitte stellen Sie sicher, diese Option zu deaktivieren, da TAC keinen Spark unterstützt Ausführung Kontext.
- Die Cluster auf Kontext zerstören beenden Checkbox beendet den Cluster wenn der Knoten zurückgesetzt wird, wenn der Destroy Spark Context-Knoten ausgeführt wird, oder wenn der Der Workflow oder die KNIME Analytics Platform wird geschlossen. Dies könnte wichtig sein, wenn Sie Ressourcen sofort nach der Verwendung freigeben. Verwenden Sie diese Funktion jedoch mit Vorsicht! Eine andere Möglichkeit ist, die automatische BeendigungFeature während der Cluster-Erstellung, wo der Cluster automatisch nach einer gewissen Inaktivitätsdauer beendet.
- Zusätzlich enthält die DB Port Registerkarte alle datenbankbezogenen Konfigurationen, die ausführlichere Erläuterungen zu den KNIME Leitfaden für die Erweiterung .

Arbeiten mit Databricks

Dieser Abschnitt beschreibt, wie mit Databricks in der KNIME Analytics Platform gearbeitet werden kann, z. Zugriff auf Daten von Databricks über KNIME und umgekehrt, wie Sie Databricks Delta verwenden und viele andere.

Ab 4.3, KNIME Analytics Plattform beschäftigt ein neues Dateimanagement den Rahmen. Weitere Informationen finden Sie in der KNIME Leitfaden für die Bearbeitung von Dateien.

Dateien

Databricks bietet zwei verschiedene Ansätze, mit Dateien zu arbeiten, die sich in einer Cloud befinden Objektspeicher:

- Unity Catalog Volumes: Unity Katalog Volumen
- Databricks Dateisystem: DBFS (abgeschrieben) Unity Katalog Volumen statt)

Mit der KNIME Analytics Platform können Sie nahtlos mit beiden arbeiten. Der einzige Unterschied ist der Connectorknoten, den Sie in Ihrem Workflow verwenden.

Databricks Unity File System Connector Node

Der Databricks Unity File System Connector-Knoten ermöglicht es Ihnen, direkt mit dem Katalog zu verbinden
Volumes ohne einen Cluster starten zu müssen, wie es bei den Databricks erstellen der Fall ist
Umweltknoten, der nützlich ist, um einfach Daten in oder aus Ihrem Arbeitsraum zu bekommen.

Vor der Verwendung des Databricks Unity File System Connector-Knotens müssen Sie mit Ihrem
Workspace mit dem Databricks Workspace Connector-Knoten. Verbindung zu Ihrem Arbeitsraum
die vollständige URL des Databricks-Workspaces eingeben, z.
http://.cloud.databricks.com/ oder https://adb-
Nummer > .azuredatabricks.net/ auf Azure und wählen Sie die entsprechende Authentifizierungsmethode aus.
Verbinden Sie den Ausgangsport des Databricks Workspace Connectors mit dem Eingangsport des
Databricks Unity File System Connector Knoten.

Der Ausgang Dateisystem Port (blau) dieses Knotens kann mit den meisten der
href = "https://docs.knime.com/2025-
12/analytics_platform_file_handling_guide/index.pdf#*read_and_write_from_or_to_a_co*
nnected_file_system" > file Handling-Knoten, die unter _IO in der
Node-Repository der KNIME Analytics Platform.

Objektspeicher mit Unity Katalog

Unity Catalog ermöglicht die Montage von Objektspeichern, wie AWS S3 Eimer , oder Azure Data Lake
Speicher Gen2 . Durch die Registrierung als neue Datenwerte im Unity Katalog können Objekte
als ob sie auf einem lokalen Dateisystem waren. Bitte überprüfen Sie die folgenden Unterlagen
von Databricks für weitere Informationen wie:

- Für AWS S3 Eimer: Externe Standorte für Databricks auf AWS verwalten
- Für Azure Data Lake Storage Gen2 Lagerung: Erstellen Sie einen externen Standort zum Verbinden Cloud-Speicher für Azure Databricks
- Für Google Cloud-Speicher: Erstellen Sie einen externen Standort, um Cloud-Speicher zu verbinden Datenbrände

Databricks Dateisystem Connector Node

Databricks Dateisystem ist depreciert, verwenden Unity Katalog Volumen statt.

Der Databricks File System Connector-Knoten ermöglicht es Ihnen, direkt mit Databricks File zu verbinden
System (DBFS) ohne einen Cluster starten zu müssen, wie es bei den Databricks ist
Umweltknoten, der nützlich ist, um einfach Daten in oder aus DBFS zu bekommen.

Im Konfigurationsdialog dieses Knotens geben Sie bitte folgende Informationen an:

- Databricks URL: Die Domain der Databricks-Bereitstellungs-URL, z. 1234-5678-Abcd.cloud.databricks.com
- Authentication: Der Zugriff auf Token oder Benutzername/Passwort als Authentifizierung Methode.

<a href="#page3">Abschnitt zu Informationen

wie man die Databricks-Bereitstellungs-URL erhält und einen Zugriffstoken generiert.

Der Ausgang DBFS-Port (blau) dieses Knotens kann mit den meisten Datei-Handling-Knoten verbunden werden die unter IO im Knoten-Repository der KNIME Analytics Platform.

