Corporation ("DataDirect") which are subject to the following terms and conditions:

This Informatica product (the "Software") includes certain drivers (the "DataDirect Drivers") from DataDirect Technologies, an operating company of Progress Software

1. WITHIN EXCLUDED SOFTWARE, THE IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NON-INFRINGEMENT.

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/oss-1.0), 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 the MIT license.


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.

Part Number: IN-BDE-1000-000-0001
# Table of Contents

**Preface** ............................................................... 6

Informatica Resources .................................................. 6
  Informatica My Support Portal ..................................... 6
  Informatica Documentation .......................................... 6
  Informatica Product Availability Matrixes ..................... 6
  Informatica Web Site ................................................. 7
  Informatica How-To Library ........................................ 7
  Informatica Knowledge Base ....................................... 7
  Informatica Support YouTube Channel ............................ 7
  Informatica Marketplace ............................................ 7
  Informatica Velocity ............................................... 7
  Informatica Global Customer Support ............................. 7

**Chapter 1: Introduction to Big Data Management Security** ............... 9

Overview ................................................................. 9
Authentication .......................................................... 10
  Kerberos Authentication ........................................... 10
Authorization ........................................................... 11
  HDFS Permissions ................................................... 11
  User Impersonation ................................................. 11
Data and Metadata Management ....................................... 12
  Cloudera Navigator and Metadata Manager .................... 12
Data Security .......................................................... 12

**Chapter 2: Running Mappings with Kerberos Authentication** .......... 14

Running Mappings with Kerberos Authentication Overview ................ 14
Requirements for the Hadoop Cluster ................................ 15
Prerequisite Tasks for Running Mappings on a Hadoop Cluster with Kerberos Authentication .................. 15
  Install the JCE Policy File for AES-256 Encryption ............ 16
  Verify Permissions on the Hadoop Warehouse .................... 16
  Configure Kerberos Security Properties in hive-site.xml .... 16
  Setting up the Kerberos Configuration File ..................... 18
Running Mappings in the Hadoop Environment when Informatica Does not Use Kerberos Authentication .......... 20
  Step 1. Create Matching Operating System Profile Names .... 21
  Step 2. Create the Principal Names and Keytab File in the AD KDC. 21
  Step 3. Specify the Kerberos Authentication Properties for the Data Integration Service ........ 21
Running Mappings in a Hadoop Environment When Informatica Uses Kerberos Authentication .................. 22
  Step 1. Set up the One-Way Cross-Realm Trust .................. 22
  Step 2. Create Matching Operating System Profile Names .... 24
Chapter 3: User Impersonation with Kerberos Authentication

User Impersonation.

User Impersonation in the Hadoop Environment.

User Impersonation in the Native Environment.

Chapter 4: Blaze Engine Security

Blaze Engine Security Overview.

Setting up a Blaze User Account.

Index.
Preface

The Big Data Management Security Guide is written for Informatica administrators. The guide contains information that you need to manage security for Big Data Management and the connection between Big Data Management and the Hadoop cluster. This book assumes that you are familiar with the Informatica domain, security for the Informatica domain, and security for Hadoop clusters.

Informatica Resources

Informatica My Support Portal


The site contains product information, user group information, newsletters, access to the Informatica customer support case management system, the Informatica How-To Library, the Informatica Knowledge Base, Informatica Product Documentation, and access to the Informatica user community.

The site contains product information, user group information, newsletters, access to the Informatica How-To Library, the Informatica Knowledge Base, Informatica Product Documentation, and access to the Informatica user community.

Informatica Documentation

The Informatica Documentation team makes every effort to create accurate, usable documentation. If you have questions, comments, or ideas about this documentation, contact the Informatica Documentation team through email at infa_documentation@informatica.com. We will use your feedback to improve our documentation. Let us know if we can contact you regarding your comments.

The Documentation team updates documentation as needed. To get the latest documentation for your product, navigate to Product Documentation from http://mysupport.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. You can access the PAMs on the Informatica My Support Portal at https://mysupport.informatica.com/community/my-support/product-availability-matrices.
Informatica Web Site

You can access the Informatica corporate web site at http://www.informatica.com. The site contains information about Informatica, its background, upcoming events, and sales offices. You will also find product and partner information. The services area of the site includes important information about technical support, training and education, and implementation services.

Informatica How-To Library

As an Informatica customer, you can access the Informatica How-To Library at http://mysupport.informatica.com. The How-To Library is a collection of resources to help you learn more about Informatica products and features. It includes articles and interactive demonstrations that provide solutions to common problems, compare features and behaviors, and guide you through performing specific real-world tasks.

Informatica Knowledge Base

As an Informatica customer, you can access the Informatica Knowledge Base at http://mysupport.informatica.com. Use the Knowledge Base to search for documented solutions to known technical issues about Informatica products. You can also find answers to frequently asked questions, technical white papers, and technical tips. If you have questions, comments, or ideas about the Knowledge Base, contact the Informatica Knowledge Base team through email at KB.Feedback@informatica.com.

Informatica Support YouTube Channel

You can access the Informatica Support YouTube channel at http://www.youtube.com/user/INFASupport. The Informatica Support YouTube channel includes videos about solutions that guide you through performing specific tasks. If you have questions, comments, or ideas about the Informatica Support YouTube channel, contact the Support YouTube team through email at supportvideos@informatica.com or send a tweet to @INFASupport.

Informatica Marketplace

The Informatica Marketplace is a forum where developers and partners can share solutions that augment, extend, or enhance data integration implementations. By leveraging any of the hundreds of solutions available on the Marketplace, you can improve your productivity and speed up time to implementation on your projects. You can access Informatica Marketplace at http://www.informaticamarketplace.com.

Informatica Velocity

You can access Informatica Velocity at http://mysupport.informatica.com. 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 have questions, comments, or ideas about Informatica Velocity, contact Informatica Professional Services at ips@informatica.com.

Informatica Global Customer Support

You can contact a Customer Support Center by telephone or through the Online Support.

Online Support requires a user name and password. You can request a user name and password at http://mysupport.informatica.com.
Introduction to Big Data Management Security

This chapter includes the following topics:

- Overview, 9
- Authentication, 10
- Authorization, 11
- Data and Metadata Management, 12
- Data Security, 12

Overview

You can configure security for Big Data Management and the Hadoop cluster to protect from threats inside and outside the network. Security for Big Data Management includes security for the Informatica domain and security for the Hadoop cluster.

Security for the Hadoop cluster includes the following areas:

Authentication

By default, Hadoop does not verify the identity of users. To use authentication, configure Kerberos for the cluster.

Big Data Management supports Hadoop clusters that use a Microsoft Active Directory (AD) Key Distribution Center (KDC) or an MIT KDC.

Authorization

After a user is authenticated, a user must be authorized to perform actions. Hadoop uses HDFS permissions to determine what a user can do to a file or directory on HDFS. For example, a user must have the correct permissions to access the directories where specific data is stored to use that data in a mapping.

Data and metadata management

Data and metadata management involves managing data to track and audit data access, update metadata, and perform data lineage. Big Data Management supports Cloudera Navigator and Metadata Manager to manage metadata and perform data lineage.
Data security

Data security involves protecting sensitive data from unauthorized access. Big Data Management supports data masking with the Data Masking transformation in the Developer tool, Dynamic Data Masking, and Persistent Data Masking.

