Informatica® PowerExchange for Amazon S3 10.2

User Guide

Informatica PowerExchange for Amazon S3 User Guide
10.2
September 2017

Publication Date: 2020-07-30

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 also connect to Amazon S3 buckets available in Virtual Private Cloud (VPC) through VPC endpoints.

You can run mappings in the native or Hadoop environment. Select the Blaze or Spark engines when you run mappings on the Hadoop environment.

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.

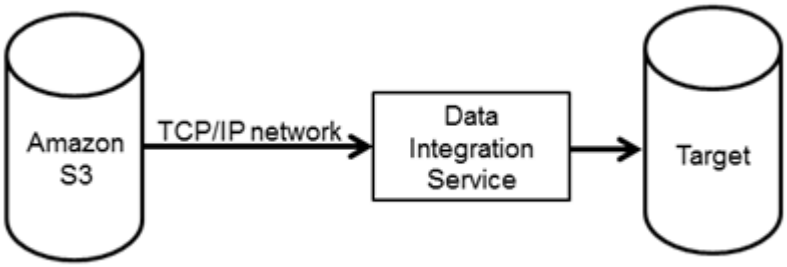
Amazon S3 stores data as objects within buckets. An object consists of a file and optionally any metadata that describes that file. To store an object in Amazon S3, you upload the file you want to store to a bucket. Buckets are the containers for objects. You can have one or more buckets. When using 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

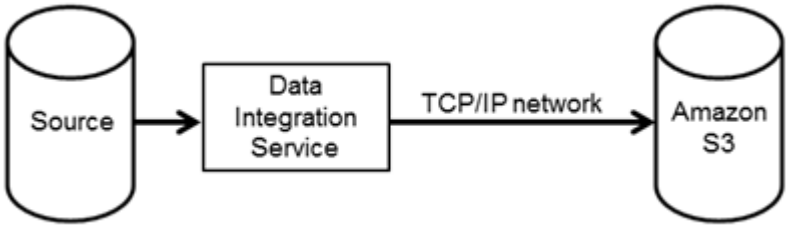
The following image shows how Informatica connects to Amazon S3 to read data:



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.
  For information about configuring user impersonation, see the Informatica Big Data Management™ Hadoop Integration Guide.
- To run mappings on MapR secure clusters, configure the MapR secure clusters on all the nodes.
  For information about configuring MapR secure clusters, see the Informatica Big Data Management™ Hadoop Integration Guide.

IAM Authentication

You can configure IAM authentication when the Data Integration Service runs on an Amazon Elastic Compute Cloud (EC2) system. Use IAM authentication for secure and controlled access to Amazon S3 resources when you run a session.

Note:

Use IAM authentication when you want to run a session on an EC2 system. Perform the following steps to configure IAM authentication:

1. Create Minimal Amazon S3 Bucket Policy. For more information, see "Create Minimal Amazon S3 Bucket Policy" on page 11.
2. Create the Amazon EC2 role. The Amazon EC2 role is used when you create an EC2 system in the S3 bucket. For more information about creating the Amazon EC2 role, see the AWS documentation.
3. Create an EC2 instance. Assign the Amazon EC2 role that you created in step #2 to the EC2 instance.
4. Install the Data Integration Service on the EC2 system.

You can use AWS IAM authentication when you run a mapping in the EMR cluster. To use AWS IAM authentication in the EMR cluster, you must create the Amazon EMR Role. Create a new Amazon EMR Role or use the default Amazon EMR Role. You must assign the Amazon ERM Role to the EMR cluster for secure access to Amazon S3 resources.

Note: Before you configure IAM Role with EMR cluster, you must install the Informatica Services on an EC2 instance with the IAM Roles assigned.

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
- DeleteObject
- ListBucket

Sample Policy

```
 { "Effect": "Allow", "Action": [ "s3:PutObject", "s3:GetObject", "s3:DeleteObject",
"s3:ListBucket", "s3:GetBucketPolicy" ], "Resource":
[ "arn:aws:s3:::<specify_bucket_name>/*", "arn:aws:s3:::<specify_bucket_name>" ] }
```

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.

You can use AWS Identity and Access Management (IAM) authentication to securely control access to Amazon S3 resources. If you have valid AWS credentials and you want to use IAM authentication, you do not have to specify the access key and secret key when you create an Amazon S3 connection.

