Informatica® PowerExchange for Amazon S3 10.1.1 Update 1

User Guide

Informatica PowerExchange for Amazon S3 User Guide
10.1.1 Update 1
December 2016

Publication Date: 2018-09-27

Table of Contents

Preface

The *PowerExchange® for Amazon S3 Guide* contains information about how to set up and use PowerExchange for Amazon S3. The guide explains how organization administrators and business users can use PowerExchange for Amazon S3 to read from and write data to Amazon S3.

This guide assumes that you have knowledge of Amazon S3 and Informatica Data Services.

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 PowerExchange for Amazon S3

This chapter includes the following topics:

PowerExchange for Amazon S3 Overview

You can use PowerExchange for Amazon S3 to read and write delimited flat file data and binary files as pass-through data from and to Amazon S3 buckets.

Amazon S3 is a cloud-based store that stores many objects in one or more buckets.

Create an Amazon S3 connection to specify the location of Amazon S3 sources and targets you want to include in a data object. You can use the Amazon S3 connection in data object read and write operations.

You can validate and run mappings in the native environment or Blaze mode.

Example

You are a medical data analyst in a medical and pharmaceutical organization who maintains patient records. A patient record can contain patient details, doctor details, treatment history, and insurance from multiple data sources.

You use PowerExchange for Amazon S3 to collate and organize the patient details from multiple input sources in Amazon S3 buckets.

Introduction to Amazon S3

Amazon Simple Storage Service (Amazon S3) is storage service in which you can copy data from source and simultaneously move data to any target. You can use Amazon S3 to store and retrieve any amount of data at any time, from anywhere on the web. You can accomplish these tasks using the AWS Management Console web interface.

Amazon S3 stores data as objects within buckets. An object consists of a file and optionally any metadata that describes that file. To store an object in Amazon S3, you upload the file you want to store to a bucket.

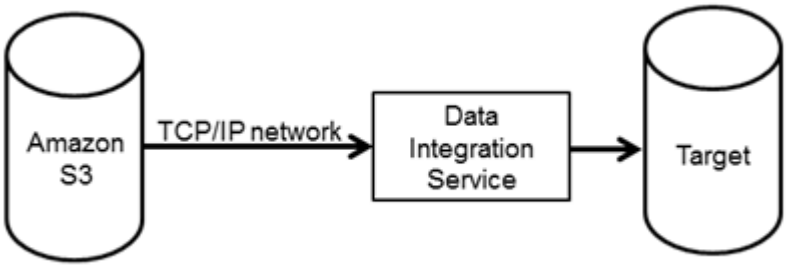
Buckets are the containers for objects. You can have one or more buckets. When using the AWS Management Console, you can create folders to group objects, and you can nest folders.

Data Integration Service and Amazon S3 Integration

The Data Integration Service uses the Amazon S3 connection to connect to Amazon S3.

Reading Amazon S3 Data

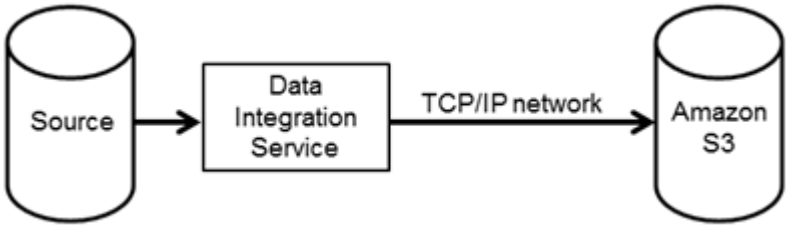
The following image shows how Informatica connects to Amazon S3 to read data:



When you run the Amazon S3 session, the Data Integration Service reads data from Amazon S3 based on the workflow and Amazon S3 connection configuration. The Data Integration Service connects and reads data from Amazon Simple Storage Service (Amazon S3) through a TCP/IP network. The Data Integration Service then stores data in a staging directory on the Data Integration Service host. Amazon S3 is a storage service in which you can copy data from source and simultaneously move data to any target. The Data Integration Service issues a copy command that copies data from Amazon S3 to the target.

