Open for Innovation
KNIME

KNIME Amazon Web Services

Benutzerhandbuch

KNIME AG, Zürich, Schweiz

Version 5.7 (letzte Aktualisierung auf)

Inhaltsverzeichnis

Überblick

KNIME Analytics Platform enthält eine Reihe von Knoten, um mit Amazon Web Services (AWSTM) Sie ermöglichen es Ihnen, Verbindungen zu Amazon-Diensten zu erstellen, wie zum Beispiel Amazon EMR , oder Amazon S3 .

Die KNIME Amazon Cloud Connectors Extension ist verfügbar auf KNIME Hubraum .

Erstellen eines Amazon EMR-Clusters

Dieser Abschnitt beschreibt eine Schritt für Schritt Anleitung zum Erstellen eines EMR-Clusters.

> Der folgende Leitfaden zielt darauf ab, einen Standard-EMR-Spark-Cluster für Tests zu erstellen und Bildungszwecke. Bitte ändern Sie die Einstellungen und Konfigurationen nach Ihren Bedürfnissen.

Vor dem Start eines EMR-Clusters sind folgende Voraussetzungen erforderlich:

- Ein Amazon AWS-Konto. Bitte folgen Sie den Anweisungen in der AWS Dokumentation .
- Ein Amazon S3 Eimer. Der Eimer wird benötigt, um Daten zwischen KNIME und Sparken und die Cluster-Log-Dateien speichern. Um einen Amazon S3 Eimer zu erstellen, folgen Sie bitte die AWS Dokumentation .

Nachdem alle Voraussetzungen erfüllt sind, können Sie den EMR-Cluster erstellen:

ANHANG In der AWS Webkonsole gehen Sie zu EMR

2. Klicken Sie auf die Schaltfläche€luster erstellen am Ende der Seite

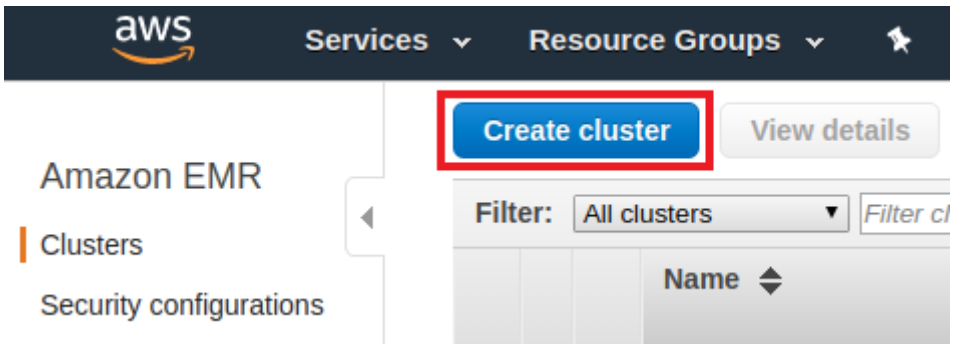

Abbildung 1. Cluster-Taste erstellen

3. Während in der Cluster-Erstellungsseite navigieren Sie zum Erweiterte Optionen

© 2025 KNIME AG. Alle Rechte vorbehalten.

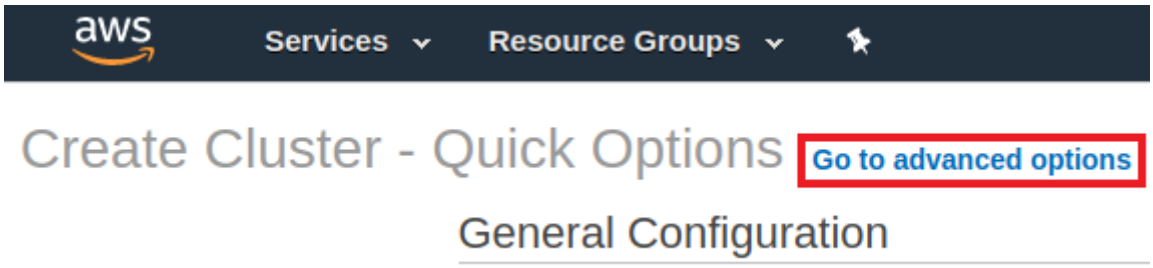

Abbildung 2. Erweiterte Optionen

L 347 vom 20.12.2013, S. 1). Unter Softwarekonfiguration , wählen Sie die zu installierende Software innerhalb des Clusters. wenn Sie wollen Livy und KNIME Spark-Knoten verwenden, installieren Livy und Spark, indem Sie die entsprechende Kontrollkästchen.

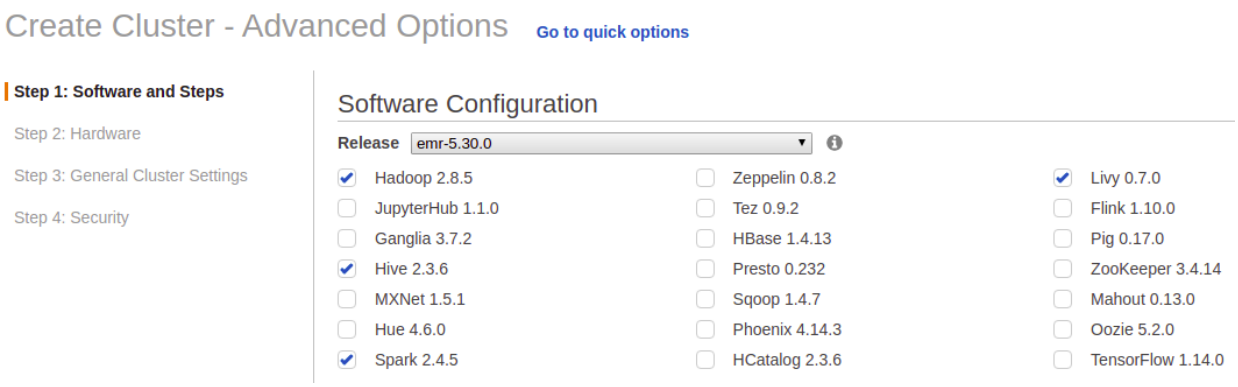

Abbildung 3. Softwarekonfiguration

Unter Softwareeinstellungen bearbeiten, Sie können die Standardkonfigurationen von Anwendungen wie Spark. Im folgenden Beispiel die Funkeneigenschaft maximierenResourceAllocation wird eingestelltwahr um den Ausführenden das Maximum zu ermöglichen Ressourcen möglich auf jedem Knoten in einem Cluster. Bitte beachten Sie, dass diese Funktion nur funktioniert auf einem reinen Spark-Cluster (ohne Hive parallel laufen).

