Informatica

Informatica® PowerExchange for Hadoop
10.1.1 Update 2

User Guide for PowerCenter

Informatica PowerExchange for Hadoop User Guide for PowerCenter
10.1.1 Update 2
March 2017

© Copyright Informatica LLC 2011, 2018

This software and documentation are provided only under a separate license agreement containing restrictions on use and disclosure. No part of this document may be reproduced or transmitted in any form, by any means (electronic, photocopying, recording or otherwise) without prior consent of Informatica LLC.

Informatica, the Informatica logo, PowerCenter, and PowerExchange are trademarks or registered trademarks of Informatica LLC in the United States and many jurisdictions throughout the world. A current list of Informatica trademarks is available on the web at https://www.informatica.com/trademarks.html. Other company and product names may be trade names or trademarks of their respective owners.

This product includes software developed by the Apache Software Foundation (http://www.apache.org/), and/or other software which is licensed under various versions of the Apache License (the "License"). You may obtain a copy of these Licenses at http://www.apache.org/licenses/. Unless required by applicable law or agreed to in writing, software distributed under these Licenses is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the Licenses for the specific language governing permissions and limitations under the Licenses.

This product includes software which was developed by Mozilla (http://www.mozilla.org/), software copyright The JBoss Group, LLC, all rights reserved; software copyright © 1999-2006 by Bruno Lowagie and Paulo Soares and other software which is licensed under various versions of the GNU Lesser General Public License Agreement, which may be found at http:// www.gnu.org/licenses/lgpl.html. The materials are provided free of charge by Informatica, "as-is", without warranty of any kind, either express or implied, including but not limited to the implied warranties of merchantability and fitness for a particular purpose.

The product includes ACE(TM) and TAO(TM) software copyrighted by Douglas C. Schmidt and his research group at Washington University, University of California, Irvine, and Vanderbilt University, Copyright (©) 1993-2006, all rights reserved.

This product includes software developed by the OpenSSL Project for use in the OpenSSL Toolkit (copyright The OpenSSL Project. All Rights Reserved) and redistribution of this software is subject to terms available at http://www.openssl.org and http://www.openssl.org/source/license.html.

This product includes Curl software which is Copyright 1996-2013, Daniel Stenberg, <daniel@haxx.se>. All Rights Reserved. Permissions and limitations regarding this software are subject to terms available at http://curl.haxx.se/docs/copyright.html. Permission to use, copy, modify, and distribute this software for any purpose with or without fee is hereby granted, provided that the above copyright notice and this permission notice appear in all copies.

The product includes software copyright 2001-2005 (©) MetaStuff, Ltd. All Rights Reserved. Permissions and limitations regarding this software are subject to terms available at http://www.dom4j.org/ license.html.

The product includes software copyright © 2004-2007, The Dojo Foundation. All Rights Reserved. Permissions and limitations regarding this software are subject to terms available at http://dojotoolkit.org/license.

This product includes ICU software which is copyright International Business Machines Corporation and others. All rights reserved. Permissions and limitations regarding this software are subject to terms available at http://source.icu-project.org/repos/icu/icu/trunk/license.html.

This product includes software copyright © 1996-2006 Per Bothner. All rights reserved. Your right to use such materials is set forth in the license which may be found at http:// www.gnu.org/software/ kawa/Software-License.html.

This product includes OSSP UUID software which is Copyright © 2002 Ralf S. Engelschall, Copyright © 2002 The OSSP Project Copyright © 2002 Cable & Wireless Deutschland. Permissions and limitations regarding this software are subject to terms available at http://www.opensource.org/licenses/mit-license.php.

This product includes software developed by Boost (http://www.boost.org/) or under the Boost software license. Permissions and limitations regarding this software are subject to terms available at http:/ /www.boost.org/LICENSE_1_0.txt.

This product includes software copyright © 1997-2007 University of Cambridge. Permissions and limitations regarding this software are subject to terms available at http:// www.pcre.org/license.txt.

This product includes software copyright © 2007 The Eclipse Foundation. All Rights Reserved. Permissions and limitations regarding this software are subject to terms available at http:// www.eclipse.org/org/documents/epl-v10.php and at http://www.eclipse.org/org/documents/edl-v10.php.