Writing Amazon S3 Data

The following image shows how Informatica connects to Amazon S3 to write data:



When you run the Amazon S3 session, the Data Integration Service writes data to Amazon S3 based on the workflow and Amazon S3 connection configuration. The Data Integration Service stores data in a staging directory on the Data Integration Service host. The Data Integration Service then connects and writes data to Amazon Simple Storage Service (Amazon S3) through a TCP/IP network. Amazon S3 is a storage service in which you can copy data from source and simultaneously move data to Amazon S3 clusters. The Data Integration Service issues a copy command that copies data from Amazon S3 to the Amazon S3 target table.

CHAPTER 2

Power Exchange for Amazon S3 Configuration Overview

This chapter includes the following topics:

Power Exchange for Amazon S3 Configuration Overview

PowerExchange for Amazon S3 installs with the Informatica services. You can enable PowerExchange for Amazon S3 with a license key.

Prerequisites

Before you can use PowerExchange for Amazon S3, perform the following tasks:

- Ensure that PowerExchange for Amazon S3 license is activated.
- Create an Access Key ID and Secret Access Key in AWS. You can provide these key values when you create an Amazon S3 connection
- Verify that you have write permissions on all the directories within the `<INFA_HOME>`directory.
- To run mappings on Hortonworks, Amazon EMR, and IBM BigInsights distributions that use non-kerberos authentication, configure user impersonation.

Configure User Impersonation

Ensure that the Hadoop administrator creates a proxy user to impersonate other users if you run a mapping on Hortonworks, Amazon EMR, or IBM BigInsights distributions that use non-kerberos authentication.

The Hadoop administrator uses Apache Ambari to configure impersonation properties for Hortonworks HDP and IBM BigInsights.

Note: If Apache Ambari is not used, the Hadoop administrator configures configure the impersonation properties in core-site.xml on each node of the Hadoop cluster and restart Hadoop services along with the cluster.

The Hadoop administrator must configure the following user impersonation properties:

hadoop.proxyuser.yarn.groups

Comma-separated list of groups that you want to allow the YARN user to impersonate on a non-secure cluster.

Set to group names of impersonation users separated by commas. If less security is preferred, use the wildcard " * " to allow impersonation from any group.

hadoop.proxyuser.yarn.hosts

Comma-separated list of hosts that you want to allow the YARN user to impersonate on a non-secure cluster.

Set to a single host name or IP address, or set to a comma-separated list. If less security is preferred, use the wildcard " * " to allow impersonation from any host.

Using Apache Ambari to Configure User Impersonation

To enable user impersonation, use Apache Ambari to add or update the `hadoop.proxyuser.yarn.groups` and `hadoop.proxyuser.yarn.hosts` properties. If the properties are already added, you must change the value for the properties to the appropriate value. For example: *

1. Log on to Apache Ambari.
2. Click **HDFS service** > **Configs** > **Advanced**.
3. Navigate to the custom core-site section.
4. Click **Add Property**.
5. Add the following properties with the appropriate value:
   - `hadoop.proxyuser.yarn.groups`
   - `hadoop.proxyuser.yarn.hosts`
6. Save and restart the Hadoop services and the Hadoop cluster.

Configuring Cloudera Clusters

You can configure additional functionality through the configuration files the in Hadoop environment.

When you use Cloudera Manager to update properties, updates are propagated to each node in the cluster. After you make updates, you can restart the Hadoop cluster and services through Cloudera Manager.

Update mapred-site.xml

Update the `mapred-site.xml` file on the Hadoop environment to configure functionality for Amazon S3 connectivity.

To enable Amazon S3 connectivity for Cloudera clusters, configure MapReduce properties in the **mapred-site.xml** file on the Hadoop cluster, and restart Hadoop services and the cluster.

1. Open the Yarn Configuration in Cloudera Manager.
2. Find the property named **NodeManager Advanced Configuration Snippet (Safety Valve) for mapred-site.xml**.

3. Click **+** and configure the following properties:

Property	Value

mapreduce.application.classpath	`$HADOOP_MAPRED_HOME/*,`
	`$HADOOP_MAPRED_HOME/lib/*,`
	`$MR2_CLASSPATH,`
	`$CDH_MR2_HOME`

mapreduce.jobhistory.intermediate-done-dir	`<Directory where the map-reduce jobs write history files>`
mapreduce.jobhistory.address	`<MapReduce JobHistory Server hostname>:<port>` **Note:** Configure this property if you want to run an Amazon S3 mappings on the Cloudera Kerberos clusters that are not enabled with NameNode high availability.