Security for the Informatica domain is separate from security for the Hadoop cluster. For a higher level of security, secure the Informatica domain and the Hadoop cluster. For more information about security for the Informatica domain, see the Informatica Security Guide.

Authentication

When the Informatica domain includes Big Data Management, users must be authenticated in the Informatica domain and the Hadoop cluster. Authentication for the Informatica domain is separate from authentication for the Hadoop cluster.

The authentication process verifies the identity of a user account.

The Informatica domain uses one of the following authentication protocols:

Native authentication

The Informatica domain stores user credentials and privileges in the domain configuration repository and performs all user authentication within the Informatica domain.

Lightweight Directory Access Protocol (LDAP)

The LDAP directory service stores user accounts and credentials that are accessed over the network.

Kerberos authentication

Kerberos is a network authentication protocol which uses tickets to authenticate users and services in a network. Users are stored in the Kerberos principal database, and tickets are issued by a KDC.

By default, Hadoop does not authenticate users. Any user can be used in the Hadoop connection. Informatica recommends that you enable authentication for the cluster. If authentication is enabled for the cluster, the cluster authenticates the user account used for the Hadoop connection between Big Data Management and the cluster. For a higher level of security, you can set up Kerberos authentication for the cluster.

For more information about how to configure authentication for the Informatica domain, see the Informatica Security Guide.

For more information about how to enable authentication for the Hadoop cluster, see the documentation for your Hadoop distribution.

Kerberos Authentication

Big Data Management and the Hadoop cluster can use Kerberos authentication to verify user accounts. You can use Kerberos authentication with the Informatica domain, with the Hadoop cluster, or with both.

Kerberos is a network authentication protocol which uses tickets to authenticate access to services and nodes in a network. Kerberos uses a Key Distribution Center (KDC) to validate the identities of users and services and to grant tickets to authenticated user and service accounts. Users and services are known as principals. The KDC has a database of principals and their associated secret keys that are used as proof of identity. Kerberos can use an LDAP directory service as a principal database.

The requirements for Kerberos authentication for the Informatica domain and for the Hadoop cluster:
Kerberos authentication for the Informatica domain

Kerberos authentication for the Informatica domain requires principals stored in a Microsoft Active Directory (AD) LDAP service. Additionally, you must use Microsoft AD for the KDC.

For more information about how to enable Kerberos authentication for the Informatica domain, see the Informatica Security Guide.

Kerberos authentication for the Hadoop cluster

Informatica supports Hadoop clusters that use an AD KDC or an MIT KDC.

When you enable Kerberos for Hadoop, each user and Hadoop service needs to be authenticated by KDC. The cluster must authenticate the Data Integration Service User and, optionally, the Blaze user.

For more information about how to configure Kerberos for Hadoop, see the documentation for your Hadoop distribution.

The configuration steps required for Big Data Management to connect to a Hadoop cluster that uses Kerberos authentication depends on whether the Informatica domain uses Kerberos.

For more information about how to configure Big Data Management to connect to a Hadoop cluster that uses Kerberos, see the "Running Mappings with Kerberos Authentication" chapter.

Authorization

Authorization controls what a user can do on a Hadoop cluster. For example, a user must be authorized to submit jobs to the Hadoop cluster.

Authorization for Big Data Management consists of HDFS permissions and user impersonation.

HDFS Permissions

HDFS permissions determine what a user can do to files and directories stored in HDFS. To access a file or directory, a user must have permission or belong to a group that has permission.

HDFS permissions are similar to permissions for UNIX or Linux systems. For example, a user requires the r permission to read a file and the w permission to write a file. When a user or application attempts to perform an action, HDFS checks if the user has permission or belongs to a group with permission to perform that action on a specific file or directory.

For more information about HDFS permissions, see the Apache Hadoop documentation or the documentation for your Hadoop distribution.

Big Data Management supports HDFS permissions without additional configuration.

User Impersonation

User impersonation allows different users to run mappings in a Hadoop cluster that uses Kerberos authentication or connect to big data sources and targets that use Kerberos authentication.

The Data Integration Service uses its credentials to impersonate the user accounts designated in the Hadoop connection to connect to the Hadoop cluster or to start the Blaze engine.

When the Data Integration Service impersonates a user account to submit a mapping, the mapping can only access Hadoop resources that the impersonated user has permissions on. Without user impersonation, the
Data Integration Service uses its credentials to submit a mapping to the Hadoop cluster. Restricted Hadoop resources might be accessible.

When the Data Integration service impersonates a user account to start the Blaze engine, the Blaze engine has the privileges and permissions of the user account used to start it.

Data and Metadata Management

Track data access for entities and manage metadata about entities to perform security audits.

Cloudera Navigator and Metadata Manager

Use Cloudera Navigator and Informatica Metadata Manager together to manage your data and metadata and audit access.

Cloudera Navigator is a data management tool for the Hadoop platform. It contains an auditing component that allows users to track data access for entities in a Hadoop cluster and a metadata component that manages metadata about the entities in a Hadoop cluster.

Metadata Manager is a web-based metadata management tool that you can use to browse and analyze metadata from metadata repositories. Metadata manager helps you understand and manage how information is derived. Metadata Manager extracts metadata about entities in a Hadoop cluster through the metadata component of Cloudera Navigator.

Use Cloudera Navigator and Metadata Manager together to perform data lineage. When you perform data lineage, you can see where data comes, where it is used, and what transformations are applied. For example, you can review who accessed sensitive data as part of a security audit to identify potential data policy violations.

For more information about Metadata Manager and Cloudera Navigator, see the Informatica Big Data Management User Guide Informatica Metadata Manager Administrator Guide.

Data Security

Data security protects sensitive data on the Hadoop cluster from unauthorized access. Big Data Management supports different methods of data masking to secure data. Data masking obscures data based on configurable rules.

For example, an analyst in the marketing department might need to use production data to conduct analysis, but a mapping developer can test a mapping with masked data. You can set data masking rules to allow the analyst to access production data and rules to allow the mapping developer to access test data that is realistic. Alternatively, an analyst may only need access to some production data and the rest of the data can be masked. You can configure data masking rules that fit your data security requirements.

You can use the following Informatica components and products to secure data on the Hadoop cluster:

Data Masking transformation

The Data Masking transformation changes sensitive production data to realistic test data for non-production environments. The Data Masking transformation modifies source data based on masking techniques that you configure for each column.
For more information about how to use the Data Masking transformation in the Hadoop environment, see the Informatica Big Data Management User Guide and the Informatica Developer Transformation Guide.

Dynamic Data Masking

When a mapping uses data from a Hadoop source, Dynamic Data Masking acts as a proxy that intercepts requests and data between the Data Integration Service and the cluster. Based on the data masking rules, Dynamic Data Masking might return the original values, masked values, or scrambled values for a mapping to use. The actual data in the cluster is not changed.

For more information about Dynamic Data Masking, see the Informatica Dynamic Data Masking Administrator Guide.

Persistent Data Masking

Persistent Data Masking allows you to mask sensitive and confidential data in non-production systems such as development, test, and training systems.

