



Informatica® Big Data Streaming
10.2.2

User Guide

© Copyright Informatica LLC 2016, 2019

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

U.S. GOVERNMENT RIGHTS Programs, software, databases, and related documentation and technical data delivered to U.S. Government customers are "commercial computer software" or "commercial technical data" pursuant to the applicable Federal Acquisition Regulation and agency-specific supplemental regulations. As such, the use, duplication, disclosure, modification, and adaptation is subject to the restrictions and license terms set forth in the applicable Government contract, and, to the extent applicable by the terms of the Government contract, the additional rights set forth in FAR 52.227-19, Commercial Computer Software License.

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

The information in this documentation is subject to change without notice. If you find any problems in this documentation, report them to us at infa_documentation@informatica.com.

Informatica products are warranted according to the terms and conditions of the agreements under which they are provided. INFORMATICA PROVIDES THE INFORMATION IN THIS DOCUMENT "AS IS" WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT.

Portions of this software and/or documentation are subject to copyright held by third parties. Required third party notices are included with the product.

Publication Date: 2019-07-09

Table of Contents

Preface	8
Object Missing.	8
Informatica Network.	8
Informatica Knowledge Base.	8
Informatica Documentation.	8
Informatica Product Availability Matrices.	9
Informatica Velocity.	9
Informatica Marketplace.	9
Informatica Global Customer Support.	9
 Chapter 1: Introduction to Big Data Streaming.....	 10
Big Data Streaming Overview.	10
Streaming Process.	11
Component Architecture.	12
Clients and Tools.	12
Application Services.	13
Repository.	13
Third-Party Applications.	13
Example	14
 Chapter 2: Big Data Streaming Administration.....	 15
Big Data Streaming Configuration Overview.	15
Verify Supported Distributions.	16
Support for Authentication Systems.	16
Configuration for Kerberized Kafka Clusters.	17
Configure the Default Realm.	17
Configure Java Authorization and Authentication Service (JAAS)	17
Configuration for Amazon Kinesis.	19
Create AWS Credentials.	19
Configuration for Amazon Kinesis Streams Sources.	19
Configuration for Amazon Kinesis Firehose Targets.	20
Configuration for Azure Event Hubs.	20
 Chapter 3: Sources in a Streaming Mapping.....	 21
Sources in a Streaming Mapping Overview.	21
Processing Hierarchical Data in Streaming Mappings.	22
Amazon Kinesis Data Objects.	23
Amazon Kinesis Data Object Overview Properties.	23
Amazon Kinesis Header Ports.	24
Amazon Kinesis Data Object Read Operation Properties.	24

Azure Event Hubs Data Objects.	28
Azure Event Hubs Data Object Overview Properties.	28
Azure Event Hubs Header Ports.	29
Azure Event Hubs Data Object Read Operation Properties.	29
HBase Data Objects.	32
HBase Data Object Read Operation Properties.	32
JMS Data Objects.	33
Integration with JMS.	33
JMS Message Structure.	33
JMS Data Object Overview Properties.	34
JMS Data Object Read Operation Properties.	35
Kafka Data Objects.	38
Kafka Data Object Overview Properties.	39
Kafka Header Ports.	40
Kafka Data Object Read Operation Properties.	41
MapR Streams Data Objects.	44
MapR Streams Object Overview Properties.	45
MapR Streams Data Object Read Operation Properties.	46
Chapter 4: Targets in a Streaming Mapping.	49
Targets in a Streaming Mapping Overview.	49
Processing Hierarchical Data in Streaming Mappings.	50
Amazon Kinesis Data Objects.	51
Amazon Kinesis Data Object Overview Properties.	52
Amazon Kinesis Data Object Write Operation Properties.	52
Azure Event Hubs Data Objects.	55
Azure Event Hubs Data Object Overview Properties.	55
Azure Event Hubs Header Ports.	56
Azure Event Hubs Data Object Write Operations.	56
Complex File Data Objects.	59
Complex File Data Object Overview Properties.	59
Compression and Decompression for Complex File Targets.	60
Complex File Data Object Write Operation Properties.	60
Complex File Execution Parameters.	63
Temporary Directory for the Complex File Target.	64
HBase Data Objects.	65
Data Object Column Configuration.	65
HBase Object Overview Properties.	66
HBase Data Object Write Operation Properties.	67
JMS Data Objects.	68
Integration with JMS.	68
JMS Message Structure.	68
JMS Data Object Overview Properties.	69

JMS Data Object Write Operation Properties.	70
Kafka Data Objects.	72
Kafka Data Object Overview Properties.	73
Kafka Header Ports.	74
Kafka Data Object Write Operation Properties.	74
Microsoft Azure Data Lake Store Data Object.	76
Microsoft Azure Data Lake Store Data Object Properties.	77
Microsoft Azure Data Lake Store Data Object Write Operation Properties.	77
MapR Streams Data Objects.	79
MapR Streams Object Overview Properties.	80
MapR Streams Data Object Write Operation Properties.	81
Relational Data Objects.	83
Relational Data Object Overview Properties.	83
Relational Data Object Write Operation Properties.	84
Chapter 5: Streaming Mappings.....	88
Streaming Mappings Overview.	88
Connections.	89
Data Objects.	89
Creating a Data Object.	90
Creating a Data Object Operation.	91
Transformations in a Streaming Mapping.	92
Stateful Computing.	94
Partitioning Configuration.	95
Example.	96
Rules in a Streaming Mapping.	96
Mapping Configurations.	97
Mapping Validation.	99
Environment Validation.	99
Data Object Validation.	99
Transformation Validation.	100
Run-time Validation.	101
Monitor Jobs.	101
Streaming Mapping Example.	102
High Availability Configuration.	102
Troubleshooting Streaming Mappings.	103
Chapter 6: Window Transformation.....	107
Window Transformation Overview.	107
Window Transformation Types.	107
Tumbling Window.	108
Sliding Window.	108
Window Transformation Port Properties.	109

Window Transformation Window Properties.	109
Tumbling Window Transformation Example.	110
Sliding Window Transformation Example.	110
Rules and Guidelines for the Window Transformations.	112

Appendix A: Connections..... 114

Connections Overview.	114
Amazon Kinesis Connection.	115
General Properties.	116
Connection Properties.	116
Creating an Amazon Kinesis Connection using Informatica Command Line Program.	117
Azure EventHub Connection.	118
General Properties.	118
Connection Properties.	119
Creating an Azure EventHub Connection using Informatica Command Line Program.	119
HBase Connection.	120
General Properties.	120
Connection Properties.	120
Creating an HBASE Connection using Informatica Command Line Program.	121
HDFS Connection.	121
General Properties.	121
Connection Properties.	122
Creating an HDFS Connection using Informatica Command Line Program.	122
Hive Connection.	122
Creating a Hive Connection using Informatica Command Line Program.	123
JMS Connection.	123
Prerequisites to Create a JMS Connection and a JMS Data Object.	123
Prerequisites to Use a JMS Connection and a JMS Data Object.	124
General Properties.	125
Connection Properties.	125
Creating a JMS Connection using Informatica Command Line Program.	126
JDBC Connection Properties.	126
Kafka Connection.	128
General Properties.	128
Kafka Broker Properties.	129
Creating a Kafka Connection using Informatica Command Line Program.	130
MapR Streams Connection.	130
General Properties.	131
Connection Properties.	131
Creating a MapR Streams Connection using Informatica Command Line Program.	131
Microsoft Azure Data Lake Store Connection.	132
Microsoft Azure Data Lake Store Connection Properties.	132

Appendix B: Sample Files.....	134
Sample Files.	134
Sample XSD File.	134
Sample JSON Schema.	135
Sample Avro Schema.	135
Sample Flat Format Schema.	136
Index.....	137

Preface

The Informatica Big Data Streaming User Guide provides information about how to configure and run streaming mappings on a Spark engine in a Hadoop environment.

Object Missing

This object is not available in the repository.

Informatica Network

The Informatica Network is the gateway to many resources, including the Informatica Knowledge Base and Informatica Global Customer Support. To enter the Informatica Network, visit <https://network.informatica.com>.

As an Informatica Network member, you have the following options:

- Search the Knowledge Base for product resources.
- View product availability information.
- Create and 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 find product resources such as how-to articles, best practices, video tutorials, and answers to frequently asked questions.

To search the Knowledge Base, visit <https://search.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

Use the Informatica Documentation Portal to explore an extensive library of documentation for current and recent product releases. To explore the Documentation Portal, visit <https://docs.informatica.com>.

Informatica maintains documentation for many products on the Informatica Knowledge Base in addition to the Documentation Portal. If you cannot find documentation for your product or product version on the Documentation Portal, search the Knowledge Base at <https://search.informatica.com>.

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

Informatica Product Availability Matrices

Product Availability Matrices (PAMs) indicate the versions of the operating systems, databases, and types of data sources and targets that a product release supports. You can browse the Informatica 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 and based on real-world experiences from hundreds of data management projects. Informatica Velocity represents the collective knowledge of Informatica consultants who work with organizations around the world to plan, develop, deploy, and maintain successful data management solutions.

You can find 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 extend and enhance your Informatica implementations. Leverage any of the hundreds of solutions from Informatica developers and partners on the Marketplace to improve your productivity and speed up time to implementation on your projects. You can find the Informatica Marketplace at <https://marketplace.informatica.com>.

Informatica Global Customer Support

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

To find your local Informatica Global Customer Support telephone number, visit the Informatica website at the following link:

<https://www.informatica.com/services-and-training/customer-success-services/contact-us.html>.

To find online support resources on the Informatica Network, visit <https://network.informatica.com> and select the eSupport option.

CHAPTER 1

Introduction to Big Data Streaming

This chapter includes the following topics:

- [Big Data Streaming Overview, 10](#)
- [Streaming Process, 11](#)
- [Component Architecture, 12](#)
- [Example , 14](#)

Big Data Streaming Overview

Use Informatica Big Data Streaming to prepare and process streams of data in real time and uncover insights in time to meet your business needs. Big Data Streaming provides pre-built connectors such as Kafka, Amazon Kinesis, HDFS, enterprise messaging systems, and data transformations to enable a code-free method of defining data integration logic.

Big Data Streaming builds on the best of open source technologies. It uses Spark Structured Streaming for stream processing, and supports other open source stream processing platforms and frameworks, such as Kafka and Hadoop. Spark Structured Streaming is a scalable and fault-tolerant open source stream processing engine built on the Spark engine.

You can create streaming mappings to stream machine, device, and social media data in the form of messages. Streaming mappings collect machine, device, and social media data in the form of messages. The mapping builds the business logic for the data and pushes the logic to the Spark engine for processing. Use a Messaging connection to get data from Apache Kafka brokers, Amazon Kinesis, and Azure Event Hubs.

The Spark engine runs the streaming mapping continuously. The Spark engine reads the data, divides the data into micro batches, processes it, updates the results to a result table, and then writes to a target.

You can stream the following types of data:

- Application and infrastructure log data
- Change data(CDC) from databases
- Clickstreams from web servers
- Geo-spatial data from devices
- Sensor data
- Time series data

- Supervisory Control And Data Acquisition (SCADA) data
- Message bus data
- Programmable logic controller (PLC) data
- Point of sale data from devices

You can stream data to different types of targets, such as Kafka, HDFS, Amazon Kinesis Firehose, HBase tables, Hive tables, JDBC-compliant databases, Microsoft Azure Event Hubs, and Azure Data Lake Store.

Big Data Streaming works with Informatica Big Data Management to provide streaming capabilities. Big Data Streaming uses Spark Structured Streaming to process streamed data. It uses YARN to manage the resources on a Spark cluster more efficiently and uses third-parties distributions to connect to and push job processing to a Hadoop environment.

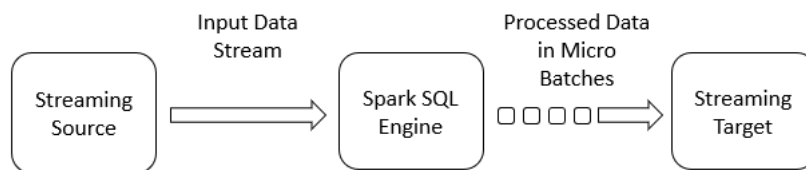
Use Informatica Developer (the Developer tool) to create streaming mappings. Use the Hadoop run-time environment and the Spark engine to run the mapping. You can configure high availability to run the streaming mappings on the Hadoop cluster.

For more information about running mappings on the Spark engine, see the *Informatica Big Data Management User Guide*.

Streaming Process

A streaming mapping receives data from unbounded data sources. An unbounded data source is one where data is continuously flowing in and there is no definite boundary. Sources stream data as events. The Spark engine processes the data and continuously updates the results to a result table.

The following image shows how the Spark engine receives data and publishes data in micro batches:



The Spark engine uses Spark Structured Streaming to process data that it receives in batches. Spark Structured Streaming receives data from streaming sources such as Kafka and divides the data into micro batches. The Spark engine continuously processes data streams as a series of small batch jobs with end-to-end latencies as low as 100 milliseconds with exactly-once fault tolerance guarantees.

For more information about Spark Structured Streaming, see the Apache Spark documentation at <https://spark.apache.org/documentation.html>.

You can perform the following high-level tasks in a streaming mapping:

1. Identify sources from which you need to stream data. You can access data that is in XML, JSON, Avro, flat, or binary format.
You can use Kafka Amazon Kinesis stream and Azure Event Hubs sources to connect to multiple big data sources.
2. Configure the mapping and mapping logic to transform the data.

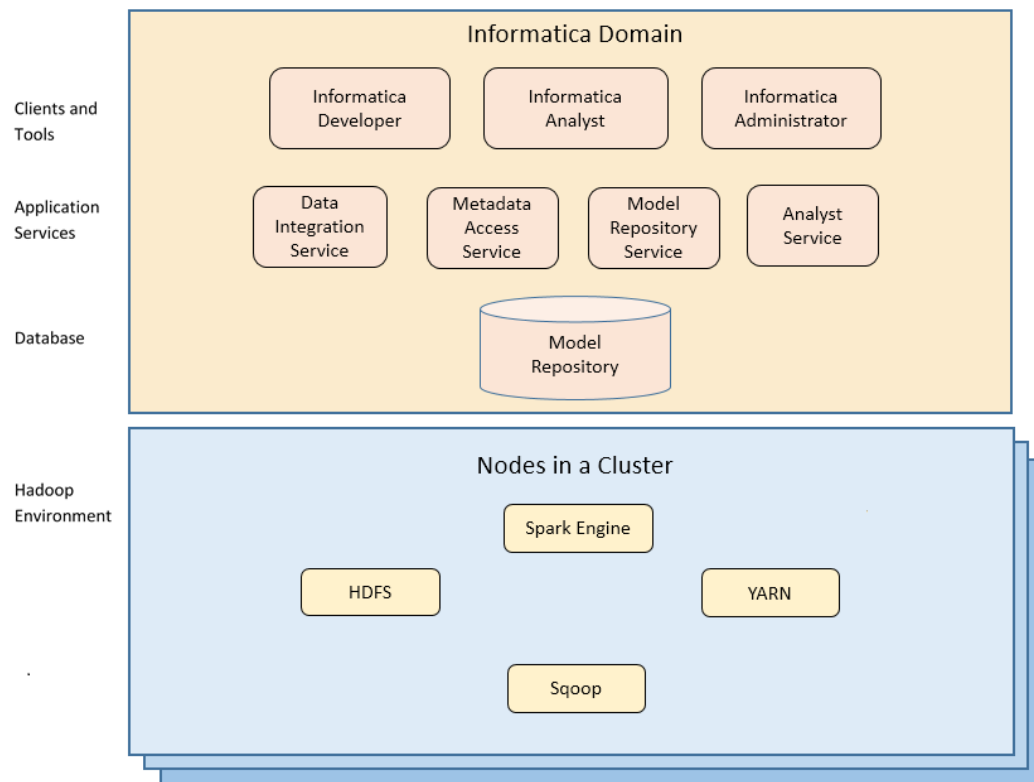
3. Run the mapping on the Spark engine in the Hadoop environment.
4. Write the data to Kafka targets, HDFS complex files, HBase, Azure Event Hubs, Azure Data Lake Store, JMS, and Kinesis Firehose delivery streams.
5. Monitor the status of your processing jobs. You can view monitoring statistics for your processing jobs in the Monitoring tool.

Note: Support for monitoring jobs is deferred. Support will be reinstated in a future release.

Component Architecture

The Big Data Streaming components for a streaming mapping include client tools, application services, repositories, and third-party tools.

The following image shows the components that Big Data Streaming uses for streaming mappings:



Clients and Tools

Based on your product license, you can use multiple Informatica tools and clients to manage streaming mappings.

Use the following tools to manage streaming mappings:

Informatica Administrator

Monitor the status of mappings on the Monitoring tab of the Administrator tool. The Monitoring tab of the Administrator tool is called the Monitoring tool.

Informatica Developer

Create and run mappings on the Spark engine from the Developer tool.

Informatica Analyst

Create rules in Informatica Analyst and run the rules as mapplets in a streaming mapping.

Application Services

Big Data Streaming uses application services in the Informatica domain to process data. The application services depend on the task you perform.

Big Data Streaming uses the following application services when you create and run streaming mappings:

Data Integration Service

The Data Integration Service processes mappings on the Spark engine in the Hadoop environment. The Data Integration Service retrieves metadata from the Model repository when you run a Developer tool mapping. The Developer tool connects to the Data Integration Service to run mappings.

Metadata Access Service

The Metadata Access Service is a user-managed service that provides metadata from a Hadoop cluster to the Developer tool at design time. HBase, HDFS, Hive, and MapR-DB connections use the Metadata Access Service when you import an object from a Hadoop cluster. Create and configure a Metadata Access Service before you create HBase, HDFS, Hive, MapR Streams, and MapR-DB connections.

Model Repository Service

The Model Repository Service manages the Model repository. The Model Repository Service connects to the Model repository when you run a mapping.

Analyst Service

The Analyst Service runs the Analyst tool in the Informatica domain. The Analyst Service manages the connections between service components and the users that have access to the Analyst tool.

Repository

Big Data Streaming includes a repository to store data related to connections and source metadata. Big Data Streaming uses application services in the Informatica domain to access data in the repository.

Big Data Streaming stores Structured Streaming mappings in the Model repository. You can manage the Model repository in the Developer tool.

Third-Party Applications

Big Data Streaming uses third-parties distributions to connect to a Spark engine on a Hadoop cluster.

Big Data Streaming pushes job processing to the Spark engine. It uses YARN to manage the resources on a Spark cluster more efficiently.

Example

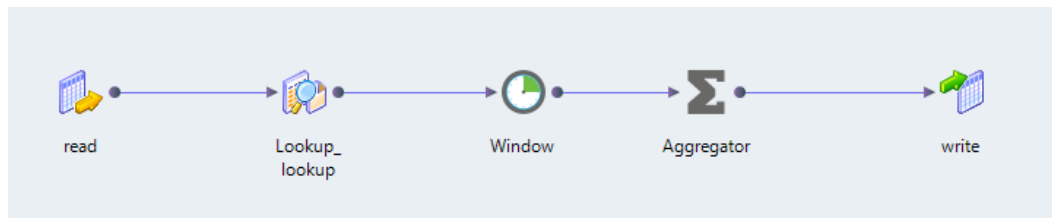
You run the IT department of a major bank that has millions of customers. You want to monitor network activity in real time. You need to collect network activity data from various sources such as firewalls or network devices to improve security and prevent attacks. The network activity data includes Denial of Service (DoS) attacks and failed login attempts made by customers. The network activity data is written to Kafka queues.

You perform the following tasks:

1. Create a streaming mapping to read data from the Kafka queues that stream data in JSON, XML, CSV, or Avro formats.
2. Add a Lookup transformation to get data from a particular customer ID.
3. Add a Window transformation to accumulate the streamed data into data groups before processing the data.
4. Process the data. Add an Aggregator transformation to perform aggregations on the data from the customer ID.
5. Monitor jobs. Monitor statistics for the mapping job on the Monitoring tab of the Administrator tool.

Note: Support for monitoring jobs is deferred. Support will be reinstated in a future release.