4. Select the **Final** check box.
5. Redeploy the client configurations.
6. Restart Hadoop services and the cluster.

Update yarn-site.xml

Update the `yarn-site.xml` file on the domain to configure functionality such as Amazon S3 for Kerberos authentication.

To run Amazon S3 mappings on Cloudera Kerberos clusters that are not enabled with high availability, you must configure properties in the `yarn-site.xml` file on the Data Integration Service node and restart the Data Integration Service.

After you configure `mapred-site.xml` on the Hadoop environment, copy the following properties to the `yarn-site.xml` file on the Data Integration Service node:

mapreduce.jobhistory.address

Location of the MapReduce JobHistory Server. The default port is 10020.

```
<property>
<name>mapreduce.jobhistory.address</name>
<value><host name>:port</value>
<description>MapReduce JobHistory Server IPC host:port</description>
</property>
```

mapreduce.jobhistory.principal

SPN for the MapReduce JobHistory server.

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

mapreduce.jobhistory.webapp.address

Web address of the MapReduce JobHistory Server. The default value is 19888.

```
<property>
<name>mapreduce.jobhistory.webapp.address</name>
<value>hostname:port</value>
<description>MapReduce JobHistory Server Web UI host:port</description>
</property>
```

Create Minimal Amazon S3 Bucket Policy

You can create a minimal Amazon S3 bucket policy to ensure that PowerExchange for Amazon S3 successfully reads and writes data from and to Amazon S3.

To restrict the user operations and user access to specific Amazon S3 buckets, assign an AWS Identity and Access Management (IAM) policy to users. Configure the IAM policy through the AWS console. To successfully read data from and write data to Amazon S3, users need the following permissions:

- PutObject
- GetObject
- GetObjectVersion
- DeleteObject
- DeleteObjectVersion
- ListBucket
- GetBucketPolicy
- ListAllMyBuckets

Sample Policy

```
{"Version": "2012-10-17", "Statement": [{ "Effect": "Allow", "Action": [ "s3:PutObject",
"s3:GetObject", "s3:GetObjectVersion", "s3:DeleteObject", "s3:DeleteObjectVersion",
"s3:ListBucket", "s3:GetBucketPolicy",  "s3:ListAllMyBuckets" ], "Resource": "*"  }]}
```

CHAPTER 3

Amazon S3 Connections

This chapter includes the following topics:

Amazon S3 Connections Overview

Amazon S3 connections enable you to read data from or write data to Amazon S3.

When you create an Amazon S3 connection, you define connection attributes. You can create an Amazon S3 connection in the Developer tool or the Administrator tool. The Developer tool stores connections in the domain configuration repository. Create and manage connections in the connection preferences.

The Developer tool uses the connection when you create data objects. The Data Integration Service uses the connection when you run mappings.

Amazon S3 Connection Properties

When you set up an Amazon S3 connection, you must configure the connection properties.

The following table describes the Amazon S3 connection properties:

Property	Description
Name	The name of the connection. The name is not case sensitive and must be unique within the domain. You can change this property after you create the connection. The name cannot exceed 128 characters, contain spaces, or contain the following special characters:~ ` ! $ % ^ & * ( ) - + = { [}] | \ : ; " ' < , > . ? /
ID	String that the Data Integration Service uses to identify the connection. The ID is not case sensitive. It must be 255 characters or less and must be unique in the domain. You cannot change this property after you create the connection. Default value is the connection name.
Description	Optional. The description of the connection. The description cannot exceed 4,000 characters.

Property	Description
Type	The Amazon S3 connection type.
Access Key	The access key ID for access to Amazon account resources.
Secret Key	The secret access key for access to Amazon account resources. The secret key is associated with the access key and uniquely identifies the account.
Folder Path	The complete path to Amazon S3 objects. The path must include the bucket name and any folder name. Do not use a slash at the end of the folder path. For example, `<bucket name>/<my folder name>`.
Master Symmetric Key	Optional. Provide a 256-bit AES encryption key in the Base64 format when you enable client-side encryption. You can generate a key using a third-party tool. **Note:** You can enable Client Side Encryption as the encryption type in the advanced properties of the data object write operation.
Region Name	Select the AWS region in which the bucket you want to access resides.

Creating an Amazon S3 Connection

Create an Amazon S3 connection before you create an Amazon S3 data object.