You can perform data masking on data that is stored in a Hadoop cluster. Additionally, you can mask data during data ingestion in the native or Hadoop environment. Masking rules can replace, scramble, or initialize data. When you create a project, you select masking rules for each table field that you want to mask. When you run the project, the Persistent Data Masking uses the masking rule technique to change the data in the Hadoop cluster. The result is realistic data that you can use for development or testing purposes that is unidentifiable.

For more information about Persistent Data Masking, see the Informatica Test Data Management User Guide and the Informatica Test Data Management Administrator Guide.
CHAPTER 2

Running Mappings with Kerberos Authentication

This chapter includes the following topics:

- Running Mappings with Kerberos Authentication Overview, 14
- Requirements for the Hadoop Cluster, 15
- Prerequisite Tasks for Running Mappings on a Hadoop Cluster with Kerberos Authentication, 15
- Running Mappings in the Hadoop Environment when Informatica Does not Use Kerberos Authentication, 20
- Running Mappings in a Hadoop Environment When Informatica Uses Kerberos Authentication, 22
- Running Mappings in the Native Environment when Informatica Uses Kerberos Authentication, 25
- Running Mappings in the Native Environment When Informatica Does not Use Kerberos Authentication, 26
- Kerberos Authentication for a Hive Connection, 26
- Kerberos Authentication for an HBase Connection, 27
- Kerberos Authentication for an HDFS Connection, 28
- Metadata Import in the Developer Tool, 28
- Create and Configure the Analyst Service, 29

Running Mappings with Kerberos Authentication Overview

You can run mappings on a Hadoop cluster that uses MIT or Microsoft Active Directory (AD) Kerberos authentication. Kerberos is a network authentication protocol that uses tickets to authenticate access to services and nodes in a network.

To run mappings on a Hadoop cluster that uses Kerberos authentication, you must configure the Informatica domain to enable mappings to run in the Hadoop cluster.

If the Informatica domain does not use Kerberos authentication, you must create a Service Principal Name (SPN) for the Data Integration Service in AD.

If the Informatica domain uses Kerberos authentication, you must configure a one-way cross-realm trust to enable the Hadoop cluster to communicate with the Informatica domain. The Informatica domain uses
Kerberos authentication on an AD service. The Hadoop cluster uses Kerberos authentication on an MIT service. The one way cross-realm trust enables the MIT service to communicate with the AD service.

Based on whether the Informatica domain uses Kerberos authentication or not, you might need to perform the following tasks to run mappings on a Hadoop cluster that uses Kerberos authentication:

- If you run mappings in a Hadoop environment, you can choose to configure user impersonation to enable other users to run mappings on the Hadoop cluster. Else, the Data Integration Service user can run mappings on the Hadoop cluster.
- If you run mappings in the native environment, you must configure the mappings to read and process data from Hive sources that use Kerberos authentication.
- If you run a mapping that has Hive sources or targets, you must enable user authentication for the mapping on the Hadoop cluster.
- If you import metadata from Hive, complex file sources, and HBase sources, you must configure the Developer tool to use Kerberos credentials to access the Hive, complex file, and HBase metadata.

Requirements for the Hadoop Cluster

To run mappings on a Hadoop cluster that uses Kerberos authentication, the Hadoop cluster must meet certain requirements.

The Hadoop cluster must have the following configuration:

- Kerberos network authentication
- HiveServer 2

Prerequisite Tasks for Running Mappings on a Hadoop Cluster with Kerberos Authentication

Before you can run mappings on a Hadoop cluster that uses Kerberos authentication, you must complete the prerequisite tasks.

To run mappings with Kerberos authentication, complete the following tasks:

- Install the JCE Policy File.
- Verify permissions on the Hadoop warehouse.
- Configure Kerberos security properties on hive-site.xml.
- Set up the Kerberos configuration file.
Install the JCE Policy File for AES-256 Encryption

If the Informatica domain uses an operating system that uses AES-256 encryption by default for tickets, you must install the JCE Policy File.

Complete the following steps to install the JCE Policy File:

1. Download the Java Cryptography Extension (JCE) Unlimited Strength Jurisdiction Policy .zip File from the Oracle Technology Network website.
2. Extract the contents of the JCE Policy File to the following location: <Informatica Big Data Server Installation Directory>/java/jre/lib/security

For JCE Policy File installation instructions, see the README.txt file included in the JCE Policy .zip file.

Verify Permissions on the Hadoop Warehouse

On the primary Name Node, run the command to verify that you have read and write permissions on the warehouse directory. Run the command at the directory that is one level higher than the directory that you want to verify permissions on.

For example, to verify permissions on /user/hive/warehouse, run the command on /user/hive.

Run the following command:

```bash
hadoop fs -ls /user/hive
```

If you do not have permissions on the Hadoop warehouse, contact an administrator to get permissions on the Hadoop warehouse.

Configure Kerberos Security Properties in hive-site.xml

Configure Kerberos security properties in the hive-site.xml file that the Data Integration Service uses when it runs mappings in a Hadoop cluster.

hive-site.xml is located in the following directory on the machine where the Data Integration runs:

```
<Informatica Installation Directory>/services/shared/hadoop/<Hadoop distribution version>/conf/
```

In hive-site.xml, configure the following Kerberos security properties:

dfs.namenode.kerberos.principal

The SPN for the Name Node.

Set this property to the value of the same property in the following file:

```
/$$Hadoop_Home$$/conf/hdfs-site.xml
```

mapreduce.jobtracker.kerberos.principal

The SPN for the JobTracker or Yarn resource manager.

Set this property to the value of the same property in the following file:

```
/$$Hadoop_Home$$/conf/mapred-site.xml
```

mapreduce.jobhistory.principal

The SPN for the MapReduce JobHistory server.

Set this property to the value of the same property in the following file:

```
/$$Hadoop_Home$$/conf/mapred-site.xml
```
hadoop.security.authentication

Authentication type. Valid values are simple, and kerberos. Simple uses no authentication.

Set this property to the value of the same property in the following file:

/$$Hadoop_Home$$/conf/core-site.xml

hadoop.security.authorization

Enable authorization for different protocols. Set the value to true.

hive.server2.enable.doAs

The authentication that the server is set to use. Set the value to true. When the value is set to true, the
user who makes calls to the server can also perform Hive operations on the server.

hive.metastore.sasl.enabled

If true, the Metastore Thrift interface is secured with Simple Authentication and Security Layer (SASL)
and clients must authenticate with Kerberos. Set the value to true.

hive.metastore.kerberos.principal

The SPN for the metastore thrift server. Replaces the string _HOST with the correct host name.

Set this property to the value of the same property in the following file:

/$$Hadoop_Home$$/conf/hive-site.xml

yarn.resourcemanager.principal

The SPN for the Yarn resource manager.