Mount AWS S3 Eimer und Azure Blob Speicher für DBFS

DBFS ermöglicht die Montage von Objektspeichern, wie AWS S3 Eimer , oder Azure Blob Lagerung . von Sie an DBFS anzubringen, können die Objekte wie auf einem lokalen Dateisystem aufgerufen werden.
Bitte überprüfen Sie die folgenden Unterlagen von Databricks für weitere Informationen wie:

- Für AWS S3 Eimer
  - Mount und unmount AWS S3 Eimer mit DBFS
  - Verschlüsseln Sie Daten beim Schreiben an S3 über DBFS
- Für Azure Blob Lagerung
  - Mount und unmount Azure Blob Aufbewahrungsbehälter an DBFS
  - Auf den Wunschzettel Daten Lake Storage Gen1 Ressource mit einem Service-Rektor und OAuth 2.0
  - Installieren Sie ein Azure Data Lake Storage Gen2 Konto mit einem Service-Prinzip und OAuth 2.0

Funk-IO-Knoten

KNIME Analytics Platform unterstützt das Lesen verschiedener Dateiformate, wie Parkett oder ORC, befinden sich in einem Einheitskatalog Volumen oder DBFS, in einem Spark DataFrame, und umgekehrt. Auch ermöglicht das Lesen und Schreiben dieser Formate direkt von/in KNIME-Tabellen mit dem Reader und Schreiberknoten.

Die KNIME Extension für Apache Spark ist auf der KNIME Hubraum . Diese Spark IO-Knoten werden dann unter Werkzeuge und Dienstleistungen Apokalypse IO im Node-Repository der KNIME Analytics Platform, mit Ausnahme von Parkett Lese- und Parkettschreiberknoten, die unter IO > Weiterlesen und IO > Schreiben jeweils.

Parken zum Parkett/ORC zum Spark

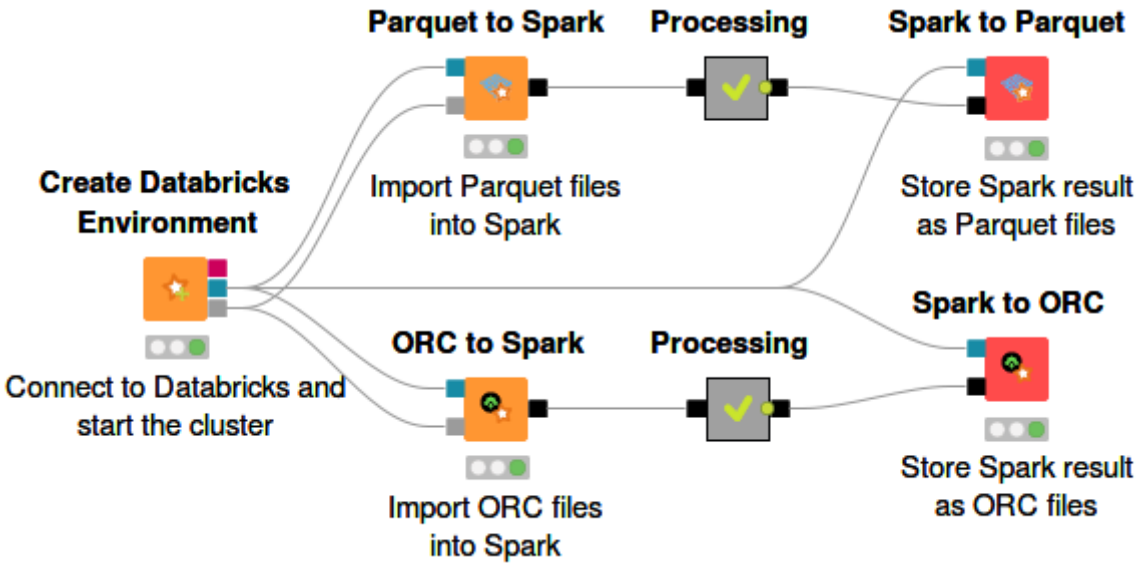

Abbildung 8. Parkett/ORC zu Spark und Spark zu Parkett/ORC Knoten

ANHANG Um Parquet-Dateien zu importieren, die sich in einem Einheitskatalog Volumen oder DBFS in einen Spark befinden DataFrame, verwenden Sie den Parkett zu Spark-Knoten, dann den Eingang DBFS-Port (blau) und der Eingang Spark Port (grau) zu den entsprechenden Ausgangsports der Databricks erstellen Abb) Im Node-Konfigurationsdialog geben Sie einfach den Pfad ein in den Ordner, in dem die Parquet-Dateien wohnen, und dann den Knoten ausführen.

Die Parkettdaten sind jetzt im Spark verfügbar und Sie können jede Anzahl von Spark nutzen Knoten zur visuellen Weiterverarbeitung der Daten.

Der ORC zum Spark-Knoten hat den gleichen Konfigurationsdialog wie der Parkett zu Spark Node.

2. Um einen Spark DataFrame zu einem Einheitskatalog Volumen oder DBFS im Parkettformat zu schreiben, verwenden Sie der Spark zu Parkett-Knoten. Der Knoten hat zwei Eingangsports. Verbindung zu einem Einheitskatalog Volume, bitte verbinden Sie das Dateisystem (blau) Port des Databricks Unity File Systems Verbindungsknoten. Um eine Verbindung zu DBFS-Verbindung herzustellen, verbinden Sie bitte den DBFS-Port (blau) von den Databricks-Umgebungsknoten erstellen und den zweiten Port zu jedem Knoten mit einem Spark Datenausgabeport (schwarz). Um den Spark zu Parkett-Knoten zu konfigurieren, öffnen Sie den Knoten

Konfigurationsdialog und geben Sie den Namen des Ordners an, der erstellt wird und in die die Parkett-Datei(en) gespeichert werden.

Unter der Partitionen Tab gibt es eine optionale Option, ob die Daten auf Basis bestimmter Spalten verteilt. Wenn die Option Überschreiben Ergebnis Partition Anzahl wird aktiviert, die Anzahl der Ausgabedateien kann angegeben werden. Dies jedoch Option wird dringend nicht empfohlen, da dies zu Leistungsproblemen führen könnte.

Der Spark zu ORC-Knoten hat den gleichen Konfigurationsdialog wie der Spark, Parkett-Knoten.

