Informatica® Big Data Management
10.1.1 Update 2

Big Data Management Security Guide

Informatica Big Data Management Big Data Management Security Guide
10.1.1 Update 2
March 2017

Publication Date: 2018-12-14

Table of Contents

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 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

Introduction to Big Data Management Security

This chapter includes the following topics:

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

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

By default, Hadoop does not verify the identity of users. To authenticate user identities, you can configure the following authentication protocols on the cluster:

- Native authentication
- Lightweight Directory Access Protocol (LDAP)
- Kerberos
- Apache Knox Gateway

Big Data Management also 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. 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.

You can run mappings on a cluster that uses one of the following security management systems for authorization:

- HDFS permissions
- User impersonation
- Apache Ranger
- Apache Sentry
- HDFS Transparent Encryption

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.

Operating system profiles

An operating system profile is a type of security that the Data Integration Service uses to run mappings. Use operating system profiles to increase security and to isolate the run-time environment for users. Big Data Management supports operating system profiles on all Hadoop distributions.

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*.

Support for Security Management Systems

Depending on the run-time engine that you use, you can run mappings on a Hadoop cluster that uses a supported security management system.

Hadoop clusters use a variety of security management systems for user authorization and authentication. The following table shows the run-time engines supported for the security management system installed on the Hadoop platform:

Hadoop Distribution	Apache Knox (authentication)	Apache Ranger (authorization)	Apache Sentry (authorization)	HDFS Transparent Encryption (authorization)	SSL/TLS protocol
Amazon EMR 5.0	Not supported	Not supported	Not supported	Not supported	Not supported
Azure HDInsight 3.5	Not supported	Not supported	Not supported	Not supported	Not supported

Hadoop Distribution	Apache Knox (authentication)	Apache Ranger (authorization)	Apache Sentry (authorization)	HDFS Transparent Encryption (authorization)	SSL/TLS protocol
Cloudera CDH 5.8, 5.9, 5.10	Not supported	Not supported	- Native run-time engine - Blaze engine - Spark engine - Hive engine	- Native run-time engine - Blaze engine - Spark engine	- Native run-time engine - Blaze engine - Spark engine - Hive engine
IBM BigInsights 4.2	- Native run-time engine - Blaze engine - Spark engine - Hive engine	- Native run-time engine - Blaze engine - Spark engine	Not supported	- Native run-time engine - Blaze engine - Spark engine	Not supported
Hortonworks HDP 2.3, 2.4, 2.5	- Native run-time engine - Blaze engine - Spark engine - Hive engine	- Native run-time engine - Blaze engine - Spark engine	Not supported	- Native run-time engine - Blaze engine - Spark engine	- Native run-time engine - Blaze engine - Spark engine - Hive engine
MapR 5.2	Not supported	Not supported	Not supported	Not supported	- Native run-time engine - Blaze engine - Spark engine - Hive engine

Authentication

When the Informatica domain includes Big Data Management, user identities 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.

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.

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.

Apache Knox Gateway

The Apache Knox Gateway is a REST API gateway that authenticates users and acts as a single access point for a Hadoop 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.

Apache Knox Gateway

The Apache Knox Gateway is a REST API gateway that authenticates users and acts as a single access point for a Hadoop cluster.

Knox creates a perimeter around a Hadoop cluster. Without Knox, users and applications must connect directly to a resource in the cluster, which requires configuration on the client machines. A direct connection to resources exposes host names and ports to all users and applications and decreases the security of the cluster.

If the cluster uses Knox, applications use REST APIs and JDBC/ODBC over HTTP to connect to Knox. Knox authenticates the user and connects to a resource.

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.

You can use the following systems to manage authorization for Big Data Management:

HDFS permissions

By default, Hadoop uses HDFS permissions to determine what a user can do to a file or directory on HDFS.

User impersonation

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

Apache Ranger

Ranger is a security plug-in that is used to authenticate users of a Hadoop cluster. Ranger manages access to files, folders, databases, tables, and columns. When a user performs an action, Ranger verifies that the user meets the policy requirements and has the correct permissions on HDFS.