When you run a mapping that reads data from an Amazon S3 source and writes data to an Amazon S3 target on the Spark engine, the mapping fails if the AWS credentials such as Access Key or Secret Key are different for source and target.

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.
Location	The domain where you want to create the connection.
Type	The Amazon S3 connection type.
Access Key	The access key ID for access to Amazon account resources. Required if you do not use AWS Identity and Access Management (IAM) authentication.
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. Required if you do not use AWS Identity and Access Management (IAM) authentication.
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>`.

Property	Description
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. Select one of the following regions: - Asia Pacific (Mumbai) - Asia Pacific (Seoul) - Asia Pacific (Singapore) - Asia Pacific (Sydney) - Asia Pacific (Tokyo) - Canada (Central) - China (Beijing) - EU (Ireland) - EU (Frankfurt) - EU (London) - South America (Sao Paulo) - US East (Ohio) - US East (N. Virginia) - US West (N. California) - US West (Oregon) Default is US East (N. Virginia).

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. PowerExchange for Amazon S3 creates the data object read operation and data object write operation for the Amazon S3 data object automatically. You can edit the advanced properties of the data object read or write operation and run a mapping.

Note: To view the list of files available in a bucket, you must select the bucket name instead of expanding the bucket name list in the **Object Explorer** view.

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.
Resource Format	You can create an Amazon S3 file data object from the following data source in Amazon S3: - Avro - Binary - Flat - Parquet You must choose the appropriate source format to read data from the source. Default is binary. The Avro and Parquet formats are not applicable on the Blaze engine.

Amazon S3 Data Object Read Operation

Create a mapping with an Amazon S3 data object read operation to read data from Amazon S3.

You can download Amazon S3 files in multiple parts, specify the location of the staging directory, and compress the data when you read data from Amazon S3.

Directory Source in Amazon S3 Sources

You can select the type of source from which you want to read data.

You can select the following type of sources from the **Source Type** option under the advanced properties for an Amazon S3 data object read operation:

- File
- Directory

Note: This option is applicable when you run a mapping in the native environment or on the Spark engine.

You must select the source file during the data object creation to select the source type as **Directory** at the run time. PowerExchange for Amazon S3 provides the option to override the value of the **Folder Path** and **File Name** properties during run time. When you select the **Source Type** option as **Directory**, the value of the **File Name** is not honored.

For read operation, if you provide the **Folder Path** value during run time, the Data Integration Service considers the value of the **Folder Path** from the data object read operation properties. If you do not provide the **Folder Path** value during run time, the Data Integration Service considers the value of the **Folder Path** that you specify during the data object creation.

Use the following rules and guidelines to select **Directory** as the source type:

- All the source files in the directory must contain the same metadata.
- All the files must have data in the same format. For example, delimiters, header fields, and escape characters must be same.

- All the files under a specified directory are parsed. The files under subdirectories are not parsed.

When you run a mapping to read multiple files and if the Amazon S3 data object is defined using file with header option on the Spark engine, the mapping runs successfully. However, the Data Integration Service does not generate a validation error for the files with no header.

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
Source Type	Select the type of source from which you want to read data. You can select the following source types: - File - Directory Default is **File**. Applicable when you run a mapping in the native environment or on the Spark engine. For more information, see "Directory Source in Amazon S3 Sources" on page 16.
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. If you do not provide the bucket name and specify the folder path starting with a slash (/) in the `/<folder_name>` format, the folder path appends with the folder path that you specified in the connection properties. For example, if you specify the `<my_bucket1>/<dir1>` folder path in the connection property and `/<dir2>` folder path in this property, the folder path appends with the folder path that you specified in the connection properties in `<my_bucket1>/<dir1>/<dir2>` format. If you specify the `<my_bucket1>/<dir1>` folder path in the connection property and `<my_bucket2>/<dir2>` folder path in this property, the Data Integration Service reads the file from the `<my_bucket2>/<dir2>` folder path that you specify in this property.
File Name	Name of the Amazon S3 file from which you want to read data.
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. Applicable to the Blaze and Spark engine. By default, the Data Integration Service downloads the file in multiple part.
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.