This product includes software licensed under the terms at http://www.tcl.tk/software/tcltk/license.html, http://www.bosrup.com/web/overlib/?License, http:// www.stlport.org/doc/ license.html, http://asm.ow2.org/license.html, http://www.cryptix.org/LICENSE.TXT, http://hsqldb.org/web/hsqlLicense.html, http:// httpunit.sourceforge.net/doc/ license.html, http://jung.sourceforge.net/license.txt , http://www.gzip.org/zlib/zlib_license.html, http://www.openldap.org/software/ release/license.html, http://www.libssh2.org, http://slf4j.org/license.html, http://www.sente.ch/software/OpenSourceLicense.html, http://fusesource.com/downloads/ license-agreements/fuse-message-broker-v-5-3- license-agreement; http://antlr.org/license.html; http://aopalliance.sourceforge.net/; http://www.bouncycastle.org/ licence.html; http://www.jgraph.com/jgraphdownload.html; http://www.jcraft.com/jsch/LICENSE.txt; http://jotm.objectweb.org/bsd_license.html; . http://www.w3.org/ Consortium/Legal/2002/copyright-software-20021231; http://www.slf4j.org/license.html; http://nanoxml.sourceforge.net/orig/copyright.html; http://www.json.org/ license.html; http://forge.ow2.org/projects/javaservice/, http://www.postgresql.org/about/licence.html, http://www.sqlite.org/copyright.html, http://www.tcl.tk/ software/tcltk/license.html, http://www.jaxen.org/faq.html, http://www.jdom.org/docs/faq.html, http://www.slf4j.org/license.html; http://www.iodbc.org/dataspace/ iodbc/wiki/iODBC/License; http://www.keplerproject.org/md5/license.html; http://www.toedter.com/en/jcalendar/license.html; http://www.edankert.com/bounce/ index.html; http://www.net-snmp.org/about/license.html; http://www.openmdx.org/#FAQ; http://www.php.net/license/3_01.txt; http://srp.stanford.edu/license.txt; http://www.schneier.com/blowfish.html; http://www.jmock.org/license.html; http://xsom.java.net; http://benalman.com/about/license/; https://github.com/CreateJS/ EaselJS/blob/master/src/easeljs/display/Bitmap.js; http://www.h2database.com/html/license.html#summary; http://jsoncpp.sourceforge.net/LICENSE; http:// jdbc.postgresql.org/license.html; http://protobuf.googlecode.com/svn/trunk/src/google/protobuf/descriptor.proto; https://github.com/rantav/hector/blob/master/ LICENSE; http://web.mit.edu/Kerberos/krb5-current/doc/mitK5license.html; http://jibx.sourceforge.net/jibx-license.html; https://github.com/lyokato/libgeohash/blob/ master/LICENSE; https://github.com/hjiang/jsonxx/blob/master/LICENSE; https://code.google.com/p/lz4/; https://github.com/jedisct1/libsodium/blob/master/ LICENSE; http://one-jar.sourceforge.net/index.php?page=documents&file=license; https://github.com/EsotericSoftware/kryo/blob/master/license.txt; http://www.scala-lang.org/license.html; https://github.com/tinkerpop/blueprints/blob/master/LICENSE.txt; http://gee.cs.oswego.edu/dl/classes/EDU/oswego/cs/dl/util/concurrent/ intro.html; https://aws.amazon.com/asl/; https://github.com/twbs/bootstrap/blob/master/LICENSE; https://sourceforge.net/p/xmlunit/code/HEAD/tree/trunk/ LICENSE.txt; https://github.com/documentcloud/underscore-contrib/blob/master/LICENSE, and https://github.com/apache/hbase/blob/master/LICENSE.txt.

This product includes software licensed under the Academic Free License (http://www.opensource.org/licenses/afl-3.0.php), the Common Development and Distribution License (http://www.opensource.org/licenses/cddl1.php) the Common Public License (http://www.opensource.org/licenses/cpl1.0.php), the Sun Binary Code License Agreement Supplemental License Terms, the BSD License (http:// www.opensource.org/licenses/bsd-license.php), the new BSD License (http:// opensource.org/licenses/BSD-3-Clause), the MIT License (http://www.opensource.org/licenses/mit-license.php), the Artistic License (http://www.opensource.org/ licenses/artistic-license-1.0) and the Initial Developer's Public License Version 1.0 (http://www.firebirdsql.org/en/initial-developer-s-public-license-version-1-0/).

This product includes software copyright © 2003-2006 Joe Walnes, 2006-2007 XStream Committers. All rights reserved. Permissions and limitations regarding this software are subject to terms available at http://xstream.codehaus.org/license.html. This product includes software developed by the Indiana University Extreme! Lab. For further information please visit http://www.extreme.indiana.edu/.

This product includes software Copyright (c) 2013 Frank Balluffi and Markus Moeller. All rights reserved. Permissions and limitations regarding this software are subject to terms of the MIT license.

See patents at https://www.informatica.com/legal/patents.html.

DISCLAIMER: Informatica LLC provides this documentation "as is" without warranty of any kind, either express or implied, including, but not limited to, the implied warranties of noninfringement, merchantability, or use for a particular purpose. Informatica LLC does not warrant that this software or documentation is error free. The information provided in this software or documentation may include technical inaccuracies or typographical errors. The information in this software and documentation is subject to change at any time without notice.