Set this property to the value of the same property in the following file:

/$$Hadoop_Home$$/conf/yarn-site.xml

The following sample code shows the values for the security properties in hive-site.xml:

<property>
    <name>dfs.namenode.kerberos.principal</name>
    <value>hdfs/_HOST@YOUR-REALM</value>
    <description>The SPN for the NameNode.</description>
</property>

<property>
    <name>mapreduce.jobtracker.kerberos.principal</name>
    <value>mapred/_HOST@YOUR-REALM</value>
    <description>The SPN for the JobTracker or Yarn resource manager.</description>
</property>

<property>
    <name>mapreduce.jobhistory.principal</name>
    <value>mapred/_HOST@YOUR-REALM</value>
    <description>The SPN for the MapReduce JobHistory server.</description>
</property>

<property>
    <name>hadoop.security.authentication</name>
    <value>kerberos</value> <!-- A value of "simple" would disable security. -->
    <description>Authentication type.</description>
</property>

<property>
    <name>hadoop.security.authorization</name>
If the Hadoop cluster uses IBM BigInsights, you must also configure the following property in the `map.reduce.jobtracker.kerberos.principal` property:

The following sample code shows the property you can configure in `hive-site.xml`:

```xml
<property>
  <name>mapreduce.jobtracker.kerberos.principal</name>
  <value>mapred/_HOST@YOUR-REALM</value>
  <description>The Kerberos principal name of the JobTracker.</description>
</property>
```

**IBM BigInsights**

If the Hadoop cluster uses IBM BigInsights, you must also configure the following property in the `hive-site.xml` file:

```
map.reduce.jobtracker.kerberos.principal
```

The Kerberos principal name of the JobTracker.

Use the following syntax for the value: `mapred/_HOST@YOUR-REALM`.

**Setting up the Kerberos Configuration File**

Set the configuration properties for the Kerberos realm that the Hadoop cluster uses to `krb5.conf` on the machine on which the Data Integration Service runs. If the Informatica domain does not use Kerberos authentication, set the properties for the MIT realm. If the Informatica domain uses Kerberos authentication, set the properties for the Active Directory realm and MIT realm.

`krb5.conf` is located in the `<Informatica Installation Directory>/java/jre/lib/security` directory.

Optionally, if the machine on which the Data Integration Service runs does not have `krb5.conf`, copy the file from the Hadoop cluster and add it to the machine on which the Data Integration Service runs.
Kerberos authentication:
The following example shows the parameters for the Hadoop domain if the Informatica domain does not use Kerberos authentication:

```
[libdefaults]
default_realm = HADOOP-AD-REALM
```

The following example shows the value if the Informatica domain does not use Kerberos authentication:

```
[krb5.conf]
default_realm = HADOOP-AD-REALM
```

The following example shows the parameters for the Hadoop realm if the Informatica domain uses Kerberos authentication:

```
[realms]
HADOOP-AD-REALM = {
    kdc = 123abcd134.hadoop-AD-realm.com
    admin_server = 123abcd124.hadoop-AD-realm.com
}
```

The following example shows the parameters for the Hadoop realm if the Informatica domain does not use Kerberos authentication:

```
[realms]
INFA-AD-REALM = {
    kdc = abc123.infa-ad-realm.com
    admin_server = abc123.infa-ad-realm.com
}
HADOOP-MIT-REALM = {
    kdc = def456.hadoop-mit-realm.com
    admin_server = def456.hadoop-mit-realm.com
}
```

In the `domain_realms` section, map the domain name or host name to a Kerberos realm name. The domain name is prefixed by a period (\(.)\). The following example shows the parameters for the Hadoop domain_realm if the Informatica domain does not use Kerberos authentication:

```
[domain realm]
.hadoop_ad_realm.com = HADOOP-AD-REALM
hadoop_ad_realm.com = HADOOP-AD-REALM
```

1. Back up `krb5.conf` before you make any changes.
2. Edit `krb5.conf`.
3. In the `libdefaults` section, set the `default_realm` property required by Informatica.
   
   The `default_realm` is the name of the service realm for the Informatica domain. If the Informatica domain does not use Kerberos authentication, then the default realm must be the name of the AD service.
   
   The following example shows the value if the Informatica domain does not use Kerberos authentication:
   
   ```
   [libdefaults]
default_realm = HADOOP-AD-REALM
   ```
   
   The following example shows the parameters for the Hadoop domain_realm if the Informatica domain does not use Kerberos authentication:
   
   ```
   [krb5.conf]
default_realm = HADOOP-AD-REALM
   ```
   
4. In the `realms` section, set or add the properties required by Informatica.
   
   The following table lists the values to which you must set properties in the `realms` section:
   
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>kdc</td>
<td>Name of the host running a KDC server for that realm.</td>
</tr>
<tr>
<td>admin_server</td>
<td>Name of the Kerberos administration server.</td>
</tr>
</tbody>
</table>

Copy `krb5.conf` to the `<Informatica Installation Directory>/java/jre/lib/security directory`.

**Note:** If you copy `krb5.conf` from the Hadoop cluster, you do not need to edit it.

The `admin_server` is the name of the Kerberos administration server.

The `kdc` is the Name of the Kerberos administration server.

The `domain_realms` section, map the domain name or host name to a Kerberos realm name. The domain name is prefixed by a period (\( .\) ).

The `admin_server` is the Name of the Kerberos administration server.

The `kdc` is the Name of the Kerberos administration server.
The following example shows the parameters for the Hadoop domain_realm if the Informatica domain uses Kerberos authentication:

```
[domain_realm]
.infa_ad_realm.com = INFA-AD-REALM
 .hadoop_ad_realm.com = HADOOP-AD-REALM
.hadoop_mit_realm.com = HADOOP-MIT-REALM

[realms]
INFA-AD-REALM = {
    kdc = abc123.infa-ad-realm.com
    admin_server = abc123.infa-ad-realm.com
}
HADOOP-MIT-REALM = {
    kdc = def456.hadoop-mit-realm.com
    admin_server = def456.hadoop-mit-realm.com
}
```

The following example shows the content of `krb5.conf` with the required properties for an Informatica domain that uses Kerberos authentication:

```
[libdefaults]
default_realm = HADOOP-AD-REALM

[realms]
HADOOP-AD-REALM = {
    kdc = 123abcd134.hadoop-ad-realm.com
    admin_server = 123abcd124.hadoop-ad-realm.com
}
```

The following example shows the content of `krb5.conf` with the required properties for an Informatica domain that does not use Kerberos authentications:

```
[libdefaults]
default_realm = INFA-AD-REALM

[realms]
INFA-AD-REALM = {
    kdc = abc123.infa-ad-realm.com
    admin_server = abc123.infa-ad-realm.com
}
```

Running Mappings in the Hadoop Environment when Informatica Does not Use Kerberos Authentication

To run mappings in a Hadoop environment that uses Kerberos authentication when the Informatica domain does not use Kerberos authentication, you must enable mappings to run in a Hadoop environment that uses Kerberos authentication. The Hadoop cluster must use Microsoft Active Directory as the KDC.

For example, HypoStores Corporation needs to run jobs that process greater than 10 terabytes of data on a Hadoop cluster that uses Kerberos authentication. HypoStores has an Informatica domain that does not use Kerberos authentication. The HypoStores administrator must enable the Informatica domain to communicate with the Hadoop cluster. Then, the administrator must enable mappings to run on the Hadoop cluster.

The HypoStores administrator must perform the following configuration tasks:

1. Create matching operating system profile user names on each Hadoop cluster node.
2. Create the principal name for the Data Integration Service in the KDC and keytab file.
3. Specify the Kerberos authentication properties for the Data Integration Service.

Step 1. Create Matching Operating System Profile Names

Create matching operating system profile user names on the machine that runs the Data Integration Service and each Hadoop cluster node to run Informatica mapping jobs.

For example, if user joe runs the Data Integration Service on a machine, you must create the user joe with the same operating system profile on each machine on which a Hadoop cluster node runs.