Property	Description
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 23.
Compression Format	Decompresses data when you read data from Amazon S3. You can decompress the data in the following formats: - None - Gzip - Bzip2 - Lzo Default is None. Applicable when you run a mapping in the native environment or on the Spark engine. The gzip compression format is applicable when you run a mapping in the native environment. **Note:** When you read an Avro file, you can decompress the file using the none compression format. When you read a Parquet file, you can decompress the file using the none, gzip, and lzo compression formats. After you decompress the Avro and Parquet files, you can read the files without using any compression format. For more information, see "Data Compression in Amazon S3 Sources and Targets" on page 22.

Column Projection Properties

The Developer tool displays the column projection properties for Avro and Parquet Amazon S3 file sources in the **Properties** view of the **Read** operation.

The following table describes the column projection properties that you configure for Avro and Parquet Amazon S3 file sources:

Property	Description
Enable Column Projection	Displays the column details of Avro or Parquet Amazon S3 files sources.
Schema Format	Displays the schema format that you selected while creating the Amazon S3 file data object. You can change the schema format and provide respective schema.
Schema	Displays the schema associated with the Avro or Parquet file. You can select a different schema. **Note:** If you disable the column projection, the schema associated with the Avro or Parquet file is removed. If you want to associate schema again with the Avro or Parquet file, enable the column projection and click **Select Schema**.
Column Mapping	Displays the mapping between input and output ports. **Note:** If you disable the column projection, the mapping between input and output ports is removed. If you want to map the input and output ports, enable the column projection and click **Select Schema** to associate a schema to the Avro or Parquet file.

Amazon S3 Data Object Write Operation

Create a mapping to write data to Amazon S3. Change the connection to an Amazon S3 connection, and define the write operation properties to write data to Amazon S3.

There is no control over the number of files created or file names written to the directory on the Spark engine. The Data Integration Service writes data to multiple files based on the source or source file size to the directory provided. You must provide the target file name and based on target file name, the Data Integration Service adds suffix characters such as MapReduce or Split information to the target file name.

If the file size is greater than 256 MB, the Data Integration Service creates multiple files inside the target folder. For example, `output.txt-m-00000`, `output.txt-m-00001`, and `output.txt-m-00002`.

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 support for various environments:

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

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

Overwriting Existing Files

You can choose to overwrite the existing files.

Select the **Overwrite File(s) If Exists** option in the Amazon S3 data object write operation properties to overwrite the existing files. By default, the value of the **Overwrite File(s) If Exists** check box is true.

If you select the **Overwrite File(s) If Exists** option, the Data Integration Service deletes the existing files with same file name and creates a new files with the same file name in the target directory.

If you do not select the **Overwrite File(s) If Exists** option, the Data Integration Service does not delete the existing files in the target directory. The Data Integration Service adds time stamp at the end of each target file name in the following format: `YYYYMMDD_HHMMSS_millisecond`. For example, the Data Integration Service renames the target file name in the following format: `output.txt-20170413_220348_164-m-00000`.

If you select the **Overwrite File(s) If Exists** option on the Spark engine, the Data Integration Service splits the existing files into multiple files with same file name. Then the Data Integration Service deletes the split files and creates new files in the target directory.

When you select the **Overwrite File(s) If Exists** option to overwrite an Avro file on the Spark engine, the Data Integration Service overwrites the existing file and appends `_avro` to the folder name. For example, `targetfile_avro`

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.

Note: By default, the Data Integration Service uploads the Amazon S3 file in multiple parts.

The following table describes the Advanced properties for an Amazon S3 data object write operation:

Property	Description
Folder Path	Bucket name that contains the Amazon S3 target file. If applicable, include the folder name that contains the target file in the `<bucket_name>/<folder_name>` format. If you do not provide the bucket name and specify the folder path starting with a slash (/) in the `/<folder_name>` format, the folder path appends with the folder path that you specified in the connection properties. For example, if you specify the `<my_bucket1>/<dir1>` folder path in the connection property and `/<dir2>` folder path in this property, the folder path appends with the folder path that you specified in the connection properties in `<my_bucket1>/<dir1>/<dir2>` format. If you specify the `<my_bucket1>/<dir1>` folder path in the connection property and `<my_bucket2>/<dir2>` folder path in this property, the Data Integration Service writes the file in the `<my_bucket2>/<dir2>` folder path that you specify in this property.
File Name	Name of the Amazon S3 file to which you want to write the source data.
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.