1. In the Developer tool, click **Window** > **Preferences**.
2. Select **Informatica** > **Connections**.
3. Expand the domain in the **Available Connections**.
4. Select the connection type **Enterprise Application** > **Amazon S3**, and click **Add**.
5. Enter a connection name and an optional description.
6. Select Amazon S3 as the connection type.
7. Click **Next**.
8. Configure the connection properties.
9. Click **Test Connection** to verify the connection to Amazon S3.
10. Click **Finish**.

CHAPTER 4

PowerExchange for Amazon S3 Data Objects

This chapter includes the following topics:

Amazon S3 Data Object Overview

An Amazon S3 data object is a physical data object that uses Amazon S3 as a source or target. An Amazon S3 data object is the physical data object that represents data based on an Amazon S3 resource.

You can configure the data object read and write operation properties that determine how data can be read from Amazon S3 sources and loaded to Amazon S3 targets.

Create an Amazon S3 data object from the Developer tool. Create a data object read operation or data object write operation for the Amazon S3 data object. You can then add the data object read or write operation to a mapping.

Data Encryption in Amazon S3 Targets

To protect data, you can enable server-side encryption or client-side encryption to encrypt data inserted in Amazon S3 buckets.

Server-side Encryption

Enable server-side encryption if you want Amazon S3 to encrypt the data while uploading the files to the buckets. To enable server-side encryption, select **Server Side Encryption** as the encryption type in the advanced properties of the data object write operation. Server-side encryption uses Amazon S3-managed keys (SSE-S3) as the encryption type.

Client-side Encryption

Enable client-side encryption if you want the Data Integration Service to encrypt the data while uploading the files to the buckets. Client-side encryption uses client-side master key as the encryption type. To enable client-side encryption, perform the following tasks:

1. Ensure that an organization administrator creates a master symmetric key, which is a 256-bit AES encryption key in Base64 format.
2. Provide the master symmetric key when you create an Amazon S3 connection.
3. Select **Client Side Encryption** as the encryption type in the advanced properties of the data object write operation.
4. Ensure that an organization administrator updates the security JAR files, required by the Amazon S3 client encryption policy, on the machine that hosts the Data Integration Service.

The following table lists the encryption type for the supported environments:

Encryption Type	Native Environment	Blaze Environment
Server-side Encryption	Yes	Yes
Client-side Encryption	Yes	No

For information about the Amazon S3 client encryption policy, see the *Amazon S3 documentation*.

Hadoop Performance Tuning Options for EMR Distribution

You can use Hadoop Performance Tuning Options to optimize performance in Amazon EMR distribution when you copy large volumes of data between Amazon S3 buckets and HDFS.

You must provide semicolon separated name-value attribute pairs for Hadoop Performance Tuning Options.

You can use the following parameters for Hadoop Performance Tuning Options:

- `mapreduce.map.java.opts`
- `fs.s3a.fast.upload`
- `fs.s3a.multipartthreshold`
- `fs.s3a.multipartsize`

- `mapreduce.map.memory.mb`

The following sample shows the recommended values for the parameters:

```
mapreduce.map.java.opts=-
Xmx4096m;fs.s3a.fast.upload=true;fs.s3a.multipart.threshold=33554432;fs.s3a.multipart.size=33
554432;mapreduce.map.memory.mb=4096
```

Amazon S3 Data Object Properties

Specify the data object properties when you create the data object.
The following table describes the properties that you configure for the Amazon S3 data objects:

Property	Description
Name	Name of the Amazon S3 data object.
Location	The project or folder in the Model Repository Service where you want to store the Amazon S3 data object.
Connection	Name of the Amazon S3 connection.

Amazon S3 Data Object Read Operation Properties

Amazon S3 data object read operation properties include run-time properties that apply to the Amazon S3 data object.

The Developer tool displays advanced properties for the Amazon S3 data object operation in the Advanced view. The following table describes the Advanced properties for an Amazon S3 data object read operation:

Property	Description
File Name	Name of the Amazon S3 resource file from which you want to read data.
Folder Path	Bucket name that contains the Amazon S3 source file. If applicable, include the folder name that contains the source file in the `bucket_name>/<folder_name>` format.
Download S3 File in Multiple Parts	Download large Amazon S3 objects in multiple parts. When the file size of an Amazon S3 object is greater than 8 MB, you can choose to download the object in multiple parts in parallel.