NOTICES

This Informatica product (the "Software") includes certain drivers (the "DataDirect Drivers") from DataDirect Technologies, an operating company of Progress Software Corporation ("DataDirect") which are subject to the following terms and conditions:

1. THE DATADIRECT DRIVERS ARE PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NON-INFRINGEMENT.
2. IN NO EVENT WILL DATADIRECT OR ITS THIRD PARTY SUPPLIERS BE LIABLE TO THE END-USER CUSTOMER FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, CONSEQUENTIAL OR OTHER DAMAGES ARISING OUT OF THE USE OF THE ODBC DRIVERS, WHETHER OR NOT INFORMED OF THE POSSIBILITIES OF DAMAGES IN ADVANCE. THESE LIMITATIONS APPLY TO ALL CAUSES OF ACTION, INCLUDING, WITHOUT LIMITATION, BREACH OF CONTRACT, BREACH OF WARRANTY, NEGLIGENCE, STRICT LIABILITY, MISREPRESENTATION AND OTHER TORTS.

The information in this documentation is subject to change without notice. If you find any problems in this documentation, please report them to us in writing at Informatica LLC 2100 Seaport Blvd. Redwood City, CA 94063.

INFORMATICA LLC PROVIDES THE INFORMATION IN THIS DOCUMENT "AS IS" WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT.

Publication Date: 2018-09-28

Table of Contents

Preface

The *PowerExchange® for Hadoop User Guide* provides information about how to build mappings to extract data from Hadoop and load data into Hadoop.

It is written for database administrators and developers. This guide assumes you have knowledge of relational database concepts and database engines, flat files, PowerCenter®, Hadoop, the Hadoop Distributed File System (HDFS), and Apache Hive. You must also be familiar with the interface requirements for other supporting applications.

Informatica Resources

Informatica Network

Informatica Network hosts Informatica Global Customer Support, the Informatica Knowledge Base, and other product resources. To access Informatica Network, visit https://network.informatica.com.

As a member, you can:

- Access all of your Informatica resources in one place.
- Search the Knowledge Base for product resources, including documentation, FAQs, and best practices.
- View product availability information.
- Review your support cases.
- Find your local Informatica User Group Network and collaborate with your peers.

Informatica Knowledge Base

Use the Informatica Knowledge Base to search Informatica Network for product resources such as documentation, how-to articles, best practices, and PAMs.

To access the Knowledge Base, visit https://kb.informatica.com. If you have questions, comments, or ideas about the Knowledge Base, contact the Informatica Knowledge Base team at KB_Feedback@informatica.com.

Informatica Documentation

To get the latest documentation for your product, browse the Informatica Knowledge Base at https://kb.informatica.com/_layouts/ProductDocumentation/Page/ProductDocumentSearch.aspx.

If you have questions, comments, or ideas about this documentation, contact the Informatica Documentation team through email at infa_documentation@informatica.com.

Informatica Product Availability Matrixes

Product Availability Matrixes (PAMs) indicate the versions of operating systems, databases, and other types of data sources and targets that a product release supports. If you are an Informatica Network member, you can access PAMs at https://network.informatica.com/community/informatica-network/product-availability-matrices.

Informatica Velocity

Informatica Velocity is a collection of tips and best practices developed by Informatica Professional Services. Developed from the real-world experience of hundreds of data management projects, Informatica Velocity represents the collective knowledge of our consultants who have worked with organizations from around the world to plan, develop, deploy, and maintain successful data management solutions.

If you are an Informatica Network member, you can access Informatica Velocity resources at http://velocity.informatica.com.

If you have questions, comments, or ideas about Informatica Velocity, contact Informatica Professional Services at ips@informatica.com.

Informatica Marketplace

The Informatica Marketplace is a forum where you can find solutions that augment, extend, or enhance your Informatica implementations. By leveraging any of the hundreds of solutions from Informatica developers and partners, you can improve your productivity and speed up time to implementation on your projects. You can access Informatica Marketplace at https://marketplace.informatica.com.

Informatica Global Customer Support

You can contact a Global Support Center by telephone or through Online Support on Informatica Network.

To find your local Informatica Global Customer Support telephone number, visit the Informatica website at the following link:
http://www.informatica.com/us/services-and-training/support-services/global-support-centers.

If you are an Informatica Network member, you can use Online Support at http://network.informatica.com.

CHAPTER 1

Understanding PowerExchange for Hadoop

This chapter includes the following topics:

PowerExchange for Hadoop Overview

PowerExchange for Hadoop integrates PowerCenter with Hadoop to extract and load data.

You can connect a flat file source to Hadoop to extract data from Hadoop Distributed File System (HDFS). You can connect a flat file target to Hadoop to load data to HDFS. You can also load data to the Hive data warehouse system.

Understanding Hadoop

Hadoop provides a framework for distributed processing of large data sets across multiple computers. It depends on applications rather than hardware for high availability.