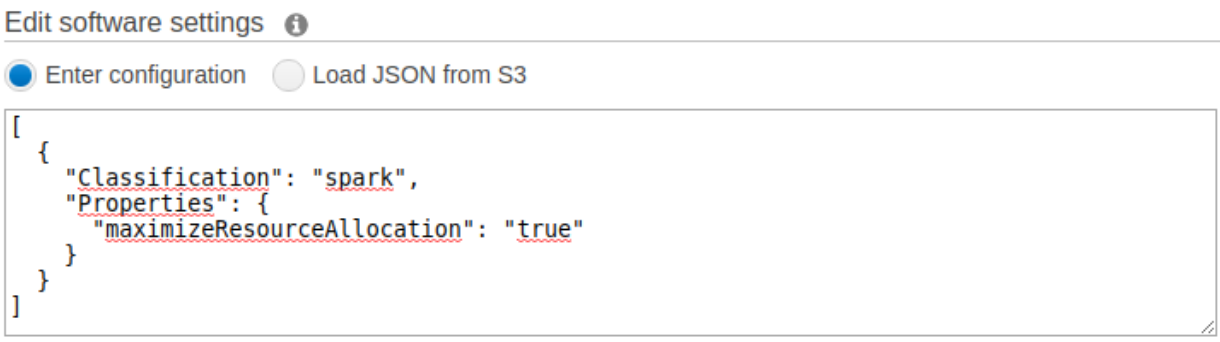

Abbildung 4. Wie maximieren Sie Ressourcen auf einem Spark-Cluster

5. Unter Hardwarekonfiguration , Sie können die EC2-Instanztypen, Anzahl EC2 angeben Instanzen zu initialisieren in jedem Knoten, und die Kaufoption, abhängig von Ihrem Budget. Für einen Standardcluster reicht es aus, die Standardkonfiguration zu verwenden. Der Rest von die Einstellungen, die Sie standardmäßig behalten können, oder sie entsprechend Ihren Bedürfnissen anpassen.

© 2025 KNIME AG. Alle Rechte vorbehalten.

Weitere Informationen zur Hardware- und Netzwerkkonfiguration finden Sie in der AWS Dokumentation . Für eine ausführlichere Anleitung über die optimale Anzahl von Beispiele und andere verwandte Dinge, bitte überprüfen Sie die entsprechenden Richtlinien in den AWS Dokumentation auch.

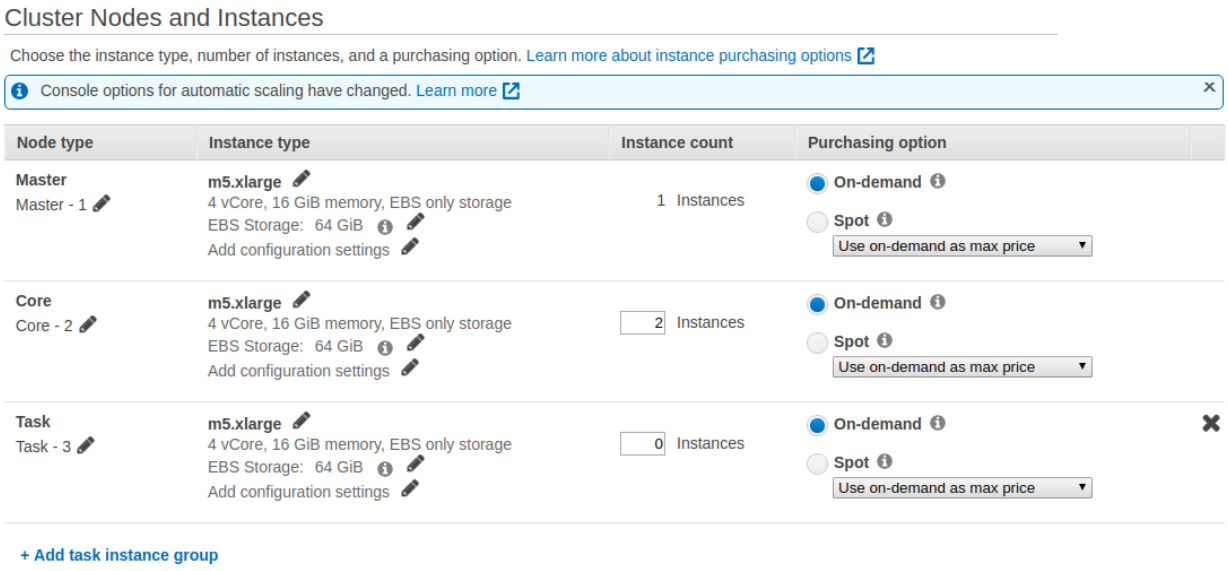

Abbildung 5. Hardwarekonfiguration

6. Unter Allgemeine Optionen Geben Sie den Clusternamen ein. Kündigungsschutz wird aktiviert standardmäßig und ist wichtig, um eine versehentliche Beendigung des Clusters zu verhindern. Zu beenden Sie müssen den Kündigungsschutz deaktivieren.

7. Unter Sicherheitsoptionen , gibt es eine Option, um das EC2-Schlüsselpaar anzugeben. Sie können fortfahren ohne EC2-Schlüsselpaar, aber wenn Sie eins haben und SSH in die EMR Cluster später, Sie können es hier zur Verfügung stellen.

   Weiter unten auf der Seite können Sie auch die Sicherheitsgruppe EG2 . Es handelt sich um eine virtuelle Firewall rund um Ihren Cluster und steuert alle Inbound- und Outbound-Verkehr Ihres Cluster-Knoten. Eine Standard-EMR-gemanagte Sicherheitsgruppe wird automatisch für Ihre neue Cluster, und Sie können die Netzwerkregeln in der Sicherheitsgruppe nach dem Cluster bearbeiten erstellt. Folgen Sie den Anweisungen in der AWS Dokumentation wie man mit EMR arbeitet Geführte Sicherheitsgruppen.