The following image shows the mapping:



CHAPTER 2

Big Data Streaming Administration

This chapter includes the following topics:

- [Big Data Streaming Configuration Overview, 15](#)
- [Verify Supported Distributions, 16](#)
- [Support for Authentication Systems, 16](#)
- [Configuration for Kerberized Kafka Clusters, 17](#)
- [Configuration for Amazon Kinesis, 19](#)
- [Configuration for Azure Event Hubs, 20](#)

Big Data Streaming Configuration Overview

After you integrate the Informatica domain with the Hadoop environment, you can use Big Data Streaming in conjunction with Big Data Management.

Before you create streaming mappings perform the following tasks:

- Verify the Hadoop distribution versions that Big Data Streaming supports.
- Before you create and use JMS connections and JMS data objects in streaming mappings, complete the required tasks in [“Configure Java Authorization and Authentication Service \(JAAS\)” on page 17](#).

Note: Support for JMS source is deferred. Support will be reinstated in a future release.

- Before you read from or write to a Kerberized Kafka cluster, complete the required tasks in [“Configuration for Kerberized Kafka Clusters” on page 17](#).

For more information about integration tasks, see the *Informatica Big Data Management Integration Guide*.

Verify Supported Distributions

The following table lists the distribution versions that Big Data Streaming supports:

Distribution	Version
Amazon EMR	5.16
Azure HDInsight	3.6.x
Cloudera CDH	5.15 5.16
Hortonworks HDP	2.6.x
MapR	Support for MapR is deferred. Support will be reinstated in a future release.

For more information about the distribution versions, see the Product Availability Matrix on Informatica Network:

<https://network.informatica.com/community/informatica-network/product-availability-matrices/overview>

Support for Authentication Systems

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 authentication. The following table shows the Spark engine support for the security management system installed on the Hadoop platform:

Hadoop Distribution	Apache Knox	Kerberos	LDAP
Amazon EMR	Not supported	Supported	Not supported
Azure HDInsight	Not supported	Not supported	Not supported
Cloudera CDH	Not supported	Supported	Not supported
Hortonworks HDP	Not supported	Supported	Not supported

Note: Sqoop cannot access Kerberos-enabled databases.

Configuration for Kerberized Kafka Clusters

To read from or write to a Kerberized Kafka cluster, configure the default realm, KDC, Hadoop connection properties, and Kafka data object read or write data operation properties.

Configure the Default Realm

Before you read from or write to a Kerberized Kafka cluster, perform the following tasks:

1. Ensure that you have the `krb5.conf` file for the Kerberized Kafka server.
2. Configure the default realm and KDC. If the default `/etc/krb5.conf` file is not configured or you want to change the configuration, add the following lines to the `/etc/krb5.conf` file:

```
[libdefaults]
default_realm = <REALM NAME>
dns_lookup_realm = false
dns_lookup_kdc = false
ticket_lifetime = 24h
renew_lifetime = 7d
forwardable = true

[realms]
<REALM NAME> = {
kdc = <Location where KDC is installed>
admin_server = <Location where KDC is installed>
}

[domain_realm]
.<domain name or hostname> = <KERBEROS DOMAIN NAME>
<domain name or hostname> = <KERBEROS DOMAIN NAME>
```

Configure Java Authorization and Authentication Service (JAAS)

To pass a static JAAS configuration file into the JVM using the `java.security.auth.login.config` property at run time, perform the following tasks:

Use a Static JAAS Configuration File

1. Ensure that you have JAAS configuration file.
For information about creating JAAS configuration and configuring Keytab for Kafka clients, see the Apache Kafka documentation at <https://kafka.apache.org/0101/documentation/#security>

For example, the JAAS configuration file can contain the following lines of configuration:

```
//Kafka Client Authentication. Used for client to kafka broker connection
KafkaClient {
com.sun.security.auth.module.Krb5LoginModule required
doNotPrompt=true
useKeyTab=true
storeKey=true
keyTab="<path to keytab file>/<keytab file name>"
principal="<principal name>"
client=true
};
```

2. Place the JAAS config file and keytab file in the same location on all the nodes of the Hadoop cluster.
Put the files in a location that is accessible to all nodes on the cluster, such as `/etc` or `/temp`.

On the **Spark Engine** tab of the Hadoop connection properties, update the `extraJavaOptions` property of the executor and the driver in the **Advanced Properties** property. Click **Edit** and update the properties in the following format:

```
spark.executor.extraJavaOptions=-Djava.security.egd=file:/dev/./urandom
-XX:MaxMetaspaceSize=256M -Djavax.security.auth.useSubjectCredsOnly=true
-Djava.security.krb5.conf=/<path to krb5.conf file>/krb5.conf
-Djava.security.auth.login.config=/<path to jaas config>/<kafka_client_jaas>.config

spark.driver.cluster.mode.extraJavaOptions=-Djava.security.egd=file:/dev/./urandom
-XX:MaxMetaspaceSize=256M -Djavax.security.auth.useSubjectCredsOnly=true
-Djava.security.krb5.conf=/<path to krb5.conf file>/krb5.conf
-Djava.security.auth.login.config=<path to jaas config>/<kafka_client_jaas>.config
```

3. Configure the following properties in the data object read or write operation:

- Data object read operation. Configure the **Consumer Configuration Properties** property in the advanced properties.
- Data object write operation. Configure the **Producer Configuration Properties** property in the advanced properties.

Specify the following value:

```
security.protocol=SASL_PLAINTEXT,sasl.kerberos.service.name=kafka,sasl.mechanism=GSSAPI
```

Embed the JAAS Configuration

To embed the JAAS configuration in the `sasl.jaas.config` configuration property, perform the following tasks:

1. On the **Spark Engine** tab of the Hadoop connection properties, update the `extraJavaOptions` property of the executor and the driver in the **Advanced Properties** property. Click **Edit** and update the properties in the following format:

```
spark.executor.extraJavaOptions=-Djava.security.egd=file:/dev/./urandom
-XX:MaxMetaspaceSize=256M -XX:+UseG1GC -XX:MaxGCPauseMillis=500 -
Djava.security.krb5.conf=<path to krb5.conf file>

spark.driver.cluster.mode.extraJavaOptions=-Djava.security.egd=file:/dev/./urandom
-XX:MaxMetaspaceSize=256M -XX:+UseG1GC -XX:MaxGCPauseMillis=500 -
Djava.security.krb5.conf=<path to krb5.conf file>
```

2. Configure the following properties in the data object read or write operation:

- Data object read operation. Configure the **Consumer Configuration Properties** property in the advanced properties.
- Data object write operation. Configure the **Producer Configuration Properties** property in the advanced properties.

Specify the following value:

```
security.protocol=SASL_PLAINTEXT,sasl.kerberos.service.name=kafka,sasl.mechanism=GSSAPI,
sasl.jaas.config=com.sun.security.auth.module.Krb5LoginModule required
useKeyTab=true
storeKey=true doNotPrompt=true serviceName="<service_name>" keyTab="<location of
keytab file>"
client=true principal="<principal_name>;"
```

Configuration for Amazon Kinesis

To create an Amazon Kinesis connection, create AWS credentials for the users. AWS credentials consist of access key ID and secret access key. Use the Amazon Kinesis connection to access Amazon Kinesis Streams as sources or Amazon Kinesis Firehose as targets.

Create AWS Credentials

Create AWS credentials for the users, such as access key ID and secret access key. An authentication type can be selected by the user during the creation of an Amazon Kinesis connection, such as AWS credential profile and cross-account IAM role. The default authentication type is AWS credential profile.

Access Key and ID

Generate an Access Key ID and Secret Access Key for the users in AWS.

AWS Credential Profile

You can define AWS credential profiles in the credentials file. Each credential profile consists of secret access key and access key ID.

Users can use the AWS credential profile names to use different AWS credentials at run time than the AWS credentials that they specify when they create an Amazon Kinesis connection with an Amazon Kinesis Streams as a source and Amazon Kinesis Firehose as a target.

Place the credentials file in the same location on all the cluster nodes. Default location is `<yarn home directory>/aws/credentials`.

Note: AWS credential profile is used at run time only.

Cross-account IAM Role

You can create an IAM role to allow a user of an AWS account to access resources of another AWS account. IAM roles allow you to define a set of permissions to access the AWS resources.

Users can share resources in one AWS account with users in a different AWS account without the need to create individual IAM users in each AWS account.

Note: Cross-account IAM role is not supported for Amazon Kinesis Firehose as target.

Configuration for Amazon Kinesis Streams Sources

To use Amazon Kinesis Streams as sources, grant required permissions to the user. To use cross-account IAM role, create an IAM role and grant access to the role.

To configure access for Amazon Kinesis Streams as a source, perform the following tasks :

- Grant consumer permissions that are part of the IAM policy to the AWS credentials that the IAM user specifies in the access key id.
For the list of permissions, see the AWS documentation at <https://docs.aws.amazon.com/streams/latest/dev/learning-kinesis-module-one-iam.html>
- Grant the following permissions to the user to fetch metadata:
 - kinesis:DescribeStream
 - kinesis:GetShardIterator
 - kinesis:GetRecords

- To use cross-account IAM role, create an IAM role in an AWS account. IAM roles are used to provide secure access to AWS resources. The role is used to establish a trusted relationship between various AWS accounts. Additional restrictions are enforced using an external ID and the IAM role can only be assumed by using the external ID if the external ID is specified for a cross-account IAM role.

Note: Multi-factor authentication is not supported.

Configuration for Amazon Kinesis Firehose Targets

To use Amazon Kinesis Firehose as targets create an AWS account, grant the required permissions and privileges to the IAM role.

To configure access for Amazon Firehose as a target, perform the following tasks:

- Create an AWS account with the required IAM permissions for the IAM user to use the AWS services such as, Kinesis Firehose, S3, Redshift, and Elastic Search.
- Grant Redshift INSERT privilege to the IAM user if the user wants to copy data from the Amazon S3 bucket to the Redshift cluster.
- Define a Firehose Delivery Stream with either S3, Redshift or Elastic Search as its destination. Configure source as Direct PUT or other sources.
- Grant required permissions to the IAM user credentials based on the target the user is writing to. For a list of permissions, see the AWS documentation at <https://docs.aws.amazon.com/firehose/latest/dev/controlling-access.html#access-to-firehose>

For more information about Amazon Kinesis connection, see [“Amazon Kinesis Connection” on page 115](#).

Configuration for Azure Event Hubs

Before you create an Azure Event hub connection, verify the following prerequisites:

1. You must have a Microsoft Azure account with a minimum role of contributor.
2. Verify that you have an Active Directory application.
3. Verify that your Active Directory application has permissions for the following API and directory:
 - Windows Azure Service Management API
 - Windows Azure Active Directory
4. Verify that the Azure Active Directory application is added with a reader role to your Azure account subscription.
5. Verify that you have an EventHub Namespace.

For more information about Azure Event Hub connection, see [“Azure EventHub Connection” on page 118](#).

CHAPTER 3

Sources in a Streaming Mapping

This chapter includes the following topics:

- [Sources in a Streaming Mapping Overview, 21](#)
- [Processing Hierarchical Data in Streaming Mappings, 22](#)
- [Amazon Kinesis Data Objects, 23](#)
- [Azure Event Hubs Data Objects, 28](#)
- [HBase Data Objects, 32](#)
- [JMS Data Objects, 33](#)
- [Kafka Data Objects, 38](#)
- [MapR Streams Data Objects, 44](#)

Sources in a Streaming Mapping Overview

You can access log file data, sensor data, Supervisory Control And Data Acquisition (SCADA) data, message bus data, Programmable logic controller (PLC) data on the Spark SQL engine in the Hadoop environment.

You can create physical data objects to access the different types of data. Based on the type of source you are reading from, you can create the following data objects:

Amazon Kinesis

A physical data object that represents data in an Amazon Kinesis Stream. Create an Amazon Kinesis data object to read from an Amazon Kinesis Stream.

Azure Event Hubs

A physical data object that represents data in Microsoft Azure Event Hubs data streaming platform and event ingestion service.

HBase

A physical data object that represents data in an HBase resource. Create an HBase data object with a read operation to perform an uncached lookup on HBase data.

JMS

A physical data object that accesses a JMS server. Create a JMS data object to read from a JMS server.

Note: Support for JMS source is deferred. Support will be reinstated in a future release.

Kafka

A physical data object that accesses a Kafka broker. Create a Kafka data object to read from a Kafka broker.

MapR Streams

A MapR Streams data object is a physical data object that represents data in a MapR Stream. Create a MapR Streams data object to read from a MapR Stream.

Note: Support for MapR Streams data objects is deferred. Support will be reinstated in a future release.

Processing Hierarchical Data in Streaming Mappings

Data objects in a streaming mapping can process hierarchical data through complex data types. If the source contains hierarchical data, you must enable the read data operation to project columns as complex data types.

The following are the complex data types:

- Array. An array is an ordered collection of elements.
- Map. A map is an unordered collection of key-value pair elements.
- Struct. A struct is a collection of elements of different data types.

The following table shows the format and complex data types that sources in a Streaming mapping support:

Format	Schema Type	Amazon Kinesis Streams	Azure Event Hub	JMS	Kafka	MapR Streams
Avro	Flat	Supported	Supported	Not supported	Supported	Supported
Avro	Hierarchical	Supported	Supported	Not supported	Supported	Supported
Binary	Binary	Supported	Supported	Supported	Supported	Supported
Flat	Flat	Supported	Supported	Supported	Supported	Not Supported
JSON	Flat	Supported	Supported	Supported	Supported	Supported
JSON	Hierarchical	Supported	Supported	Supported	Supported	Supported
XML	Flat	Supported	Supported	Supported	Supported	Supported
XML	Hierarchical	Supported	Supported	Supported	Supported	Supported

For more information about processing hierarchical data, see the *Informatica Big Data Management User Guide*.

Amazon Kinesis Data Objects

An Amazon Kinesis data object is a physical data object that represents data in an Amazon Kinesis Data Stream. After you create an Amazon Kinesis connection, create an Amazon Kinesis data object to read from Amazon Kinesis Data Streams.

Kinesis Data Streams is a real-time data stream processing option that Amazon Kinesis offers within the AWS ecosystem. It is a customizable option for users who want to build custom applications to process and analyze streaming data. You must manually provision enough capacity to meet system needs.

When you configure the Amazon Kinesis data object, specify the name of the Amazon Kinesis Data Stream that you read from. After you create the data object, create a read operation to read data from an Amazon Kinesis Data Stream. You can then add the data object read operation as a source in streaming mappings.

When you configure the data operation properties, specify the format in which the data object reads data. When you read from Amazon Kinesis Data Stream sources, you can read data in JSON, XML, Avro, Flat, or binary format. When you specify XML format, you must provide a XSD file. When you specify Avro format, provide a sample Avro schema in a .avsc file. When you specify JSON or Flat format, you must provide a sample file.

You can associate the data object with an intelligent structure model and directly parse input from text, CSV, XML, or JSON input files, as well as PDF forms, Microsoft Word tables, or Microsoft Excel.

You can pass any payload format directly from source to target in Streaming mappings. You can project columns in binary format pass a payload from source to target in its original form or to pass a payload format that is not supported.

Streaming mappings can read, process, and write hierarchical data. You can use array, struct, and map complex data types to process the hierarchical data. You assign complex data types to ports in a mapping to flow hierarchical data. Ports that flow hierarchical data are called complex ports.

Note: You cannot run a mapping with an Amazon Kinesis data object on a MapR distribution.

For more information about processing hierarchical data, see the *Informatica Big Data Management User Guide*.

For more information about Kinesis Data Streams, see the Amazon Web Services documentation.

RELATED TOPICS:

- [“Creating a Data Object” on page 90](#)

Amazon Kinesis Data Object Overview Properties

Overview properties include general properties that apply to the Amazon Kinesis data object. The Developer tool displays overview properties of the data object in the **Overview** view.

You can configure the following overview properties for Amazon Kinesis data objects:

General

You can configure the following general properties for the Amazon Kinesis data object:

- **Name.** Name of the Amazon Kinesis operation.
- **Description.** Description of the Amazon Kinesis data object.
- **Native Name.** Name of the Amazon Kinesis data object.
- **Path Information.** The path of the data object in Amazon Kinesis. For example, `/DeliveryStreams/router1`

Column

You can configure the name, native name, data type, precision, access type, scale, and description of the columns in the Amazon Kinesis resource.

Advanced

The following are the advanced properties for the Amazon Kinesis data object:

- Amazon Resource Name. The Kinesis resource that the Amazon Kinesis data object is reading from or writing to.
- Type. The type of delivery stream that the Amazon Kinesis data object is reading from or writing to. The delivery stream is either Kinesis Stream or Firehose DeliveryStream
- Number of Shards. Specify the number of shards that the Kinesis Stream is composed of. This property is not applicable for Firehose DeliveryStream.

Amazon Kinesis Header Ports

An Amazon Kinesis data object contains a default header port that represents metadata associated with events.

An Amazon Kinesis data object contains the following header port:

timestamp

Time at which an event is generated. You can accumulate streamed data into data groups and then process the data groups based on the timestamp values.

Amazon Kinesis Data Object Read Operation Properties

The Data Integration Service uses read operation properties when it reads data from Amazon Kinesis Streams.

General Properties

The Developer tool displays general properties for Amazon Kinesis sources in the **Read** view.

The following table describes the general properties for the Amazon Kinesis data object read operation:

Property	Description
Name	The name of the Amazon Kinesis read operation. You can edit the name in the Overview view. When you use the Amazon Kinesis stream as a source in a mapping, you can edit the name in the mapping.
Description	The description of the Amazon Kinesis data object operation.

Ports Properties

Ports properties for a physical data object include port names and port attributes such as data type and precision.

The following table describes the ports properties that you configure for Amazon Kinesis stream sources:

Property	Description
Name	The name of the source.
Type	The native data type of the source.
Precision	The maximum number of significant digits for numeric data types, or the maximum number of characters for string data types.
Detail	The detail of the data type.
Scale	The scale of the data type.
Description	The description of the resource.

Sources Properties

The sources properties list the resources of the Amazon Kinesis data object.

The following table describes the sources property that you can configure for Amazon Kinesis Streams sources:

Property	Description
Sources	The sources which the Amazon Kinesis data object reads from. You can add or remove sources.

Run-time Properties

The run-time properties include properties that the Data Integration Service uses when reading data from the source at run time.

The run-time property for Amazon Kinesis Stream source includes the name of the Amazon Kinesis connection.

Advanced Properties

The following table describes the advanced properties for Amazon Kinesis Stream sources:

Property	Description
Operation Type	Specifies the type of data object operation. This is a read-only property.
Guaranteed Processing	Ensures that the mapping processes messages that the sources publish and delivers them to the targets at least once. In the event of a failure, there could be potential duplicates but the messages are processed successfully. If the external source or the target is not available, the mapping execution stops to avoid any data loss. By default, the Guaranteed Processing property is selected.
Degree of Parallelism	The number of processes that run in parallel within a shard. Specify a value that is less than or equal to the number of shards.

Column Projections Properties

The following table describes the columns projection properties that you configure for Amazon Kinesis Stream sources:

Property	Description
Column Name	The name field that contains data. This property is read-only.
Type	The native data type of the resource. This property is read-only.
Enable Column Projection	Indicates that you use a schema to read the data that the source streams. By default, the data is streamed in binary format. To change the format in which the data is processed, select this option and specify the schema format.
Schema Format	The format in which the source processes data. You can select one of the following formats: <ul style="list-style-type: none">- XML- JSON- Avro- Flat
Use Schema	Specify the XSD schema for the XML format, the sample JSON for the JSON format. Specify a .avsc file for the Avro format or a sample file for the Flat format.
Use Intelligent Structure Model	Displays the intelligent structure model associated with the complex file. You can select a different model. Note: If you disable the column projection, the intelligent structure model associated with the data object is removed. If you want to associate an intelligent structure model with the data object again, enable the column projection and click Select Model . For more information on intelligent structure models, see the <i>Informatica Big Data Management User Guide</i> .

Property	Description
Column Mapping	The mapping of source data to the data object. Click View to see the mapping.
Project Column as Complex Data Type	Project columns as complex data type for sources with hierarchical data. Select this option if the source has hierarchical data. For more information on hierarchical data, see the <i>Informatica Big Data Management User Guide</i> .

Configuring Scheme for Flat Files

Configure schema for flat files when you configure column projection properties.