Open a UNIX shell and enter the following UNIX command to create a user with the user name joe.

```
useradd joe
```

Step 2. Create the Principal Names and Keytab File in the AD KDC

Create an SPN in the KDC database for Microsoft Active Directory service that matches the user name of the user that runs the Data Integration Service. Create a keytab file for the SPN on the machine where the KDC runs. Then, copy the keytab file to the machine where the Data Integration Service runs.

To create an SPN and Keytab file in the Active Directory server, complete the following steps:

Create a user in the Microsoft Active Directory Service.

Login to the machine on which the Microsoft Active Directory Service runs and create a user with the same name as the user you created in “Step 1. Create Matching Operating System Profile Names” on page 21.

Create an SPN associated with the user.

Use the following guidelines when you create the SPN and keytab files:

- The user principal name (UPN) must be the same as the SPN.
- Enable delegation in Microsoft Active Directory.
- Use the ktpass utility to create an SPN associated with the user and generate the keytab file.

For example, enter the following command:

```
ktpass -out infa.hadoop.keytab -mapuser joe -pass tempBG82008 -princ joe/ domain12345INFA-AD-REALM -crypto all
```

The `-out` parameter specifies the name and path of the keytab file. The `-mapuser` parameter is the user to which the SPN is associated. The `-pass` parameter is the password for the SPN in the generated keytab. The `-princ` parameter is the SPN.

Step 3. Specify the Kerberos Authentication Properties for the Data Integration Service

In the Data Integration Service properties, configure the properties that enable the Data Integration Service to connect to a Hadoop cluster that uses Kerberos authentication. Use the Administrator tool to set the Data Integration Service properties.

Configure the following properties:

**Hadoop Kerberos Service Principal Name**

Service Principal Name (SPN) of the Data Integration Service to connect to a Hadoop cluster that uses Kerberos authentication.
Running Mappings in a Hadoop Environment When Informatica Uses Kerberos Authentication

To run mappings in a Hadoop environment that uses Kerberos authentication when the Informatica domain also uses Kerberos authentication, you must configure a one-way cross-realm trust between the Informatica domain and the Hadoop cluster. The one-way cross-realm trust enables the Informatica domain to communicate with the Hadoop cluster.

The Informatica domain uses Kerberos authentication on a Microsoft Active Directory service. The Hadoop cluster uses Kerberos authentication on an MIT Kerberos service. You set up a one-way cross-realm trust so that the KDC for the MIT Kerberos service can communicate with the KDC for the Active Directory service. After you set up the cross-realm trust, you must configure the Informatica domain to enable mappings to run in the Hadoop cluster.

For example, HypoStores Corporation needs to run jobs that process greater than 10 terabytes of data on a Hadoop cluster that uses Kerberos authentication. HypoStores has an Informatica domain that uses Kerberos authentication. The HypoStores administrator must set up a one way cross-realm trust to enable the Informatica domain to communicate with the Hadoop cluster. Then, the administrator must configure the Informatica domain to run mappings on the Hadoop cluster.

The HypoStores administrator must perform the following configuration tasks:

1. Set up the one-way cross-realm trust.
2. Create matching operating system profile user names on each Hadoop cluster node.
3. Create the Service Principal Name and Keytab File in the Active Directory Server.
4. Specify the Kerberos authentication properties for the Data Integration Service.

Step 1. Set up the One-Way Cross-Realm Trust

Set up a one-way cross-realm trust to enable the KDC for the MIT Kerberos server to communicate with the KDC for the Active Directory server.

When you set up the one-way cross-realm trust, the Hadoop cluster can authenticate the Active Directory principals.
To set up the cross-realm trust, you must complete the following steps:

1. Configure the Active Directory server to add the local MIT realm trust.
2. Configure the MIT server to add the cross-realm principal.
3. Translate principal names from the Active Directory realm to the MIT realm.

**Configure the Microsoft Active Directory Server**

Add the MIT KDC host name and local realm trust to the Active Directory server. To configure the Active Directory server, complete the following steps:

1. Enter the following command to add the MIT KDC host name:
   ```bash
   ksetup /addkdc <mit_real_name> <kdc_hostname>
   ```
   For example, enter the command to add the following values:
   ```bash
   ksetup /addkdc HADOOP-MIT-REALM def456.hadoop-mit-realm.com
   ```
2. Enter the following command to add the local realm trust to Active Directory:
   ```bash
   netdom trust <mit_real_name> /Domain:<ad_real_name> /add /realm /password:<TrustPassword>
   ```
   For example, enter the command to add the following values:
   ```bash
   ```
3. Enter the following commands based on your Microsoft Windows environment to set the proper encryption type:
   - For Microsoft Windows 2008, enter the following command:
     ```bash
     ksetup /SetEncTypeAttr <mit_real_name> <enc_type>
     ```
   - For Microsoft Windows 2003, enter the following command:
     ```bash
     ktpass /MITRealmName <mit_real_name> /TrustEncryp <enc_type>
     ```
   **Note:** The `enc_type` parameter specifies AES, DES, or RC4 encryption. To find the value for `enc_type`, see the documentation for your version of Windows Active Directory. The encryption type you specify must be supported on both versions of Windows that use Active Directory and the MIT server.

**Configure the MIT Server**

Configure the MIT server to add the cross-realm krbtgt principal. The krbtgt principal is the principal name that a Kerberos KDC uses for a Windows domain.

Enter the following command in the kadmin.local or kadmin shell to add the cross-realm krbtgt principal:

```bash
kadmin: addprinc -e "<enc_type_list>" krbtgt/<mit_real_name>@<MY-AD-REALM.COM>
```

The `enc_type_list` parameter specifies the types of encryption that this cross-realm krbtgt principal will support. The krbtgt principal can support either AES, DES, or RC4 encryption. You can specify multiple encryption types. However, at least one of the encryption types must correspond to the encryption type found in the tickets granted by the KDC in the remote realm.

For example, enter the following value:

```bash
kadmin: addprinc -e "rc4-hmac: normal des3-hmac-shal: normal" krbtgt/HADOOP-MIT-REALM@INFA-AD-REALM
```
Translate Principal Names from the Active Directory Realm to the MIT Realm

To translate the principal names from the Active Directory realm into local names within the Hadoop cluster, you must configure the hadoop.security.auth_to_local property in the core-site.xml file on all the machines in the Hadoop cluster.

For example, set the following property in core-site.xml on all the machines in the Hadoop cluster:

```xml
<property>
  <name>hadoop.security.auth_to_local</name>
  <value>
    RULE:[1:$1@$0](^.*@INFA-AD-REALM$)/s/(.*)@INFA-AD-REALM$/\1/g
    RULE:[2:$1@$0](^.*@INFA-AD-REALM$)/s/(.*)@INFA-AD-REALM$/\1/g
    DEFAULT
  </value>
</property>
```

Step 2. Create Matching Operating System Profile Names

Create matching operating system profile user names on the machine that runs the Data Integration Service and each Hadoop cluster node to run Informatica mapping jobs.

For example, if user joe runs the Data Integration Service on a machine, you must create the user joe with the same operating system profile on each machine on which a Hadoop cluster node runs.