Apache Sentry

Sentry is a security plug-in that is used to enforce role-based authorization for data and metadata on a Hadoop cluster. Sentry can secure data and metadata at the table and column level. For example, Sentry can restrict access to columns that contain sensitive data and prevent unauthorized users from accessing the data.

HDFS Transparent Encryption

Hadoop implements transparent data encryption in HDFS directories.

Fine-Grained SQL Authorization

SQL standards-based authorization enables database administrators to impose row-level authorization on Hive tables and views. A more fine-grained level of SQL standards-based authorization enables administrators to impose row and column level authorization. You can configure a Hive connection to observe row and column level SQL standards-based authorization.

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.

Blaze Engine Security

Secure the Blaze engine with a designated user account for the Blaze engine.

The Blaze engine runs on the Hadoop cluster as a service. Informatica recommends that you create a user account on the cluster for the Blaze engine. A designated user account isolates the Blaze engine from other services on the cluster. You can grant the Blaze user the minimum required privileges and permissions that Blaze requires to run. If you do not use a designated user for Blaze, the Data Integration Service user starts the Blaze engine on the Hadoop cluster.

SQL Standards-Based Authorization on Hive Source Rows and Columns

SQL standards-based authorization enables database administrators to impose row-level authorization on Hive tables and views. A more fine-grained level of SQL standards-based authorization enables administrators to impose row and column level authorization when you read data from a Hive source. You can configure a Hive connection to observe row and column level SQL standards-based authorization.

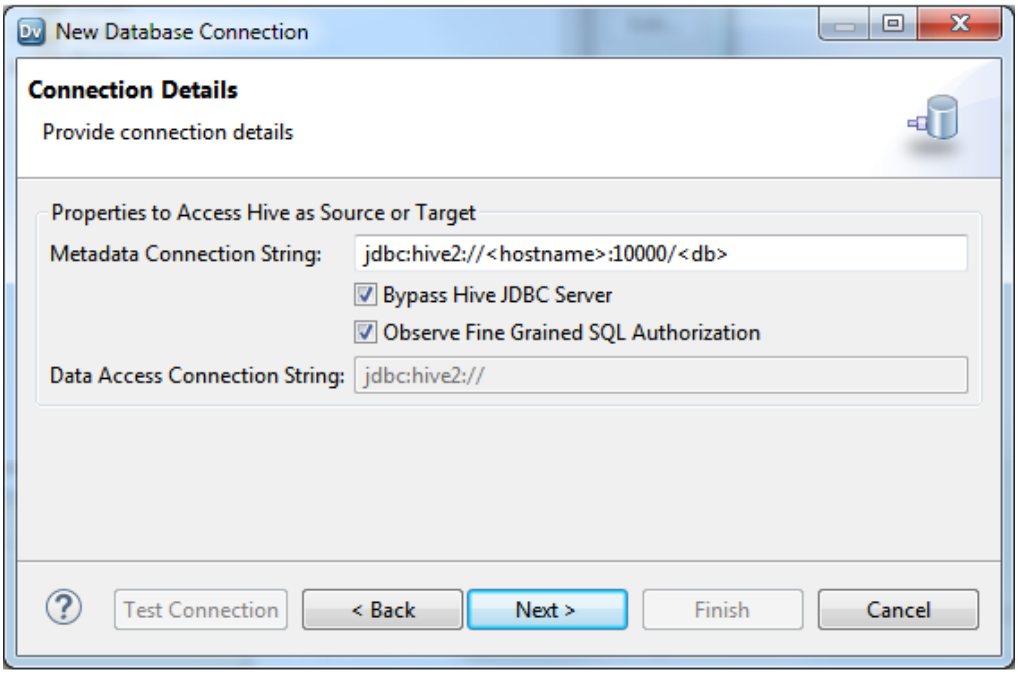
When you select the option to observe fine-grained SQL authentication in a Hive source, the mapping observes row and column-level restrictions on data access in the source. If you do not select the option, the Blaze run-time engine ignores the restrictions, and results include restricted data.

