Open for Innovation
KNIME

KNIME Big Data Extensions Admin Guide

KNIME AG, Zurich, Switzerland
Version 5.2 (last updated on 2023-08-15)

Table of Contents

Overview

KNIME Big Data Extensions integrate Apache Spark and the Apache Hadoop ecosystem with KNIME Analytics Platform.

This guide is aimed at IT professionals who need to integrate KNIME Analytics Platform with an existing Hadoop/Spark environment.

The steps in this guide are required so that users of KNIME Analytics Platform run Spark workflows. Note that running Spark workflows on KNIME Server requires **additional** steps outlined in Secured Cluster Connection Guide for KNIME Server.

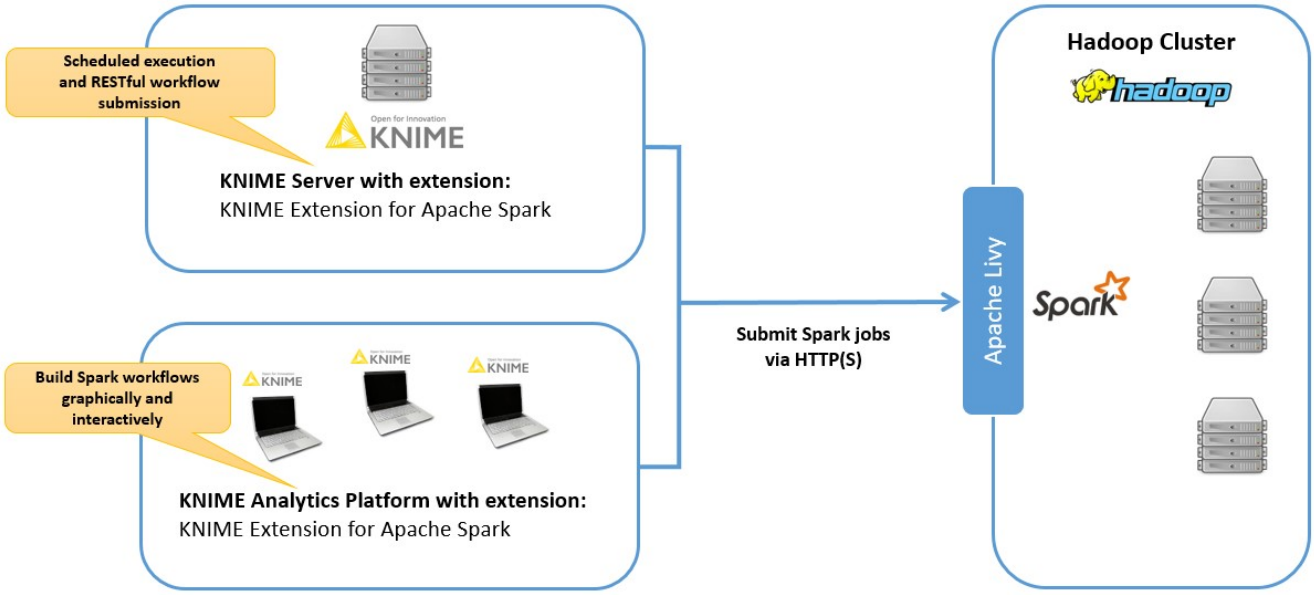

Figure 1. Overall architecture

KNIME Extension for Apache Spark requires Apache Livy as REST service to be installed on an edge/fronted node of the cluster. See the requirements in the compatibility lists below and howto install Livy.

General Compatibility

KNIME Extension for Apache Spark is compatible with

- Spark 2.x - 3.4
- Livy 0.4 - 0.7

Cloudera CDP Compatibility

KNIME Extension for Apache Spark is compatible with

- Spark 3.3 on CDP 7.1.8 as provided by Cloudera CDS 3.3
- Spark 3.2 on CDP 7.1.7 as provided by Cloudera CDS 3.2
- Spark 3.1 on CDP 7.1.6 as provided by Cloudera CDS 3.1
- Spark 3.0 on CDP 7.1.5 as provided by Cloudera CDS 3.0
- Spark 2.4 as included in CDP 7

Cloudera CDH Compatibility

KNIME Extension for Apache Spark is compatible with

- Spark 2.x on CDH 5 as provided by Cloudera CDS
- Spark 2.x as included in CDH 6

Cloudera CDH 5/6 does not include Livy, therefore KNIME provides CSDs/parcels for Livy (see Cloudera CDH).

Cloudera HDP Compatibility

KNIME Extension for Apache Spark is compatible with

- Spark 2.x as included in HDP 2.6.3 - 2.6.5
- Spark 2.x as included in HDP 3.0.0 - 3.1.5

Amazon EMR Compatibility

KNIME Extension for Apache Spark is compatible with

- EMR 6.x with Spark 3.x and Livy 0.6 - 0.7
- EMR 5.9+ with Spark 2.x and Livy 0.4 - 0.7

Apache Spark update policy

So far we did release new versions of Apache Spark once the new version was supported by

H2O Sparkling Water to provide our users with a consistent user experience. However, in the past, the H2O Sparkling Water release occasionally took quite some time to be published which is why we decided to decouple the release of the KNIME Extension for Apache Spark and the KNIME Databricks Integration, from the KNIME H2O Sparkling Water Integration and the KNIME Extension for Local Big Data Environments.

This way we can support the latest Spark version quicker while supporting all big data nodes with the local big data environment.