Property	Description
Staging Directory	Amazon S3 staging directory. Applicable to the native environment. Ensure that the user has write permissions on the directory. In addition, ensure that there is sufficient space to enable staging of the entire file. Default staging directory is the `/temp` directory on the machine that hosts the Data Integration Service.
Hadoop Performance Tuning Options	Applicable to the Amazon EMR cluster. Provide semicolon separated name-value attribute pairs to optimize performance when you copy large volumes of data between Amazon S3 and HDFS . For more information see," Hadoop Performance Tuning Options for EMR Distribution" on page 17.

Amazon S3 Data Object Write Operation Properties

Amazon S3 data object write operation properties include run-time properties that apply to the Amazon S3 data object.

The Developer tool displays advanced properties for the Amazon S3 data object operation in the Advanced view. The following table describes the Advanced properties for an Amazon S3 data object write operation:

Property	Description
File Name	Name of the Amazon S3 resource file to which you want to write the source data.
Folder Path	Bucket name that contains the target file. If applicable, include the folder name that contains the target file in the `bucket_name>/<folder_name>` format.
Encryption Type	Method you want to use to encrypt data. Select one of the following values:- - None. The data is not encrypted. - Client Side Encryption. The Data Integration Service uses the master symmetric key you specify in the Amazon S3 connection properties to encrypt data. - Server Side Encryption. Amazon S3 encrypts data while uploading the files to Amazon buckets.
Staging Directory	Amazon S3 staging directory. Applicable to the native environment. Ensure that the user has write permissions on the directory. In addition, ensure that there is sufficient space to enable staging of the entire file. Default staging directory is the `/temp` directory on the machine that hosts the Data Integration Service.
File Merge	Applicable to the Hadoop environment. Enable File Merge to merge the target files into a single file.
Hadoop Performance Tuning Options	Applicable to the Amazon EMR cluster. Provide semicolon separated name-value attribute pairs to optimize performance when you copy large volumes of data between Amazon S3 and HDFS. For more information see," Hadoop Performance Tuning Options for EMR Distribution" on page 17.

Creating an Amazon S3 Data Object

Creata an Amazon S3 data object to add to a mapping.

1. Select a project or folder in the **Object Explorer** view.
2. Click **File** > **New** > **Data Object**.
3. Select **Amazon S3 Data Object** and click **Next**.

   The **Amazon S3 Data Object** dialog box appears.
4. Enter a name for the data object.
5. Click **Browse** next to the **Location** option and select the target project or folder.
6. Click **Browse** next to the **Connection** option and select the Amazon S3 connection from which you want to import the Amazon S3 object.
7. To add a resource, click **Add** next to the **Selected Resources** option.

   The **Add Resource** dialog box appears.
8. Select the check box next to the Amazon S3 object you want to add and click **OK**.
9. Click **Finish**.
   The data object appears under Data Objects in the project or folder in the **Object Explorer** view.

Creating a Data Object Operation

You can create the data object read or write operation for an Amaazon S3 data object. You can then add the Amazon S3 data object operation to a mapping.

1. Select the data object in the **Object Explorer** view.
2. Right-click and select **New** > **Data Object Operation**.

   The **Data Object Operation** dialog box appears.
3. Enter a name for the data object operation.
4. Select the type of data object operation. You can choose to create a read or write operation.
5. Click **Add**.

   The **Select Resources** dialog box appears.
6. Select the Amazon S3 data object for which you want to create the data object operation and click **OK**.
7. Click **Finish**.

The Developer tool creates the data object operation for the selected data object.

Configuring Column Projection

After you create a data object operation, you can project the columns as a binary data type, sample the metadata of an Amazon S3 file and project the columns, or manually project the columns.

Projecting Binary Columns

Perform the following steps to project columns as a binary data type:

1. Go to **Column Projection** tab.
2. Clear the **Enable Column Projection** field.

   The columns appear as a binary data type.

Sampling Metadata

Perform the following steps to sample metadata file and project columns:

1. Go to **Column Projection** tab.
2. Click **Edit Column Projection**.
3. Select **Reconfigure**.

   The **Column Projection** page appears.
4. Choose **Sample Metadata File**.

   You can click **Browse** and navigate to the directory that contains the file.
5. Select a code page in **Code page** field.

   The page matches the code page of the data that you want to process.

   Note: The **Delimited** and **Fixed-width** format properties are not applicable for PowerExchange for Amazon S3.
6. Click **Next**.
7. Configure the format properties.