   - [ ] Wenn erforderlich, fügen Sie Ihre IP in die Inbound Regeln für den Zugriff auf den Cluster.
   - [ ] Um einige AWS-Dienste von der KNIME Analytics Platform zugänglich zu machen, Sie müssen bestimmte Ports des EMR-Masterknotens aktivieren. Zum Beispiel Hive ist über den Hafen 10000 erreichbar.

8. Klicken Sie auf Cluster erstellen und der Cluster wird gestartet. Es könnte ein paar Minuten dauern, bis alles die Ressourcen sind verfügbar. Sie wissen, dass der Cluster bereit ist, wenn es ein Wartezeichen gibt

© 2025 KNIME AG. Alle Rechte vorbehalten.

text decoration: underline;

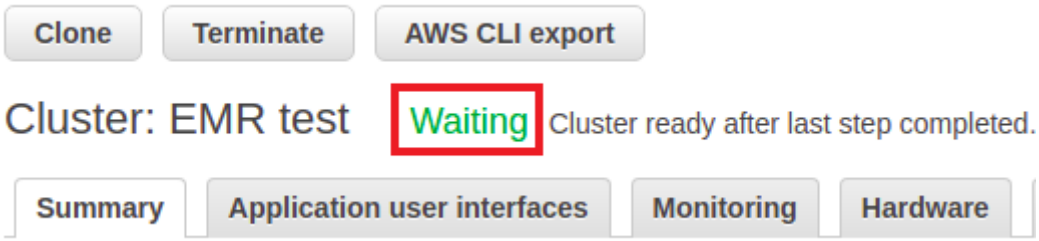

Abbildung 6. Cluster ist bereit

© 2025 KNIME AG. Alle Rechte vorbehalten.

Anschluss an S3

Sie benötigen den Amazon Authentication Node und Amazon S3 Connector Node, um einen Verbindung zu Amazon S3 innerhalb der KNIME Analytics Platform. Für weitere Details, bitte die neue KNIME Leitfaden für die Bearbeitung von Dateien

Sie können überprüfen, ob eine Verbindung erfolgreich durch Klicken Sie auf Prüfanschluss Schaltfläche im Konfigurationsdialog der Amazon Authentication Node. Ein neues Pop-up-Fenster zeigt die Verbindungsinformationen im Format S3://accessKeyId@region und ob eine Verbindung erfolgreich erstellt wird.

Nachdem die Verbindung zu Amazon S3 hergestellt ist, können Sie dann eine Vielzahl von KNIME verwenden Datei-Handling-Knoten, um Dateien auf Amazon S3 zu verwalten (siehe <a href="#page6" style="color: #ff6600; text-de)

Die KNIME-Dateihandling-Knoten sind im Knoten-Repository unter IO .

Weitere Informationen zu Amazon S3 finden Sie unter: AWS Dokumentation .

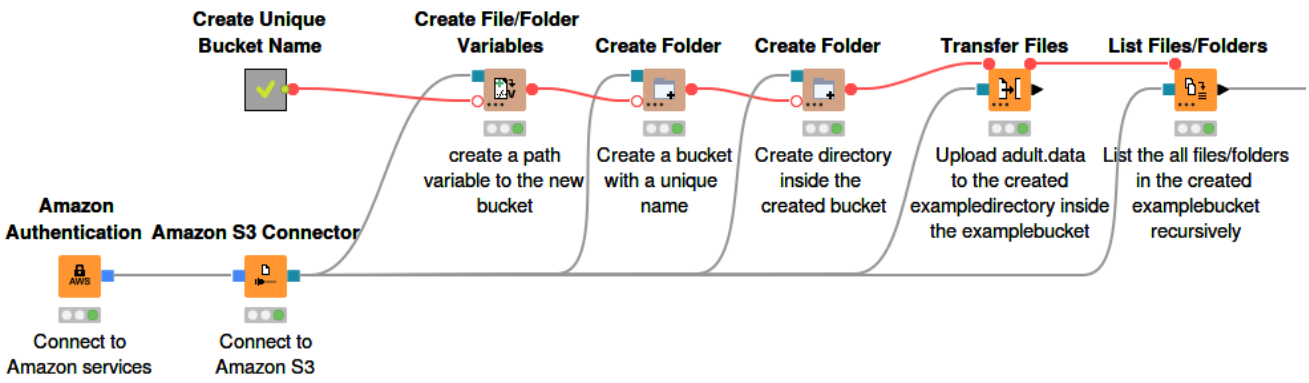

Abbildung 7. Beispiel Nutzung von Amazon Authentication Node und Amazon S3 Connector Node

© 2025 KNIME AG. Alle Rechte vorbehalten.

Apache Hive

Dieser Abschnitt beschreibt, wie eine Verbindung zu Hive auf EMR in KNIME Analytics Plattform.