- On the **Column Projection** tab, enable column projection and select the flat schema format.
The page displays the column projection properties page.
- On the column projection properties page, configure the following properties:
 - Sample Metadata File. Select a sample file.
 - Code page. Select the UTF-8 code page.
 - Format. Format in which the source processes data. Default value is Delimited. You cannot change it.
- Click **Next**.
- In the delimited format properties page, configure the following properties:

Property	Description
Delimiters	Specify the character that separates entries in the file. Default is a comma (.). You can only specify one delimiter at a time. If you select Other and specify a custom delimiter, you can only specify a single-character delimiter.
Text Qualifier	Specify the character used to enclose text that should be treated as one entry. Use a text qualifier to disregard the delimiter character within text. Default is No quotes. You can only specify an escape character of one character.
Preview Options	Specify the escape character. The row delimiter is not applicable as only one row is created at a time.
Maximum rows to preview	Specify the rows of data you want to preview.

- Click **Next** to preview the flat file data object.
If required, you can change the column attributes. The data type timestampWithTZ format is not supported.
- Click **Finish**.
The data object opens in the editor.

Azure Event Hubs Data Objects

An Azure Event Hubs data object is a physical data object that represents an event hub object in Microsoft Azure Event Hubs data streaming platform and event ingestion service. After you create an Azure Event Hubs connection, create an Azure Event Hubs data object to read from event data from event hub event.

Azure Event Hubs is a highly scalable data streaming platform and event ingestion service, that receives and processes events. Azure Event Hubs can process and store events or data produced by distributed software and devices.

When you configure the Azure Event Hubs data object, specify the name of the event hub that you read from. After you create the data object, create a read operation to read event data from an event hub. You can then add the data object read operation as a source in streaming mappings.

When you configure the data operation properties, specify the format in which the Azure Event Hubs data object reads data. You can specify XML, JSON, Avro, or flat as format. When you specify XML format, you must provide a .xsd file. When you specify Avro format, provide a sample Avro schema in a .avsc file. When you specify JSON or flat format, you must provide a sample file.

You can associate the data object with an intelligent structure model and directly parse input from .txt, .csv, .xml, or JSON input files.

You can pass any payload format directly from source to target in Streaming mappings. You can project columns in binary format pass a payload from source to target in its original form or to pass a payload format that is not supported.

Streaming mappings can read, process, and write hierarchical data. You can use array, struct, and map complex data types to process the hierarchical data. You assign complex data types to ports in a mapping to flow hierarchical data. Ports that flow hierarchical data are called complex ports.

For more information about processing hierarchical data, see the *Informatica Big Data Management User Guide*.

RELATED TOPICS:

- [“Creating a Data Object” on page 90](#)

Azure Event Hubs Data Object Overview Properties

Overview properties include general properties that apply to the Azure Event Hubs data object. The Developer tool displays overview properties of the data object in the **Overview** view.

You can configure the following overview properties for Azure Event Hubs data objects:

General

You can configure the following general properties for the Azure Event Hubs data object:

- **Name.** Name of the Azure Event Hubs operation.
- **Description.** Description of the Azure Event Hubs data object.
- **Native Name.** Name of the Azure Event Hubs data object.
- **Path Information.** The path of the data object in Azure Event Hubs. For example, `/EventHubs/avroevents`

Column

You can configure the name, native name, data type, precision, access type, scale, and description of the columns in the Azure Event Hubs resource.

Advanced

The following are the advanced properties for the Azure Event Hubs data object:

- Location. The location of the Azure Event Hubs.
- Date of Creation. The date of creation of the Azure Event Hubs.
- Partition Count. The number of partitions that the Event Hub has when you import the data object.

Azure Event Hubs Header Ports

An Azure Event Hubs data object contains default header port that represent metadata associated with events.

An Azure Event Hubs data object contains the following header ports:

partitionKey

Partition keys to separate events into different partitions. Events with same partition key are sent to the same partition on the event hub.

timestamp

Time at which an event is generated. You can accumulate streamed data into data groups and then process the data groups based on the timestamp values.

Azure Event Hubs Data Object Read Operation Properties

The Data Integration Service uses read operation properties when it reads data from an Azure Event Hubs.

General Properties

The Developer tool displays general properties for Azure Event Hubs sources in the **Read** view.

The following table describes the general properties for the Azure Event Hubs data object read operation:

Property	Description
Name	The name of the Azure Event Hubs data object You can edit the name in the Overview view. When you use the Azure Event Hubs as a source in a mapping, you can edit the name in the mapping.
Description	The description of the Azure Event Hubs data object operation.

Ports Properties

Ports properties for a physical data object include port names and port attributes such as data type and precision.

The following table describes the ports properties that you configure for Azure Event Hubs sources:

Property	Description
Name	The name of the source.
Type	The native data type of the source.
Precision	The maximum number of significant digits for numeric data types, or the maximum number of characters for string data types.
Detail	The detail of the data type.
Scale	The scale of the data type.
Description	The description of the resource.

Sources Properties

The sources properties list the resources of the Azure Event Hubs data object.

The following table describes the sources property that you can configure for Azure Event Hubs events:

Property	Description
Sources	The sources which the Azure Event Hubs data object reads from. You can add or remove sources.

Run-time Properties

The run-time properties include properties that the Data Integration Service uses when reading data from the source at run time.

The run-time property for Azure Event Hubs event includes the name of the Azure Event Hubs connection.

Advanced Properties

The following table describes the advanced properties for Azure Event Hubs sources:

Property	Description
Operation Type	Specifies the type of data object operation. This is a read-only property.
Consumer Group	The name of the event hub Consumer Group that you read events from.
Max Rate	The maximum number of events that are consumed in a single batch for each partition.

Property	Description
Shared Access Policy Name	The name of the event hub Shared Access Policy. To read from an event hub, you must have Listen permission. If you specify a value for this property, it overwrites the value configured in the Azure Event Hubs connection.
Shared Access Policy Primary Key	The primary key of the event hub Shared Access Policy. If you specify a value for this property, it overwrites the value configured in the Azure Event Hubs connection.
Guaranteed Processing	Ensures that the mapping processes messages that the sources publish and delivers them to the targets at least once. In the event of a failure, there could be potential duplicates but the messages are processed successfully. If the external source or the target is not available, the mapping execution stops to avoid any data loss. By default, the Guaranteed Processing property is selected.
Start Position Offset	The time from which the Azure Event Hubs data object starts reading messages from an event hub. You can select one of the following options: <ul style="list-style-type: none"> - CUSTOM. Read messages from a specific time. - EARLIEST. Read the earliest messages available on the event hub. - LATEST. Read the latest messages from an event hub.
Custom Start Position Enqueue Time	Required if you set the Start Position Offset property to custom. A UTC timezone datetime value in the ISO8601 format from which the Azure Event Hubs data object starts reading events from an event hub. Specify time in the following format: YYYY-MM-DDThh:mm:ss.sssZ

Column Projections Properties

The following table describes the columns projection properties that you configure for Azure Event Hubs sources:

Property	Description
Column Name	The name field that contains data. This property is read-only.
Type	The native data type of the resource. This property is read-only.
Enable Column Projection	Indicates that you use a schema to read the data that the source streams. By default, the data is streamed in binary format. To change the format in which the data is processed, select this option and specify the schema format.
Schema Format	The format in which the source processes data. You can select one of the following formats: <ul style="list-style-type: none"> - XML - JSON - Flat - Avro

Property	Description
Use Schema	Specify the XSD schema for the XML format, the sample JSON for the JSON format. Specify a .avsc file for the Avro format.
Use Intelligent Structure Model	Displays the intelligent structure model associated with the complex file. You can select a different model. Note: If you disable the column projection, the intelligent structure model associated with the data object is removed. If you want to associate an intelligent structure model with the data object again, enable the column projection and then click Select Model . For more information on intelligent structure models, see the <i>Informatica Big Data Management User Guide</i> .
Column Mapping	The mapping of source data to the data object. Click View to see the mapping.
Project Column as Complex Data Type	Project columns as complex data type for sources with hierarchical data. Select this option if the source has hierarchical data. For more information on hierarchical data, see the <i>Informatica Big Data Management User Guide</i> .

HBase Data Objects

An HBase data object is a physical data object that represents data based on an HBase resource. After you create an HBase connection, create an HBase data object and a read data object operation.

Create a data object read operation to when you want to perform an uncached lookup on HBase data.

HBase Data Object Read Operation Properties

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

The Developer tool displays advanced properties for the HBase data object operation in the Advanced view.

The following table describes the Advanced properties for an HBase data object read operation:

Property	Description
Operation Type	The type of data object operation. This is a read-only property.
Date Time Format	Format of the columns of the date data type. Specify the date and time formats by using any of the Java date and time pattern strings.
default Column Data Type	Streaming mappings do not support this property.
default Precision	Streaming mappings do not support this property.
default Scale	Streaming mappings do not support this property.

Property	Description
Default Column Family	Streaming mappings do not support this property.
Control File Location	Streaming mappings do not support this property.

JMS Data Objects

A JMS data object is a physical data object that accesses a JMS server. After you configure a JMS connection, create a JMS data object to read from JMS sources.

Note: Support for JMS source is deferred. Support will be reinstated in a future release.

JMS providers are message-oriented middleware systems that send JMS messages. The JMS data object connects to a JMS provider to read or write data.

The JMS data object can read JMS messages from a JMS provider message queue. When you configure a JMS data object, configure properties to reflect the message structure of the JMS messages. The input ports and output ports are JMS message headers.

When you configure the read data operation properties, specify the format in which the JMS data object reads data. You can specify XML, JSON, or Flat as format. When you specify XML format, you must provide an XSD file. When you specify JSON, you must provide a sample file.

You can pass any payload format directly from source to target in Streaming mappings. You can project columns in binary format pass a payload from source to target in its original form or to pass a payload format that is not supported.

Streaming mappings can read, process, and write hierarchical data. You can use array, struct, and map complex data types to process the hierarchical data. You assign complex data types to ports in a mapping to flow hierarchical data. Ports that flow hierarchical data are called complex ports.

For more information about processing hierarchical data, see the *Informatica Big Data Management User Guide*.

RELATED TOPICS:

- [“Creating a Data Object” on page 90](#)

Integration with JMS

You manually create JMS source and target data objects to reflect the message structure of JMS messages.

The JMS data object can read messages of type `TextMessage`. This type of message contains a string object. `TextMessages` can contain XML or JSON message data.

JMS Message Structure

JMS messages contain the following components:

- Header
- Properties
- Body

Header Fields

JMS messages contain a fixed number of header fields. Each JMS message uses these fields regardless of message type. Every JMS source and target definition includes a pre-defined set of header fields.

The following table describes the JMS message header fields:

Header Field	Description
JMSDestination	Destination to which the message is sent. JMS destinations can be a message queue or a recipient who listens for messages based on the message topic.
JMSDeliveryMode	Delivery mode of the message. The delivery mode can be persistent or non-persistent.
JMSMessageID	Unique identification value for the message.
JMSTimestamp	Time at which the message was handed off to the provider to be sent to the destination.
JMSCorrelationID	Links one message with another. For example, JMSCorrelationID can link a response message with the corresponding request message.
JMSReplyTo	Destination to which a reply message can be sent.
JMSRedelivered	Indicates that a message might have been delivered previously, but not acknowledged.
JMSType	Type of message based on a description of the message. For example, if a message contains a stock trade, the message type might be stock trade.
JMSExpiration	Amount of time in milliseconds the message remains valid. The messages remain in memory during this period.
JMSPriority	Priority of the message from 0-9. 0 is the lowest priority. 9 is the highest.

Property Fields

JMS source and target definitions can optionally include message property fields. Property fields contain additional message header information. JMS providers use properties in a JMS message to give provider-specific information. Applications that use a JMS provider can add property fields with application-specific information to a message.

Body Fields

JMS source and target definitions can optionally include a message body. The body contains one or more fields. Only certain types of JMS messages contain a body.

JMS Data Object Overview Properties

The Data Integration Service uses overview properties when it reads data from a JMS source.

Overview properties include general properties that apply to the JMS data object. They also include object properties that apply to the resources in the JMS data object. The Developer tool displays overview properties for JMS messages in the Overview view.

General Properties

The following table describes the general properties that you configure for JMS data objects:

Property	Description
Name	The name of the JMS data object.
Description	The description of the JMS data object.
Connection	The name of the JMS connection.

Objects Properties

The following table describes the objects properties that you configure for JMS data objects:

Property	Description
Name	The name of the topic or queue of the JMS source.
Description	The description of the JMS source.
Native Name	The native name of JMS source.
Path Information	The type and name of the topic or topic pattern of the JMS source.

JMS Data Object Read Operation Properties

The Data Integration Service uses read operation properties when it reads data from a JMS source. You can edit the format, run-time, and advanced properties.

General Properties

The Developer tool displays general properties for JMS sources in the **Read** view.

The following table describes the general properties that you view for JMS sources:

Property	Description
Name	The name of the JMS read operation. This property is read-only. You can edit the name in the Overview view. When you use the JMS source as a source in a mapping, you can edit the name in the mapping.
Description	The description of the JMS source.

Ports Properties

Ports properties for a physical data object include port names and port attributes such as data type and precision.

The following table describes the ports properties that you configure for JMS sources:

Property	Description
Name	The name of the JMS source.
Type	The native data type of the source.
Precision	The maximum number of significant digits for numeric data types, or the maximum number of characters for string data types.
Scale	The scale of the data type.
Description	The description of the resource.

Sources Properties

The sources properties list the resources of the JMS data object. You can add or remove resources in the data object.

Run-time Properties

The run-time properties displays the name of the connection.

The following table describes the run-time property that you configure for JMS sources:

Property	Description
Connection	Name of the JMS connection.

Advanced Properties

The Developer tool displays the advanced properties for JMS sources in the Output transformation in the **Read** view.

You can configure the following advanced properties for JMS sources:

Property	Description
Operation Type	Specifies the type of data object operation. This is a read-only property.
JMS Destination	Name of the queue or topic to which the JMS provider publishes messages. The data object subscribes to JMS messages from this queue or topic.
Client ID	Client identifier to identify the connection and set up a durable connections.

Property	Description
Durable Subscription Name	Name of the durable subscription that can receive messages sent while the subscribers are not active. Durable subscriptions provide the flexibility and reliability of queues, but still allow clients to send messages to many recipients.
JMS Message Selector	Criteria for filtering message header or message properties, to limit which JMS messages the data object receives.
Guaranteed Processing	Ensures that the mapping processes messages that the sources publish and delivers them to the targets at least once. In the event of a failure, there could be potential duplicates but the messages are processed successfully. If the external source or the target is not available, the mapping execution stops to avoid any data loss. By default, the Guaranteed Processing property is selected.

Column Projections Properties

The following table describes the columns projection properties that you configure for JMS sources:

Property	Description
Column Name	The name field that contains data. This property is read-only.
Type	The native data type of the resource. This property is read-only.
Enable Column Projection	Indicates that you use a schema to read the data that the source streams. By default, the data is streamed in binary format. To change the format in which the data is streamed, select this option and specify the schema format.
Schema Format	The format in which the source streams data. You can select one of the following formats: <ul style="list-style-type: none"> - XML - JSON - Flat
Use Schema	Specify the XSD schema for the XML format or the sample JSON for the JSON format. For the Flat file format, configure the schema to associate a flat file to the source. When you provide a sample file, the Data Integration Service uses UTF-8 code page when reading the data.
Column Mapping	The mapping of source data to the data object. Click View to see the mapping.
Project Column as Complex Data Type	Project columns as complex data type for hierarchical data. For more information, see the <i>Informatica Big Data Management User Guide</i> .

Configuring a Schema for Flat Files

Configure schema for flat files when you configure column projection properties.

1. On the **Column Projection** tab, enable column projection and select the flat schema format.
The page displays the column projection properties page.

2. On the column projection properties page, configure the following properties:
 - Sample Metadata File. Select a sample file.
 - Code page. Select the UTF-8 code page.
 - Format. Format in which the source processes data. Default value is Delimited. You cannot change it.
3. Click **Next**.
4. In the delimited format properties page, configure the following properties:

Property	Description
Delimiters	Specify the character that separates entries in the file. Default is a comma (.). You can only specify one delimiter at a time. If you select Other and specify a custom delimiter, you can only specify a single-character delimiter.
Text Qualifier	Specify the character used to enclose text that should be treated as one entry. Use a text qualifier to disregard the delimiter character within text. Default is No quotes. You can only specify an escape character of one character.
Preview Options	Specify the escape character. The row delimiter is not applicable as only one row is created at a time.
Maximum rows to preview	Specify the rows of data you want to preview.

5. Click **Next** to preview the flat file data object.
If required, you can change the column attributes. The data type timestampWithTZ format is not supported.
6. Click **Finish**.
The data object opens in the editor.

Kafka Data Objects

A Kafka data object is a physical data object that represents data in a Kafka stream. After you configure a Messaging connection, create a Kafka data object to read from Apache Kafka brokers.

Kafka runs as a cluster comprised of one or more servers each of which is called a broker. Kafka brokers stream data in the form of messages. These messages are published to a topic.

Kafka topics are divided into partitions. Spark Structured Streaming can read the partitions of the topics in parallel. This gives better throughput and could be used to scale the number of messages processed. Message ordering is guaranteed only within partitions. For optimal performance you should have multiple partitions. You can create or import a Kafka data object.

When you configure the Kafka data object, specify the topic name that you read from or write to. You can specify the topic name or use a regular expression for the topic name pattern only when you read from Kafka. To subscribe to multiple topics that match a pattern, you can specify a regular expression. When you run the application on the cluster, the pattern matching is done against topics before the application runs. If you add a topic with a similar pattern when the application is already running, the application will not read from the topic.

After you create a Kafka data object, create a read operation. You can use the Kafka data object read operation as a source in streaming mappings. If you want to configure high availability for the mapping, ensure that the Kafka cluster is highly available. You can also read from a Kerberised Kafka cluster.

When you configure the data operation read properties, you can specify the time from which the Kafka source starts reading Kafka messages from a Kafka topic.

You can associate the data object with an intelligent structure model and directly parse input from text, CSV, XML, or JSON input files, as well as PDF forms, Microsoft Word tables, or Microsoft Excel.

When you configure the data operation properties, specify the format in which the Kafka data object reads data. You can specify XML, JSON, Avro, or Flat as format. When you specify XML format, you must provide a XSD file. When you specify Avro format, provide a sample Avro schema in a .avsc file. When you specify JSON or Flat format, you must provide a sample file.

You can pass any payload format directly from source to target in Streaming mappings. You can project columns in binary format pass a payload from source to target in its original form or to pass a payload format that is not supported.

Streaming mappings can read, process, and write hierarchical data. You can use array, struct, and map complex data types to process the hierarchical data. You assign complex data types to ports in a mapping to flow hierarchical data. Ports that flow hierarchical data are called complex ports.

For more information about processing hierarchical data, see the *Informatica Big Data Management User Guide*.

For more information about Kafka clusters, Kafka brokers, and partitions, see <http://kafka.apache.org/11/documentation.html>.

For more information about how to use topic patterns in Kafka data objects, see <https://kb.informatica.com/h2l/HowTo%20Library/1/1132-HowtoUseTopicPatternsInKafkaDataObjects-H2L.pdf>.

RELATED TOPICS:

- [“Creating a Data Object” on page 90](#)

Kafka Data Object Overview Properties

The Data Integration Service uses overview properties when it reads data from or writes data to a Kafka broker.

Overview properties include general properties that apply to the Kafka data object. They also include object properties that apply to the resources in the Kafka data object. The Developer tool displays overview properties for Kafka messages in the **Overview** view.

General Properties

The following table describes the general properties that you configure for Kafka data objects:

Property	Description
Name	The name of the Kafka read operation.
Description	The description of the Kafka data object.
Connection	The name of the Kafka connection.

Objects Properties

The following table describes the objects properties that you configure for Kafka data objects:

Property	Description
Name	The name of the topic or topic pattern of the Kafka data object.
Description	The description of the Kafka data object.
Native Name	The native name of Kafka data object.
Path Information	The type and name of the topic or topic pattern of the Kafka data object.

Column Properties

The following table describes the column properties that you configure for Kafka data objects:

Property	Description
Name	The name of the Kafka data object.
Native Name	The native name of the Kafka data object.
Type	The native data type of the Kafka data object.
Precision	The maximum number of significant digits for numeric data types, or the maximum number of characters for string data types.
Scale	The scale of the data type.
Description	The description of the Kafka data object.
Access Type	The type of access the port or column has.

Kafka Header Ports

The Kafka data objects contains default header ports that represent metadata associated with events.