Bankett/ORC Leser und Schriftsteller

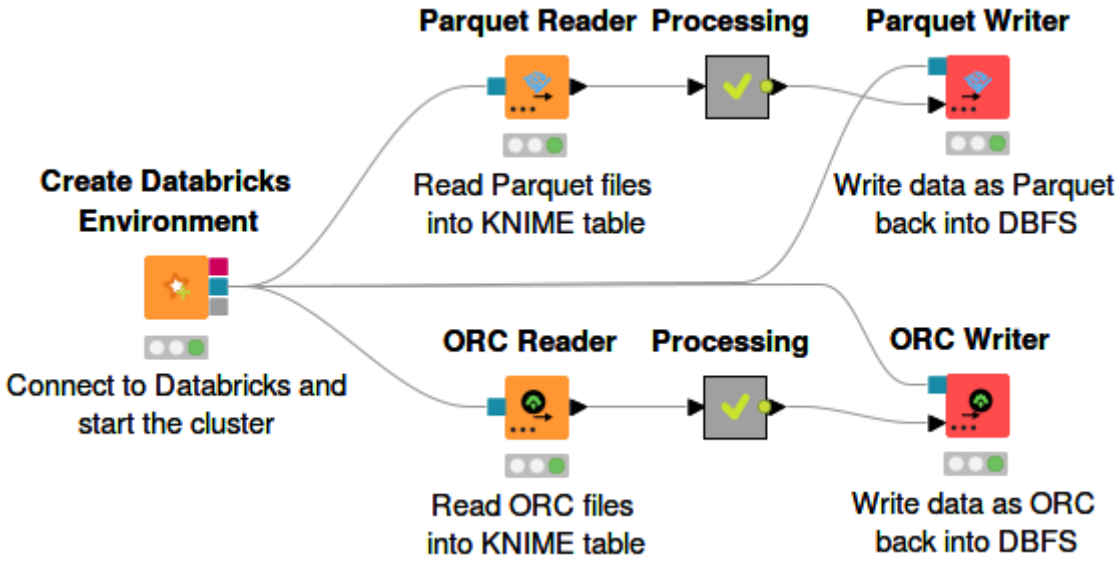

Abbildung 9. Bankett/ORC Lese- und Schreiberknoten

ANHANG Zum Import von Daten im Parkettformat aus einem Einheitskatalogvolumen oder DBFS direkt in KNIME Tabellen, verwenden Sie den Parquet Reader Knoten. Der Knotenkonfigurationsdialog ist einfach, Sie müssen nur den DBFS-Pfad eingeben, wo die Parkettdatei liegt. Unter der Typ Mapping Tab, die Zuordnung von Parkett-Datentypen zu KNIME-Typen muss angegeben werden.

Die Parkettdaten sind jetzt lokal verfügbar und Sie können alle Standard KNIME-Knoten nutzen die weitere Datenverarbeitung visuell durchzuführen.

Die ORC Leseknoten hat den gleichen Konfigurationsdialog wie der Parkett Leseknoten.

2. Um eine KNIME-Tabelle in eine Parkett-Datei in einem Einheitskatalog Volumen oder DBFS zu schreiben, verwenden Sie die Parquet Writer Node. Um mit einem Einheitskatalog Volumen zu verbinden, verbinden Sie bitte die Datei System (blau) Port des Databricks Unity File System Connector Knotens. Zu verbinden

DBFS, bitte verbinden Sie den DBFS (blau) Port mit dem DBFS Port des Databricks erstellen Umweltknoten. Geben Sie im Node-Konfigurationsdialog den Standort auf DBFS ein, wo Sie möchten die Parkett-Datei schreiben und unter der Typ Mapping Tab, die Mapping von KNIME-Datentypen zu Parkett-Datentypen.

Der ORC Writer-Knoten hat den gleichen Konfigurationsdialog wie der Parkett Schreiberknoten.

Für weitere Informationen über die Typ Mapping Tab, bitte überprüfen Sie die Datenbank Dokumentation .

Databricks SQL Waren

Verbinden Sie mit einem Databricks SQL Warehouse

Verwenden Sie die Databricks SQL Warehouse Connector Node eine Verbindung zu einem Databricks SQL Waren . Sobald Sie angeschlossen sind, können Sie KNIME Datenbankintegration visuell fügen Sie Aussagen zusammen und schieben Sie ihre Ausführung in das Databricks Warehouse. In zusätzlich können Sie den Knoten verwenden, um Ihre Daten direkt im Warehouse zu manipulieren.