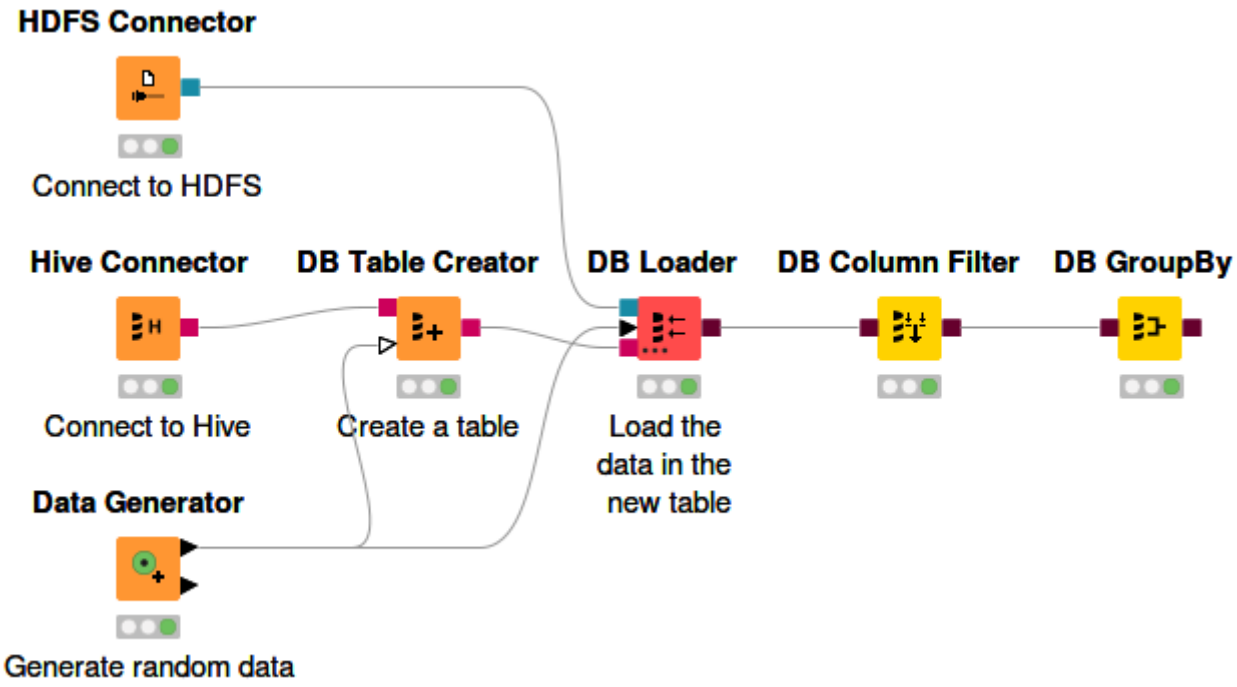

Abbildung 8. Verbinden Sie mit Hive und erstellen Sie einen Hive Tisch

In Abbildung 8, ein Beispiel Workflow zeigt, wie man sich mit Hive verbinden und einen Hive-Tisch erstellen kann.

Registrieren Amazon JDBC Hive driver

Um den Amazon JDBC Hive Treiber in der KNIME Analytics Platform zu registrieren:

1. ANHANG Den Treiber herunterladen AWS Website
2. Extraktion der .zip Datei und die gewünschte Treiberversion
3. Folgen Sie der Anleitung in der Datenbankdokumentation über die Registrierung eines externen JDBC Treiber in KNIME.

Weitere Informationen zum Amazon JDBC Hive Treiber finden Sie unter AWS Dokumentation.

© 2025 KNIME AG. Alle Rechte vorbehalten.

Hive Connector

Der Hive Connector-Knoten erzeugt eine Verbindung über JDBC zu einer Hive-Datenbank. Der Ausgang dieser Knoten ist eine Datenbankverbindung, die mit dem Standard verwendet werden kann KNIME Datenbank Knoten .

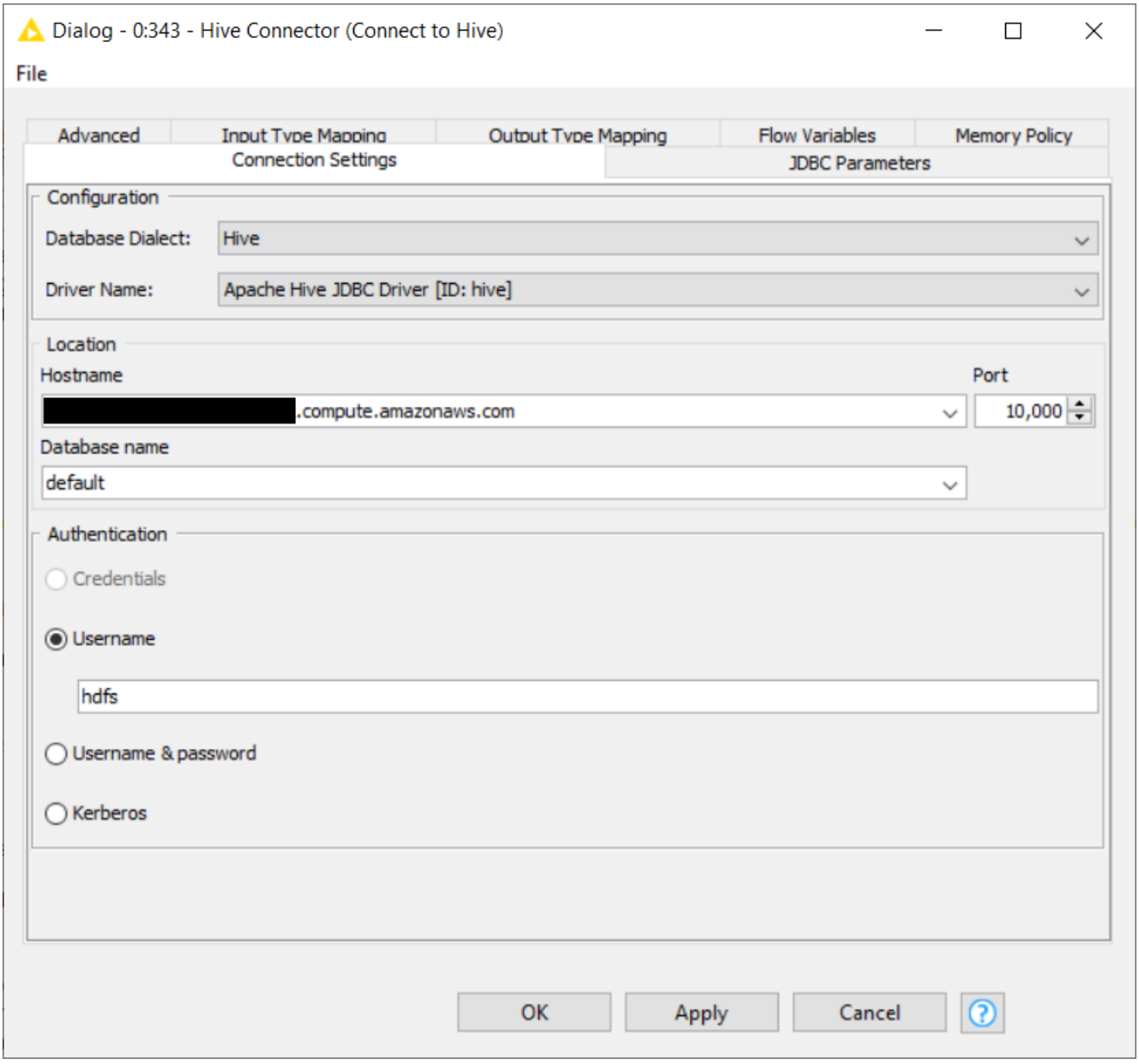

Abbildung 9. Hive Connector Konfigurationsdialog

Im Dialogfeld Knotenkonfiguration müssen Sie festlegen:

- Datenbankdialekt und Treibername. Der Fahrername ist der Name des Fahrers, wenn über die Registrierung des Hive-Treibers).
- Server-Hostname (oder IP-Adresse), Port und Datenbankname
- Authentifizierungsmechanismus. Standardmäßig der Benutzername Hdfs kann als Benutzername verwendet werden ohne Passwort.