Property	Description
Delimiters	Character used to separate columns of data. If you enter a delimiter that is the same as the escape character or the text qualifier, you might receive unexpected results. Amazon S3 reader and writer support Delimiters.
Text Qualifier	Quote character that defines the boundaries of text strings. If you select a quote character, the Developer tool ignores delimiters within pairs of quotes. Amazon S3 reader supports Text Qualifier.
Import Column Names From First Line	If selected, the Developer tool uses data in the first row for column names. Select this option if column names appear in the first row. The Developer tool prefixes"FIELD_" to field names that are not valid. Amazon S3 reader and writer support Import Column Names From First Line.
Escape Character	Character immediately preceding a column delimiter character embedded in an unquoted string, or immediately preceding the quote character in a quoted string. When you specify an escape character, the Data Integration Service reads the delimiter character as a regular character.

   Note: The **Start import at line**, **Treat consecutive delimters as one**, and **Retain escape character in data** properties in the **Column Projection** dialog box are not applicable for PowerExchange for Amazon S3.
8. Click **Next** to preview the flat file data object.

   You must change the data types to string manually.
9. Click **Finish**.

Projecting Columns Manually

After sampling the metadata, you can manually edit the projected columns.

Perform the following steps to project columns manually:

1. Go to **Column Projection** tab.
2. Click **Edit Column Projection**.
3. Click **New** icon and add fields manually.

CHAPTER 5

PowerExchange for Amazon S3 Mappings

This chapter includes the following topics:

PowerExchange for Amazon S3 Mappings Overview

After you create an Amazon S3 data object read or write operation, you can create a mapping.

You can create an Informatica mapping containing an Amazon S3 data object read operation as the input, and a relational or flat file data object operation as the target. You can create an Informatica mapping containing objects such as a relational or flat file data object operation as the input, transformations, and an Amazon S3 data object write operation as the output to load data to Amazon S3 buckets.

Validate and run the mapping. You can deploy the mapping and run it or add the mapping to a Mapping task in a workflow.

Mapping Validation and Run-time Environments

You can validate and run mappings in the native environment or in Blaze mode.

The Data Integration Service validates whether the mapping can run in the selected environment. You must validate the mapping for an environment before you run the mapping in that environment.

Native environment

You can configure the mappings to run in the native or Hadoop run-time environment. When you run mappings in the native environment, the Data Integration Service processes the mapping and runs the mapping from the Developer tool.

Blaze Mode

When you run mappings in the Blaze mode, the Data Integration Service pushes the mapping to a Hadoop cluster and processes the mapping on a Blaze engine. The Data Integration Service generates an execution plan to run mappings on the Blaze engine.

The Blaze engine execution plan simplifies the mapping into segments. The plan contains tasks to start the mapping, run the mapping, and create and cleanup the temporary tables and file required to run the mapping. The plan contains multiple tasklets and the task recovery strategy. The plan also contains pre and post grid task preparation commands for each mapping before running the main mapping on a Hadoop cluster. A pre-grid task can include a task such as copying data to HDFS. A post-grid task can include tasks such as cleaning up temporary files or copying data from HDFS.
You can view the plan in the Developer tool before you run the mapping and in the Administrator tool after you run the mapping. In the Developer tool, the Blaze engine execution plan appears as a workflow. You can click on each component in the workflow to get the details. In the Administrator tool, the Blaze engine execution plan appears as a script.

For more information about the Hadoop environment and Blaze mode, see the *Informatica Big Data Management™ User Guide*.

APPENDIX A

Amazon S3 Datatype Reference

This appendix includes the following topics:

Datatype Reference Overview

When you run the session to read data from or write data to Amazon S3, the Data Integration Service converts the transformation data types to comparable native Amazon S3 data types.

Amazon S3 and Transformation Datatypes

The following table lists the Amazon S3 data types that the Data Integration Service supports and the corresponding transformation data types:

Amazon S3 Data Type	Transformation Data Type	Description
BIGINT	Bigint	Precision of 19 digits, scale of 0
NUMBER	Decimal	For transformations that support precision up to 28 digits, the precision is 1 to 28 digits, and the scale is 0 to 28. If you specify the precision greater than the maximum number of digits, the Data Integration Service converts decimal values to double in high precision mode.
STRING	String	1 to 104,857,600 characters
NSTRING	String	1 to 104,857,600 characters

INDEX