Hadoop applications use HDFS as the primary storage system. HDFS replicates data blocks and distributes them across nodes in a cluster.

Hive is a data warehouse system for Hadoop. You can use Hive to add structure to datasets stored in file systems that are compatible with Hadoop.

PowerCenter and Hadoop Integration

PowerExchange for Hadoop accesses Hadoop to extract data from HDFS or load data to HDFS or Hive.

To extract data from HDFS, a PowerExchange for Hadoop mapping contains a flat file source. To load data to HDFS or Hive, a PowerExchange for Hadoop mapping contains a flat file target.

In the Workflow Manager, you specify the HDFS flat file reader to extract data from HDFS. You specify the HDFS flat file writer to load data to HDFS or Hive. You select a Hadoop HDFS connection object to access the HDFS database tier of Hadoop.

The Integration Service communicates with Hadoop through the Java Native Interface (JNI). JNI is a programming framework that enables Java code running in a Java Virtual Machine (JVM) to call or be called.

CHAPTER 2

PowerExchange for Hadoop Configuration

This chapter includes the following topics:

PowerExchange for Hadoop Configuration Overview

PowerExchange for Hadoop installs with PowerCenter.

Complete the following tasks to install or upgrade PowerExchange for Hadoop:

1. Install or upgrade PowerCenter.
2. If the Hadoop cluster runs MapR, configure PowerCenter to connect to a Hadoop cluster on MapR.
3. If you are upgrading from PowerCenter, recreate HDFS connections for updated Hadoop distribution versions.

Configure PowerCenter for MapR

You can enable PowerCenter to run mappings on a Hadoop cluster on MapR.

Perform the following tasks:

- Set environment variables for MapR.
- Copy MapR distribution files.
- Configure the PowerCenter Integration Service.

Environment Variables for MapR Distribution

When you use MapR distribution to access Hadoop sources and targets, you must configure environment variables.

For MapR Ticket Cluster

Configure the following MapR environment variables:

- Set environment variable MAPR_HOME to the following path: `<Informatica Installation Directory>/source/services/shared/hadoop/mapr<version>`.
- On the Linux operating system, change environment variable LD_LIBRARY_PATH to include the following path: `<Informatica Installation Directory>/source/services/shared/hadoop/mapr<version>`.
- Set the MapR Container Location Database name variable CLDB in the following file: `<Informatica Installation Directory>/source/services/shared/hadoop/mapr<version>/conf/mapr-clusters.conf`.
- Set environment variable JAVA_OPTS to: `-Dmapr.library.flatclass -Dhadoop.login=maprsasl`
- Set environment variable MAPR_TICKETFILE_LOCATION to the location of the MapR ticket file. Optional.

For MapR Kerberos Cluster

Configure the following MapR environment variables:

- Set environment variable MAPR_HOME to the following path: `<Informatica Installation Directory>/source/services/shared/hadoop/mapr<version>`.
- On the Linux operating system, change environment variable LD_LIBRARY_PATH to include the following path: `<Informatica Installation Directory>/source/services/shared/hadoop/mapr<version>`.
- Set the MapR Container Location Database name variable CLDB in the following file: `<Informatica Installation Directory>/source/services/shared/hadoop/mapr<version>/conf/mapr-clusters.conf`.
- Set environment variable JAVA_OPTS to: `-Dmapr.library.flatclass -Dhadoop.login=hybrid`
- Optional. Set environment variable MAPR_TICKETFILE_LOCATION to the location of the MapR ticket file.
- Optional. Set environment variable KRB5_CONFIG to the following path: `<Informatica Installation Directory_DMAPR>/java/jre/lib/security/krb5.conf`

For MapR non-Secure Cluster

Configure the following MapR environment variables:

- Set environment variable MAPR_HOME to the following path: `<Informatica Installation Directory>/source/services/shared/hadoop/mapr<version>`.
- On the Linux operating system, change environment variable LD_LIBRARY_PATH to include the following path: `<Informatica Installation Directory>/source/services/shared/hadoop/mapr<version>`.
- Set the MapR Container Location Database name variable CLDB in the following file: `<Informatica Installation Directory>/source/services/shared/hadoop/mapr<version>/conf/mapr-clusters.conf`.

Copy MapR Distribution Files for PowerCenter Mappings in the Native Environment

When you use the MapR distribution to run mappings in a native environment, you must copy MapR files to the machine on which the PowerCenter Integration Service runs.

Perform the following steps:

1. Go to the following directory on any node in the cluster: `<MapR installation directory>/conf`
   For example, go to the following directory: `/opt/mapr/conf`.
2. Find the following files:
   - mapr-cluster.conf
   - mapr.login.conf
3. Copy the files to the following directory on the machine on which the PowerCenter Integration Service runs:
   `<Informatica installation directory>/source/services/shared/hadoop/mapr<version>/conf`
4. Log in to the Administrator tool.
5. In the Domain Navigator, select the PowerCenter Integration Service.
6. Recycle the Service.
   Click **Actions** > **Recycle Service**.