© 2025 KNIME AG. Alle Rechte vorbehalten.

Für weitere Informationen über die erweiterten Optionen im Connector-Knoten,

bitte überprüfen KNIME Datenbankdokumentation .

HDFS

Zum Hochladen oder Arbeiten mit Remote-Dateien auf dem EMR-Cluster wird empfohlen, die HDFS zu verwenden

Abbildung 8

) Amazon EMR 5.x kann verwenden Hdfs oder hatte er als HDFS-Administrator

Benutzer.

© 2025 KNIME AG. Alle Rechte vorbehalten.

Amazona

Dieser Abschnitt beschreibt Amazon Athena und wie man mit ihm verbinden, sowie eine Athena erstellen Tabelle über die KNIME Analytics Platform.

Amazona ist ein Abfragedienst, in dem Benutzer SQL-Abfragen gegen ihre Daten ausführen können die sich auf Amazon S3 befinden. In Athena enthalten Datenbanken und Tabellen im Grunde die Metadaten für die zugrunde liegenden Quelldaten. Für jeden Datensatz muss eine entsprechende Tabelle sein erstellt in Athena. Die Metadaten enthalten Informationen wie den Standort des Datensatzes in Amazon S3 und die Struktur der Daten, z.B. Spaltennamen, Datentypen und so weiter.

Der KNIME Amazon Athena Connector Erweiterung ist verfügbar auf KNIME Hubraum .

Es ist sehr wichtig zu beachten, dass Athena nur Ihre Daten auf S3 liest, können Sie nicht hinzufügen oder ändern.

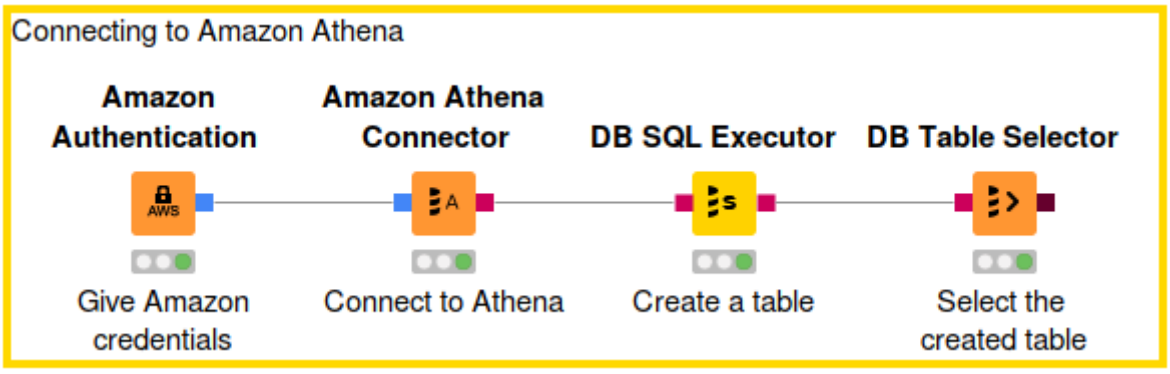

Abbildung 10. Verbinden Sie mit Athena und erstellen Sie eine Athena Tabelle

Verbindung mit Amazon Athena

Aufgrund der Lizenzbeschränkung müssen Sie den Athena JDBC Treiber herunterladen Amazon und registrieren Sie es einmal vor der Verbindung mit Athena. Um den Treiber herunterzuladen Bitte klicken Sie Hier. und die neueste Version des JDBC Treibers herunterladen **ohne** das AWS SDK z. AthenaJDBC42-2.0.35.1001.jar . Einmal heruntergeladen registrieren Sie sich Fahrer über die KNIME Präferenzseite mit Athena als Datenbanktyp in der Leitfaden für die Erweiterung .

Verbindung mit Amazon Athena über die KNIME Analytics Plattform:

ANHANG Verwenden Sie den Amazon Authentication-Knoten, um eine Verbindung zu AWS-Diensten zu erstellen. In der Knoten-Konfigurationsdialog geben Sie bitte die AWS-Zugriffsschlüssel-ID und den geheimen Zugriffsschlüssel an. Weitere Informationen zu AWS-Zugriffsschlüsseln finden Sie in der AWS Dokumentation .

© 2025 KNIME AG. Alle Rechte vorbehalten.

2. Der Amazon Athena Connector-Knoten schafft eine Verbindung zu Athena durch den gebauten
   in Athena JDBC Fahrer. Bitte geben Sie folgende Informationen im Knoten an
   Konfigurationsdialog:

   a. Der Hostname des Athena-Servers. Es hat das Format
      athena..amazonaws.com . Zum Beispiel: athena.eu-west-
      1.amazonaws.com .

   B. Name des S3-Installationsverzeichnisses, um das Abfrageergebnis zu speichern. Zum Beispiel S3://aws-
      athena-query-results-eu-west-1/ .