Lesen und Schreiben von Databricks SQL Warehouse

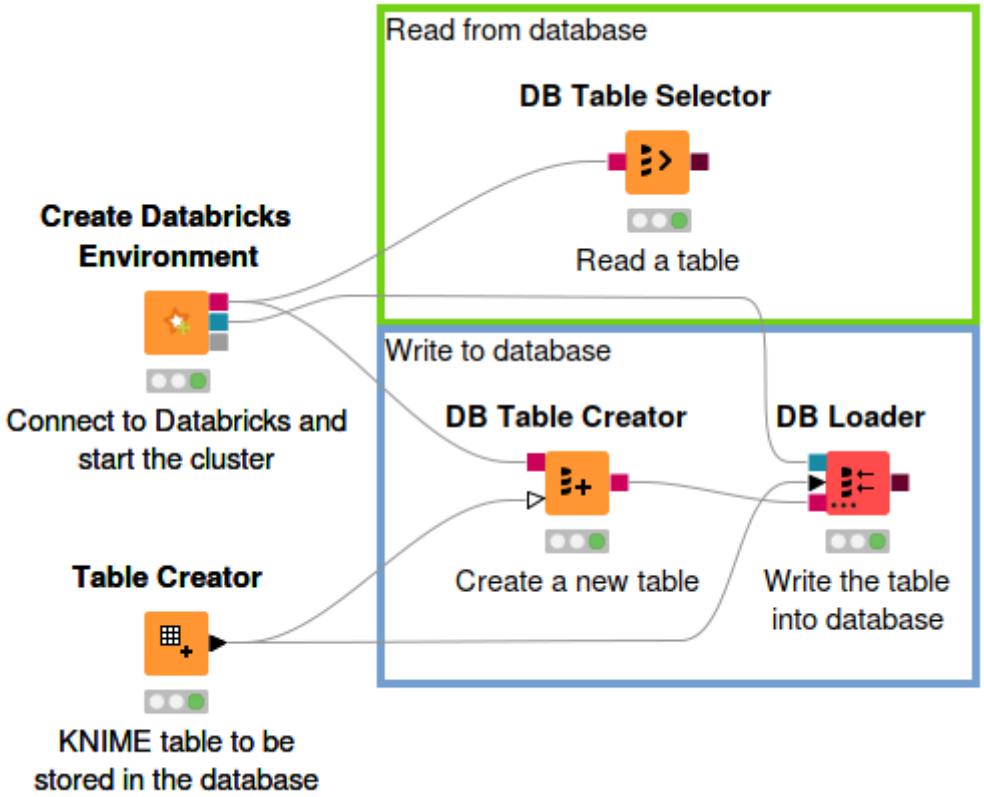

Abbildung 10. Wie man aus/zu Databricks Datenbank liest und schreibt

Um eine KNIME-Tabelle in Databricks-Datenbank zu speichern:

ANHANG Verwenden Sie den DB Table Creator-Knoten. Der Knoten hat zwei Eingangsports. Verbindung der DB (rot) Port zum DB-Port des Databricks SQL Warehouse Connector-Knotens, und der zweite Port zur Ziel-KNIME-Tabelle. Geben Sie im Knotenkonfigurationsdialog das Schema ein und den Tischnamen. Seien Sie vorsichtig, wenn Sie spezielle Zeichen im Tabellennamen verwenden, z. underscore (_) wird nicht unterstützt. Die Ausführung dieses Knotens wird eine leere Tabelle in der Datenbank mit der gleichen Tabellenspezifikation wie die Eingabe KNIME-Tabelle.

> Der DB Table Creator Knoten bietet viele weitere Funktionalitäten. Für mehr Informationen zum Knoten, bitte überprüfen Sie die Datenbankdokumentation .

2. Den DB Loader-Knoten an den DB Table Creator-Knoten anhängen. Dieser Knoten hat drei Eingänge Häfen. Verbinden Sie den zweiten Datenport mit der Ziel-KNIME-Tabelle, dem DB (rot) Port mit dem ausgeben DB-Port des DB-Tabellen-Knotens, und der Datei-Handling (blau)-Port entweder zu die Databricks Unity File System Connector oder Databricks Dateisystem Connector Knoten. Ausführen dieses Knotens lädt den Inhalt der KNIME-Tabelle auf die neu erstellte Tabelle in der Datenbank.

> Weitere Informationen zum DB Loader-Knoten finden Sie unter: Datenbankdokumentation .

Um eine Tabelle aus einer Databricks-Datenbank zu lesen, verwenden Sie den DB Table Selector-Knoten, wo die Eingabe
Der DB (rote) Port ist mit dem DB-Port des Databricks SQL Warehouse Connector-Knotens verbunden.

Anstelle des Databricks SQL Warehouse Connector-Knotens können Sie auch den
erster DB-Port (rot) des Databricks-Umweltknotens erstellen.

Weitere Informationen zu anderen KNIME-Datenbankknoten finden Sie in der
Datenbankdokumentation .

Databricks Delta

Databricks Delta eine Speicherschicht zwischen dem Databricks File System (DBFS) und Apache
Spark API. Es bietet zusätzliche Funktionen wie ACID-Transaktionen auf Spark, Schema
Durchsetzung, Zeitreise und viele andere.

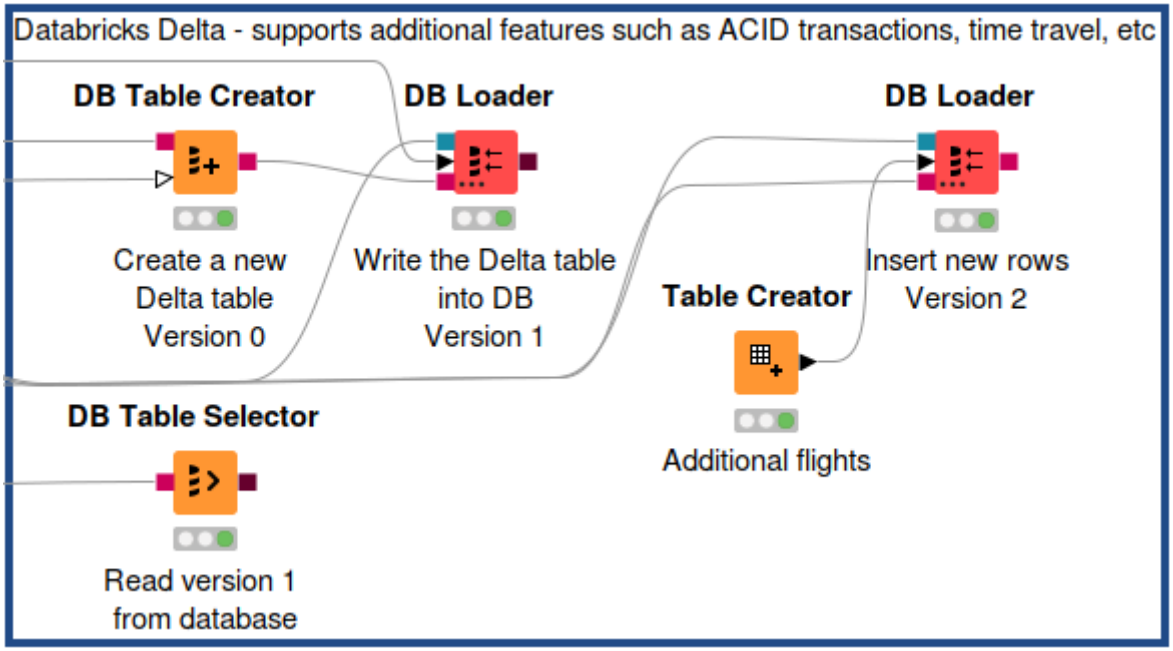

Abbildung 11. Databricks Delta auf KNIME Analytics Plattform

Um eine Delta-Tabelle in der KNIME Analytics Platform mit dem DB Table Creator-Knoten zu erstellen:

ANHANG Verbinden Sie den ersten Port mit dem DB-Port (rot) des Databricks SQL Warehouse Connector
Knoten und der zweite Port zur Ziel-KNIME-Tabelle

2. Geben Sie im Konfigurationsdialog den Tabellennamen und das Schema wie üblich ein und konfigurieren Sie
die anderen Einstellungen wie nach Ihren Bedürfnissen. Um diesen Tisch zu einem Delta-Tisch zu machen,
Einsatz DELTA Erklärung nach Zusätzliche Optionen Tab (siehe <a href="#page15" style="co)

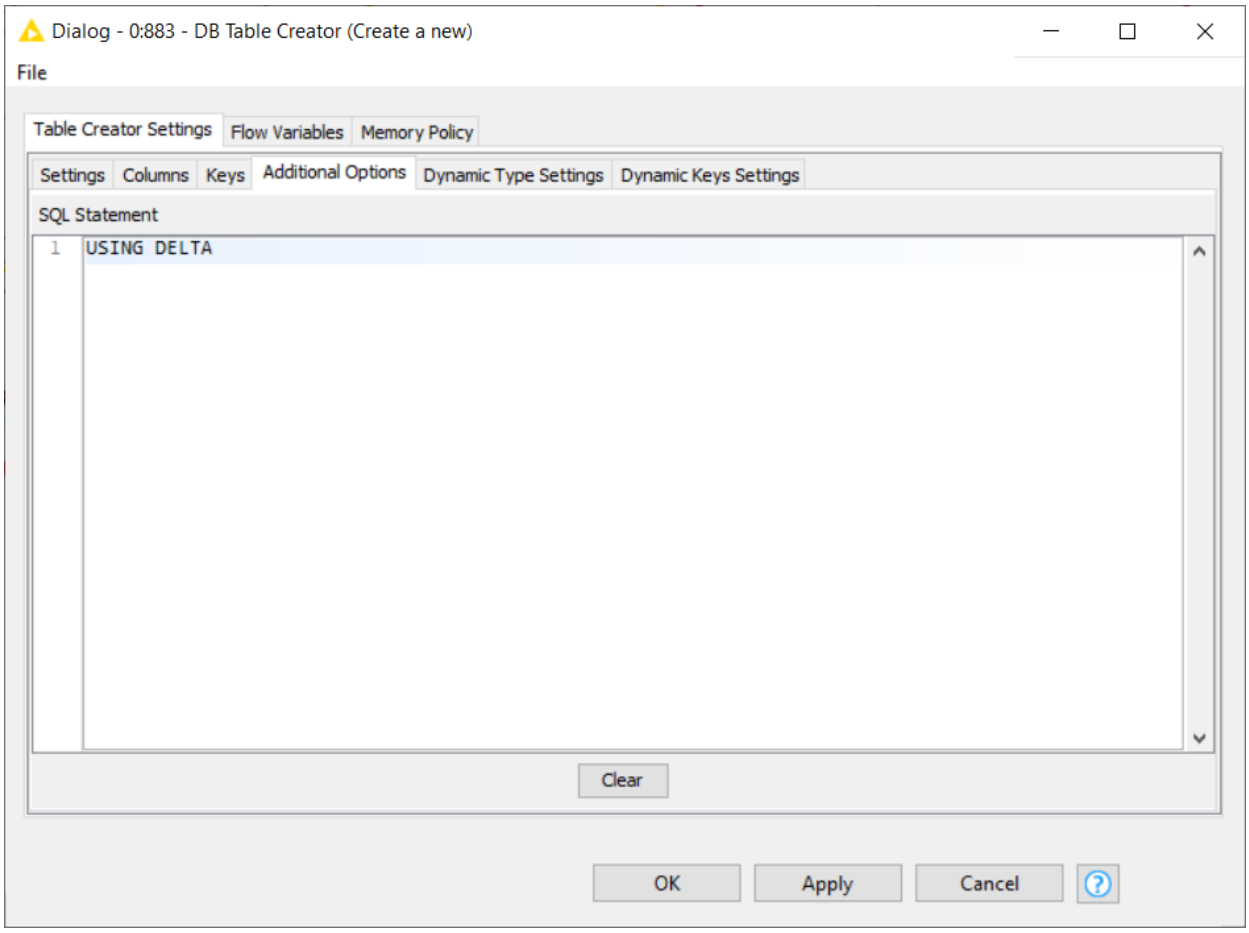

Abbildung 12. Zusätzliche Optionen Registerkarte innerhalb des DB Tabelle Creator-Knotenkonfigurationsdialogs

3. Ausführen des Knotens und eine leere Delta-Tabelle wird mit der gleichen Tabellenspezifikation erstellt

als Eingabe-KNIME-Tabelle. Füllen Sie die Tabelle mit Daten, z.B. dem DB Loader-Knoten (siehe
Abschnitt <a href="#page12" style="color: #ff6600; text-decoration: underline;">Lesen und Schreibe)

Time Travel Funktion

Databricks Delta bietet viele zusätzliche Funktionen, um die Datensicherheit zu verbessern, wie zum Beispiel Zeit

Reisen. Zeitreisen ist eine Datenversions-Fähigkeit, mit der Sie einen älteren Snapshot eines

Delta-Tabelle (Rollback).