Configuring Hive Connection for Sources that Use Fine-Grained SQL Authorization

You can configure a Hive database connection to observe fine-grained SQL authorization on source tables.

1. Create a Hive connection, or edit an existing Hive connection.
2. Select the **Observe Fine Grained SQL Authorization** option.

   The following image shows the **Connection Details** configuration dialog box in the Developer tool:



   Note: You can also configure the **Observe Fine Grained SQL Authorization** option for a Hive connection in the Administrator tool. See the *Informatica 10.1.1 Administration Guide*.

3. Click **Next**, and then **OK**.

With the **Observe Fine Grained SQL Authorization** option, the mapping observes row and column-level restrictions on data access in the source.

Apache Ranger KMS

Apache Ranger KMS is an open source key management service that supports HDFS data at rest encryption.

Ranger KMS enables the following functions:

Key management

You can create, update, or delete encryption key zones that control access to functionality.

Access control policies

You can administer access control policies for encryption keys. You can create or edit keys to control access by users to functionality.

You can run mappings on a Hortonworks HDP cluster that uses Ranger KMS.

Configuring Apache Ranger KMS

Use the Apache Ambari console to configure Apache KMS for access control.

1. Create a user for Informatica.

   The user is the same as the Data Integration Service user or the Kerberos SPN user.
2. Add the Informatica user to a new KMS repository, or to an existing KMS repository.
3. Grant permissions to the Informatica user.
4. Create and configure an encryption key.
5. Create an encryption zone that uses the encryption key you created.

   For example:

   ```
   hdfs dfs -mkdir /zone_encr_infa
   hdfs crypto -createZone -keyName infa_key -path /zone_encr_infa
   ```
6. In the Ambari cluster administration utility, browse to the Custom KMS Site page and add the following properties:

   ```
   hadoop.kms.proxyuser.<user>.groups=*
   hadoop.kms.proxyuser.<user>.hosts=*
   hadoop.kms.proxyuser.<user>.users=*
   ```

   where <user> is the Informatica user name you configured in Step 1.
7. Browse to **Ambari Agent** > **HDFS** > **Custom Core Site** and update the following properties

   ```
   hadoop.kms.proxyuser.<user>.hosts
   hadoop.kms.proxyuser.<user>.groups
   ```
8. Browse to **Ambari Agent** > **Ranger KMS** > **Configs** and search for *proxyuser* in the Ranger KMS Configurations area. To register all Hadoop system users with Ranger KMS, add the following properties:

   ```
   hadoop.kms.proxyuser.HTTP.hosts=*
   hadoop.kms.proxyuser.HTTP.users=*
   hadoop.kms.proxyuser.hive.hosts=*
   hadoop.kms.proxyuser.hive.users=*
   hadoop.kms.proxyuser.keyadmin.hosts=*
   hadoop.kms.proxyuser.keyadmin.users=*
   hadoop.kms.proxyuser.nn.hosts=*
   hadoop.kms.proxyuser.nn.users=*
   hadoop.kms.proxyuser.rm.hosts=*
   hadoop.kms.proxyuser.rm.users=*
   hadoop.kms.proxyuser.yarn.hosts=*
   hadoop.kms.proxyuser.yarn.users=*
   ```

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 Masking

Data masking obscures data based on configurable rules. It prevents unauthorized users from reading sensitive data on the Hadoop cluster. Big Data Management supports different methods of data masking to secure data.

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 bout 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*.

Operating System Profiles

Use operating system profiles to increase security and to isolate the run-time environment for users. You can create and manage operating system profiles on the Security tab of the Administrator tool.

By default, the Data Integration Service process runs all mappings using the permissions of the operating system user that starts Informatica Services. When you configure the Data Integration Service to use operating system profiles, the Data Integration Service process runs mappings with the permission of the operating system user you define in the operating system profile. In the Hadoop run-time environment, the Data Integration Service pushes the processing to the Hadoop cluster and the Big Data Management engines run mappings with the operating system profile.

For more information about operating system profiles, see the *Informatica Security Guide*.