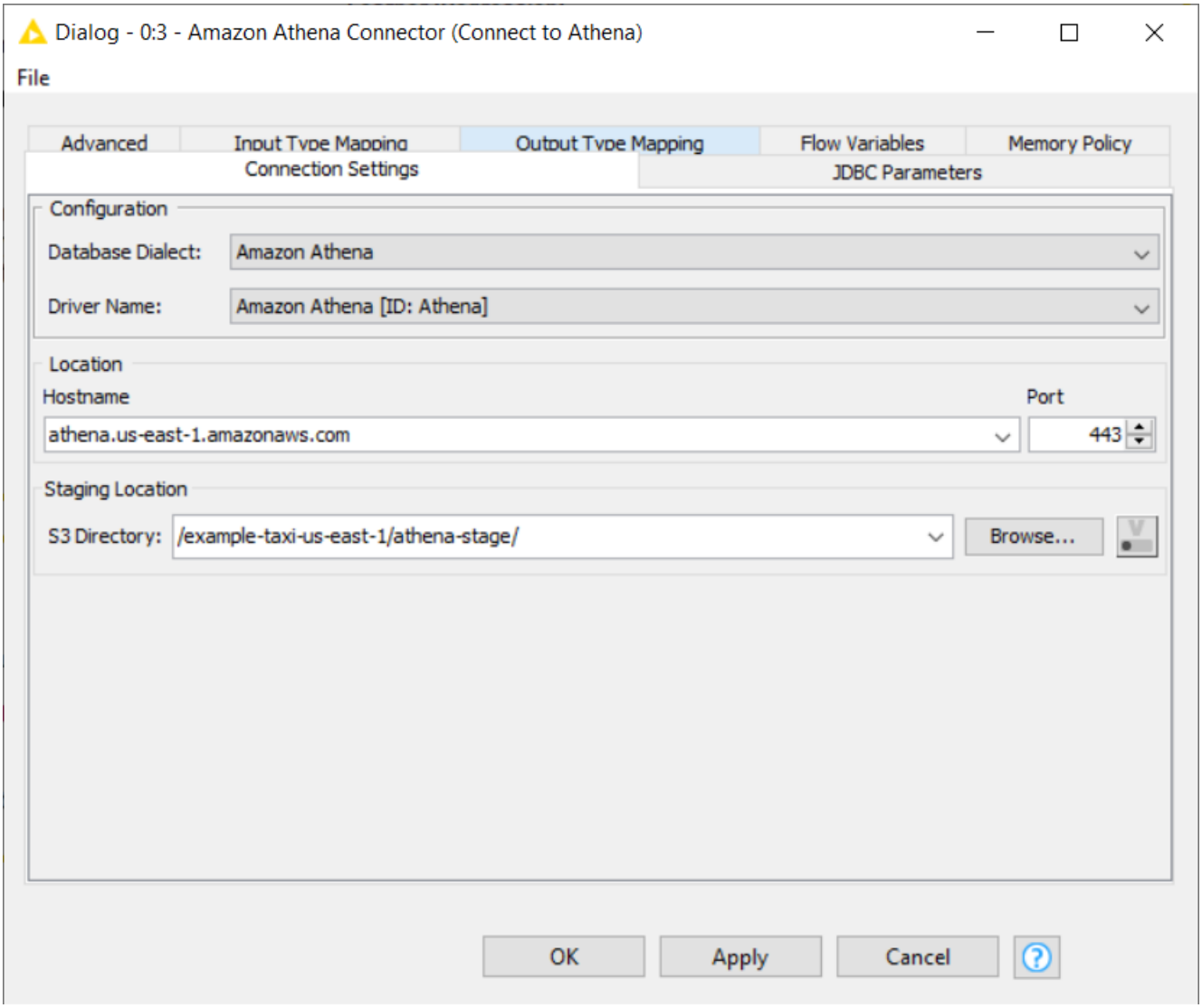

Abbildung 11. Athena Connector Node

Nach der Ausführung dieses Knotens wird eine Verbindung zu Athena hergestellt. Aber bevor du kannst
Startabfragen von Daten in S3 müssen Sie die entsprechende Athena-Tabelle erstellen.

© 2025 KNIME AG. Alle Rechte vorbehalten.

Erstellen Sie eine Athena Tabelle

Die Erstellung einer Athena-Tabelle in der KNIME Analytics Platform erfordert eine SQL-Anweisung, in der Sie müssen Sie selbst bauen BESCHREIBUNG Erklärung. Das folgende Beispiel zeigt a KREIE TABELLE Erklärung zur Erstellung einer Tabelle für den Amazon CloudFront-Log-Datensatz, der Teil des das öffentliche Beispiel Athena Datensatz zur Verfügung gestellt bei s3://athena-examples-REGION>/cloudfront/plaintext/ . Nach dem eigenen Bau BESCHREIBUNG Erklärung, kopieren Sie die Erklärung zum Knotenkonfigurationsdialog des DB SQL Executor node.

```
KREATE EXTERNAL TABELLE IF NICHT EXISTS cloudfront_logs (
`Date` DATE
Zeit STRING,
Standort STRING,
Bytes INT,
Anfrage IP STRING,
Methode STRING,
Host STRING,
Uri STRING,
Status INT,
Schiedsrichter STRING,
STRING,
Browser STRING,
Browserversion STRING
) ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe '
MIT ERZEUGNISSE (
"input.regex" = "^(?!#)([^ ]+)\s+([^ ]+)\s+([^ ]+)\s+([^ ]+)\s+([^ ]+)\s+([^ ]+)\s+([^
]+)\s+([^ ]+)\s+([^ ]+)\s+([^ ]+)\s+([^
]+)\s+[^\(]+[\(]([^\;]+)*\%20([^\/]+)[\/](.*)$"
) LOCATION 's3://athena-examples-/cloudfront/plaintext/';
```

Sobald der DB SQL Executor-Knoten ausgeführt wird, die entsprechende Athena-Tabelle, die enthält Metadaten der Datendateien werden erstellt. Jetzt können Sie die Dateien mit dem Standard KNIME abfragen Datenbankknoten .

> Wenn Sie nicht mit SQL vertraut sind und es lieber interaktiv machen, können Sie auch die Tabelle mit der Athena Web-Konsole erstellen. So können Sie sogar lassen AWS Kleber Crawler um das Dateischema zu erkennen (Spaltennamen, Spaltentypen, unter andere Dinge) automatisch statt sie manuell einzugeben. Folgen Sie der Tutorial in der Athena Dokumentation für eine ausführlichere Erklärung.

Ein Beispiel-Workflow, um die Verwendung des Athena Connector-Knotens zu demonstrieren, um eine Verbindung mit Amazon Athena aus der KNIME Analytics Platform ist auf KNIME Hubraum (siehe 10.
<a href="#page1
)

© 2025 KNIME AG. Alle Rechte vorbehalten.

Spark Jobs auf einem EMR-Cluster ausführen

Dieser Abschnitt beschreibt, wie ein Spark-Job auf einem EMR-Cluster von innen konfiguriert und ausgeführt werden kann KNIME Analytics Platform. Vor dem Betrieb eines Spark-Jobs auf einem EMR-Cluster hat ein Spark-Kontext zu erstellen. Um einen Spark-Kontext über Livy zu erstellen, verwenden Sie den Create Spark Context (Livy)-Knoten.

Spark Context (Livy) Node erstellen

Der Create Spark Context (Livy)-Knoten erstellt einen Spark-Kontext über Apokalypse . Der Knoten hat einen Remote-Anschlussport (blau) als Eingabe. Die Idee ist, dass dieser Knoten Zugang zu einem Remote-Dateisystem, um temporäre Dateien zwischen KNIME und dem Spark-Kontext zu speichern.