The following table lists the header ports of the Kafka data object:

Header ports	Data type	Description
partitionId	integer	Partition ID of the data. When the Kafka topic is divided into partitions, a sequential ID is assigned to each event and is unique to each event within the partition. Use the partition ID to identify the event that you want to consume from a particular partition.
key	binary	Key value associated with an event. You can group events based on the key value and then process the data.
TopicName	string	Name of the Kafka topic from where you receive events.
timestamp	Date/Time	Time at which an event is generated. You can accumulate streamed data into data groups and then process the data groups based on the timestamp values.

Kafka Data Object Read Operation Properties

The Data Integration Service uses read operation properties when it reads data from a Kafka broker.

General Properties

The Developer tool displays general properties for Kafka sources in the **Read** view.

The following table describes the general properties that you view for Kafka sources:

Property	Description
Name	The name of the Kafka broker read operation. You can edit the name in the Overview view. When you use the Kafka broker as a source in a mapping, you can edit the name in the mapping.
Description	The description of the Kafka broker.

Ports Properties

Ports properties for a physical data object include port names and port attributes such as data type and precision.

The following table describes the ports properties that you configure for Kafka broker sources:

Property	Description
Name	The name of the resource.
Type	The native data type of the resource.
Precision	The maximum number of significant digits for numeric data types, or the maximum number of characters for string data types.
Scale	The scale of the data type.
Description	The description of the resource.

Run-time Properties

The run-time properties displays the name of the connection.

The following table describes the run-time property that you configure for Kafka sources:

Property	Description
Connection	Name of the Kafka connection.

Advanced Properties

The Developer tool displays the advanced properties for Kafka sources in the Output transformation in the Read view.

The following table describes the advanced properties that you can configure for Kafka sources:

Property	Description
Operation Type	Specifies the type of data object operation. This is a read-only property.
Guaranteed Processing	Ensures that the mapping processes messages that the sources publish and delivers them to the targets at least once. In the event of a failure, there could be potential duplicates but the messages are processed successfully. If the external source or the target is not available, the mapping execution stops to avoid any data loss. By default, the Guaranteed Processing property is selected.
Start Position Offset	The time from which the Kafka source starts reading messages from a Kafka topic when you run a mapping. You can select one of the following options: <ul style="list-style-type: none">- Custom. Read messages from a specific time.- Earliest. Read the earliest messages available on the Kafka topic.- Latest. Read messages received by the Kafka topic after the mapping has been deployed. This property is applicable for Kafka versions 0.10.1.0 and later.
Custom Start Position Timestamp	The time in GMT from which the Kafka source starts reading Kafka messages from a Kafka topic. Specify a time in the following format: dd-MM-yyyy HH:mm:ss.SSS The milliseconds are optional. This property is applicable for Kafka versions 0.10.1.0 and above.
Consumer Configuration Properties	The configuration properties for the consumer. If the Kafka data object is reading data from a Kafka cluster that is configured for Kerberos authentication, include the following property: <code>security.protocol=SASL_PLAINTEXT,sasl.kerberos.service.name=kafka,sasl.mechanism=GSSAPI</code>

Sources Properties

The sources properties list the resources of the Kafka data object.

The following table describes the sources property that you can configure for Kafka sources:

Property	Description
Sources	The sources which the Kafka data object reads from. You can add or remove sources.

Column Projection Properties

The Developer tool displays the column projection properties in the Properties view of the Read operation.

To specify column projection properties, double click on the read operation and select the data object. The following table describes the columns projection properties that you configure for Kafka sources:

Property	Description
Column Name	The name field that contains data. This property is read-only.
Type	The native data type of the source. This property is read-only.
Enable Column Projection	Indicates that you use a schema to read the data that the source streams. By default, the data is streamed in binary format. To change the format in which the data is streamed, select this option and specify the schema format.
Schema Format	The format in which the source streams data. Select one of the following formats: <ul style="list-style-type: none">- XML- JSON- Flat- Avro
Use Schema	Specify the XSD schema for the XML format, a sample file for JSON, or .avsc file for Avro format. For the Flat file format, configure the schema to associate a flat file to the Kafka source. When you provide a sample file, the Data Integration Service uses UTF-8 code page when reading the data.
Use Intelligent Structure Model	Displays the intelligent structure model associated with the complex file. You can select a different model. Note: If you disable the column projection, the intelligent structure model associated with the data object is removed. If you want to associate an intelligent structure model again with the data object, enable the column projection and click Select Model. For more information on intelligent structure models, see the <i>Informatica Big Data Management User Guide</i> .
Column Mapping	The mapping of source data to the data object. Click View to see the mapping.
Project Column as Complex Data Type	Project columns as complex data type for hierarchical data. For more information, see the <i>Informatica Big Data Management User Guide</i> .

Configuring Schema for Flat Files

Configure schema for flat files when you configure column projection properties.

1. On the **Column Projection** tab, enable column projection and select the flat schema format.
The page displays the column projection properties page.
2. On the column projection properties page, configure the following properties:
 - Sample Metadata File. Select a sample file.
 - Code page. Select the UTF-8 code page.

- **Format.** Format in which the source processes data. Default value is Delimited. You cannot change it.
3. Click **Next**.
 4. In the delimited format properties page, configure the following properties:

Property	Description
Delimiters	Specify the character that separates entries in the file. Default is a comma (.). You can only specify one delimiter at a time. If you select Other and specify a custom delimiter, you can only specify a single-character delimiter.
Text Qualifier	Specify the character used to enclose text that should be treated as one entry. Use a text qualifier to disregard the delimiter character within text. Default is No quotes. You can only specify an escape character of one character.
Preview Options	Specify the escape character. The row delimiter is not applicable as only one row is created at a time.
Maximum rows to preview	Specify the rows of data you want to preview.

5. Click **Next** to preview the flat file data object.
If required, you can change the column attributes. The data type timestampWithTZ format is not supported.
6. Click **Finish**.
The data object opens in the editor.

MapR Streams Data Objects

A MapR Streams data object is a physical data object that represents data in a MapR Stream. After you create a MapR Streams connection, create a MapR Streams data object to read from MapR Streams.

Note: Support for MapR Streams data objects is deferred. Support will be reinstated in a future release.

Before you create and use MapR Stream data objects in Streaming mappings, complete the required prerequisites.

For more information about the prerequisite tasks, see the *Informatica Big Data Management Integration Guide*.

When you configure the MapR Streams data object, specify the stream name that you read from in the following format:

```
/pathname:topic name
```

You can specify the stream name or use a regular expression for the stream name pattern only when you read from MapR Streams. The regular expression that you specify applies to the topic name and not the path name. To subscribe to multiple topics that match a pattern, you can specify a regular expression. When you run the application on the cluster, the pattern matching is done against topics before the application runs. If you add a topic with a similar pattern when the application is already running, the application will not read from the topic.

After you create a MapR Streams data object, create a read data object operation. You can then add the data object read operation as a source in streaming mappings.

You can associate the data object with an intelligent structure model and directly parse input from text, CSV, XML, or JSON input files, as well as PDF forms, Microsoft Word tables, or Microsoft Excel.

When you configure the data operation properties, specify the format in which the MapR Streams data object reads data. You can specify XML, JSON, or Avro as format. When you specify XML format, you must provide a XSD file. When you specify Avro format, provide a sample Avro schema in a .avsc file. When you specify JSON or Flat format, you must provide a sample file.

You can pass any payload format directly from source to target in Streaming mappings. You can project columns in binary format pass a payload from source to target in its original form or to pass a payload format that is not supported.

Streaming mappings can read, process, and write hierarchical data. You can use array, struct, and map complex data types to process the hierarchical data. You assign complex data types to ports in a mapping to flow hierarchical data. Ports that flow hierarchical data are called complex ports.

For more information about processing hierarchical data, see the *Informatica Big Data Management User Guide*.

For more information about how to use topic patterns in MapRStreams data objects, see <https://kb.informatica.com/h2l/HowTo%20Library/1/1149-HowtoUseTopicPatternsinMapRStreamsDataObjects-H2L.pdf>.

RELATED TOPICS:

- [“Creating a Data Object” on page 90](#)

MapR Streams Object Overview Properties

Overview properties include general properties that apply to the MapR Streams data object. The Developer tool displays overview properties in the Overview view.

General

Property	Description
Name	Name of the MapR Streams data object.
Description	Description of the MapR Streams data object.
Native Name	Native name of the MapR Stream.
Path Information	The path of the MapR Stream.

Column Properties

The following table describes the column properties that you configure for MapR Streams data objects:

Property	Description
Name	The name of the MapR Streams data object.
Native Name	The native name of the MapR Streams data object.

Property	Description
Type	The native data type of the MapR Streams data object.
Precision	The maximum number of significant digits for numeric data types, or the maximum number of characters for string data types.
Scale	The scale of the data type.
Description	The description of the MapR Streams data object.
Access Type	The access type of the port or column.

MapR Streams Data Object Read Operation Properties

The Data Integration Service uses read operation properties when it reads data from MapR Streams.

General Properties

The Developer tool displays general properties for MapR Streams sources in the **Read** view.

The following table describes the general properties for the MapRStreams data object read operation:

Property	Description
Name	The name of the MapRStreams read operation. This property is read-only. You can edit the name in the Overview view. When you use the MapR Streams as a source in a mapping, you can edit the name in the mapping.
Description	The description of the MapRStreams data object operation.

Ports Properties

Ports properties for a physical data object include port names and port attributes such as data type and precision.

The following table describes the ports properties that you configure for MapR Stream sources:

Property	Description
Name	The name of the resource.
Type	The native data type of the resource.
Precision	The maximum number of significant digits for numeric data types, or the maximum number of characters for string data types.
Scale	The scale of the data type.
Description	The description of the resource.

Sources Properties

The sources properties list the resources of the MapRStreams data object.

The following table describes the sources property that you can configure for MapR Stream sources:

Property	Description
Sources	The sources which the MapRStreams data object reads from. You can add or remove sources.

Run-time Properties

The run-time property for MapR Stream source includes the name of the MapRStream connection.

Advanced Properties

The Developer tool displays the advanced properties for MapR Stream sources in the Output transformation in the **Read** view.

The following table describes the advanced properties for MapR Stream sources:

Property	Description
Operation Type	Specifies the type of data object operation. This is a read-only property.
Guaranteed Processing	Ensures that the mapping processes messages that the sources publish and delivers them to the targets at least once. In the event of a failure, there could be potential duplicates but the messages are processed successfully. If the external source or the target is not available, the mapping execution stops to avoid any data loss. Select this option to avoid data loss.

Column Projections Properties

The following table describes the columns projection properties that you configure for MapR Stream sources:

Property	Description
Column Name	The name field that contains data. This property is read-only.
Type	The native data type of the resource. This property is read-only.
Enable Column Projection	Indicates that you use a schema to read the data that the source streams. By default, the data is streamed in binary format. To change the format in which the data is streamed, select this option and specify the schema format.
Schema Format	The format in which the source streams data. You can select one of the following formats: <ul style="list-style-type: none">- XML- JSON- Avro

Property	Description
Use Schema	Specify the XSD schema for the XML format, a sample JSON for the JSON format, or .avsc file for the Avro format.
Use Intelligent Structure Model	<p>Displays the intelligent structure model associated with the complex file. You can select a different model.</p> <p>Note: If you disable the column projection, the intelligent structure model associated with the data object is removed. If you want to associate an intelligent structure model again with the data object, enable the column projection and click Select Model.</p> <p>For more information on intelligent structure models, see the <i>Informatica Big Data Management User Guide</i>.</p>
Column Mapping	The mapping of source data to the data object. Click View to see the mapping.
Project Column as Complex Data Type	<p>Project columns as complex data type for hierarchical data.</p> <p>For more information, see the <i>Informatica Big Data Management User Guide</i>.</p>

CHAPTER 4

Targets in a Streaming Mapping

This chapter includes the following topics:

- [Targets in a Streaming Mapping Overview, 49](#)
- [Processing Hierarchical Data in Streaming Mappings, 50](#)
- [Amazon Kinesis Data Objects, 51](#)
- [Azure Event Hubs Data Objects, 55](#)
- [Complex File Data Objects, 59](#)
- [HBase Data Objects, 65](#)
- [JMS Data Objects, 68](#)
- [Kafka Data Objects, 72](#)
- [Microsoft Azure Data Lake Store Data Object, 76](#)
- [MapR Streams Data Objects, 79](#)
- [Relational Data Objects, 83](#)

Targets in a Streaming Mapping Overview

You can access log file data, sensor data, Supervisory Control And Data Acquisition (SCADA) data, message bus data, Programmable logic controller (PLC) data on the Spark engine in the Hadoop environment.

You can create physical data objects to access the different types of data. Based on the type of target you are writing to, you can create the following data objects:

Amazon Kinesis

A physical data object that represents data in an Amazon Kinesis Firehose Delivery Stream. Create an Amazon Kinesis data object to write to an Amazon Kinesis Firehose Delivery Stream.

Azure Event Hub

A physical data object that represents data in Microsoft Azure Event Hubs data streaming platform and event ingestion service. Create an Azure Event Hub data object to connect to an Event Hub target.

Complex file

A representation of a file in the Hadoop file system. Create a complex file data object to write data to an HDFS sequence file or binary file.

For more information about complex file data objects, see the *Informatica Big Data Management User Guide*.

HBase

A physical data object that represents data in an HBase resource. Create an HBase data object to connect to an HBase data target.

JMS

A physical data object that accesses a JMS server. Create a JMS data object to write to a JMS server.

Kafka

A physical data object that accesses a Kafka broker. Create a Kafka data object to write to a Kafka broker.

MapR Streams

A MapR Streams data object is a physical data object that represents data in a MapR Stream. Create a MapR Streams data object to write to a MapR Stream.

Note: Support for MapR Streams data objects is deferred. Support will be reinstated in a future release.

Microsoft Azure Data Lake

A Microsoft Azure Data Lake Store data object is a physical data object that represents a Microsoft Azure Data Lake Store table. Create an Azure Data Lake object to write to a Microsoft Azure Data Lake Store table.

Relational

A physical data object that you can use to access a relational table. You can create a relational object to connect to a Hive or JDBC-compliant database.

For more information about relational data objects, see the *Informatica Developer Tool Guide*.

Processing Hierarchical Data in Streaming Mappings

Data objects in a streaming mapping can process hierarchical data through complex data types. If you want to project data as complex data types, you must enable the write data operation to project data as complex data types.

The following table shows the format and complex data types that targets in a Streaming mapping support:

Format	Schema Type	Amazon Kinesis Firehose	Azure Data Lake Store	Azure Event Hub	Complex File	JMS	Kafka	MapR Streams
Avro	Flat	Not supported	Supported	Supported	Supported	Not supported	Supported	Supported
Avro	Hierarchical	Not supported	Supported	Supported	Supported	Not supported	Supported	Supported
Binary	Binary	Supported	Supported	Supported	Supported	Supported	Supported	Supported
Flat	Flat	Not Supported	Not Supported	Supported	Not supported	Supported	Supported	Not Supported

Format	Schema Type	Amazon Kinesis Firehose	Azure Data Lake Store	Azure Event Hub	Complex File	JMS	Kafka	MapR Streams
JSON	Flat	Supported	Supported	Supported	Supported	Supported	Supported	Supported
JSON	Hierarchical	Supported	Supported	Supported	Supported	Supported	Supported	Supported
XML	Flat	Not supported	Supported	Supported	Supported	Supported	Supported	Supported
XML	Hierarchical	Not supported	Supported	Supported	Supported	Supported	Supported	Supported

For more information about processing hierarchical data, see the *Informatica Big Data Management User Guide*.

Amazon Kinesis Data Objects

An Amazon Kinesis data object is a physical data object that represents data in an Amazon Kinesis Data Firehose Delivery Stream. After you create an Amazon Kinesis connection, create an Amazon Kinesis data object to write to Amazon Kinesis Data Firehose.

Kinesis Data Firehose is a real-time data stream processing option that Amazon Kinesis offers within the AWS ecosystem. Kinesis Data Firehose allows batching, encrypting, and compressing of data. Kinesis Data Firehose can automatically scale to meet system needs.

When you configure the Amazon Kinesis data object, specify the name of the Data Firehose Delivery Stream that you write to. You can specify the Kinesis Data Firehose Delivery Stream name or use a regular expression for the stream name pattern. If the input has multiple partitions, you can create multiple Kinesis Data Firehose Delivery Streams to the same target and send the data from these partitions to the individual delivery streams based on the pattern you specify in the stream name.

After you create the data object, create a data object write operation to write data to an Amazon Kinesis Data Firehose Delivery Stream. You can then add the data object write operation as a target in Streaming mappings.

When you configure the data operation properties, specify the format in which the data object writes data. When you write to Amazon Data Firehose targets, you can specify JSON or binary as the format.

When you specify JSON format, you must provide a sample file.

You can pass any payload format directly from source to target in Streaming mappings. You can project columns in binary format pass a payload from source to target in its original form or to pass a payload format that is not supported.

Streaming mappings can read, process, and write hierarchical data. You can use array, struct, and map complex data types to process the hierarchical data. You assign complex data types to ports in a mapping to flow hierarchical data. Ports that flow hierarchical data are called complex ports.

When you run a mapping to write data to an Amazon Kinesis Data Firehose Delivery Stream, the data object uses the AWS Firehose SDK to write data.

Note: You cannot run a mapping with an Amazon Kinesis data object on MapR and Azure HDInsight distributions.

For more information about processing hierarchical data, see the *Informatica Big Data Management User Guide*.

For more information about Kinesis Data Firehose, see the Amazon Web Services documentation.

Amazon Kinesis Data Object Overview Properties

Overview properties include general properties that apply to the Amazon Kinesis data object. The Developer tool displays overview properties of the data object in the **Overview** view.

You can configure the following overview properties for Amazon Kinesis data objects:

General

You can configure the following general properties for the Amazon Kinesis data object:

- **Name.** Name of the Amazon Kinesis operation.
- **Description.** Description of the Amazon Kinesis data object.
- **Native Name.** Name of the Amazon Kinesis data object.
- **Path Information.** The path of the data object in Amazon Kinesis. For example, `/DeliveryStreams/router1`

Column

You can configure the name, native name, data type, precision, access type, scale, and description of the columns in the Amazon Kinesis resource.

Advanced

The following are the advanced properties for the Amazon Kinesis data object:

- **Amazon Resource Name.** The Kinesis resource that the Amazon Kinesis data object is reading from or writing to.
- **Type.** The type of delivery stream that the Amazon Kinesis data object is reading from or writing to. The delivery stream is either Kinesis Stream or Firehose DeliveryStream
- **Number of Shards.** Specify the number of shards that the Kinesis Stream is composed of. This property is not applicable for Firehose DeliveryStream.

Amazon Kinesis Data Object Write Operation Properties

The Data Integration Service uses write operation properties when it writes data to Amazon Kinesis Firehose.

General Properties

The Developer tool displays general properties for Amazon Kinesis targets in the **Write** view.

The following table describes the general properties that you view for Amazon Kinesis targets:

Property	Description
Name	The name of the Amazon Simple Storage Service (Amazon S3), Amazon Redshift tables, or Amazon Elasticsearch Service (Amazon ES) . This property is read-only.
Description	The description of the target.

Ports Properties

Ports properties for a physical data object include port names and port attributes such as data type and precision.

The following table describes the ports properties that you configure for Amazon Kinesis targets:

Property	Description
Name	The name of the target.
Type	The native data type of the target.
Precision	The maximum number of significant digits for numeric data types, or the maximum number of characters for string data types.
Detail	The detail of the data type.
Scale	The scale of the data type.
Description	The description of the target.

Target Properties

The targets properties list the targets of the Amazon Kinesis data object.

The following table describes the sources property that you can configure for Kinesis Firehose targets:

Property	Description
Target	The target which the Amazon Kinesis data object writes to. You can add or remove targets.

Run-time Properties

The run-time properties include properties that the Data Integration Service uses when writing data to the target at run time, such as reject file names and directories.

The run-time property for Amazon Kinesis targets includes the name of the Amazon Kinesis connection.

Advanced Properties

Advanced properties include tracing level, row order, and retry attempt properties.