Zugriff auf die Versionshistorie (Metadaten) in einer Delta-Tabelle auf dem Databricks Web UI:

ANHANG Navigieren Sie zum Daten Tab in der linken Scheibe

2. Wählen Sie die Datenbank und den Delta-Tabellennamen

3. Die Metadaten und eine Vorschau der Tabelle werden angezeigt. Wenn die Tabelle tatsächlich eine Delta-Tabelle ist,

es wird zusätzlich GeschichteTab neben dem Details Tab (siehe <a href="#page16" style="color: #f)

L 347 vom 20.12.2013, S. 1). Unter de GeschichteTab, es gibt die Versionierungsliste der Tabelle, zusammen mit der

Zeitstempel, Betriebsarten und andere Informationen.

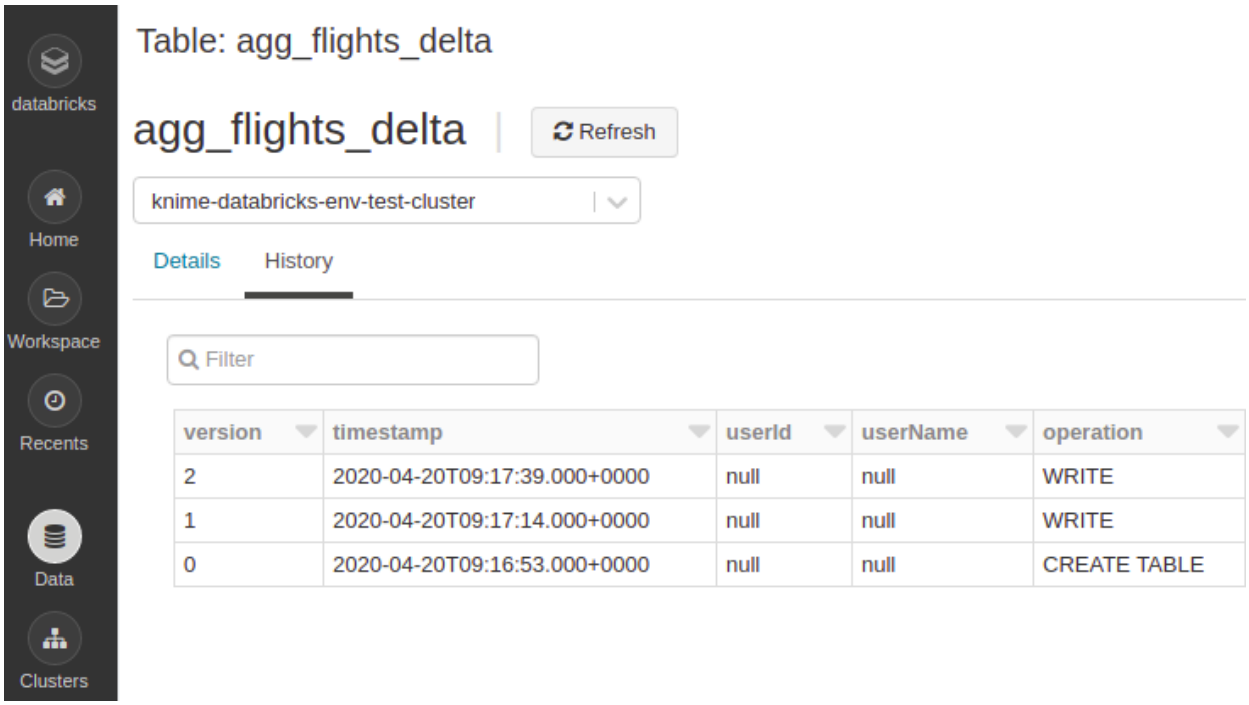

Abbildung 13. Delta-Tabellen-Versionsgeschichte

Alternativ können Sie auch auf die Versionsgeschichte einer Delta-Tabelle direkt in KNIME zugreifen
Analyseplattform:

ANHANG Verwenden Sie den DB Query Reader-Knoten. Schließen Sie den Eingang DB-Port (rot) des DB Query Readers an
Knoten zum DB-Port des Databricks SQL Warehouse Connector Knotens.

2. Geben Sie im Node-Konfigurationsdialog die folgende SQL-Anweisung ein:

```
DESCRIBE HISTORY
```

wenn ist der Name der Tabelle, auf deren Versionshistorie Sie zugreifen möchten.

3. Führen Sie den Knoten aus. Dann klicken Sie mit der rechten Maustaste auf den Knoten, wählen KNIME Datentabelle um die
Versionshistorietabelle (ähnlich der Tabelle in <a href="#page16" style="color: #ff6600; text-decoratio)

Weitere Informationen zu Delta-Tabellen-Metadaten finden Sie in der Datenbrände
Dokumentation .

Neben der Versionsgeschichte, Zugriff auf ältere Versionen einer Delta-Tabelle in KNIME Analytics
Plattform ist auch sehr einfach:

ANHANG Verwenden Sie einen DB Table Selector-Knoten. Schließen Sie den Eingangsport mit dem DB-Port (rot) des
Databricks SQL Warehouse Connector Node.

2. Geben Sie im Konfigurationsdialog das Schema und den Delta-Tabellennamen ein. Dann aktivieren Sie die
Benutzerdefinierte Abfrage Checkbox. Ein Textbereich wird angezeigt, wo Sie Ihren eigenen SQL schreiben können

Erklärung.

a. Um ältere Versionen mit der Versionsnummer zu erreichen, geben Sie folgendes SQL ein

Erklärung:

```
SELECT * FROM #table # VERSION AS of
```

Ort ist die Version der Tabelle, auf die Sie zugreifen möchten. Überprüfung Abbildung 14 ein Beispiel einer Versionsnummer zu sehen.