Eine breite Palette von Dateisystemen werden unterstützt, wie HDFS, webHDFS, httpFS, Amazon S3, Azure Blob Store und Google Cloud Storage. Bitte beachten Sie jedoch, dass die Verwendung, z.B. HDFS ist kompliziert auf einem Remote-Cluster, weil sich der Speicher auf dem Cluster befindet, daher jede Daten, die gespeichert werden, werden verloren gehen, sobald der Cluster beendet ist.

- Die empfohlene und einfache Möglichkeit ist Amazon S3 zu verwenden. Bitte überprüfen Sie die <a href="#page6" style="color: #ff6600; text-decoration: none;">über den Aufbau einer Verbindung zu Amazon S3.
- Die anderen Anschlussknoten sind unter IO > Anschlüsse innerhalb des Knotens Repository.

Öffnen Sie den Knotenkonfigurationsdialog des Knotens Spark Context (Livy) erstellen. In diesem Fenster Sie haben einige Informationen, die wichtigsten sind:

- Die Spark-Version. Die Version muss die gleiche sein wie die von Livy. Andernfalls der Knoten wird scheitern. Sie finden die Spark-Version in der Cluster-Zusammenfassungsseite oder in der Softwarekonfiguration Schritt während der Cluster-Erstellung (siehe <a href="#page3" style="color: #ff6600; te) auf dem Amazon EMR Web-Konsole.
- Die Livy URL inklusive Protokoll und Port z. http://localhost:8998 . Sie können die URL in der Cluster-Zusammenfassungsseite auf der Amazon EMR-Webkonsole (siehe <a href="#page14" style=) Dann befestigen Sie einfach den Standardport 8998 am Ende der URL.

© 2025 KNIME AG. Alle Rechte vorbehalten.

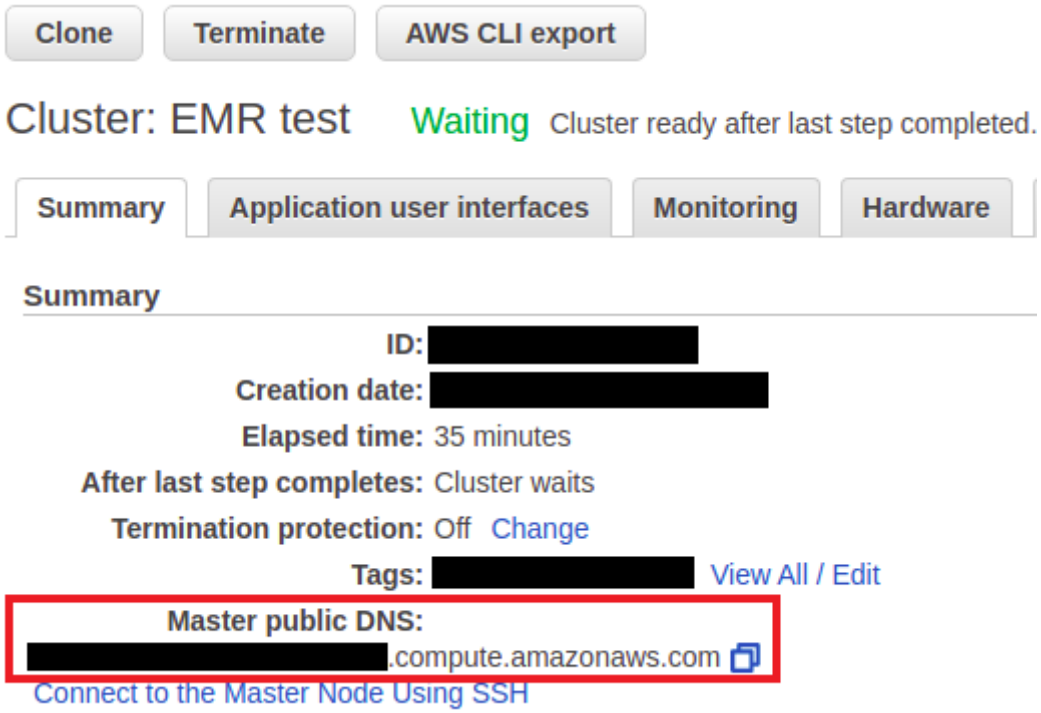

Abbildung 12. Die Livy URL auf der Cluster-Zusammenfassungsseite

• Wählen Sie die Authentifizierungsmethode aus. In der Regel ist keine Authentifizierung erforderlich, aber wenn Sie haben beispielsweise eine Kerberos-Authentifizierung auf dem Cluster eingerichtet, können sie auch in KNIME. Wenn das der Fall ist, müssen Sie Kerberos in KNIME Analytics einrichten Plattform zuerst. Bitte überprüfen Sie die KNIME Dokumentation von Kerberos für weitere Details.

• Unter Spark Ausführende Ressourcen Abschnitt, kann die Ressourcen manuell eingestellt werden, d.h. Menge Speicher und Anzahl der Kerne, für jeden Spark-Executor. Es gibt drei Mögliche Spark-Executor-Zuordnungsstrategien, standardmäßig, fest und dynamisch.

• Unter Erweiterte Tab, gibt es eine Option, um den Einlegebereich für Spark Jobs einzustellen. Für Amazon S3, es ist zwingend erforderlich, ein Inserate-Verzeichnis bereitzustellen. Zusätzlich gibt es auch eine Option, um die standardmäßigen Spark-Treiberressourcen (die Menge an Speicher und Kerne) zu überschreiben der Spark-Treiber-Prozess wird Zuweisung, und um benutzerdefinierte Spark-Einstellungen .

© 2025 KNIME AG. Alle Rechte vorbehalten.