Property	Description
File Merge	Enable File Merge to merge the target files into a single file. Applicable when you run a mapping on the Blaze engine.
Hadoop Performance Tuning Options	Provide semicolon separated name-value attribute pairs to optimize performance when you copy large volumes of data between Amazon S3 and HDFS. Applicable to the Amazon EMR cluster. For more information, see " Hadoop Performance Tuning Options for EMR Distribution" on page 23.
Compression Format	Compresses data when you write data to Amazon S3. You can compress the data in the following formats: - None - Deflate - Gzip - Bzip2 - Lzo - Snappy Default is None. Applicable when you run a mapping in the native environment or on the Spark engine. The gzip compression format is applicable when you run a mapping in the native environment. **Note:** When you write an Avro file, you can compress the file using the none, deflate, and snappy compression formats. When you read a Parquet file, you can compress the file using the none, gzip, lzo, and snappy compression formats. For more information, see "Data Compression in Amazon S3 Sources and Targets" on page 22.
Overwrite File(s) If Exists	You can choose to overwrite the existing files. Select the check box if you want to overwrite the existing files. Default is true. For more information, see "Overwriting Existing Files" on page 19.

Column Projection Properties

The Developer tool displays the column projection properties for Avro and Parquet Amazon S3 file targets in the **Properties** view of the **Write** operation.

The following table describes the column projection properties that you configure for Avro and Parquet Amazon S3 file targets:

Property	Description
Enable Column Projection	Displays the column details of Avro or Parquet Amazon S3 files targets.
Schema Format	Displays the schema format that you selected while creating the Amazon S3 file data object. You can change the schema format and provide respective schema.
Schema	Displays the schema associated with the Avro or Parquet file. You can select a different schema. **Note:** If you disable the column projection, the schema associated with the Avro or Parquet file is removed. If you want to associate schema again with the Avro or Parquet file, enable the column projection and click **Select Schema**.
Column Mapping	Displays the mapping between input and output ports. **Note:** If you disable the column projection, the mapping between input and output ports is removed. If you want to map the input and output ports, enable the column projection and click **Select Schema** to associate a schema to the Avro or Parquet file.

Data Compression in Amazon S3 Sources and Targets

You can decompress the data when you read data from Amazon S3 or compress data when you write data to Amazon S3.

Configure the compression format in the **Compression Format** option under the advanced properties for an Amazon S3 data object read and write operation. The source or target file in Amazon S3 contains the same extension that you select in the **Compression Format** option. When you perform a read operation, the Data Integration Service decompresses the data and then sends the data to Amazon S3 bucket. When you perform a write operation, the Data Integration Service compresses the data.

Note: Data Compression is applicable when you run a mapping in the native environment or on the Spark engine.

The following table lists the compression formats for the support for various operations and file formats in the native environment or on the Spark engine:

Compression format	Read	Write	Avro File	Parquet File
None	Yes	Yes	Yes	Yes
Deflate	No	Yes	Yes	No
Gzip	Yes	Yes	No	Yes
Bzip2	Yes	Yes	No	No
Lzo	Yes	Yes	No	Yes
Snappy	No	Yes	Yes	Yes

Note: After you compress the Avro and Parquet files, you can read the files without using any compression format.

You can compress a flat file in the none and gzip compression format when you run a mapping in the native environment. You can compress a flat file in the none, gzip, bzip2, and lzo compression format when you run a mapping on the Spark engine.

To read a compressed file from Amazon S3 on the Spark engine, the compressed file must have specific extensions. If the extensions used to read the compressed file are not specific or not valid, the Data Integration Service does not process the file. The following table describes the extensions that are appended based on the compression format that you use:

Compression Format	File Name Extension
Gzip	.GZ
Deflate	.deflate
Bzip2	.BZ2

Compression Format	File Name Extension
Lzo	.LZO
Snappy	.snappy

Configuring Lzo Compression Format

To write the `.jar` files in the lzo compression format on the Spark engine, you must copy the `.jar` files for the lzo compression format in the `lib` folder of the distribution directory to the Data Integration Service.

Perform the following steps to copy the `.jar` files from the distribution directory to the Data Integration Service:

1. Copy the `lzo.jar` file from the cluster to the `<Informatica installation directory>/<distribution>/lib` directory on the Data Integration Service.
2. Copy the lzo native binaries from the cluster to the `<Informatica installation directory>/<distribution>/lib/native` directory on the Data Integration Service.

Hadoop Performance Tuning Options for EMR Distribution

You can use Hadoop Performance Tuning Options to optimize the performance in the 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.