The following table describes the advanced properties for Amazon Kinesis targets:

Property	Description
Operation Type	Specifies the type of data object operation. This is a read-only property.
Record Delimiter	The record delimiter that is inserted into the Kinesis Firehose delivery stream.
Maximum Error Retry Attempts	The number of times that the Data Integration Service attempts to reconnect to the target.
Response Wait Time (milliseconds)	The number of milliseconds that the Data Integration Service waits for a response to send a batch request.
Retry Attempt Delay Time (milliseconds)	The number of milliseconds that the Data Integration Service waits before it retries to send data to the Kinesis Firehose delivery stream.
Runtime Properties	The runtime properties for the connection pool and AWS client configuration. Specify the properties as key -value pairs. For example: <code>key1=value1, key2=value2</code>

Column Projection Properties

The following table describes the columns projection properties that you configure for Amazon Kinesis Firehose targets:

Property	Description
Column Name	The name field that contains data. This property is read-only.
Type	The native data type of the resource. This property is read-only.
Enable Column Projection	Indicates that you use a schema to write data to the target. By default, the data is streamed in binary format. To change the format in which the data is streamed, select this option and specify the schema format.
Schema Format	The format in which data is written to the target. You can select one of the following formats: <ul style="list-style-type: none">- XML- JSON- Flat- Avro
Use Schema	Specify the XSD schema for the XML format, a sample file for JSON, or .avsc file for Avro format. For the Flat file format, configure the schema to associate a flat file to the source.

Property	Description
Column Mapping	The mapping of source data to the data object. Click View to see the mapping.
Project Column as Complex Data Type	Project columns as complex data type for hierarchical data. For more information on hierarchical data, see the <i>Informatica Big Data Management User Guide</i> .

Azure Event Hubs Data Objects

An Azure Event Hubs data object is a physical data object that represents data in Microsoft Azure Event Hubs data streaming platform and event ingestion service. After you create an Azure Event Hubs connection, create an Azure Event Hubs data object to write to an event hub event.

Azure Event Hubs is a highly scalable data streaming platform and event ingestion service, that receives and processes events. Azure Event Hubs can process and store events or data produced by distributed software and devices.

When you configure the Azure Event Hubs data object, specify the name of the event that you write to. After you create the data object, create a data object write operation to write data to an event hub. You can then add the data object write operation as a target in streaming mappings.

Configure partition keys to separate events into different partitions. Events with the same partition key are sent to the same partition on the event hub. Events that are sent in a batch do not need to have the same partition key.

When you configure the data operation properties, specify the format in which the Azure Event Hubs data object writes data. You can specify XML, JSON, Avro, or Flat as format. When you specify XML format, you must provide a XSD file. When you specify Avro format, provide a sample Avro schema in a .avsc file. When you specify JSON or Flat format, you must provide a sample file.

You can pass any payload format directly from source to target in Streaming mappings. You can project columns in binary format pass a payload from source to target in its original form or to pass a payload format that is not supported.

Streaming mappings can read, process, and write hierarchical data. You can use array, struct, and map complex data types to process the hierarchical data. You assign complex data types to ports in a mapping to flow hierarchical data. Ports that flow hierarchical data are called complex ports.

For more information about processing hierarchical data, see the *Informatica Big Data Management User Guide*.

Azure Event Hubs Data Object Overview Properties

Overview properties include general properties that apply to the Azure Event Hubs data object. The Developer tool displays overview properties of the data object in the **Overview** view.

You can configure the following overview properties for Azure Event Hubs data objects:

General

You can configure the following general properties for the Azure Event Hubs data object:

- Name. Name of the Azure Event Hubs operation.
- Description. Description of the Azure Event Hubs data object.
- Native Name. Name of the Azure Event Hubs data object.
- Path Information. The path of the data object in Azure Event Hubs. For example, `/EventHubs/avroevents`

Column

You can configure the name, native name, data type, precision, access type, scale, and description of the columns in the Azure Event Hubs resource.

Advanced

The following are the advanced properties for the Azure Event Hubs data object:

- Location. The location of the Azure Event Hubs.
- Date of Creation. The date of creation of the Azure Event Hubs.
- Partition Count. The number of partitions that the Event Hub has when you import the data object.

Azure Event Hubs Header Ports

An Azure Event Hubs data object contains default header port that represent metadata associated with events.

An Azure Event Hubs data object contains the following header ports:

partitionKey

Partition keys to separate events into different partitions. Events with same partition key are sent to the same partition on the event hub.

timestamp

Time at which an event is generated. You can accumulate streamed data into data groups and then process the data groups based on the timestamp values.

Azure Event Hubs Data Object Write Operations

The Data Integration Service uses write operation properties when it writes data to an Azure Event Hubs.

General Properties

The Developer tool displays general properties for Azure Event Hub sources in the **Write** view.

The following table describes the general properties for the Azure Event Hub data object write operation:

Property	Description
Name	The name of the Azure Event Hub data object. This property is read-only. You can edit the name in the Overview view. When you use the Azure Event Hub as a target in a mapping, you can edit the name in the mapping.
Description	The description of the Azure Event Hub data object operation.

Ports Properties

Ports properties for a physical data object include port names and port attributes such as data type and precision.

The following table describes the ports properties that you configure for Azure EventHub targets:

Property	Description
Name	The name of the target.
Type	The native data type of the target.
Precision	The maximum number of significant digits for numeric data types, or the maximum number of characters for string data types.
Detail	The detail of the data type.
Scale	The scale of the data type.
Description	The description of the resource.

Target Properties

The targets properties list the targets of the Azure EventHub data object.

The following table describes the sources property that you can configure for Azure Event Hub targets:

Property	Description
Target	The target which the Azure EventHub data object writes to. You can add or remove targets.

Run-time Properties

The run-time properties include properties that the Data Integration Service uses when writing data to the target at run time, such as reject file names and directories.

The run-time property for Azure Event Hub targets includes the name of the Azure EventHub connection.

Advanced Properties

The following table describes the advanced properties for Azure Event Hub targets:

Property	Description
Operation Type	Specifies the type of data object operation. This is a read-only property.
Batch Size	The batch size of the events. Specify an integer value for the batch size.
Shared Access Policy Name	The name of the Event Hub Namespace Shared Access Policy. To write to Event Hubs, you must have Send permissions. If you specify a value for this property, it overwrites the value configured in the Azure EventHub connection.
Shared Access Policy Primary Key	The primary key of the Event Hub Namespace Shared Access Policy. If you specify a value for this property, it overwrites the value configured in the Azure EventHub connection.
Publisher Properties	The Event Hub publisher configuration properties. Specify properties as key-value pairs. For example, key1=value1,key2=value2

Column Projections Properties

The Developer tool displays the column projection properties in the **Properties** view of the write operation.

To specify column projection properties, double click on the write operation and select the data object. The following table describes the columns projection properties that you configure for Azure EventHub targets:

Property	Description
Column Name	The field in the target that the data object writes to. This property is read-only.
Type	The native data type of the target. This property is read-only.
Enable Column Projection	Indicates that you use a schema to publish the data to the target. By default, the data is streamed in binary format. To change the streaming format, select this option and specify the schema format.
Schema Format	The format in which you stream data to the target. You can select one of the following formats: <ul style="list-style-type: none">- XML- JSON- Flat- Avro
Use Schema	Specify the XSD schema for the XML format or the sample file for JSON or Avro format. For the Flat file format, configure the schema to associate a flat file.

Property	Description
Column Mapping	The mapping of data object to the target. Click View to see the mapping.
Project as Hierarchical Type	Project columns as complex data type for hierarchical data. For more information, see the <i>Informatica Big Data Management User Guide</i> .

Complex File Data Objects

Create a complex file data object with an HDFS connection to write data to HDFS sequence files or binary files.

When you create a complex file data object, a read and write operation is created. To use the complex file data object as a target in streaming mappings, configure the complex file data object write operation properties. You can select the mapping environment and run the mappings on the Spark engine of the Hadoop environment.

When you configure the data operation properties, specify the format in which the complex file data object writes data to the HDFS sequence file. You can also specify binary as format.

You can pass any payload format directly from source to target in Streaming mappings. You can project columns in binary format pass a payload from source to target in its original form or to pass a payload format that is not supported.

Streaming mappings can read, process, and write hierarchical data. You can use array, struct, and map complex data types to process the hierarchical data. You assign complex data types to ports in a mapping to flow hierarchical data. Ports that flow hierarchical data are called complex ports.

For more information about processing hierarchical data, see the *Informatica Big Data Management User Guide*.

Complex File Data Object Overview Properties

The Data Integration Service uses overview properties when it reads data from or writes data to a complex file.

Overview properties include general properties that apply to the complex file data object. They also include object properties that apply to the resources in the complex file data object. The Developer tool displays overview properties for complex files in the **Overview** view.

General Properties

The following table describes the general properties that you configure for complex files:

Property	Description
Name	The name of the complex file operation.
Description	The description of the complex file data object.

Property	Description
Native Name	Name of the HDFS connection.
Path Information	The path on the Hadoop file system.

Column

You can configure the name, native name, data type, precision, access type, scale, and description of the columns in the Azure Event Hubs resource.

Compression and Decompression for Complex File Targets

You can write compressed files, specify compression formats, and decompress files. You can use compression formats such as Bzip2 and Lz4, or specify a custom compression format.

You can compress sequence files at a record level or at a block level.

For information about how Hadoop processes compressed and uncompressed files, see the Hadoop documentation.

The following table describes the compression formats:

Compression Options	Description
None	The file is not compressed.
Auto	The Data Integration Service detects the compression format of the file based on the file extension.
DEFLATE	The DEFLATE compression format that uses a combination of the LZ77 algorithm and Huffman coding.
Gzip	The GNU zip compression format that uses the DEFLATE algorithm.
Bzip2	The Bzip2 compression format that uses the Burrows–Wheeler algorithm.
Lzo	The Lzo compression format that uses the Lempel–Ziv–Oberhumer algorithm. In a streaming mapping, the compression format is LZ4. The LZ4 compression format uses the LZ77 algorithm.
Snappy	The LZ77-type compression format with a fixed, byte-oriented encoding.
Custom	Custom compression format. If you select this option, you must specify the fully qualified class name implementing the <code>CompressionCodec</code> interface in the Custom Compression Codec field.

Complex File Data Object Write Operation Properties

The Data Integration Service uses write operation properties when it writes data to a complex file. Select the Input transformation to edit the general, ports, targets, and run-time properties.

General Properties

The Developer tool displays general properties for complex file targets in the **Write** view.

The following table describes the general properties that you configure for complex file targets:

Property	Description
Name	The name of the complex file. You can edit the name in the Overview view. When you use the complex file as a target in a mapping, you can edit the name in the mapping.
Description	The description of the complex file.

Ports Properties

Port properties for a physical data object include port names and port attributes such as data type and precision.

The following table describes the ports properties that you configure for complex file targets:

Property	Description
Name	The name of the resource.
Type	The native data type of the resource.
Precision	The maximum number of significant digits for numeric data types, or the maximum number of characters for string data types.
Scale	The scale for each column. Scale is the maximum number of digits that a column can accommodate to the right of the decimal point. Applies to decimal columns. The scale values you configure depend on the data type.
Detail	The detail of the data object.
Description	The description of the resource.

Target Properties

The target properties list the targets of the complex file data object.

The following table describes the target properties that you configure for complex file targets in a streaming mapping:

Property	Description
Target	The target which the complex data object writes to. You can add or remove targets.

Run-time Properties

The run-time properties include the name of the connection that the Data Integration Service uses to write data to the HDFS sequence file or binary file.

You can configure dynamic partitioning or fixed partitioning.

Advanced Properties

The Developer tool displays the advanced properties for complex file targets in the Input transformation in the **Write** view.

The following table describes the advanced properties that you configure for complex file targets in a streaming mapping:

Property	Description
Operation Type	Indicates the type of data object operation. This is a read-only property.
File Directory	The location of the complex file target. At run time, the Data Integration Service creates temporary directories in the specified file directory to manage the target files. If the directory is in HDFS, enter the path without the node URI. For example, <code>/user/lib/testdir</code> specifies the location of a directory in HDFS. The path must be 512 characters or less.
Overwrite Target	Streaming mappings do not support this property.
File Name	The name of the output file. Spark appends the file name with a unique identifier before it writes the file to HDFS.
File Format	The file format. Select one of the following file formats: <ul style="list-style-type: none">- Binary. Select Binary to read any file format.- Sequence. Select Sequence File Format for target files of a specific format that contain key and value pairs.
Output Format	The class name for files of the output format. If you select Output Format in the File Format field, you must specify the fully qualified class name implementing the <code>OutputFormat</code> interface.
Output Key Class	The class name for the output key. By default, the output key class is <code>NullWritable</code> .
Output Value Class	The class name for the output value. By default, the output value class is <code>Text</code> .
Compression Format	Optional. The compression format for binary files. Select one of the following options: <ul style="list-style-type: none">- None- Auto- DEFLATE- gzip- bzip2- LZ0- Snappy- Custom

Property	Description
Custom Compression Codec	Required for custom compression. Specify the fully qualified class name implementing the <code>CompressionCodec</code> interface.
Sequence File Compression Type	Optional. The compression format for sequence files. Select one of the following options: <ul style="list-style-type: none"> - None - Record - Block

Column Projection Properties

The following table describes the columns projection properties that you configure for complex file targets:

Property	Description
Column Name	The name of the column in the source table that contains data.
Type	The native data type of the resource.
Enable Column Projection	Indicates that you use a schema to publish the data to the target. By default, the columns are projected as binary data type. To change the format in which the data is projected, select this option and specify the schema format. When you create a complex file for an XML schema, select Binary as the resource format. After you create the object, enable column projection and manually select the schema format as XML.
Schema Format	The format in which you stream data to the target. Select one of the following formats: <ul style="list-style-type: none"> - XML - JSON - Avro
Use Schema	Specify the XSD schema for the XML format, the sample JSON for the JSON format, or sample Avro file for the Avro format.
Column Mapping	Click View to see the mapping of the data object to target mapping.
Project as Hierarchical Type	Project columns as complex data type for hierarchical data. For more information on hierarchical data, see the <i>Informatica Big Data Management User Guide</i> .

Complex File Execution Parameters

When you write to an HDFS complex file, you can configure how the complex file data object writes to the file. Specify these properties in the execution parameters property of the streaming mapping.

Use execution parameters to configure the following properties:

Rollover properties

When you write to an HDFS complex file, the file rollover process closes the current file that is being written to and creates a new file on the basis of file size or time. When you write to the HDFS file, you

can configure a time-based rollover or size-based rollover. You can use the following optional execution parameters to configure rollover:

- `rolloverTime`. You can configure a rollover of the HDFS file when a certain period of time has elapsed. Specify rollover time in hours. For example, you can specify a value of 1.
- `rolloverSize`. You can configure a rollover of the HDFS target file when the target file reaches a certain size. Specify the size in GB. The default rollover size is 1 GB.

The default is size-based rollover. You can implement both rollover schemes for a target file, in which case, the event that occurs first triggers a rollover. For example, if you set rollover time to 1 hour and rollover size to 1 GB, the target service rolls the file over when the file reaches a size of 1 GB even if the 1-hour period has not elapsed.

Pool properties

You can configure the maximum pool size that one Spark executor can have to write to a file. Use the `pool.maxTotal` execution parameter to specify the pool size. Default pool size is 8.

Retry Interval

You can specify the time interval for which Spark tries to create the target file or write to it if it fails to do so the first time. Spark tries a maximum of three times during the time interval that you specify. Use the `retryTimeout` execution parameter to specify the timeout in milliseconds. Default is 30,000 milliseconds.

Temporary Directory for the Complex File Target

At run time, the Data Integration Service creates a temporary directory in the location that you set for the **File Directory** property that you configure for the complex file targets in a streaming mapping. A temporary directory is created for each mapping. A temporary directory contains the target files to which the data is currently written.

Based on the rollover limit, data is written to the target file in the temporary directory. After a target file reaches the specified rollover limit, the target file is closed and moved to the specified target file directory from the temporary directory.

For example, if you set the **File Directory** property to `hdfs://demo.domain.com:8020/tmp/batch/`, you can find the target file to which data is currently written in the following directory: `hdfs://demo.domain.com:8020/tmp/batch/active/586bdfd67170008d125f52f68c1f02/`. The `hdfs://demo.domain.com:8020/tmp/batch/` directory contains the target files that are moved after the rollover limit is reached.

If a mapping fails, all the target files are closed and moved to the specified target file directory from the temporary directory, and the temporary directory will be deleted. In the preceding example, the temporary directory `586bdfd67170008d125f52f68c1f02` will be deleted.

If you use Big Data Management to process batch data, you can use the batch complex files that are created in the file directory as the source objects in Big Data Management.

HBase Data Objects

An HBase data object is a physical data object that represents data in an HBase resource. After you create an HBase connection, create an HBase data object with a write data operation to write data to an HBase table.

When you create an HBase data object, you can select an HBase table and view all the column families in the table. You can specify the column names in the column family if you know the column name and data type, or you can search the rows in the HBase table and specify the columns.

You can write to a column family or to a single binary column. When you create the data object, specify the column families to which you can write or choose to write all the data as a single stream of binary data.

Data Object Column Configuration

When you want to write data to columns in a column family, you can specify the columns when you create the HBase data object.

You can write data to columns in one of the following ways:

- Add the columns in the column families.
- Search for the columns names in the column family and add the columns.
- Get all the columns in a column family as a single stream of binary data.

When you manually create the data object, the column name should be of the following format:

`columnFamily__columnQualifier`

Add Columns

When you create a data object, you can specify the columns in one or more column families in an HBase table.

When you add an HBase table as the resource for an HBase data object, all the column families in the HBase table appear. If you know the details of the columns in the column families, you can select a column family and add the column details. In the **Column Families** dialog box, select the column family to which you want to add the columns. Column details include column name, data type, precision, and scale.

Although data is stored in binary format in HBase tables, you can specify the associated data type of the column to transform the data. To avoid data errors or incorrect data, verify that you specify the correct data type for the columns.

Verify that you specify valid column details when you add columns to avoid unexpected run-time behaviors. If you do not specify a value for a column when you write data to an HBase table, the Data Integration Service specifies a null value for the column at run time.

If the HBase table has more than one column family, you can add column details for multiple column families when you create the data object. Select one column family at a time and add the columns details. The column family name is the prefix for all the columns in the column family for unique identification.

Search and Add Columns

When you create a data object, you can search the rows in an HBase table to identify the column in the table and select the columns you want to add.

When you do not know the columns in an HBase table, you can search the rows in the table to identify all the columns and the occurrence percentage of the column. You can infer if the column name is valid based on

the number of times the column occurs in the table. For example, if column name eName occurs rarely while column name empName occurs in a majority of rows, you can infer the column name as empName.

When you search and add columns, you can specify the maximum number of rows to search and the occurrence percentage value for a column. If you specify the maximum numbers of rows as 100 and the column occurrence percent as 90, all columns that appear at least 90 times in 100 rows appear in the results. You can select the columns in the results to add the columns to the data object.

Get All Columns

Binary data or data that can be converted to a byte array can be stored in an HBase column. You can read from and write to an HBase tables in bytes.

When you create a data object, you can choose to get all the columns in a column family as a single stream of binary data.

Use the HBase data object as a target to write data in all the columns in the source data object as a single column of binary data in the target HBase table.

The Data Integration Service generates the data in the binary column based on the protobuf format. Protobuf format is an open source format to describe the data structure of binary data. The protobuf schema is described as messages.

HBase Object Overview Properties

The Data Integration Service uses overview properties when it writes data to an HBase resource.

Overview properties include general properties that apply to the HBase data object. They also include object properties that apply to the resources in the HBase data object. The Developer tool displays overview properties for HBase resources in the Overview view.

General Properties

The following table describes the general properties that you configure for the HBase data objects:

Property	Description
Name	Name of the HBase data object.
Location	The project or folder in the Model repository where you want to store the HBase data object.
Native Name	Native name of the HBase data object.
Path	Path to the HBase data object.

Add Column Properties

In the **Column Families** dialog box, select the column family to which you want to add the columns. The following table describes the column properties that you configure when you associate columns with column families:

Property	Description
Name	Name of the column in the column family.
Type	Data type of the column.

Property	Description
Precision	Precision of the data.
Scale	Scale of the data.

HBase Data Object Write Operation Properties

The Data Integration Service uses write operation properties when it writes data to an HBase resource.

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

Advanced Properties

The Developer tool displays the advanced properties for HBase targets in the **Advanced** view.