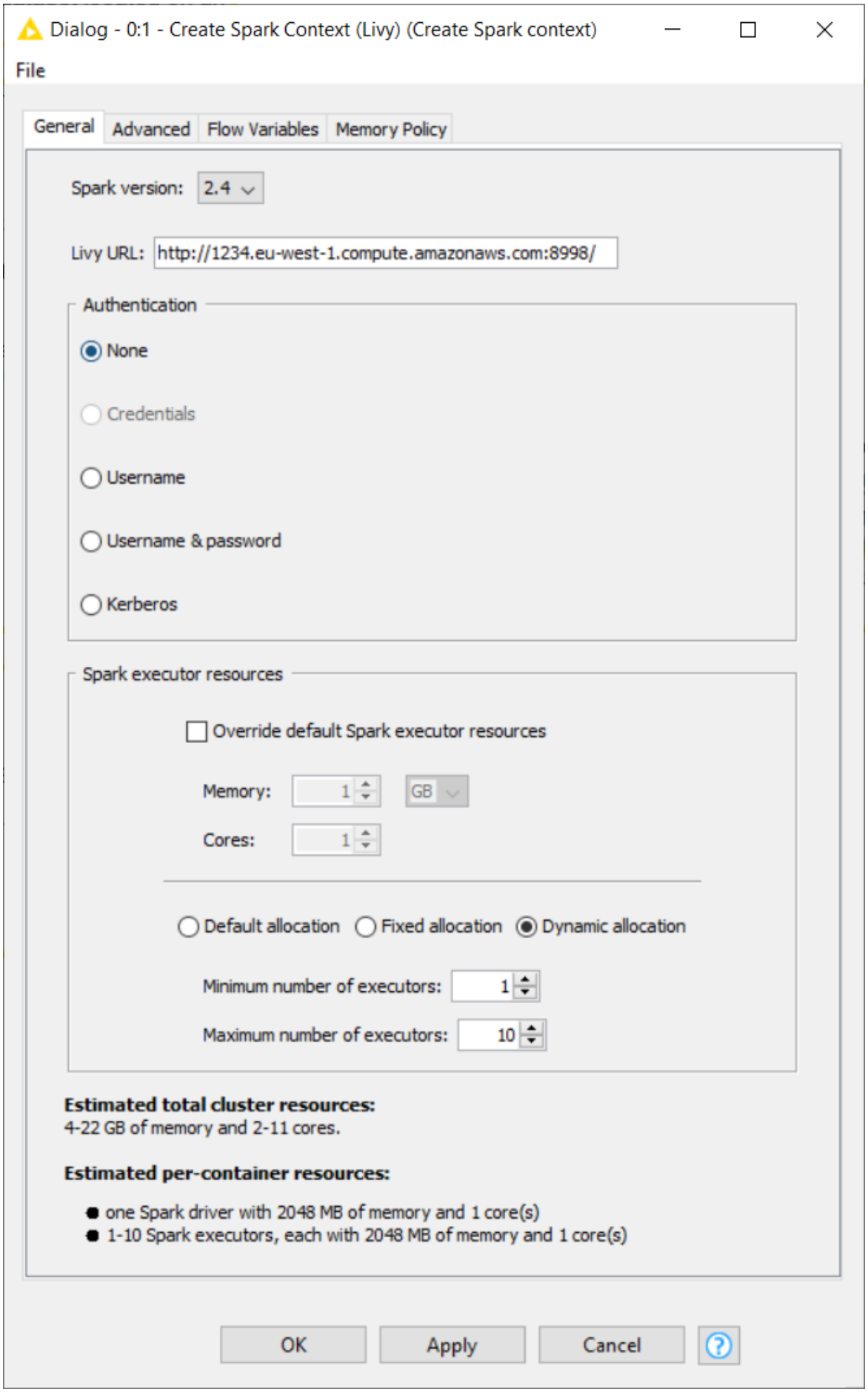

Abbildung 13. Erstellen Sie Spark Context (Livy) Node

© 2025 KNIME AG. Alle Rechte vorbehalten.

Nachdem der Create Spark Context (Livy)-Knoten ausgeführt wird, wird der Ausgang Spark-Knoten (grau) den neu erstellten Spark-Kontext enthalten. Es ermöglicht die Ausführung von Spark-Jobs über KNIME Spark Knoten .

Für eine ausführlichere Erklärung zum Lesen und Schreiben von Daten zwischen Remote-Dateisystem und Spark DataFrame über die KNIME Analytics Platform, bitte Schauen Sie sich KNIME Dokumentation von Databricks .

Abbildung 14 zeigt ein einfaches Beispiel, bei dem ein Random Forest-Algorithmus verwendet wird, um eine Prädiktionsmodell auf einem Datensatz, alle auf einem EMR-Spark-Cluster ausgeführt.

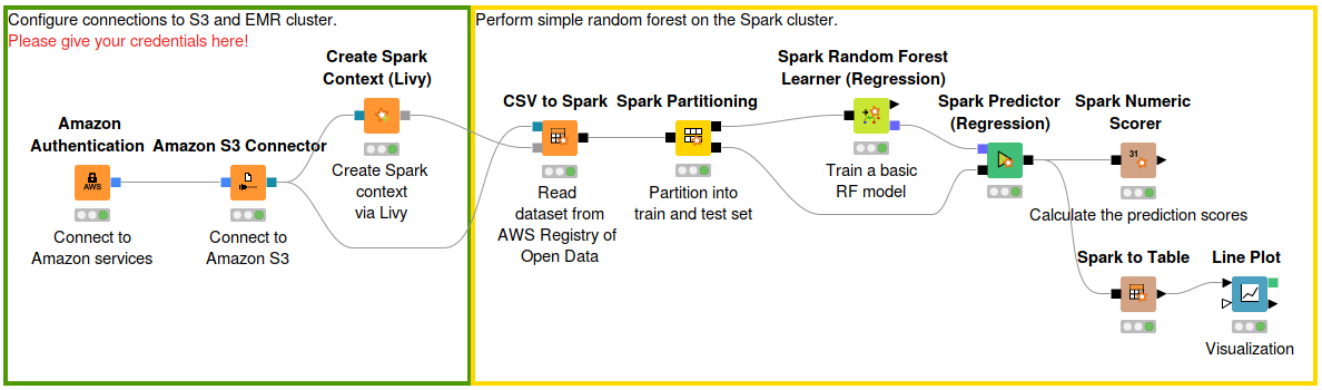

Abbildung 14. Trainieren Sie ein maschinelles Lernmodell auf einem Spark EMR-Cluster

Ein Beispiel-Workflow, um die Verwendung von Amazon EMR aus KNIME zu demonstrieren Analytics Platform ist auf KNIME Hubraum .

© 2025 KNIME AG. Alle Rechte vorbehalten.

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

Die Marke KNIME® und das Logo OPEN FOR INNOVATION® werden von der KNIME AG unter Lizenz verwendet von der KNIME GmbH, und sind in den USA registriert. KNIME® ist auch in Deutschland registriert.