Configure the PowerCenter Integration Service

To enable support for MapR, configure the PowerCenter Integration Service.

Perform the following steps:

1. Log in to the Administrator tool.
2. In the Domain Navigator, select the PowerCenter Integration Service.
3. Click the **Processes** view.
4. Add the environment variables as required. For more information about environment variables, see "Environment Variables for MapR Distribution" on page 11.
5. Add the following custom property:

   JVMClassPath

   Use the following value: `<Informatica Installation Directory>/source/services/shared/hadoop/mapr<version>/*:<Informatica Installation Directory>/source/services/shared/hadoop/*`
6. Recycle the service.

   Click **Actions** > **Recycle Service**.

Registering the Plug-in

A plug-in is an XML file that defines the functionality of PowerExchange for Hadoop. To register the plug-in, the repository must be running in exclusive mode. Use Informatica Administrator (the Administrator tool) or the pmrep RegisterPlugin command to register the plug-in.

Register the plug-in if you are upgrading from PowerCenter 10.1 release.

The plug-in file for PowerExchange for Hadoop is pmhdfs.xml. When you install the Repository component, the installer copies pmhdfs.xml to the following directory:

```
<Informatica Installation Directory>/server/bin/native
```

Note: If you do not have the correct privileges to register the plug-in, contact the user who manages the PowerCenter Repository Service.

Update the Repository Plug-in

If you upgraded an existing repository, you must update the repository plug-in to enable PowerExchange for HDFS to run on the Hadoop distribution. If you created a new repository, skip this task.

1. Ensure that the Repository service is running in exclusive mode.
2. On the server machine, open the command console.
3. Run cd <Informatica installation directory>/server/bin
4. Run ./pmrep connect -r <repo_name> -d <domain_name> -n <username> -x <password>
5. Run ./pmrep registerplugin -i native/pmhdfs.xml -e -N true
6. Set the Repository service to normal mode.
7. Open the PowerCenter Workflow manager on the client machine.

   The distribution appears in the Connection Object menu.

HDFS Connections for Hadoop Sources and Targets

Use a HDFS connection object to access a Hadoop source or target.

When you upgrade PowerExchange for Hadoop, you must recreate HDFS connections to access Hadoop source or target. Use the Namenode URI property to recreate the HDFS connections.

Configure PowerCenter for Kerberos-Enabled Hadoop Cluster

You can configure PowerCenter and the PowerCenter Integration Service to read data from and write data to a Hadoop cluster that uses Kerberos authentication for Cloudera, Hortonworks, IBM BigInsights, and MapR distributions.

Perform the following steps:

1. Create a directory where PowerCenter Administrator user has the read access. For example: <infa_home>/pwx-hadoop/conf
2. Copy the following files from Hadoop cluster to directory created in step 1:
   - `/etc/hadoop/conf/core-site.xml`

- /etc/hadoop/conf/mapred-site.xml
- /etc/hadoop/conf/hdfs-site.xml
- /etc/hive/conf/hive-site.xml

3. Ensure that the PowerCenter Administrator user exists on all Hadoop cluster nodes and has the same UID and run kinit to create Kerberos ticket cache file on all nodes.
4. Run the kinit on the Informatica node where PowerCenter Integration Service is running to create the Kerberos ticket cache file. For example: `/tmp/krb5cc_<UID>`
5. Edit the `core-site.xml` in the directory created in step 1 and add the following parameter:

```
<property>
<name>hadoop.security.kerberos.ticket.cache.path</name>
<value>/tmp/REPLACE_WTH_CACHE_FILENAME</value>
<description>Path to the Kerberos ticket cache. </description>
</property>
```

6. In Informatica Administrator, go to Integration Service Processes tab and add the environment variable "CLASSPATH" with the value of the directory created in step 1.
7. Restart the PowerCenter Integration Service.
8. Create the HDFS connection and assign to source or target and run the workflow.

High Availability

You can configure PowerCenter to read from and write to a highly available Hadoop cluster.

A highly available Hadoop cluster can provide uninterrupted access to the NameNode in theHadoop cluster. The NameNode tracks file data across the cluster.

You can configure PowerCenter to communicate with a highly available Hadoop cluster on the following Hadoop distributions:

- Cloudera CDH
- Hortonworks HDP
- IBM BigInsights
- MapR

Configure PowerCenter for a Highly Available Cloudera CDH NameNode

You can configure PowerCenter and the PowerCenter Integration Service to connect to a Cloudera CDH cluster that uses a highly available NameNode.