Open a UNIX shell and enter the following UNIX command to create a user with the user name joe.

```shell
useradd joe
```

Step 3. Create an SPN and Keytab File in the Active Directory Server

Create an SPN in the KDC database for Microsoft Active Directory service that matches the user name of the user that runs the Data Integration Service. Create a keytab file for the SPN on the machine on which the KDC server runs. Then, copy the keytab file to the machine on which the Data Integration Service runs.

You do not need to use the Informatica Kerberos SPN Format Generator to generate a list of SPNs and keytab file names. You can create your own SPN and keytab file name.

To create an SPN and Keytab file in the Active Directory server, complete the following steps:

Create a user in the Microsoft Active Directory Service.

Login to the machine on which the Microsoft Active Directory Service runs and create a user with the same name as the user you created in "Step 2. Create Matching Operating System Profile Names" on page 24.

Create an SPN associated with the user.

Use the following guidelines when you create the SPN and keytab files:

- The user principal name (UPN) must be the same as the SPN.
- Enable delegation in Microsoft Active Directory.
- Use the ktpass utility to create an SPN associated with the user and generate the keytab file.

For example, enter the following command:

```shell
ktpass -out infa_hadoop.keytab -mapuser joe -pass tempBG2008 -princ joe/dcmain12345@INFA-AD-REALM -crypto all
```

**Note:** The -out parameter specifies the name and path of the keytab file. The -mapuser parameter is the user to which the SPN is associated. The -pass parameter is the password for the SPN in the generated keytab. The -princ parameter is the SPN.
Step 4. Specify the Kerberos Authentication Properties for the Data Integration Service

In the Data Integration Service properties, configure the properties that enable the Data Integration Service to connect to a Hadoop cluster that uses Kerberos authentication. Use the Administrator tool to set the Data Integration Service properties.

Configure the following properties:

**Hadoop Kerberos Service Principal Name**

Service Principal Name (SPN) of the Data Integration Service to connect to a Hadoop cluster that uses Kerberos authentication.

Enter the property in the following format:

```
<princ_name>/<DIS domain name>@<realm name>
```

For example, enter the following value:

```
joe/domain12345@INFA-AD-REALM
```

**Hadoop Kerberos Keytab**

Path and file name of the Kerberos keytab file on the machine on which the Data Integration Service runs.

Enter the following property:

```
<keytab_path>
```

For example, enter the following value for the property:

```
<Informatica Installation Directory>/isp/config/keys/infa_hadoop.keytab
```

Running Mappings in the Native Environment when Informatica Uses Kerberos Authentication

To read and process data from Hive, HBase, or HDFS sources that use Kerberos authentication, you must configure Kerberos authentication for mappings in the native environment.

To read and process data from Hive, HBase, or HDFS sources, perform the following steps:

1. Complete the prerequisite tasks for running mappings on a Hadoop cluster with Kerberos authentication.
2. Complete the tasks for running mappings in the Hadoop environment when Informatica uses Kerberos authentication.
3. Create matching operating system profile user names on the machine that runs the Data Integration Service and each Hadoop cluster node used to run Informatica mapping jobs.
4. Create an AD user that matches the operating system profile user you created in step 3.
5. Create an SPN associated with the user.

Use the following guidelines when you create the SPN and keytab files:

- The UPN must be the same as the SPN.
- Enable delegation in AD.
- Use the ktpass utility to create an SPN associated with the user and generate the keytabs file.
Running Mappings in the Native Environment When Informatica Does not Use Kerberos Authentication

To read and process data from Hive, HBase, or HDFS sources that use Kerberos authentication, you must configure Kerberos authentication in the native environment.

**Note:** If Informatica does not use Kerberos and the Hadoop cluster uses an MIT KDC with a one-way trust to AD, use the steps described in Running Mappings in a Hive Environment when Informatica Uses Kerberos Authentication.

To read and process data from Hive, HBase, or HDFS sources that use Kerberos authentication, perform the following steps:

1. Complete the prerequisite tasks for running mappings on a Hadoop cluster with Kerberos authentication.
2. Create matching operating system profile user names on the machine that runs the Data Integration Service and each Hadoop cluster node used to run Informatica mapping jobs.
3. Create an AD user that matches the operating system profile user you created in step 2.
4. Create an SPN associated with the user.

Use the following guidelines when you create the SPN and keytab files:
- The UPN must be the same as the SPN.
- Enable delegation in AD.
- Use the ktpass utility to create an SPN associated with the user and generate the keytab file.

For example, enter the following command:

```
    ktpass -out infa_hadoop.keytab -mapuser joe -pass tempBG@2008 -princ joe/
domain12345@HADOOP-AD-REALM -crypto all
```

The `-out` parameter specifies the name and path of the keytab file. The `-mapuser` parameter is the user to which the SPN is associated. The `-pass` parameter is the password for the SPN in the generated keytab. The `-princ` parameter is the SPN.

Kerberos Authentication for a Hive Connection

When you run a mapping that has Hive sources or targets, you must enable user authentication for the mapping on the Hadoop cluster. You must configure connection properties for the Hive connection in the Administrator or Developer tool to access Hive as a source or target.

To enable a user account to access a Hive source or target, configure the following Hive connection property:
Metadata Connection String

The JDBC connection URI used to access the metadata from the Hadoop server.

You must add the SPN to the connection string.

The connection string must be in the following format:

```java
jdbc:hive2://<host name>:<port>/<db>;principal=hive_princ_name
```

Where

- **host name** is the name or IP address of the machine that hosts the Hive server.
- **port** is the port on which the Hive server is listening.
- **db** is the name of the database to which you want to connect.
- **hive_princ_name** is the SPN of the HiveServer2 service that runs on the NameNode. Use the value set in this file:

  ```
  /etc/hive/conf/hive-site.xml
  ```

To enable a user account to run mappings with Hive sources in the native environment, configure the following properties in the Hive connection:

**Bypass Hive JDBC Server**

JDBC driver mode. Select the check box to use JDBC embedded mode.

Data Access Connection String

The connection string used to access data from the Hadoop data store.

The connection string must be in the following format:

```java
jdbc:hive2://<host name>:<port>/<db>;principal=hive_princ_name
```

Where

- **host name** is name or IP address of the machine that hosts the Hive server.
- **port** is the port on which the Hive server is listening.
- **db** is the name of the database to which you want to connect.
- **hive_princ_name** is the SPN of the HiveServer2 service that runs on the NameNode. Use the value set in this file:

  ```
  /etc/hive/conf/hive-site.xml
  ```

Kerberos Authentication for an HBase Connection

The Data Integration Service can connect to an HBase source or target in a Hadoop cluster that uses Kerberos authentication. To access HBase as a source or target, configure the HBase connection properties in the Administrator or Developer tool.

To enable a user account to access a HBase source or target, configure the following HBase connection properties:

**Zookeeper Host(s)**

Name of the machine that hosts the zookeeper.

**Zookeeper Port**

Port number where the zookeeper service is configured.
Enable Kerberos Connection

Select the check box to connect to a secure HBase.

HBase Master Principal

SPN of the HBase Master Service.

Use the value set in this file: /etc/hbase/conf/hbase-site.xml

HBase Region Server Principal

SPN of the HBase Region Server service.

Use the value set in this file: /etc/hbase/conf/hbase-site.xml

In the native or Hadoop environment, the SPN of the Data Integration Service user can access HBase sources or targets or run mappings on the Hadoop cluster.