The following table describes the advanced properties that you can configure for an HBase data object in a streaming mapping:

Property	Description
Operation Type	The type of data object operation. This is a read-only property.
Date Time Format	Format of the columns of the date data type. Specify the date and time formats by using any of the Java date and time pattern strings.
Auto Flush	Optional. Indicates whether you want to enable Auto Flush. You can set auto flush to the following values: <ul style="list-style-type: none"> - Enable Auto Flush to set the value to true. The Data Integration Service runs each Put operation immediately as it receives them. The service does not buffer or delay the Put operations. Operations are not retried on failure. When you enable auto flush, the operations are slow as you cannot run operations in bulk. However, you do not lose data as the Data Integration Service writes the data immediately. - Disable Auto Flush to set the auto flush value to false. When you disable auto flush, the Data Integration Service accepts multiple Put operations before making a remote procedure call to perform the write operations. If the Data integration Service stops working before it flushes any pending data writes to HBase, that data is lost. Disable auto flush if you need to optimize performance. Default is disabled.
default Column Data Type	Streaming mappings do not support this property.
default Precision	Streaming mappings do not support this property.
default Scale	Streaming mappings do not support this property.
Default Column Family	Streaming mappings do not support this property.
Control File Location	Streaming mappings do not support this property.

JMS Data Objects

A JMS data object is a physical data object that accesses a JMS server. After you configure a JMS connection, create a JMS data object to write to JMS targets.

JMS providers are message-oriented middleware systems that send JMS messages. The JMS data object connects to a JMS provider to write data.

The JMS data object can write JMS messages to a JMS provider. When you configure a JMS data object, configure properties to reflect the message structure of the JMS messages. The input ports and output ports are JMS message headers.

When you configure the write data operation properties, specify the format in which the JMS data object writes data. You can specify XML, JSON, or Flat as format. When you specify XML format, you must provide an XSD file. When you specify JSON or Flat format, you must provide a sample file.

You can pass any payload format directly from source to target in Streaming mappings. You can project columns in binary format pass a payload from source to target in its original form or to pass a payload format that is not supported.

Streaming mappings can read, process, and write hierarchical data. You can use array, struct, and map complex data types to process the hierarchical data. You assign complex data types to ports in a mapping to flow hierarchical data. Ports that flow hierarchical data are called complex ports.

For more information about processing hierarchical data, see the *Informatica Big Data Management User Guide*.

Integration with JMS

You manually create JMS source and target data objects to reflect the message structure of JMS messages.

The JMS data object can read messages of type `TextMessage`. This type of message contains a string object. `TextMessages` can contain XML or JSON message data.

JMS Message Structure

JMS messages contain the following components:

- Header
- Properties
- Body

Header Fields

JMS messages contain a fixed number of header fields. Each JMS message uses these fields regardless of message type. Every JMS source and target definition includes a pre-defined set of header fields.

The following table describes the JMS message header fields:

Header Field	Description
JMSDestination	Destination to which the message is sent. JMS destinations can be a message queue or a recipient who listens for messages based on the message topic.
JMSDeliveryMode	Delivery mode of the message. The delivery mode can be persistent or non-persistent.

Header Field	Description
JMSMessageID	Unique identification value for the message.
JMSTimestamp	Time at which the message was handed off to the provider to be sent to the destination.
JMSCorrelationID	Links one message with another. For example, JMSCorrelationID can link a response message with the corresponding request message.
JMSReplyTo	Destination to which a reply message can be sent.
JMSRedelivered	Indicates that a message might have been delivered previously, but not acknowledged.
JMSType	Type of message based on a description of the message. For example, if a message contains a stock trade, the message type might be stock trade.
JMSExpiration	Amount of time in milliseconds the message remains valid. The messages remain in memory during this period.
JMSPriority	Priority of the message from 0-9. 0 is the lowest priority. 9 is the highest.

Property Fields

JMS source and target definitions can optionally include message property fields. Property fields contain additional message header information. JMS providers use properties in a JMS message to give provider-specific information. Applications that use a JMS provider can add property fields with application-specific information to a message.

Body Fields

JMS source and target definitions can optionally include a message body. The body contains one or more fields. Only certain types of JMS messages contain a body.

JMS Data Object Overview Properties

The Data Integration Service uses overview properties when it reads data from a JMS source.

Overview properties include general properties that apply to the JMS data object. They also include object properties that apply to the resources in the JMS data object. The Developer tool displays overview properties for JMS messages in the Overview view.

General Properties

The following table describes the general properties that you configure for JMS data objects:

Property	Description
Name	The name of the JMS data object.
Description	The description of the JMS data object.
Connection	The name of the JMS connection.

Objects Properties

The following table describes the objects properties that you configure for JMS data objects:

Property	Description
Name	The name of the topic or queue of the JMS source.
Description	The description of the JMS source.
Native Name	The native name of JMS source.
Path Information	The type and name of the topic or topic pattern of the JMS source.

JMS Data Object Write Operation Properties

The Data Integration Service uses write operation properties when it writes data to a JMS source. You can edit the format, run-time, and advanced properties.

General Properties

The Developer tool displays general properties for JMS targets in the **Write** view.

The following table describes the general properties that you view for JMS targets:

Property	Description
Name	The name of the write operation of JMS. This property is read-only. You can edit the name in the Overview view. When you use the JMS target as a target in a mapping, you can edit the name in the mapping.
Description	The description of the JMS target.

Ports Properties

Ports properties for a physical data object include port names and port attributes such as data type and precision.

The following table describes the ports properties that you configure for JMS targets:

Property	Description
Name	The name of the JMS target.
Type	The native data type of the target.
Precision	The maximum number of significant digits for numeric data types, or the maximum number of characters for string data types.
Scale	The scale of the data type.
Description	The description of the resource.

Target Properties

The target properties list the resources of the JMS data object. You can add or remove resources in the data object.

Run-time Properties

The run-time properties displays the name of the connection.

The following table describes the run-time property that you configure for JMS targets:

Property	Description
Connection	Name of the JMS connection.

Advanced Properties

The Developer tool displays the advanced properties for JMS targets in the Input transformation in the **Write** view.

You can configure the following advanced properties for JMS targets:

Property	Description
Operation Type	Specifies the type of data object operation. This is a read-only property.
Destination	Name of the queue or topic to which the JMS provider publishes messages. The data object subscribes to JMS messages from this queue or topic.
Message Type	The format in which the message is written to the target. Specify one of the following formats: <ul style="list-style-type: none">- Text- Binary

Column Projections Properties

The following table describes the columns projection properties that you configure for JMS targets:

Property	Description
Column Name	The name field that contains data. This property is read-only.
Type	The native data type of the resource. This property is read-only.
Enable Column Projection	Indicates that you use a schema to publish the data to the target. By default, the data is streamed in binary format. To change the format in which the data is streamed, select this option and specify the schema format.

Property	Description
Schema Format	The format in which you stream data to the target. You can select one of the following formats: <ul style="list-style-type: none"> - XML - JSON - Flat
Use Schema	Specify the XSD schema for the XML format, and the sample JSON for the JSON format. For the Flat file format, configure the schema to associate a flat file.
Column Mapping	The mapping of source data to the data object. Click View to see the mapping.

Kafka Data Objects

A Kafka data object is a physical data object that represents data in a Kafka stream. After you configure a Messaging connection, create a Kafka data object to write to Apache Kafka brokers.

Kafka runs as a cluster comprised of one or more servers each of which is called a broker. Kafka brokers stream data in the form of messages. These messages are published to a topic. When you write data to a Kafka messaging stream, specify the name of the topic that you publish to. You can also write to a Kerberised Kafka cluster.

Kafka topics are divided into partitions. Spark Structured Streaming can read the partitions of the topics in parallel. This gives better throughput and could be used to scale the number of messages processed. Message ordering is guaranteed only within partitions. For optimal performance you should have multiple partitions.

After you create a Kafka data object, create a write operation. You can use the Kafka data object write operation as a target in Streaming mappings. If you want to configure high availability for the mapping, ensure that the Kafka cluster is highly available.

When you configure the data operation properties, specify the format in which the Kafka data object writes data. You can specify XML, JSON, Avro, or Flat as format. When you specify XML format, you must provide a XSD file. When you specify Avro format, provide a sample Avro schema in a .avsc file. When you specify JSON or Flat format, you must provide a sample file.

You can pass any payload format directly from source to target in Streaming mappings. You can project columns in binary format pass a payload from source to target in its original form or to pass a payload format that is not supported.

Streaming mappings can read, process, and write hierarchical data. You can use array, struct, and map complex data types to process the hierarchical data. You assign complex data types to ports in a mapping to flow hierarchical data. Ports that flow hierarchical data are called complex ports.

For more information about processing hierarchical data, see the *Informatica Big Data Management User Guide*.

For more information about Kafka clusters, Kafka brokers, and partitions see <http://kafka.apache.org/11/documentation.html>.

Kafka Data Object Overview Properties

The Data Integration Service uses overview properties when it reads data from or writes data to a Kafka broker.

Overview properties include general properties that apply to the Kafka data object. They also include object properties that apply to the resources in the Kafka data object. The Developer tool displays overview properties for Kafka messages in the **Overview** view.

General Properties

The following table describes the general properties that you configure for Kafka data objects:

Property	Description
Name	The name of the Kafka read operation.
Description	The description of the Kafka data object.
Connection	The name of the Kafka connection.

Objects Properties

The following table describes the objects properties that you configure for Kafka data objects:

Property	Description
Name	The name of the topic or topic pattern of the Kafka data object.
Description	The description of the Kafka data object.
Native Name	The native name of Kafka data object.
Path Information	The type and name of the topic or topic pattern of the Kafka data object.

Column Properties

The following table describes the column properties that you configure for Kafka data objects:

Property	Description
Name	The name of the Kafka data object.
Native Name	The native name of the Kafka data object.
Type	The native data type of the Kafka data object.
Precision	The maximum number of significant digits for numeric data types, or the maximum number of characters for string data types.
Scale	The scale of the data type.
Description	The description of the Kafka data object.
Access Type	The type of access the port or column has.

Kafka Header Ports

The Kafka data object contains default header ports that represent metadata associated with events.

The following table lists the header ports of the Kafka data object:

Header ports	Data type	Description
partitionId	integer	Partition ID of the data. When the Kafka topic is divided into partitions, a sequential ID is assigned to each event and is unique to each event within the partition. Use the partition ID to identify the event that you want to consume from a particular partition.
key	binary	Key value associated with an event. You can group events based on the key value and then process the data.
TopicName	string	Name of the Kafka topic from where you receive events.

Kafka Data Object Write Operation Properties

The Data Integration Service uses write operation properties when it writes data to a Kafka broker.

General Properties

The Developer tool displays general properties for Kafka targets in the **Write** view.

The following table describes the general properties that you view for Kafka targets:

Property	Description
Name	The name of the write operation of Kafka broker.
Description	The description of the Kafka broker.

Ports Properties

Ports properties for a physical data object include port names and port attributes such as data type and precision.

The following table describes the ports properties that you configure for Kafka broker sources:

Property	Description
Name	The name of the resource.
Type	The native data type of the resource.
Precision	The maximum number of significant digits for numeric data types, or the maximum number of characters for string data types.
Scale	The scale of the data type.
Description	The description of the resource.

Run-time Properties

The run-time properties displays the name of the connection.

The following table describes the run-time property that you configure for Kafka targets:

Property	Description
Connection	Name of the Kafka connection.

Target Properties

The targets properties list the targets of the Kafka data object.

The following table describes the sources property that you can configure for Kafka targets:

Property	Description
Target	The target which the Kafka data object writes to. You can add or remove targets.

Advanced Properties

The Developer tool displays the advanced properties for Kafka targets in the Input transformation in the **Write** view.

The following table describes the advanced properties that you configure for Kafka targets:

Property	Description
Operation Type	Specifies the type of data object operation. This is a read-only property.
Metadata Fetch Timeout in milliseconds	The time after which the metadata is not fetched.
Batch Flush Time in milliseconds	The interval after which the data is published to the target.
Batch Flush Size in bytes	The batch size of the events after which the data is written to the target.
Producer Configuration Properties	The configuration properties for the producer. If the Kafka data object is writing data to a Kafka cluster that is configured for Kerberos authentication, include the following property: <code>security.protocol=SASL_PLAINTEXT,sasl.kerberos.service.name=kafka,sasl.mechanism=GSSAPI</code>

For more information about Kafka broker properties, see <http://kafka.apache.org/082/documentation.html>.

Column Projections Properties

The Developer tool displays the column projection properties in the **Properties** view of the write operation.

To specify column projection properties, double click on the write operation and select the data object. The following table describes the columns projection properties that you configure for Kafka targets:

Property	Description
Column Name	The field in the target that the data object writes to. This property is read-only.
Type	The native data type of the target. This property is read-only.
Enable Column Projection	Indicates that you use a schema to publish the data to the target. By default, the data is streamed in binary format. To change the streaming format, select this option and specify the schema format.
Schema Format	The format in which you stream data to the target. You can select one of the following formats: <ul style="list-style-type: none">- XML- JSON- Flat- Avro
Use Schema	Specify the XSD schema for the XML format, acsample file for JSON, or .avsc file for Avro format. For the Flat file format, configure the schema to associate a flat file to the Kafka target. When you provide a sample file, the Data Integration Service uses UTF-8 code page when writing the data.
Column Mapping	The mapping of data object to the target. Click View to see the mapping.
Project Column as Complex Data Type	Project columns as complex data type for hierarchical data. For more information, see the <i>Informatica Big Data Management User Guide</i> .

Microsoft Azure Data Lake Store Data Object

A Microsoft Azure Data Lake Store data object is a physical data object that represents data in a Microsoft Azure Data Lake Store table. After you create an Azure Data Lake Store connection, create a Microsoft Azure Data Lake store data object write operation to write to a Microsoft Azure Data Lake Store table.

You can use Microsoft Azure Data Lake Store to store data irrespective of size, structure, and format. Use Microsoft Azure Data Lake Store to process large volumes of data to achieve faster business outcomes.

When you configure the data operation properties, specify the format in which the data object writes data. You can specify XML, JSON, or Avro as format. When you specify XML format, you must provide an XSD file. When you specify Avro format, provide a sample Avro schema in a .avsc file. When you specify JSON, you must provide a sample file.

You cannot run a mapping with a Microsoft Data Lake Store data object on a MapR distribution.

You can pass any payload format directly from source to target in Streaming mappings. You can project columns in binary format pass a payload from source to target in its original form or to pass a payload format that is not supported.

Streaming mappings can read, process, and write hierarchical data. You can use array, struct, and map complex data types to process the hierarchical data. You assign complex data types to ports in a mapping to flow hierarchical data. Ports that flow hierarchical data are called complex ports.

For more information about processing hierarchical data, see the *Informatica Big Data Management User Guide*.

Microsoft Azure Data Lake Store Data Object Properties

The Microsoft Azure Data Lake Store Overview view displays general information about the Microsoft Azure Data Lake Store data object and the object properties that apply to the Microsoft Azure Data Lake Store table you import.

General Properties

You can configure the following properties for a Microsoft Azure Data Lake Store data object:

- Name. Name of the Microsoft Azure Data Lake Store data object.
- Description. Description of the Microsoft Azure Data Lake Store data object.
- Connection. Name of the Microsoft Azure Data Lake Store connection.

Microsoft Azure Data Lake Store Data Object Write Operation Properties

The Data Integration Service writes data to a Microsoft Azure Data Lake Store object based on the data object write operation. The Developer tool displays the data object write operation properties for the Microsoft Azure Data Lake Store data object in the Data Object Operation section.

You can view the data object write operation from the Input and Target properties.

Input properties

Represent data that the Data Integration Service reads from a Microsoft Azure Data Lake Store directory server. Select the input properties to edit the port properties and specify the advanced properties of the data object write operation.

Target properties

Represent data that the Data Integration Service writes to Microsoft Azure Data Lake Store. Select the target properties to view data, such as the name and description of the Microsoft Azure Data Lake Store object.

Input Properties of the Data Object Write Operation

Input properties represent data that the Data Integration Service writes to a Microsoft Azure Data Lake Store directory server. Select the input properties to edit the port properties of the data object write operation. You can also specify advanced data object write operation properties to write data to Microsoft Azure Data Lake Store objects.

The input properties of the data object write operation include general properties that apply to the data object write operation. Input properties also include port, source, and advanced properties that apply to the data object write operation.

You can view and change the input properties of the data object write operation from the **General**, **Ports**, **Targets**, **run-time**, and **Advanced** tabs.

Ports Properties - Input Write

The input ports properties list the data types, precision, and scale of the data object write operation.

The following table describes the input ports properties that you must configure in the data object write operation:

Property	Description
Name	Name of the port.
Type	Data type of the port.
Precision	Maximum number of significant digits for numeric data types, or maximum number of characters for string data types. For numeric data types, precision includes scale.
Scale	Maximum number of digits after the decimal point for numeric values.
Description	Description of the port.

Run-time Properties

The run-time properties display the name of the connection used for write transformation.

The following table describes the run-time properties that you configure for a Microsoft Azure Data Lake Store write operation:

Property	Description
Connection	Name of the Microsoft Azure Data Lake Store connection.

Advanced Properties

You can use the advanced properties to specify data object write operation properties to write data to a Microsoft Azure Data Lake Store server.

The following table describes the advanced properties that you configure for a Microsoft Azure Data Lake Store write operation:

Property	Description
Tracing Level	By default, the tracing level for every transformation is Normal. Change the tracing level to a Verbose setting when you need to troubleshoot a transformation that is not behaving as expected. Set the tracing level to Terse when you want the minimum amount of detail to appear in the log.
Directory path override	Overrides the default directory path. Note: You must manually create a temporary active directory within the target file directory of Azure Data Lake Store connection. The temporary active directory contains the target files to which the data is currently written. After the target file reaches the specified rollover limit, the target file is closed and moved to the specified target file directory from the temporary active directory.
File name override	Overrides the default file name.

Property	Description
Compress newly created File	Streaming mappings do not support this property.
Maintain row order	Streaming mappings do not support this property.
If file exists	Streaming mappings do not support this property.

Column Projection Properties

The Developer tool displays the column projection properties in the **Properties** view of the write operation.

To specify column projection properties, double click on the write operation and select the data object. The following table describes the columns projection properties that you configure for Azure Data Lake Store targets:

Property	Description
Column Name	The field in the target that the data object writes to. This property is read-only.
Type	The native data type of the target. This property is read-only.
Enable Column Projection	Indicates that you use a schema to publish the data to the target. By default, the data is streamed in binary format. To change the streaming format, select this option and specify the schema format.
Schema Format	The format in which you stream data to the target. You can select one of the following formats: <ul style="list-style-type: none"> - XML - JSON - Avro
Schema	Specify the XSD schema for the XML format, a sample file for JSON, or .avsc file for Avro format.
Column Mapping	The mapping of data object to the target. Click View to see the mapping.
Project as Hierarchical Type	Project columns as complex data type for hierarchical data. For more information, see the <i>Informatica Big Data Management User Guide</i> .

MapR Streams Data Objects

A MapR Streams data object is a physical data object that represents data in a MapR Stream. After you create a MapR Streams connection, create a MapR Streams data object to write data to MapR Streams.

Note: Support for MapR Streams data objects is deferred. Support will be reinstated in a future release.

Before you create and use MapR Stream data objects in Streaming mappings, complete the required prerequisites.

For more information about the prerequisite tasks, see the *Informatica Big Data Management Integration Guide*.

When you write data to a MapR stream, specify the name of the stream that you publish to.

After you create a MapR Streams data object, create a write data object operation. You can then add the data object write operation as a target in Streaming mappings.

When you configure the data operation properties, specify the format in which the MapR Streams data object writes data. You can specify XML, JSON, or Avro as format. When you specify XML format, you must provide an XSD file. When you specify Avro format, provide a sample Avro schema in a .avsc file. When you specify JSON, you must provide a sample file.

You can pass any payload format directly from source to target in Streaming mappings. You can project columns in binary format pass a payload from source to target in its original form or to pass a payload format that is not supported.

Streaming mappings can read, process, and write hierarchical data. You can use array, struct, and map complex data types to process the hierarchical data. You assign complex data types to ports in a mapping to flow hierarchical data. Ports that flow hierarchical data are called complex ports.

For more information about processing hierarchical data, see the *Informatica Big Data Management User Guide*.