B. Um ältere Versionen mit Zeitstempeln zu erreichen, geben Sie die folgende SQL-Anweisung ein:

```
SELECT * FROM #table # TIMESTAMP AS of
```

Ort ist das Zeitstempelformat. Die unterstützten Zeitstempelformat, bitte überprüfen Sie die Dokumentation von Databricks .

3. Führen Sie den Knoten aus. Dann klicken Sie mit der rechten Maustaste auf den Knoten, wählen DB Daten , und klicken Cache no. of Zeilen um den Tisch zu sehen.

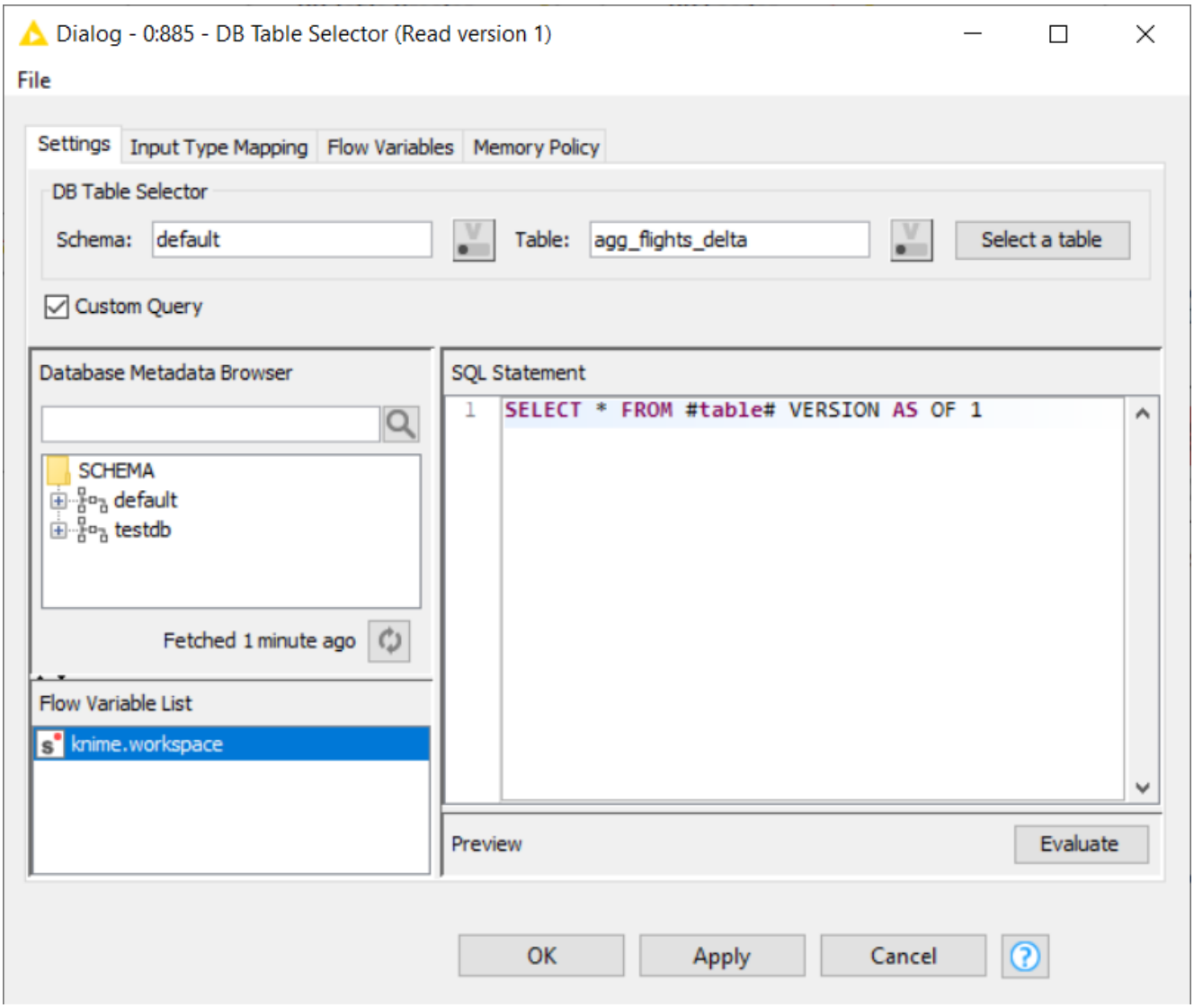

Abbildung 14. Konfigurationsdialog des DB-Tabellenauswahlknotens

Spark zu Hive / Hive zu Spark

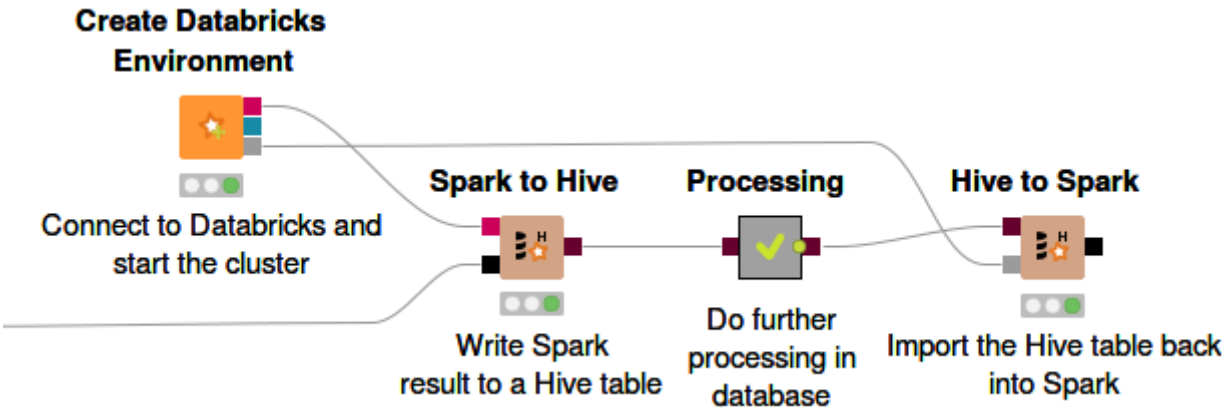

Abbildung 15. Beispiel Nutzung von Spark zu Hive und Hive zu Spark-Knoten

Es ist möglich, einen Spark DataFrame direkt in einer Hive-Datenbank mit dem Spark nach Hive zu speichern