In fact, in this case, the nodes of the KNIME H2O Sparkling Water Integration will work in the local big data environment but not in your big data environment if the latest Spark version is selected.

Once the latest version of Spark is supported by H2O Sparkling Water we will update the KNIME H2O Sparkling Water Integration and the KNIME Extension for Local Big Data Environments.

When updating the H2O Sparkling Water extension we will also update the KNIME H2O Machine Learning Integration since they both need to run with the same version.

Apache Livy setup

Cloudera CDP

Cloudera Runtime 7.0 - 7.1 includes Spark 2.4 and Livy as Service. Cloudera povides Spark 3.x as a Custom Service Descriptor that can coexists with the included Spark version. See *Installing CDS Powered by Apache Spark* in the CDS 3.3 or CDS 3.2 or CDS 3.1 or CDS 3.0 Cloudera documentation for more information. Note that Livy in the Spark 3.x CSD uses 28998 instead of the usual 8998 as default port.

If you plan to run Spark workflows on KNIME Server: Please consult the Secured Cluster Connection Guide for KNIME Server to allow KNIME Server to impersonate users.

Cloudera CDH

For Cloudera CDH, KNIME provides a CSD and parcel so that Livy can be installed as an add-on service. The current version of Livy for CDH provided by KNIME is `0.5.0.knime3`.

The following steps describe how to install Livy as managed service through Cloudera Manager using a parcel. If in doubt, please also consider the official Cloudera documentation on handling parcels.

Prerequisites

- A cluster with CDH 5.8 and newer, or CDH 6.0 and newer
  - *Only On CDH 5*: Spark 2.2 or higher as an add-on service (provided by Cloudera CDS)
- Root shell access (e.g. via SSH) on the machine where Cloudera Manager is installed.
- Full administrative access on the Cloudera Manager WebUI.

Installation steps

In a root shell on the machine where Cloudera Manager is installed:

1. Download a matching CSD from CSDs for Cloudera CDH to `/opt/cloudera/csd/` on the machine, where Cloudera Manager is installed.
2. **Only if Cloudera Manager cannot access the public internet:** Download/copy the matching `.parcel` and `.sha1` file from Parcels for Cloudera CDH to `/opt/cloudera/parcel-repo`.
3. Restart Cloudera Manager from the command line, for example with:

```
systemctl restart cloudera-scm-server
```

In the Cloudera Manager WebUI:

1. Navigate to the Parcel manager and locate the `LIVY` parcel.
2. Download (unless already done manually), Distribute and Activate the `LIVY` parcel.
3. Add the Livy Service to your cluster (see the official Cloudera documentation on adding services).
4. Navigate to the HDFS service configuration and add the following settings to the *Cluster-wide Advanced Configuration Snippet (Safety Valve) for core-site.xml*:
   - `hadoop.proxyuser.livy.hosts=*`
   - `hadoop.proxyuser.livy.groups=*`

5. *If your cluster is using HDFS Transparent Encryption:* Navigate to the KMS service configuration and add the following settings to the *Key Management Server Advanced Configuration Snippet (Safety Valve) for kms-site.xml*:
    - `hadoop.kms.proxyuser.livy.hosts=*`
    - `hadoop.kms.proxyuser.livy.groups=*`
6. *If you plan to run Spark workflows on KNIME Server:* Please consult the Secured Cluster Connection Guide for KNIME Server to allow KNIME Server to impersonate users.
7. Restart all services affected by your configuration changes.

Cloudera HDP

HDP already includes compatible versions of Apache Livy and Spark 2 (see Cloudera HDP Compatibility). Please follow the respective Hortonworks documentation to install Spark with the *Livy for Spark2 Server* component:

- Installing Spark Using Ambari (HDP 2.6.5)
- Install Spark Using Ambari (HDP 3.1)

KNIME Extension for Apache Spark only supports *Livy for Spark2 Server* which uses Spark 2. The *Livy for Spark Server* component is not supported, since it is based on Spark 1.

Amazon EMR

Amazon EMR already includes compatible versions of Apache Livy and Spark 2 (see Amazon EMR Compatibility), simply make sure to select *Livy* in the software configuration of your cluster.

Downloads

Apache Livy downloads

CSDs for Cloudera CDH

- CSD for CDH 5

- CSD for CDH 6

Parcels for Cloudera CDH

Download links for CDH 5:

RHEL/CentOS	RHEL 7: parcel / sha	RHEL 6: parcel / sha	RHEL 5: parcel / sha
SLES	SLES 12: parcel / sha	SLES 11: parcel / sha	
Ubuntu	Ubuntu 16 (Xenial): parcel / sha	Ubuntu 14 (Trusty): parcel / sha	Ubuntu 12 (Precise): parcel / sha
Debian	Debian 8 (Jessie): parcel / sha	Debian 7 (Wheezy): parcel / sha	

Download links for CDH 6

RHEL/CentOS	RHEL 7: parcel / sha	RHEL 6: parcel / sha	RHEL 5: parcel / sha
SLES	SLES 12: parcel / sha	SLES 11: parcel / sha	
Ubuntu	Ubuntu 16 (Xenial): parcel / sha	Ubuntu 14 (Trusty): parcel / sha	Ubuntu 12 (Precise): parcel / sha
Debian	Debian 8 (Jessie): parcel / sha	Debian 7 (Wheezy): parcel / sha	

KNIME AG
Talacker 50
8001 Zurich, Switzerland
www.knime.com
info@knime.com