CHAPTER 2

Running Mappings with Kerberos Authentication

This chapter includes the following topics:

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 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. Otherwise, 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 want to run a mapping with the Blaze runtime engine, you configure settings on the Informatica domain. To run a mapping on a cluster with High Availability, you configure additional settings.
- 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

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 19.

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 tempBG@2008 -princ joe/
  domain12345@INFA-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

When you run the Hadoop Configuration Manager, you enter values for the following properties that enable the Data Integration Service to connect to a Hadoop cluster that uses Kerberos authentication:

Hadoop Kerberos Service Principal Name

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

Hadoop Kerberos Keytab

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

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:

   ```
   ksetup /addkdc <mit_realm_name> <kdc_hostname>
   ```

   For example, enter the command to add the following values:

   ```
   ksetup /addkdc HADOOP-MIT-REALM def456.hadoop-mit-realm.com
   ```

2. Enter the following command to add the local realm trust to Active Directory:

   ```
   netdom trust <mit_realm_name> /Domain:<ad_realm_name> /add /realm /
   passwordt:<TrustPassword>
   ```

   For example, enter the command to add the following values:

   ```
   netdom trust HADOOP-MIT-REALM /Domain:INFA-AD-REALM /add /realm /passwordt:trust1234
   ```

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:

   ```
   ksetup /SetEncTypeAttr <mit_realm_name> <enc_type>
   ```

   For Microsoft Windows 2003, enter the following command:

   ```
   ktpass /MITRealmName <mit_realm_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:

```
kadmin:  addprinc -e "<enc_type_list>" krbtgt/<mit_realm_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:

```
kadmin:  addprinc -e "rc4-hmac:normal des3-hmac-sha1: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:

```
<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.

```
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 22.

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 tempBG@2008 -princ joe/
  domain12345@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.
     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.

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 for mappings 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.

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>\clients\java\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>`.

CHAPTER 3

User Impersonation with Kerberos Authentication

This chapter includes the following topics:

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

To run mappings on the Blaze, Spark or Hive run-time engines, you must create a proxy user for the user who will impersonate other users.

You can use the Ambari configuration manager to configure impersonation properties on Hadoop distributions that use Ambari. If you do not use Apache Ambari, you must configure the impersonation properties in core-site.xml on the Hadoop cluster and restart Hadoop services and the cluster.

Configure the following user impersonation properties:

hadoop.proxyuser.<impersonation_user>.groups

Allows impersonation from any group.

Note: The user name that you specify in this property depends on whether the cluster uses Kerberos authentication.

When the cluster uses Kerberos, use the following example to help configure the property:

```
<property>
    <name>hadoop.proxyuser.<SPN_user>.groups</name>
    <value><user_name></value>
    <description>Allows impersonation from any group.</description>
</property>
```

where <SPN_user> is the Service Principal Name that corresponds to the keytab file that the cluster uses to authenticate the client.

When the cluster does not use Kerberos, use the following example to help configure the property:

```
<property>
    <name>hadoop.proxyuser.<domain_starter_user>.groups</name>
    <value><user_name></value>
    <description>Allows impersonation from any group.</description>
</property>
```

where <domain_starter_user> is the user account that is used to start the Informatica domain. This is the same as the Informatica administrator user name.

hadoop.proxyuser.<impersonation_user>.hosts

Allows impersonation from any host.

Note: The user name that you specify in this property depends on whether the cluster uses Kerberos authentication.

When the cluster uses Kerberos, use the following example to help configure the property:

```
<property>
    <name>hadoop.proxyuser.<SPN_user>.hosts</name>
    <value><user_name></value>
    <description>Allows impersonation from any group.</description>
</property>
```

where <SPN_user> is the Service Principal Name that corresponds to the keytab file that the cluster uses to authenticate the client.

When the cluster does not use Kerberos, use the following example to help configure the property:

```
<property>
    <name>hadoop.proxyuser.<domain_starter_user>.hosts</name>
    <value><user_name></value>
    <description>Allows impersonation from any group.</description>
</property>
```

where <domain_starter_user> is the user account that is used to start the Informatica domain. This is the same as the Informatica administrator user name.

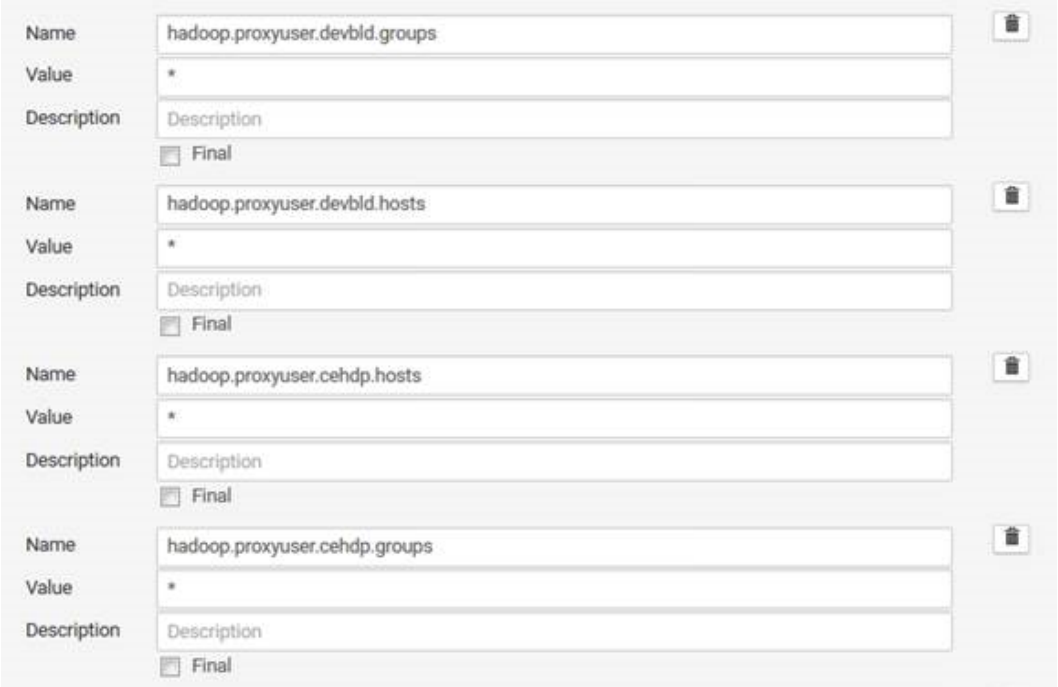
Using Apache Ambari to Configure User Impersonation

To enable user impersonation, use Apache Ambari to add or update the `hadoop.proxyuser.<impersonation_user>.groups` and `hadoop.proxyuser.<impersonation_user>.hosts` properties. If the properties are already added, you must change the value for the properties to `*` (asterisk).

1. Start Apache Ambari.
2. Click **HDFS service** > **Configs** > **Advanced**.

3. Navigate to the custom core-site section.
4. Add or update the user impersonation properties.
   - To add a property, click **Add Property** and enter the name of the property as the key and the value as * (asterisk).
   - To update a property, set the property value to * (asterisk).
5. Save and restart the Hadoop services and the Hadoop cluster.

The following image shows the proxy users devbld and cdchp configured in the Apache Ambari console.



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:

```
<property>
               <name>hadoop.proxyuser.bob.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.bob.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
```

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 Job Options** property. Use the Administrator tool to configure the execution options for the Data Integration Service.

1. Click **Edit** to edit the **Launch Job Options** property in the execution options for the Data Integration Service properties.
2. Choose the launch job option.
   - If you configure the Data Integration Service to launch jobs as a separate process, 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>/java/jre/lib/security`.

- If you configure the Data Integration Service to launch jobs in the service process, you must specify the location of krb5.conf in the **Java Command Line Options** property in the Advanced Properties of the Data Integration Service process. Use the following syntax:

  ```
  -Djava.security.krb5.conf=<Informatica installation directory>/java/jre/lib/
  security/krb5.conf
  ```

  .

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

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.

To configure the Informatica domain to use Blaze to run mappings on a Kerberos-enabled cluster, see the *Informatica Big Data Management 10.1.1 Update 2 Installation and Configuration Guide*.

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