To enable another user to access HBase sources or targets or run mappings on the Hadoop cluster, you must configure user impersonation in the native or Hadoop environment.

Kerberos Authentication for an HDFS Connection

The Data Integration Service can connect to an HDFS source or target in a Hadoop cluster that uses Kerberos authentication. You must configure connection properties for the HDFS connection in the Administrator or Developer tool to access HDFS as a source or target.

If the HDFS source or target in a Hadoop cluster uses Kerberos authentication, the Data Integration Service uses the user name of the HDFS connection to connect to the HDFS source or target.

To use another user name, you must configure user impersonation to run mappings in the native or Hadoop environment.

Metadata Import in the Developer Tool

To import metadata from Hive, HBase, and Complex File sources, configure the Developer tool to get Kerberos credentials to access Hive, HBase and complex file metadata.

To configure the Developer tool for metadata import, complete the following steps:

1. Copy hive-site.xml from the machine on which the Data Integration Service runs to a Developer Tool client installation directory. hive-site.xml is located in the following directory:

   \<Informatica Installation Directory>\services\shared\hadoop\<Hadoop distribution version>\conf/.

   Copy hive-site.xml to the following location:

   \<Informatica Installation Directory>\clients\hadoop\<Hadoop_distribution_version>\conf

2. Copy krb5.conf from \<Informatica Installation Directory>\services\shared\security to C:/Windows.

3. Rename krb5.conf to krb5.ini.
4. In krb5.ini, verify the value of the forwardable option to determine how to use the kinit command. If forwardable=true, use the kinit command with the -f option. If forwardable=false, or if the option is not specified, use the kinit command without the -f option.

5. Run the command from the command prompt of the machine on which the Developer tool runs to generate the Kerberos credentials file. For example, run the following command: kinit joe/domain12345@MY-REALM.

   Note: You can run the kinit utility from the following location: <Informatica Installation Directory>/bin/kinit.exe

6. Launch the Developer tool and import the Hive, HBase, and complex file sources.

Create and Configure the Analyst Service

To use the Analyst Service with a Hadoop cluster that uses Kerberos authentication, create the Analyst Service and configure it to use the Kerberos ticket for the Data Integration Service.

Perform the following steps:

1. Verify that the Data Integration Service is configured for Kerberos.
   For more information, see the Informatica Big Data Management Security Guide.

2. Create an Analyst Service.
   For more information about how to create the Analyst Service, see the Informatica Application Services Guide.

3. Log in to the Administrator tool.

4. In the Domain Navigator, select the Analyst Service.

5. In the Processes tab, edit the Advanced Properties.

6. Add the following value to the JVM Command Line Options field:
   DINFA_HADOOP_DIST_DIR=<Informatica installation directory>/services/shared/hadoop/
   <hadoop_distribution>.
User Impersonation with Kerberos Authentication

This chapter includes the following topics:

- User Impersonation, 30
- Create a Proxy Directory for Clusters that Run MapR, 31
- User Impersonation in the Hadoop Environment, 31
- User Impersonation in the Native Environment, 33

User Impersonation

You can enable different users to run mappings in a Hadoop cluster that uses Kerberos authentication or connect to big data sources and targets that use Kerberos authentication. To enable different users to run mappings or connect to big data sources and targets, you must configure user impersonation.

You can configure user impersonation for the native or Hadoop environment.

Before you configure user impersonation, you must complete the following prerequisites:

- Complete the prerequisite tasks for running mappings on a Hadoop cluster with Kerberos authentication.
- Configure Kerberos authentication for the native or Hadoop environment.
- If the Hadoop cluster uses MapR, create a proxy directory for the user who will impersonate other users.

If the Hadoop cluster does not use Kerberos authentication, you can specify a user name in the Hadoop connection to enable the Data Integration Service to impersonate that user.

If the Hadoop cluster uses Kerberos authentication, you must specify a user name in the Hadoop connection.
Create a Proxy Directory for Clusters that Run MapR

If the Hadoop cluster runs MapR, you must create a proxy directory for the user who will impersonate other users.

To enable user impersonation for the native and Hadoop environments, perform the following steps:

1. Go to the following directory on the machine on which the Data Integration Service runs:
   `<Informatica installation directory>/services/shared/hadoop/mapr_<version>/conf`
2. Create a directory named "proxy".
   Run the following command:
   ```
   mkdir <Informatica installation directory>/services/shared/hadoop/mapr_<version>/
   conf/proxy
   ```
3. Change the permissions for the proxy directory to `-rwxr-xr-x`.
   Run the following command:
   ```
   chmod 755 <Informatica installation directory>/services/shared/hadoop/mapr_<version>/
   conf/proxy
   ```
4. Verify the following details for the user that you want to impersonate with the Data Integration Service user:
   - Exists on the machine on which the Data Integration Service runs
   - Exists on every node in the Hadoop cluster
   - Has the same user-id and group-id on machine on which the Data Integration Service runs as well as the Hadoop cluster.
5. Create a file for the Data Integration Service user that impersonates other users.
   Run the following command:
   ```
   touch <Informatica installation directory>/services/shared/hadoop/mapr_<version>/
   conf/proxy/<username>
   ```
   For example, to create a file for the Data Integration Service user named user1 that is used to impersonate other users, run the following command:
   ```
   touch $INFA_HOME/services/shared/hadoop/mapr_<version>/conf/proxy/user1
   ```
6. Log in to the Administrator tool.
7. In the Domain Navigator, select the Data Integration Service.
8. Recycle the Data Integration Service.
   Click Actions > Recycle Service.

User Impersonation in the Hadoop Environment

To enable different users to run mapping and workflow jobs on a Hadoop cluster that uses Kerberos authentication, you must configure user impersonation in the Hadoop environment.

For example, the HypoStores administrator wants to enable user Bob to run mappings and workflows on the Hadoop cluster that uses Kerberos authentication.
To enable user impersonation, you must complete the following steps:

1. Enable the SPN of the Data Integration Service to impersonate another user named Bob to run Hadoop jobs.
2. Specify Bob as the user name for the Data Integration Service to impersonate in the Hadoop connection or Hive connection.

Note: If you create a Hadoop connection, you must use user impersonation.

Step 1. Enable the SPN of the Data Integration Service to Impersonate Another User

To run mapping and workflow jobs on the Hadoop cluster, enable the SPN of the Data Integration Service to impersonate another user.

Configure user impersonation properties in core-site.xml on the Name Node on the Hadoop cluster.

core-site.xml is located in the following directory:

/etc/hadoop/conf/core-site.xml

Configure the following properties in core-site.xml:

- hadoop.proxyuser.<superuser>.groups
  Enables the superuser to impersonate any member in the specified groups of users.

- hadoop.proxyuser.<superuser>.hosts
  Enables the superuser to connect from specified hosts to impersonate a user.

For example, set the values for the following properties in core-site.xml:

```xml
<property>
  <name>hadoop.proxyuser.<superuser>.groups</name>
  <value>group1,group2</value>
  <description>Allow the superuser <DIS_user> to impersonate any members of the group group1 and group2</description>
</property>

<property>
  <name>hadoop.proxyuser.<superuser>.hosts</name>
  <value>host1,host2</value>
  <description>The superuser can connect only from host1 and host2 to impersonate a user</description>
</property>
```

Step 2. Specify a User Name in the Hadoop Connection

In the Developer tool, specify a user name in the Hadoop connection for the Data Integration Service to impersonate when it runs jobs on the Hadoop cluster.