Knoten. Der Knoten hat zwei Eingangsports. Verbinden Sie den DB-Port (rot) mit dem DB-Port des Erstellens Databricks-Umweltknoten und der zweite Spark-Datenport (schwarz) zu jedem Knoten mit einem Funkdatenausgabeport. Dieser Knoten ist sehr nützlich, um Spark-Ergebnis dauerhaft in einem Datenbank.

Andererseits wird der Hive to Spark node verwendet, um eine Hive-Tabelle zurück in einen Spark zu importieren DataFrame. Der Knoten hat zwei Eingangsports. Verbinden Sie den Hive-Port (braun) mit dem Ziel Hive Tabelle, und der Spark-Port (grau) zum Spark-Port des Databricks-Umweltknotens erstellen.

Datenaufbereitung und -analyse

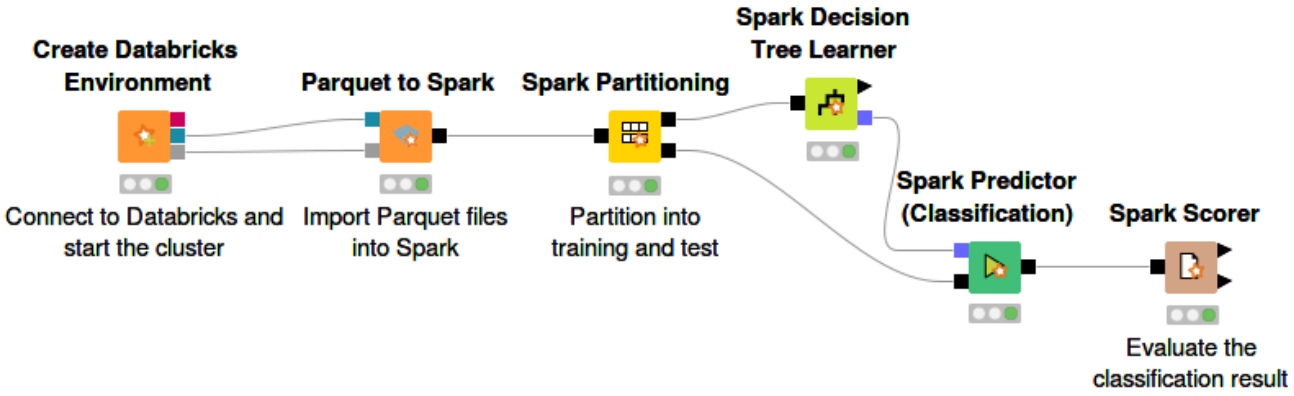

Abbildung 16. Beispiel der maschinellen Lernanwendung mit Funkknoten auf Databricks

Die Databricks-Integrationsknoten vermischen sich nahtlos mit den anderen KNIME-Knoten, die Sie können über die KNIME Analytics Platform eine Vielzahl von Aufgaben auf Databrick-Clustern ausführen, z.B. die Ausführung von Spark-Jobs über die KNIME Spark-Knoten, während die Berechnungsprozess in den Databricks-Cluster. Jede Datenvorverarbeitung und -analyse kann einfach mit den Spark-Knoten gemacht, ohne dass eine einzelne Zeile des Codes geschrieben werden muss.

Für fortgeschrittene Benutzer gibt es eine Möglichkeit, die Skripting-Knoten zu verwenden, um benutzerdefinierte Spark-Jobs zu schreiben, wie die PySpark Script-Knoten, Spark DataFrame Java Snippet-Knoten oder die Spark SQL Suchknoten. Diese Scripting-Knoten erlauben neben den Standard KNIME Spark-Knoten eine detailliertere Kontrolle über die gesamte Datenwissenschaft Pipeline.

Die Scripting-Knoten sind unter Werkzeuge und Dienstleistungen Apokalypse Sonstiges im Knoten-Repository der KNIME Analytics Platform.

Weitere Informationen zu den Spark-Knoten finden Sie in der KNIME Erweiterung für Apache Spark Produktseite .

Ein Beispiel-Workflow, um die Nutzung des Databricks-Umweltknotens erstellen zu zeigen um eine Databricks zu verbinden Cluster aus der KNIME Analytics Platform ist auf der KNIME Hubraum .

Databricks Modelle

Mit der Arbeit Databricks Modelle wie große Sprachmodelle (LLM) und Einbettungen zuerst

Sie müssen die KNIME AI Extension. Einmal installiert können Sie die Databricks LLM

Auswählerknoten um mit einem Databricks-Chat-Modell zu verbinden. Für weitere Details siehe Datenbrände

Dokumentation.

So erstellen Sie Ihre eigene Vektordatenbank in KNIME mit einer Einbettung von Databricks

Sie können Databricks Embedding Model Selector Node. Sobald Sie angeschlossen sind, können Sie

den verschiedenen Vektorspeicherknoten in KNIME, um Ihre Vektordatenbank zu verwalten.

Weitere Informationen zu den KI-Fähigkeiten in KNIME finden Sie in KNIME AI Extension Guide .

KNIME AG
Talacker 50
8001 Zürich, Schweiz
www.knime.com
Info@knime.com