Perform the following steps:

1. Go to the following directory on any node in the Hadoop cluster: `/etc/hadoop/conf`
2. Locate the following files:
   - core-site.xml

- hdfs-site.xml

3. Copy the files to the following directory on the machine where the PowerCenter Integration Service runs: `<INFA_HOME>/services/shared/hadoop/cdh<version>/conf`
4. In the HDFS connection, use the value for the `fs.default.name` property for the NameNode URI.

   You can find the value for the `fs.default.name` property in `core-site.xml`.

Configure PowerCenter for a Highly Available Hortonworks HDP NameNode

You can configure PowerCenter and the PowerCenter Integration Service to connect to a Hortonworks HDP cluster that uses a highly available NameNode.

Perform the following steps:

1. Go to the following directory on any node in the Hadoop cluster: `/etc/hadoop/conf`
2. Locate the following files:
   - core-site.xml
   - hdfs-site.xml
3. Copy the files to the following directory on the machine where the PowerCenter Integration Service runs: `<INFA_HOME>/services/shared/hadoop/hw<version>/conf`
4. In the HDFS connection, use the value for the `fs.default.name` property for the NameNode URI.

   You can find the value for the `fs.default.name` property in `core-site.xml`.

Configure PowerCenter for a Highly Available IBM BigInsights NameNode

You can configure PowerCenter and the PowerCenter Integration Service to connect to a IBM BigInsights cluster that uses a highly available NameNode.

Perform the following steps:

1. Go to the following directory on any node in the Hadoop cluster: `/data/ibm/biginsights/hadoop-conf`
2. Locate the following files:
   - core-site.xml
   - hdfs-site.xml
3. Copy the files to the following directory on the machine where the PowerCenter Integration Service runs: `<INFA_HOME>/services/shared/hadoop/ibmbi<version>/conf`
4. In the HDFS connection, use the value for the `fs.default.name` property for the NameNode URI.

   You can find the value for the `fs.default.name` property in `core-site.xml`.

Configure PowerCenter for a Highly Available MapR NameNode

You can configure PowerCenter and the PowerCenter Integration Service to connect to a MapR cluster that uses a highly available NameNode.

Perform the following steps:

1. Go to the following directory on any node in the Hadoop cluster: `/opt/mapr/conf`

2. Locate the following files:
   - core-site.xml
   - hdfs-site.xml
3. Copy the files to the following directory on the machine where the PowerCenter Integration Service runs: `<INFA_HOME>/services/shared/hadoop/mapr<version>/conf`
4. In the HDFS connection, use the value for the `fs.default.name` property for the NameNode URI.

   You can find the value for the `fs.default.name` property in `core-site.xml`.

CHAPTER 3

PowerExchange for Hadoop Sources and Targets

This chapter includes the following topic:

PowerExchange for Hadoop Sources and Targets Overview

You include a flat file source definition in a mapping to extract Hadoop data. You include a delimited flat file target definition in a mapping to load data into HDFS or to a Hive table.

You can import a flat file definition or manually create one. To load data to a Hadoop target, the flat file definition must be delimited.

CHAPTER 4

PowerExchange for Hadoop Sessions

This chapter includes the following topics:

PowerExchange for Hadoop Sessions Overview

After you create a PowerExchange for Hadoop mapping in the Designer, you create a PowerExchange for Hadoop session in the Workflow Manager to read, transform, and write Hadoop data.

Before you create a session, configure a Hadoop HDFS application connection to connect to the HDFS host. When the Integration Service extracts or loads Hadoop data, it connects to a Hadoop cluster through the HDFS host that runs the name node service for a Hadoop cluster.

If the mapping contains a flat file source, you can configure the session to extract data from HDFS. If the mapping contains a flat file target, you can configure the session to load data to HDFS or a Hive table.

When the Integration Service loads data to a Hive table, it first loads data to HDFS. The Integration Service then generates an SQL statement to create the Hive table and load the data from HDFS to the table.

PowerExchange for Hadoop Connections

Use a Hadoop HDFS application connection object for each Hadoop source or target that you want to access.

The following table describes the properties that you configure for a Hadoop HDFS application connection:

Property	Description
Name	The connection name used by the Workflow Manager. Connection name cannot contain spaces or other special characters, except for the underscore character.
User Name	The name of the user in the Hadoop group that is used to access the HDFS host.
Password	Password to access the HDFS host. Reserved for future use.
NameNode URI	The URI to access HDFS. Use the following format to specify the NameNode URI for Hadoop distributions: `hdfs://<namenode>:<port>` Where - <namenode> is the host name or IP address of the NameNode. - <port> is the port that the NameNode listens for remote procedure calls (RPC). Use one of the following formats to specify the NameNode URI in MapR distribution: - maprfs:/// - maprfs:///mapr/my.cluster.com/ Where `my.cluster.com` is the cluster name that you specify in the mapr-clusters.conf file.
Hive Driver Name	The name of the Hive driver. By default, the driver name is: `org.apache.hive.jdbc.HiveDriver`

Hive URL	The URL to the Hive host.
	For MapR Ticket cluster, specify the URL in the following format:
	`jdbc:hive2://hostname:portnumber/default;auth=MAPRSASL /default;auth=MAPRSASL`
	For MapR Kerberos cluster, specify the URL in the following format:
	`jdbc:hive2://hostname:portnumber/default;auth=MAPRSASL /default;principal=<spn>`

Hive User Name	The Hive user name. Reserved for future use.
Hive Password	The password for the Hive user. Reserved for future use.
Hadoop Distribution	The name of the Hadoop distribution. Default is Cloudera cdh5.4.