MapR Streams Object Overview Properties

Overview properties include general properties that apply to the MapR Streams data object. The Developer tool displays overview properties in the Overview view.

General

Property	Description
Name	Name of the MapR Streams data object.
Description	Description of the MapR Streams data object.
Native Name	Native name of the MapR Stream.
Path Information	The path of the MapR Stream.

Column Properties

The following table describes the column properties that you configure for MapR Streams data objects:

Property	Description
Name	The name of the MapR Streams data object.
Native Name	The native name of the MapR Streams data object.
Type	The native data type of the MapR Streams data object.

Property	Description
Precision	The maximum number of significant digits for numeric data types, or the maximum number of characters for string data types.
Scale	The scale of the data type.
Description	The description of the MapR Streams data object.
Access Type	The access type of the port or column.

MapR Streams Data Object Write Operation Properties

The Data Integration Service uses write operation properties when it writes data to MapR Streams.

General Properties

The Developer tool displays general properties for MapR Stream targets in the **Write** view.

The following table describes the general properties that you view for MapR Stream targets:

Property	Description
Name	The name of the MapR Stream target. This property is read-only.
Description	The description of the MapR Stream target.

Port Properties

Ports properties for a physical data object include port names and port attributes such as data type and precision.

The following table describes the ports properties that you configure for MapR Stream targets:

Property	Description
Name	The name of the target.
Type	The native data type of the target.
Precision	The maximum number of significant digits for numeric data types, or the maximum number of characters for string data types.
Scale	The scale of the data type.
Description	The description of the target.

Target Properties

The targets properties list the targets of the MapR Streams data object.

The following table describes the sources property that you can configure for MapR Stream targets:

Property	Description
Target	The target which the MapR Streams data object writes to. You can add or remove targets.

Run-time Properties

The run-time properties displays the name of the connection.

The run-time property for MapR Stream target includes the name of the MapR Stream connection.

Column Projection Properties

The following table describes the columns projection properties that you configure for MapR Stream targets:

Property	Description
Column Name	The name field that contains data. This property is read-only.
Type	The native data type of the resource. This property is read-only.
Enable Column Projection	Indicates that you use a schema to write data to the target. By default, the data is streamed in binary format. To change the format in which the data is streamed, select this option and specify the schema format.
Schema Format	The format in which the source streams data. You can select one of the following formats: <ul style="list-style-type: none">- XML- JSON- Avro
Use Schema	Specify the XSD schema for the XML format, the sample JSON for the JSON format, or the sample .avsc file for the Avro format.
Column Mapping	The mapping of source data to the data object. Click View to see the mapping.
Project Column as Complex Data Type	Project columns as complex data type for hierarchical data. For more information, see the <i>Informatica Big Data Management User Guide</i> .

Relational Data Objects

Create a relational data object to write to Hive tables or JDBC-compliant database. To write to Hive tables, create a relational data object with a Hive connection. To write to a JDBC-compliant database, create a relational data object with a JDBC connection.

To use the relational data object as a target in streaming mappings, configure the relational data object write operation properties. You can select the mapping environment and run the mappings on the Spark engine of the Hadoop environment.

Hive targets

When you write to a Hive target in a streaming mapping, you write to a Hive table. You can write to the following types of tables:

- Managed or internal tables. When you write to a managed table or an internal table, Hive writes data to the Hive warehouse directory. If you enable the **Truncate target table** property in the Advanced properties while writing to the Hive table, the data in the table is overwritten. If you do not select this property, data is appended.
- External tables. When you write to an external table, you must truncate the target table to overwrite data. You can write to external partitioned tables but you cannot truncate external partitioned tables.

Truncation of tables happens only once in the beginning when you write data.

JDBC targets

You can include a JDBC-compliant database as a target in an Informatica mapping. Use the JDBC drivers in the JDBC connection to configure the JDBC parameters. JDBC drivers use the values that you configure in the **User Name** and **Password** fields of the JDBC connection. If you want to write to a JDBC target, you must create a target table .

Relational Data Object Overview Properties

The Data Integration Service uses overview properties when it writes data to a relational data object.

The Overview properties include general properties that apply to the relational data object. They also include column properties that apply to the resources in the relational data object.

General Properties

The following table describes the general properties that you configure for relational data objects:

Property	Description
Name	Name of the relational data object.
Description	Description of the relational data object.

Column Properties

The following table describes the column properties that you can view for relational data objects:

Property	Description
Name	Name of the column.
Native Type	Native data type of the column.
Precision	Maximum number of significant digits for numeric data types, or maximum number of characters for string data types. For numeric data types, precision includes scale.
Scale	Maximum number of digits after the decimal point for numeric values.
Description	Description of the column.
Nullable	Indicates if the column can contain null values.
Primary Key	Identifies a primary key column in a table that you import from a database source.

Advanced Properties

Advanced properties include run-time and other properties that apply to the relational data object. The Developer tool displays advanced properties for relational data object in the **Advanced** view.

Property	Description
Connection	Name of the Hive connection.
Owner	Name of the resource owner. This property is not applicable for Hive sources and targets.
Resource	Name of the resource.
Database Type	Type of the source. This property is read-only.
Resource Type	Type of the resource. This property is read-only.

Relational Data Object Write Operation Properties

The Data Integration Service uses write operation properties to write data to Hive or JDBC-compliant database.

The data object write operation properties include general, ports, run-time, target, and advanced properties.

General Properties

The general properties include the properties for name, description, and metadata synchronization.

The following table describes the general properties that you configure for the relational data object:

Property	Description
Name	Name of the relational data object. You can edit the name in the Overview view. When you use the relational file as a source in a mapping, you can edit the name within the mapping.
Description	Description of the relational data object.
When column metadata changes	Indicates whether object metadata is synchronized with the source. Select one of the following options: <ul style="list-style-type: none">- Synchronize output ports. The Developer tool reimports the object metadata from the source.- Do not synchronize. Object metadata may vary from the source.

Data Object Properties

On the Data Object tab, you can specify or change the target, and make relational and customized data object targets dynamic.

The following table describes the data object properties that you configure for Hive targets in a streaming mapping:

Property	Description
Specify By	To specify target columns and metadata, select one of the following options: <ul style="list-style-type: none">- Value. The write operation uses the associated data object to specify target columns and metadata.- Parameter. The write operation uses a parameter to specify target columns and metadata. Default is the Value option.
Data Object	If you created the target from an existing data object, the field displays the name of the object. Click Browse to change the data object to associate with the data object.
Parameter	Choose or create a parameter to associate with the target.
At runtime, get data object columns from data source	When you enable this option, the Data Integration Service fetches metadata and data definition changes from target tables to the data object.

Ports Properties

Ports properties include column names and column attributes such as data type and precision.

The following table describes the ports properties that you configure for relational targets:

Property	Description
Name	Name of the column.
Type	Native data type of the column.
Precision	Maximum number of significant digits for numeric data types, or maximum number of characters for string data types. For numeric data types, precision includes scale.
Scale	Maximum number of digits after the decimal point for numeric values.
Description	Description of the column.
Column	Name of the column in the resource.
Resource	Name of the resource.

Run-time Properties

The following table describes the run-time properties that you configure for Hive targets:

Property	Description
Connection	Name of the Hive connection.
Owner	Name of the Hive database.
Resource	Name of the resource.

Advanced Properties

The following table describes the advanced properties that you configure for Hive targets in streaming mappings:

Property	Description
Tracing level	Controls the amount of detail in the mapping log file.
Target Schema Strategy	Type of target schema strategy for the target table. You can select one of the following target schema strategies: <ul style="list-style-type: none">- RETAIN. The Data Integration Service retains the existing target schema.- CREATE. The Data Integration Service drops the target table at run time and replaces it with a table based on a target table that you identify.- APPLYNEWCOLUMNS. The Data Integration Service alters the target table by applying new columns from the associated data object or mapping flow to the target table.- APPLYNEWSHEMA. The Data Integration Service alters the target table and applies the new schema from the associated data object or mapping flow to the target table.- FAIL. The Data Integration Service fails the mapping if the target schema from the mapping flow is different from the schema of the target table.- Assign Parameter. You can assign a parameter to represent the value for the target schema strategy and then change the parameter at run time.
DDL Query	The DDL query based on which the Data Integration Service creates or replace the target table at run time. You cannot create a custom DDL query that creates or replaces a Hive table at run time in streaming mappings.
Truncate target table	Truncates the target before loading data. Note: If the mapping target is a Hive partition table, you can choose to truncate the target table only with Hive version 0.11.
Truncate Hive Target Partition	Truncates the external table before loading data. Streaming mappings do not support this property. The property is ignored.
PreSQL	The SQL command the Data Integration Service runs against the target database before it reads the source. The Developer tool does not validate the SQL.
PostSQL	The SQL command that the Data Integration Service runs against the target database after it writes to the target. The Developer tool does not validate the SQL. Streaming mappings do not support Post-SQL queries. Post-SQL queries in streaming mappings are ignored.
Maintain Row Order	Maintain row order while writing data to the target. Select this option if the Data Integration Service should not perform any optimization that can change the row order. When the Data Integration Service performs optimizations, it might lose the row order that was established earlier in the mapping. You can establish row order in a mapping with a sorted flat file source, a sorted relational source, or a Sorter transformation. When you configure a target to maintain row order, the Data Integration Service does not perform optimizations for the target.

CHAPTER 5

Streaming Mappings

This chapter includes the following topics:

- [Streaming Mappings Overview, 88](#)
- [Connections, 89](#)
- [Data Objects, 89](#)
- [Transformations in a Streaming Mapping, 92](#)
- [Stateful Computing, 94](#)
- [Rules in a Streaming Mapping, 96](#)
- [Mapping Configurations, 97](#)
- [Mapping Validation, 99](#)
- [Monitor Jobs, 101](#)
- [Streaming Mapping Example, 102](#)
- [High Availability Configuration, 102](#)
- [Troubleshooting Streaming Mappings, 103](#)

Streaming Mappings Overview

Use the Developer tool to create and run streaming mappings in the Hadoop run-time environment and process data that is in JSON, XML, CSV, or Avro format.

Develop a mapping to read, transform, and write data according to your business needs. When you create a streaming mapping, select the Hadoop environment and the Spark engine. When you run a streaming mapping, the Data Integration Service pushes the processing to nodes on a Spark engine in the Hadoop cluster.

Use the following steps as a guideline when you develop a streaming mapping:

1. Create connections that you want to use to access streaming data.
2. Create input, output, and reusable objects that you want to use in the mapping. Create physical data objects to use as mapping input or output.
3. Create reusable transformations that you want to use.
4. Create rules.
5. Create the streaming mapping.
6. Add objects to the mapping.

7. Link ports between mapping objects to create a flow of data from sources to targets, through transformations.
8. Configure the mapping
9. Validate the mapping to identify errors.
10. Save the mapping and run it to see the mapping output.

Connections

A connection is a repository object that defines a connection in the domain configuration repository.

Create a connection to import data objects, preview data, and run mappings.

For more information about the connections that you can use in streaming mappings, see the Appendix, "Connections", in this guide.

Data Objects

Based on the type of source you are reading from or target you are writing to you can create physical data objects to access the different types of data.

The following table lists the data objects that you can include in streaming mappings and the read and write data object operations for each:

Data Object	Source	Data Object Operation	Target	Data Object Operation
Amazon Kinesis	Amazon Kinesis Stream	Read	Amazon Kinesis Firehose Delivery Stream	Write
Azure Data Lake Store	-	-	Microsoft Azure Data Lake Store	Write
Azure Event Hub	Azure Event Hub Service	Read	Azure Event Hub Service	Write
Complex File	-	-	HDFS sequence file or binary file	Write
HBase	-	-	HBase tables	Write
JMS	JMS server	Read	JMS server	Write
Kafka	Kafka broker	Read	Kafka broker	Write
MapR Streams	MapR Stream	Read	MapR Stream	Write
Relational	-	-	Hive table JDBC-compliant database	Write

RELATED TOPICS:

- [“Complex File Data Objects” on page 59](#)
- [“HBase Data Objects” on page 65](#)
- [“JMS Data Objects” on page 68](#)
- [“Kafka Data Objects” on page 72](#)
- [“Microsoft Azure Data Lake Store Data Object” on page 76](#)
- [“MapR Streams Data Objects” on page 79](#)
- [“Relational Data Objects” on page 83](#)
- [“Amazon Kinesis Data Objects” on page 23](#)
- [“Azure Event Hubs Data Objects” on page 28](#)
- [“JMS Data Objects” on page 33](#)
- [“Kafka Data Objects” on page 38](#)
- [“MapR Streams Data Objects” on page 44](#)

Creating a Data Object

Create a data object to add to a streaming mapping.

1. Select a project or folder in the **Object Explorer** view.
2. Click **File > New > Data Object**.
3. Select the data object that you want to add to the streaming mapping and click **Next**.
The 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 connection that you want to use.
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 resource 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.

RELATED TOPICS:

- [“Amazon Kinesis Data Objects” on page 23](#)
- [“Azure Event Hubs Data Objects” on page 28](#)
- [“JMS Data Objects” on page 33](#)
- [“Kafka Data Objects” on page 38](#)
- [“MapR Streams Data Objects” on page 44](#)
- [“Complex File Data Objects” on page 59](#)
- [“HBase Data Objects” on page 65](#)
- [“JMS Data Objects” on page 68](#)
- [“Kafka Data Objects” on page 72](#)
- [“Microsoft Azure Data Lake Store Data Object” on page 76](#)

- [“MapR Streams Data Objects” on page 79](#)
- [“Relational Data Objects” on page 83](#)

Creating a Data Object Operation

You can create the data object read or write operation for a data object depending on the source that you read from or target that you write to. You can then add the object operation to a streaming 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 StreamRead or StreamWrite operation.
The Data Integration Service uses read operation properties when it reads data and write operation properties when it writes data.
5. Click **Add**.
The **Select Resources** dialog box appears.
6. Select the 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.

RELATED TOPICS:

- [“Amazon Kinesis Data Object Read Operation Properties” on page 24](#)
- [“Amazon Kinesis Data Object Write Operation Properties” on page 52](#)
- [“Azure Event Hubs Data Object Read Operation Properties” on page 29](#)
- [“Azure Event Hubs Data Object Write Operations” on page 56](#)
- [“JMS Data Object Read Operation Properties” on page 35](#)
- [“JMS Data Object Write Operation Properties” on page 70](#)
- [“Kafka Data Object Read Operation Properties” on page 41](#)
- [“Kafka Data Object Write Operation Properties” on page 74](#)
- [“MapR Streams Data Object Read Operation Properties” on page 46](#)
- [“MapR Streams Data Object Write Operation Properties” on page 81](#)
- [“Complex File Data Object Write Operation Properties” on page 60](#)
- [“HBase Data Object Write Operation Properties” on page 67](#)
- [“Microsoft Azure Data Lake Store Data Object Write Operation Properties” on page 77](#)
- [“Relational Data Object Write Operation Properties” on page 84](#)

Transformations in a Streaming Mapping

Informatica Developer provides a set of transformations that perform specific functions. Some restrictions and guidelines apply to processing transformations in a streaming mapping.

You can use the following transformations in a streaming mapping:

Aggregator

Mapping validation fails when the

- The transformation contains stateful variable ports.
- The transformation contains unsupported functions in an expression.

Data Masking

Supported without restrictions.

Note: Support for the Data Masking transformation is deferred. Support will be reinstated in a future release.

Expression

Mapping validation fails when the transformation contains unsupported functions in an expression.

If an expression results in numerical errors, such as division by zero or SQRT of a negative number, it returns an infinite or an NaN value. In the native environment, the expression returns null values and the rows do not appear in the output.

Filter

Supported without restrictions.

Java

The following restrictions apply to the Java transformation:

- The value Transaction for transformation scope is not valid.
- The transformation is always stateless.
- The Partitionable field is ignored.

You can use all data types to process hierarchical data.

For more information about Java transformation support on the Spark engine, see the *Informatica Big Data Management User Guide*.

Joiner

Mapping validation fails in the following situations:

- Case sensitivity is disabled.

Lookup

Use a Lookup transformation to look up data in a flat file, HDFS, Hive, relational, and HBase data.

Mapping validation fails in the following situations:

- Case sensitivity is disabled.
- The lookup is a data object.

The mapping fails in the following situations:

- The transformation is unconnected.

You cannot use a float data type to look up data in a Hive table as comparing equality of floating point numbers is unsafe.

When you configure the transformation to return the first, last, or any value on multiple matches, the Data Integration Service returns any value.

Note: Informatica recommends that you select the **Ignore null values that match** property in Lookup transformation advanced properties to avoid cross join of DataFrames.

To use a Lookup transformation on uncached HBase tables, perform the following steps:

1. Create an HBase data object. When you add an HBase table as the resource for a HBase data object, include the ROW ID column.
2. Create a HBase read data operation and import it into the streaming mapping.
3. When you import the data operation to the mapping, select the **Lookup** option.
4. In the Lookup tab, configure the following options:
 - Lookup column. Specify an equality condition on ROW ID
 - Operator. Specify =
5. Verify that format for any date value in the HBase tables is of a valid Java date format. Specify this format in the **Date Time Format** property of the **Advanced Properties** tab of the data object read operation.

Note: If an HBase lookup does not result in a match, it generates a row with null values for all columns. You can add a Filter transformation after the Lookup transformation to filter out null rows.

Mapping validation fails in the following situations:

- If you do not include ROW ID in the condition
- If you specify any operator other than =
- If you include multiple conditions in the transformation.
- If you select column of type date from input columns.
- If you look up binary data.

Normalizer

Supported without restrictions.

Python

When you close a Jep instance, you might not be able to call CPython modules.

Rank

Mapping validation fails if case-sensitivity is disabled.

Note: Support for the Rank transformation is deferred. Support will be reinstated in a future release.

Router

Supported without restrictions.

Sorter

To use the Sorter transformation in a Streaming mapping, configure the following properties:

- Advanced properties of the data object write properties. Enable the **Maintain Row Order** field.
- Custom properties of the Data Integration Service. Set the `ExecutionContextOptions.Infa.HonorTargetOrdering` property to true if there are one or more transformations between the Sorter transformation and the target.

Mapping validation fails in the following situations:

- Case sensitivity is disabled.

The Data Integration Service logs a warning and ignores the Sorter transformation in the following situations:

- There is a type mismatch in between the target and the Sorter transformation sort keys.
- The transformation contains sort keys that are not connected to the target.
- The transformation is not directly upstream from the Write transformation.

The Data Integration Service treats null values as high even if you configure the transformation to treat null values as low.

Note: Support for the Sorter transformation is deferred. Support will be reinstated in a future release.

Union

Supported without restrictions.

Window

The Window transformation is supported with restrictions. The Window transformation applies only to streaming mappings.

For more information about the restrictions, see the [“Rules and Guidelines for the Window Transformations” on page 112](#).

Transformations that are not listed here are not supported.

For more information about the transformations, see the *Informatica Developer Transformation Guide*.

For more information about transformation support on the Spark engine, see the *Informatica Big Data Management User Guide*.

Stateful Computing

Stateful computing involves storing and retrieving state while evaluating expressions in an Expression transformation.

You can use variable port definitions in Expression transformations to perform stateful computing in a streaming mapping.

Use variable port definitions to perform the following tasks:

- Calculate and store the state in stateful variables when an event arrives.
- Use the stored state and input to compute the output.

When you configure the Expression transformation, configure the partition keys on the **Windowing** tab. When you define variable ports in an Expression transformation, you can optionally specify a partition key that uses

one or more input ports. The partition keys determine which columns to group the data by while performing stateful computing. The stored state will have one value for every stateful variable for each value of the partition key.

The evaluation of the ports is ordered. The output ports are computed after the variable ports are computed and contain updated values of variables.