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 parameter:

```
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
```

Creating an Amazon S3 Data Operations

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

Note: PowerExchange for Amazon S3 supports only UTF-8 encoding to read or write data.

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. In the Resource Format list, select any of the following formats:
   - Binary: to read any resource format.
   - Flat: to read a flat resource.
   - Avro: to read an Avro resource.
   - Parquet: to read a Parquet resource.
6. Click **Browse** next to the **Location** option and select the target project or folder.
7. Click **Browse** next to the **Connection** option and select the Amazon S3 connection from which you want to import the Amazon S3 object.
8. To add a resource, click **Add** next to the **Selected Resources** option.

   The **Add Resource** dialog box appears.
9. Select the check box next to the Amazon S3 object you want to add and click **OK**.
10. Click **Next**.
11. Choose **Sample Metadata File**.

    You can click **Browse** and navigate to the directory that contains the file.

    Note: The **Delimited** and **Fixed-width** format properties are not applicable for PowerExchange for Amazon S3.
12. Click **Next**.
13. 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.

Property	Description
Row Delimiter	Specify a line break character. Select from the list or enter a character. Preface an octal code with a backslash (\). To use a single character, enter the character. The Data Integration Service uses only the first character when the entry is not preceded by a backslash. The character must be a single-byte character, and no other character in the code page can contain that byte. Default is line-feed, \012 LF (\n).
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.

14. Click **Next** to preview the flat file data object.
15. Click **Finish**.

    The data object appears under the Physical Data Objects category in the project or folder in the **Object Explorer** view. A read and write operation is created for the data object. Depending on whether you want to use the Amazon S3 data object as a source or target, you can edit the read or write operation properties.

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.

Filtering Metadata

You can filter the metadata to optimize the search performance.

1. Select a project or folder in the **Object Explorer** view.
2. Select an Amazon S3 data object and click **Add**.
3. Click **Next**.
4. Click **Add** next to the **Selected Resources** option.

   The **Add Resource** dialog box appears.
5. Select the bucket or the folder from where you want to search the data.
6. Type the name of the file or any regular expressions in the **Name** field to search for the metadata available in the selected bucket or the folder in the following format: `abc*` or `[0-9]*`.
7. Click **Go**.

   The list of all the file names starting with the alphabet or the number that you entered in the **Name** field is displayed.

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, Blaze, or Spark engine.

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 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 Engine

When you run mappings on the Blaze engine, 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.

Spark Engine

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

For more information about the Hadoop environment, Blaze, and Spark engines, see the *Informatica Big Data Management™ Administrator 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 Data Types

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

Avro Amazon S3 File Data Types and Transformation Data Types

Avro Amazon S3 file data types map to transformation data types that the Data Integration Service uses to move data across platforms.

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

Amazon S3 File Data Type	Transformation Data Type	Range and Description
Boolean	Integer	TRUE (1) or FALSE (0)
Bytes	Binary	Precision 4000
Double	Double	Precision 15
Float	Double	Precision 15
Int	Integer	-2,147,483,648 to 2,147,483,647 Precision 10, scale 0
Long	Bigint	-9,223,372,036,854,775,808 to 9,223,372,036,854,775,807 Precision 19, scale 0
Null	Integer	-2,147,483,648 to 2,147,483,647 Precision 10, scale 0
String	String	-1 to 104,857,600 characters

Parquet Amazon S3 File Data Types and Transformation Data Types

Amazon S3 file data types map to transformation data types that the Data Integration Service uses to move data across platforms.

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

Amazon S3 File Data Type	Transformation Data Type	Range and Description
Boolean	Integer	TRUE (1) or FALSE (0)
Byte_Array	Binary	Arbitrarily long byte array

Amazon S3 File Data Type	Transformation Data Type	Range and Description
Double	Double	Precision 15
Float	Double	Precision 15 **Note:** Applicable when you run a mapping on the Spark engine.
Int32	Integer	-2,147,483,648 to +2,147,483,647
Int64	Bigint	-9,223,372,036,854,775,808 to +9,223,372,036,854,775,807 8-byte signed integer
Int96	Binary	12-byte signed integer **Note:** Applicable when you run a mapping on the Spark engine.

The Parquet schema that you specify to read or write a Parquet file must be in smaller case. Parquet does not support case-sensitive schema.

The Developer tool does not support the following Parquet data types:

- Int96 (TIMESTAMP_MILLIS)
- Date
- Timestamp

INDEX