Sessions with Hadoop Sources

You can configure a session to extract data from HDFS.

When you configure a session for a Hadoop source, you select the HDFS Flat File reader file type and a Hadoop HDFS application connection object. You can stage the source data and configure partitioning.

Staging HDFS Source Data

You can optionally stage an HDFS source. The Integration Service stages source files on the local machine and then loads data from the staged file or files into the target.

Stage an HDFS source when you want the Integration Service to read the source files and then close the connection before continuing to process the data.

Configure staging for an HDFS source by setting HDFS flat file reader properties in a session.

You can configure the following types of staging:

Direct

Use direct staging when you want to read data from a source file. The Integration Service reads data from the source file and stages the data on the local machine before passing to downstream transformations.

For example, if you stage a file named `source.csv` from the Hadoop source location to the following directory:

```
c:\staged_files\source_stage.csv
```

The Integration Service stages `source.csv` as `source_stage.csv` in the `c:\staged_files` directory. Then, the Integration Service loads data from the `source_stage.csv` file into the Hadoop target file as specified by the output file path in the HDFS flat file writer properties.

Indirect

Use indirect staging when you want to read data from multiple source files. The Integration Service reads data from multiple files in the source. It creates an indirect file that contains the names of the source files. It then stages the indirect file and the files read from the source in the local staging directory before passing to downstream transformations.

For example, you stage the files named `source1.csv` and `source2.csv` from the Hadoop source location to the following directory:

```
c:\staged_files\source_stage_list.txt
```

The Integration Service creates an indirect file named `source_stage_list.txt` that contains the following entries:

```
source1.csv
source2.csv
```

The Integration Service stages the indirect file and the source files. In the `c:\staged_files` directory, you would see the following files:

```
source_stage_list.txt
source1.csv
source2.csv
```

Then, the Integration Service loads data from the staged source files into the Hadoop target path as specified by the output file path in the HDFS flat file writer properties.

To configure staging, set the following session properties:

Property	Set to
Is Staged	Enabled.
File Path	For direct staging, enter the name and path of the source file. For indirect staging, enter the directory path to the source files.
Staged File Name	Name and path on the local machine used to stage the source data.

Hadoop Source Connections

To extract data from a Hadoop source, you configure the session properties to select the HDFS flat file reader file type and a Hadoop HDFS application connection object.

In the sources node section of the **Mapping** tab in the session properties, select **HDFS Flat File Reader** as the reader type for the source. Then, select a Hadoop HDFS application connection object for the source.

When you select HDFS flat file reader type, you can select a code page of a delimited file from the codepage drop-down list. You cannot set the code page to use a user-defined variable or a session parameter file.

Session Properties for a Hadoop Source

You can configure a session for a Hadoop source to set staging and partitioning properties.

The following table describes the properties that you can configure for a Hadoop source:

Session Property	Description
Is Staged	Before reading the source file, the Integration Service stages the remote file or files locally.
Staged File Name	The local file name directory where the Integration Service stages files. If you use direct staging, the Integration Service stages the source file using this file name. If you use indirect staging, the Integration Service uses the file name to create the indirect file.
Concurrent read partitioning	Order in which multiple partitions read input rows from a source file. You can choose one of the following options: - **Optimize throughput**. The Integration Service does not preserve row order when multiple partitions read from a single file source. Use this option if the order in which multiple partitions read from a file source is not important. - **Keep relative input row order**. The Integration Service preserves the input row order for the rows read by each partition. Use this option to preserve the sort order of the input rows read by each partition. - **Keep absolute input row order**. The Integration Service preserves the input row order for all rows read by all partitions. Use this option to preserve the sort order of the input rows each time the session runs. In a pass-through mapping with passive transformations, the order of the rows written to the target will be in the same order as the input rows.
File Path	The Hadoop directory path of the flat file source. The path can be relative or absolute. If relative, it is relative to the home directory of the Hadoop user. For example, you might specify the following value in the **File Path** property: `\home\foo\`

Sessions with Hadoop Targets

You can configure a session to load data to HDFS or a Hive table.

When you configure a session for a Hadoop target, you select the HDFS Flat File writer file type and a Hadoop HDFS application connection object.