If you do not specify a user name, the Hadoop cluster authenticates jobs based on the SPN of the Data Integration Service.

For example, if Bob is the name of the user that you entered in core-site.xml, enter Bob as the user name.
User Impersonation in the Native Environment

To enable different users to run mappings that read or processes data from big data sources or targets that use Kerberos authentication, configure user impersonation for the native environment.

For example, the HypoStores administrator wants to enable user Bob to run mappings that read and process data from Hive and HBase sources and HDFS targets that use Kerberos authentication.

To enable user impersonation, you must complete the following steps:

1. Login to the Kerberos realm.
2. Specify Kerberos authentication properties for the Data Integration Service.
3. Configure the execution options for the Data Integration Service.
4. Specify the URL for the Hadoop cluster in the Hive, HBase, or HDFS connection.
5. Configure the mapping impersonation property that enables user Bob to run the mapping in the native environment.

**Step 1. Login to the Kerberos Realm**

Use the SPN and keytab of the Data Integration Service user to login to the Kerberos realm on the machine that hosts the KDC server.

**Step 2. Specify the Kerberos Authentication Properties for the Data Integration Service**

In the Data Integration Service properties, configure the properties that enable the Data Integration Service to connect to a Hive, HBase, or HDFS sources and targets that use Kerberos authentication.

Configure the following properties:

**Hadoop Kerberos Service Principal Name**

Service Principal Name (SPN) of the Data Integration Service to connect to a Hadoop cluster that uses Kerberos authentication.

Enter the property in the following format:

<princ_name>

For example, enter the following value:

joe/domain12345@MY-REALM

**Hadoop Kerberos Keytab**

Path to the Kerberos keytab file on the machine on which the Data Integration Service runs.

Enter the property in the following format:

<keytab_path>

For example, enter the following path:

<Informatica Big Data Management Server Installation Directory>/isp/config/keys/infa_hadoop.keytab
Step 3. Configure the Execution Options for the Data Integration Service

To determine whether the Data Integration Service runs jobs in separate operating system processes or in one operating system process, configure the Launch Jobs as Separate Processes property. Use the Administrator tool to configure the execution options for the Data Integration Service.

1. Click Edit to edit the Launch Jobs as Separate Processes property in the execution options for the Data Integration Service properties.
2. Choose to enable or disable the property.
   - If you enable the property, you must specify the location of the krb5.conf in the Java Virtual Manager (JVM) Options as a custom property in the Data Integration Service process. krb5.conf is located in the following directory:\Informatica Installation Directory\services\shared\security.
   - If you disable the property, you must specify the Java Command Line Options property in the Advanced Properties of the Data Integration Service process.

Step 4. Specify the URL for the Hadoop Cluster in the Connection Properties

In the Administrator or Developer tool, specify the URL for the Hadoop cluster on which the Hive, HBase, or HDFS source or target resides. Configure the Hive, HBase, or HDFS connection properties to specify the URL for the Hadoop cluster.

In the Hive connection, configure the properties to access Hive as a source or a target.

In the HBase connection, configure the Kerberos authentication properties.

In the HDFS connection, configure the NameNode URI property.

Step 5. Configure the Mapping Impersonation Property

In the Developer tool, configure the mapping impersonation property in the native environment that enables another user to run mappings that read or process data from big data sources that use Kerberos authentication.

1. Launch the Developer tool and open the mapping that you want to run.
   The mapping opens in the editor.
2. Click the Run-time tab.
3. Select Native as the execution environment.
4. To enable another user to run the mapping, click Mapping Impersonation User Name and enter the value in the following format:
   <Hadoop service name>/<Hostname>@<YOUR-REALM>.
   Where
   - Hadoop service name is the name of the Hadoop service on which the Hive, HBase, or HDFS source or target resides.
   - Hostname is the name or IP address of the machine on which the Hadoop service runs. The hostname is optional.
   - YOUR-REALM is the Kerberos realm.
   The following special characters can only be used as delimiters: ':' and '@'.
5. Right-click an empty area in the editor and click Run Mapping.
Chapter 4

Blaze Engine Security

This chapter includes the following topics:

- Blaze Engine Security Overview, 35
- Setting up a Blaze User Account, 35

Blaze Engine Security Overview

The Blaze engine runs on the Hadoop cluster as a YARN application. When the Blaze engine starts, it has the privileges and permissions of the user account used to start it. You can designate a user account to start the Blaze engine or use the Data Integration Service user account. Informatica recommends that you create a user account on the Hadoop cluster for Blaze.

A designated user account isolates the Blaze engine from other services on the Hadoop cluster. Grant the user account the minimum required privileges and permissions. When you limit the privileges and permissions of a user account, you limit the attack surface that is available to unauthorized users.

If there is not a specific user account for the Blaze engine, the Blaze engine uses the Data Integration Service user account. The Data Integration Service user account has permissions that the Blaze engine does not need. For example, the Data Integration Service user account has permission to impersonate other users. Blaze does not need this permission.

When you submit a job to the Hadoop cluster, the Blaze engine uses the mapping impersonation user to run the job. If there is not a mapping impersonation user specified, the Blaze engine uses the Data Integration Service user.

Setting up a Blaze User Account

Create a user account for the Blaze engine. Use that user account to start the Blaze engine.

Perform the following steps:

1. On every node in the Hadoop cluster, create an operating system user account.
   For example, to create a user account named "blaze", run the following command on every node in the cluster:
   ```
   useradd blaze
   ```
2. Give the user the minimum required privileges and permissions.
For example, the user account for the Blaze engine must be able to read from HDFS.

3. In the Developer tool, create a Hadoop connection.
   
   **Note:** You must use user impersonation for the Hadoop connection if the Hadoop cluster uses Kerberos authentication.

   For more information about how to create a Hadoop connection, see the *Big Data Management User Guide*.

4. In the connection properties, use the user account you created in step 1 for the Blaze Service User Name.
INDEX

A
authentication
infrastructure security 10
Kerberos 10
authorization
HDFS permissions 11
infrastructure security 11

C
cloudera navigator
data management 12
cross-realm trust
Kerberos authentication 22

D
data management
cloudera navigator 12
Metadata Manager 12
data security
Dynamic Data Masking 12
Persistent Data Masking 12
Secure@Source 12

data management
cloudera navigator 12
Metadata Manager 12
data security
Dynamic Data Masking 12
Persistent Data Masking 12
Secure@Source 12

H
HDFS permissions
authorization 11

I
infrastructure security
authentication 10

Kerberos authentication
authenticating a Hive connection 26
authenticating an HBase connection 27
authenticating an HDFS connection 28
cross-realm trust 22
Hadoop warehouse permissions 16
impersonating another user 32
Informatica domain with Kerberos authentication 22
Informatica domain without Kerberos authentication 20
JCE Policy File 16
Kerberos security properties 16
mappings in a native environment 25, 26
metadata import 28
operating system profile names 21
overview 14
prerequisites 15
requirements 15
user impersonation 30
user impersonation in the native environment 33

M
Metadata Manager
data management 12

U
user impersonation
Hadoop environment 31
impersonating another user 32
user name 32