When you configure the session to load data to HDFS, you can configure partitioning, file header, and output options. When you configure the session to load data to a Hive table, you can configure partitioning and output options.

When the Integration Service loads data to a Hive table, it generates a relational table in the Hive database. You can overwite the Hive table data when you run the session again.

Hadoop Target Connections

To load data to a Hadoop target, you configure the session properties to select the HDFS flat file writer file type and a Hadoop HDFS application connection object.

In the targets node section of the **Mapping** tab in the session properties, select **HDFS Flat File Writer** as the writer type for the target. Then, select a Hadoop HDFS application connection object for the target.

When you select HDFS flat file writer type, you can select a code page of a delimited file from the codepage drop-down list. You cannot set the code page to use a user-defined variable or a session parameter file.

Hadoop Target Partitioning

When you configure a session to load data to a Hadoop target, you can write the target output to a separate file for each partition or to a merge file that contains the target output for all partitions.

You can select the following merge types for Hadoop target partitioning:

No Merge

The Integration Service generates one target file for each partition. If you stage the files, the Integration Service transfers the target files to the remote location at the end of the session. If you do not stage the files, the Integration Service generates the target files at the remote location.

Sequential Merge

The Integration Service creates one output file for each partition. At the end of the session, the Integration Service merges the individual output files into a merge file, deletes the individual output files, and transfers the merge file to the remote location.

If you set the merge type to sequential, you need to define the merge file path and the output file path in the session properties. The merge file path determines the final Hadoop target location where the Integration Service creates the merge file. The Integration Service creates the merge file from the intermediate merge file output in the location defined for the output file path.

Session Properties for a Hadoop Target

You can configure a session for a Hadoop target to load data to HDFS or a Hive table. When you load data to a Hadoop target, you can set partitioning properties and file paths. When you load data to HDFS, you can also set header options.

The following table describes the properties that you can configure for a Hadoop target:

Session Property	Description
Merge Type	Type of merge that the Integration Service performs on the data for partitioned targets. You can choose one of the following merge types: - No Merge - Sequential Merge
Append if Exists	Appends data to a file. If the merge file path refers to a directory, this option applies to the files in the directory.
Header Options	Creates a header row in the flat file when loading data to HDFS.
Auto generate partition file names	Generates partition file names.
Merge File Path	If you choose a sequential merge type, defines the final Hadoop target location where the Integration Service creates the merge file. The Integration Service creates the merge file from the intermediate merge file output in the location defined in the output file path.
Generate And Load Hive Table	Generates a relational table in the Hive database. The Integration Service loads data into the Hive table from the HDFS flat file target.
Overwrite Hive Table	Overwrites the data in the Hive table.
Hive Table Name	Hive table name. Default is the target instance name.
Externally Managed Hive Table	Loads Hive table data to the location defined in the output file path.
Output File Path	Defines the absolute or relative directory path or file path on the HDFS host where the Integration Service writes the HDFS data. A relative path is relative to the home directory of the Hadoop user. If you choose a sequential merge type, defines where the Integration Service writes intermediate output before it writes to the final Hadoop target location as defined by the merge file path. If you choose to generate partition file names, this path can be a directory path.
Reject File Path	The path to the reject file. By default, the Integration Service writes all reject files to service process variable directory, $PMBadFileDir.

CHAPTER 5

Data Type Reference

This chapter includes the following topics:

Data Type Reference Overview

PowerCenter uses the following data types for Hadoop data objects:

- Flat file data types. Flat file data types appear in the physical data object column properties.
- Transformation data types. Set of data types that appear in the transformations. They are internal data types based on ANSI SQL-92 generic data types, which the PowerCenter Integration Service uses to move data across platforms. Transformation data types appear in all transformations in a mapping.

When the PowerCenter Integration Service reads source data, it converts the native data types to the comparable transformation data types before transforming the data. When the PowerCenter Integration Service writes to a target, it converts the transformation data types to the comparable native data types.

Flat File and Transformation Data Types

The following table lists the flat file data types that the PowerCenter Integration Service supports and the corresponding transformation data types:

Flat File Data type	Transformation Data type	Range
Bigint	Bigint	Precision of 19 digits, scale of 0
Datetime	Date/Time	Jan 1, 0001 A.D. to Dec 31, 9999 A.D. (precision to the nanosecond)
Double	Double	Precision of 15 digits
Int	Integer	-2,147,483,648 to 2,147,483,647
Nstring	String	1 to 104,857,600 characters

Flat File Data type	Transformation Data type	Range
Number	Decimal	Precision 1 to 28, scale 0 to 28
String	String	1 to 104,857,600 characters

When the PowerCenter Integration Service reads non-numeric data in a numeric column from a flat file, it drops the row and writes a message in the log. Also, when the PowerCenter Integration Service reads non-datetime data in a datetime column from a flat file, it drops the row and writes a message in the log.

INDEX