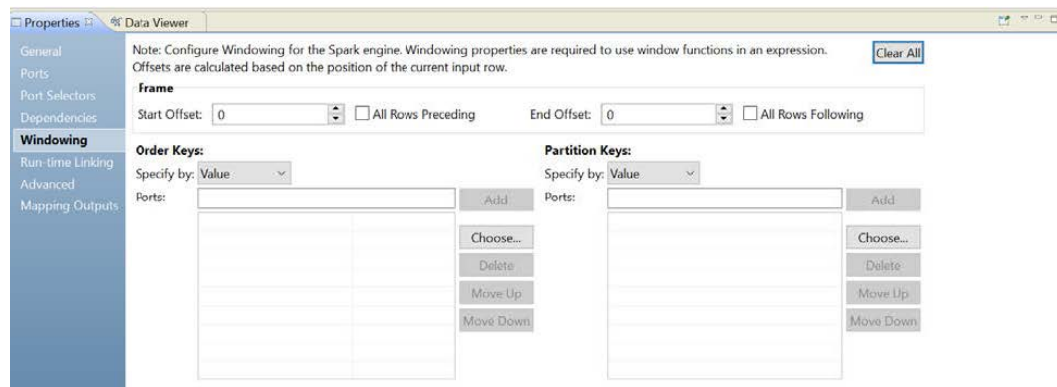
When you configure windowing properties, you define a stateful expression in the Expression transformation. Streaming mappings support all the expression transform function except Window functions and aggregate functions.

For more information about guidelines for configuring variable ports, see the *Informatica Developer Transformation Guide*.

Partitioning Configuration

Configure partitioning by specifying one or more columns under partition keys on the **Windowing** tab.

The following image shows the Windowing tab:



Optionally, configure the partition keys to separate the input rows into different partitions. Configure the partition keys to define partition boundaries, rather than performing the calculation across all inputs. If you do not define partition keys, all the data is included in the same partition. The variable values stored will be global and every row can read and update the same set of variables.

You can specify the partition keys by value or parameter. Select **Value** to use port names.

The following table lists the data type support for variable ports:

Data Type	Initial State Value
String	EMPTY_STRING
Integer	0
Double	0.0
Long	0
Text	EMPTY_STRING
Decimal	0.0

Example

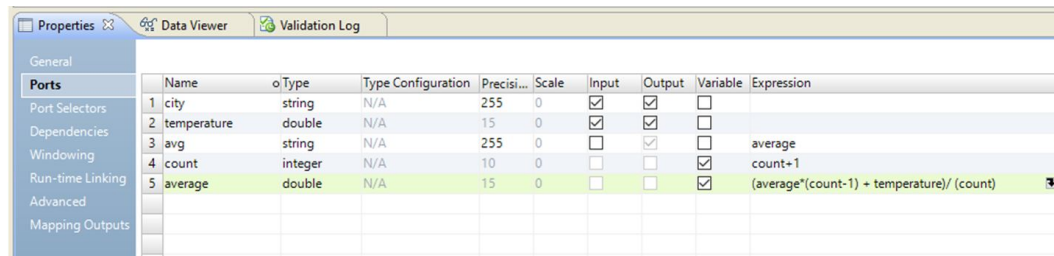
You want to compute average temperature of cities in a region. Add an Expression transformation to your streaming mapping.

Configure the following ports in the transformation:

- city. Input port of string data type.
- temperature. Input port of double data type.
- avg. Output port that displays the average temperature.
- count. Variable port of integer data type with expression 'count+1'.
- average. Variable port of double data type with expression $(\text{average} * (\text{count}-1) + \text{temperature}) / (\text{count})$

You partition the data by "city". The "average" corresponds to previously stored value for the "average" and is computed with each update in the value of "count". Since "count" is already incremented when "average" is evaluated, specify "count-1" in the expression to get the new average.

The following image shows an Expression transformation:



	Name	Type	Type Configuration	Precision	Scale	Input	Output	Variable	Expression
1	city	string	N/A	255	0	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
2	temperature	double	N/A	15	0	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
3	avg	string	N/A	255	0	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	average
4	count	integer	N/A	10	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	count+1
5	average	double	N/A	15	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	$(\text{average} * (\text{count}-1) + \text{temperature}) / (\text{count})$

Rules in a Streaming Mapping

A rule expresses the business logic that defines conditions applied to source data. You can add expression rules to streaming mapping to cleanse, change, or validate data. Create expression rules in the Analyst tool.

You might want to use a rule in different circumstances. You can add a rule to cleanse one or more data columns.

Rules that you create in the Analyst tool appear as mapplets in the Developer tool. You can use a mapplet in a mapping or validate the mapplet as a rule.

In a streaming mapping, you can use the following functions in expression rules:

- LastDay
- Add to Date
- Concat
- Date Diff
- Date Time
- Greatest
- Choose
- Least
- Length

- Null
- Reverse
- Truncate
- Convert to Data

For more information about rules, see the *Informatica Profile Guide*.

For more information about mapplets, see the *Informatica Mapping Guide*.

Mapping Configurations

To configure a mapping, configure the connection and run-time properties for the mapping.

When you configure the mapping, configure the following properties:

Validation Environment

The environment in which the validations are done. Select Hadoop in the validation environment and select the Spark engine. The Data Integration Service pushes the mapping logic to the Spark engine..

Execution Environment

The environment in which the mappings are executed. Select Hadoop as the execution environment.

Hadoop

Specify the following properties for the Spark engine:

- **Connection.** Select the connection to the Spark engine used for pushdown of processing. Select **Connection** and browse for a connection or select a connection parameter.
- **Runtime Properties.** An optional list of configuration parameters to apply to the Spark engine. You can change the default Spark configuration properties values, such as `spark.executor.memory` or `spark.driver.cores`.

Use the following format:

```
<property1>=<value>
```

- `<property1>` is a Spark configuration property.
- `<value>` is the value of the property.

To enter multiple properties, separate each name-value pair with the following text: `& :`

To use a JMS source or Amazon Kinesis Streams source in the mapping, configure two or more executors for the mapping. For example, use the following configuration:

```
spark.executor.instances=2 &: spark.executor.cores=2 &: spark.driver.cores=1
```

To use an AWS credential profile, configure the following properties for the mapping:

- `spark.yarn.appMasterEnv.AWS_CREDENTIAL_PROFILES_FILE=<absolute path to the credentials file>/credentials`
- `spark.executorEnv.AWS_CREDENTIAL_PROFILES_FILE=<absolute path to the credentials file>/credentials`
- `spark.driverEnv.AWS_CREDENTIAL_PROFILES_FILE=<absolute path to the credentials file>/credentials`

In the case of a mapping failure, to enable the mapping to start reading data from the time of failure, configure the `infaspark.checkpoint.directory` property. Specify a directory that is created under the directory you specify in the **State Store** property such as `infaspark.checkpoint.directory <directory>`.

For example: `infaspark.checkpoint.directory restart1`.

Note: Ensure that the checkpoint directory path does not start with a forward slash (/).

Source Configuration

Specify the following properties to configure how the data is processed:

- **Maximum Rows Read.** Specify the maximum number of rows that are read before the mapping stops running. Default is `Read All Rows`.
- **Maximum Runtime Interval.** Specify the maximum time to run the mapping before it stops. If you set values for this property and the **Maximum Rows Read** property, the mapping stops running after one of the criteria is met. Default is `Run Indefinitely`. A value of `Run Indefinitely` enables the mapping to run without stopping.
- **State Store.** Specify the HDFS location on the cluster to store information about the state of the Spark Job. Default is `/tmp`.
You can configure the state store as part of the configuration of execution options for the Data Integration Service.

You can use these properties to test the mapping.

Streaming Properties

Specify the following streaming properties:

- **Batch interval.** The Spark engine processes the streaming data from sources and publishes the data in batches. The batch interval is number of seconds after which a batch is submitted for processing.
- **Cache refresh interval.** You can cache a large lookup source or small lookup tables. When you cache the lookup source, the Data Integration Service queries the lookup cache instead of querying the lookup source for each input row. You can configure the interval for refreshing the cache used in a relational Lookup transformation.

Run Configurations

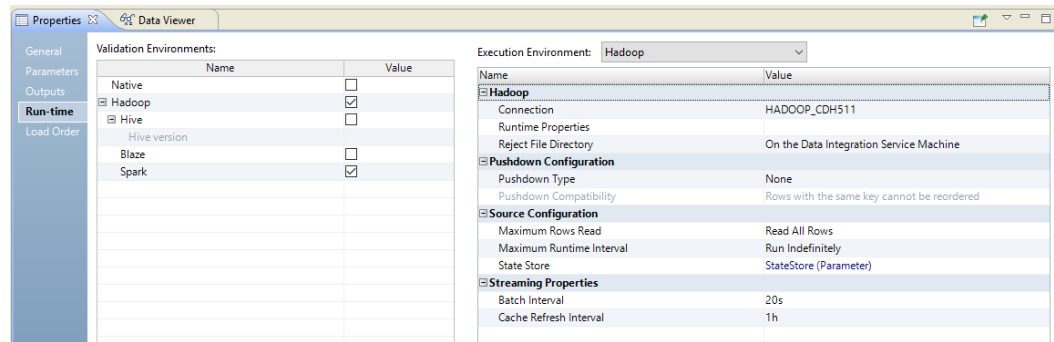
The Developer tool applies configuration properties when you run streaming mappings. Set configuration properties for streaming mappings in the **Run** dialog box.

Configure the following source properties:

- **Read all rows.** Reads all rows from the source.
- **Read up to how many rows.** The maximum number of rows to read from the source if you do not read all rows.
- **Maximum runtime interval.** The maximum time to run the mapping before it stops. If you set values for this property and the **Maximum Rows Read** property, the mapping stops running after one of the criteria is met.

When you run the mapping, the Data Integration Service converts the mapping to a Scala program and package it in a JAR file and sends it to the Hadoop cluster. You can view the details in the Spark execution plan in the Developer tool or Administrator tool.

The following image shows the connection and run-time properties:



Mapping Validation

When you develop a streaming mapping, you must configure it so that the Data Integration Service can read and process the entire mapping. The Developer tool marks a mapping as not valid when it detects errors that will prevent the Data Integration Service from running the mapping.

The Developer tool considers the following types of validation:

- Environment
- Data object
- Transformation
- Run-time

Environment Validation

The Developer tool performs environment validation each time you validate a streaming mapping.

The Developer tool generates an error in the following scenarios:

- The mapping has Native environment.
- The mapping does not have a Spark validation environment and Hadoop execution environment.
- The mapping has a Hadoop execution environment and has Native, Blaze, Hive on MapReduce, and Spark validation environment.

Data Object Validation

When you validate a mapping, the Developer tool verifies the source and target data objects that are part of the streaming mapping.

The Developer tool generates an error in the following scenarios:

- The mapping contains a source data object other than Kafka or JMS, and target data object other than Kafka, complex file, or relational data object.
- The read and write properties of the Kafka, JMS, and complex file data objects are not specified correctly.
- If you add a Kafka or a JMS data object to an LDO mapping, REST mapping, or a mapplet.

Transformation Validation

When you validate a mapping, the Developer tool performs validation on the transformations that are part of the streaming mapping.

The Developer tool performs the following validations:

- You must use a Window transformation directly upstream from an Aggregator transformation in a streaming mapping.
- You cannot use multiple Aggregator transformations in the same streaming pipeline.
- If you connect the Window port from a Window transformation to an Aggregator transformation, you cannot connect the Window port to any downstream transformation.
- If a mapplet contains an Aggregator transformation, you must include a Window transformation directly upstream from the mapplet.
- You must use a Window transformation directly upstream from a Joiner transformation in a streaming mapping.
- You cannot use multiple Joiner transformations in the same streaming pipeline.
- A mapping contains only those transformations supported for streaming mappings.
- You cannot use an Aggregator transformation and a passive Lookup transformation that is configured with an inequality lookup condition in the same pipeline.
- You cannot use more than one passive Lookup transformation that is configured with an inequality lookup condition in the same pipeline.
- You cannot use an Aggregator transformation anywhere before a Joiner transformation in a streaming mapping.
- If you connect the Window port from a Window transformation to a Joiner transformation, you cannot connect the Window port to any downstream transformation.
- If a mapplet contains a Joiner transformation, you must include a Window transformation directly upstream from the mapplet.
- The window port from the Window transformation cannot be connected to more than one downstream transformation.
- A Window transformation is added between a streaming source and an Aggregator, or Joiner transformation.
- A Window transformation has at least one upstream streaming source.
- All Window transformations have a slide interval that is a multiple of the mapping batch interval.
- The slide interval of a sliding Window transformation is less than window size.
- The format of the parameter of the window size has the TimeDuration parameter type.
- The window size and slide interval of a Window transformation are greater than 0.
- The downstream Window transformation in the pipelines leading to a Joiner transformation has the same slide intervals.
- A Window transformation is not added to a Logical Data Object mapping, REST mapping, or a mapplet.
- If one pipeline leading to a Union transformation has a Window transformation, all streaming pipelines have a Window transformation. All downstream Window transformations in the pipelines leading to the Union transformations have the same slide intervals.
- A Union transformation does not merge data from streaming and non-streaming pipelines.
- A Joiner transformation does not join data from streaming and non-streaming pipelines.

Run-time Validation

The Developer Tool performs validations each time you run a streaming mapping.

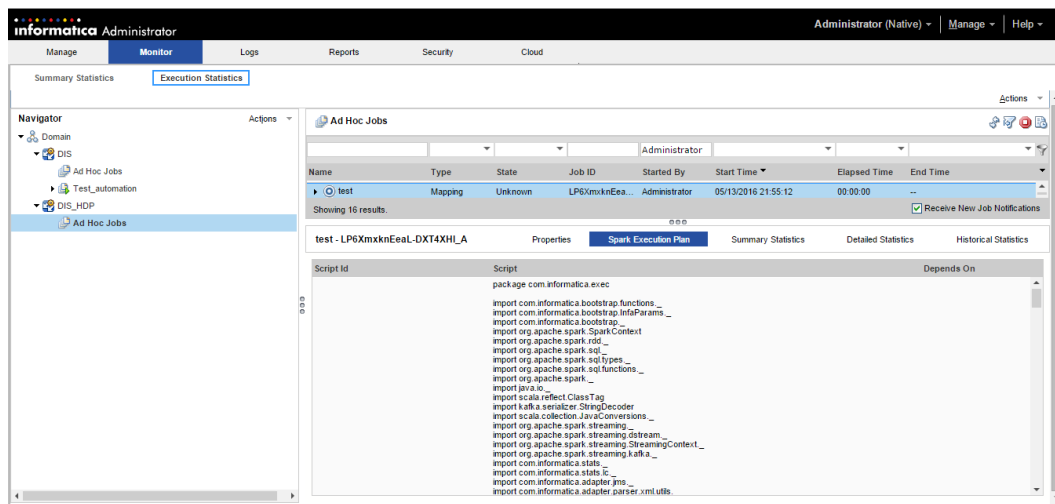
The Developer tool generates an error in the following scenarios:

- The state store is not configured when you specify the source configuration properties for the mapping.
- Either the Maximum Rows Read or the Maximum Runtime Interval property is not configured when you specify the source configuration properties for the mapping.
- The Maximum Runtime Interval does not have the correct time format.
- The Batch Interval does not have the correct time format.
- If the default value of the Batch Interval and Maximum Runtime Interval properties are not specified.

Monitor Jobs

You can monitor statistics and view log events for a streaming mapping job in the **Monitoring** tab of the Administrator tool.

The following image shows the **Monitor** tab in the Administrator tool:



Note: Support for monitoring jobs is deferred. Support will be reinstated in a future release.

Use the **Execution Statistics** view of the **Monitor** tab to monitor properties, run-time statistics, and run-time reports.

The **Spark Execution Plan** view appears when you run a streaming mapping with the Spark engine in the Hadoop environment and displays the execution plan for the Spark engine mapping.

View the streaming mapping statistics in the **Spark Execution Plan** view of the **Execution Statistics** view.

Note: If a failover occurs, the statistics might not be accurate.

Streaming Mapping Example

You run the IT department of a major bank that has millions of customers. You want to monitor network activity in real time. You need to collect network activity data from various sources such as firewalls or network devices to improve security and prevent attacks. The network activity data includes Denial of Service (DoS) attacks and failed login attempts made by customers. The network activity data is written to Kafka queues.

Create a Streaming mapping to read the network activity data and write the data to HDFS.

You can use the following objects in the Streaming mapping:

Kafka data object

The input file is a Kafka queue that contains the network activity data.

Create a Kafka data object. Configure a Kafka connection and specify the queue that contains the network activity data as a resource for the data object. Create a data object read operation and configure the properties. Drag the data object into the mapping as a source data object.

Transformations

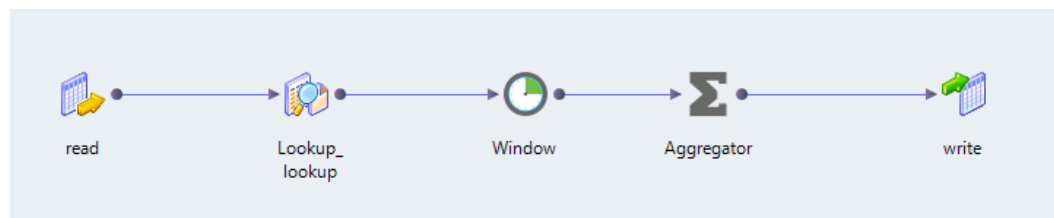
Add a Lookup transformation to get data from a particular customer ID. Add a Window transformation to accumulate the streamed data into data groups before processing the data.

HDFS complex file data object

Create a complex file data object. Configure an HDFS connection to write to an HDFS sequence file. Create the data object write operation and configure the properties. Drag the data object into the mapping as a target data object.

Link ports between mapping objects to create a flow of data.

The following image shows the sample mapping:



When you run the mapping, the data is read from the Kafka queue and written to the HDFS sequence file.

High Availability Configuration

To configure high availability for the streaming mapping, configure a state store directory for the source and guaranteed processing of the messages streamed by the source. Also configure the Spark execution parameters to enable the mapping to run without failing.

To configure high availability, perform the following configurations:

State store configuration

Configure a state store directory. Spark uses the state store directory to store the checkpoint information at regular intervals during the execution of the mapping. If a failure occurs, Spark restarts processing by reading from this state store directory.

Execution parameters

To ensure that the mapping runs without failing, configure the maximum number of tries to submit the mapping to Spark for processing. Configure the `spark.yarn.maxAppAttempts` and `yarn.resourcemanager.am.max-attempts` execution parameters when you configure the mapping properties. The values that you specify for both parameters must be equal and less than the values configured on the CDH or HortonWorks configuration.

Troubleshooting Streaming Mappings

When I run a streaming mapping, the mapping fails, and I see the following errors in the application logs of the Hadoop cluster:

```
User class threw exception: org.apache.spark.SparkException: Job aborted due to stage failure:
Task 0 in stage 1.0 failed 4 times, most recent failure: Lost task 0.3 in stage 1.0 (TID 4, localhost):
java.lang.Exception: Retry Failed: Total 3 attempts made at interval 10000ms
at
com.informatica.adapter.streaming.hdfs.common.RetryHandler.errorOccured(RetryHandler.java:74)
at
com.informatica.adapter.streaming.hdfs.HDFSMessageSender.sendMessage(HDFSMessageSender.java:55)
at com.informatica.bootstrap.InfaStreaming$$anonfun$writeToHdfsPathRealtime$1$$anonfun$apply$5.apply(InfaStreaming.scala:144)
at com.informatica.bootstrap.InfaStreaming$$anonfun$writeToHdfsPathRealtime$1$$anonfun$apply$5.apply(InfaStreaming.scala:132)
at org.apache.spark.rdd.RDD$$anonfun$foreachPartition$1$$anonfun$apply$28.apply(RDD.scala:902)
at org.apache.spark.rdd.RDD$$anonfun$foreachPartition$1$$anonfun$apply$28.apply(RDD.scala:902)
at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkContext.scala:1916)
at org.apache.spark.SparkContext$$anonfun$runJob$5.apply(SparkContext.scala:1916)
at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:70)
at org.apache.spark.scheduler.Task.run(Task.scala:86)
at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:274)
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:617)
at java.lang.Thread.run(Thread.java:745)
```

This error occurs if the HDFS NameNode is configured incorrectly.

To resolve this error, ensure that you specify the NameNode URI correctly in the HDFS connection and that the NameNode is up and running.

When I try to run streaming mappings concurrently, a few of the mappings fail and I get the following error in the Data Integration Service logs:

```
Caused by: java.lang.OutOfMemoryError: Java heap space
at java.util.Arrays.copyOf(Arrays.java:3332)
at java.lang.AbstractStringBuilder.ensureCapacityInternal(AbstractStringBuilder.java:124)
at java.lang.AbstractStringBuilder.append(AbstractStringBuilder.java:448)
at java.lang.StringBuilder.append(StringBuilder.java:136)
```

This error occurs when the Data Integration Service does not have sufficient memory to run concurrent mappings. The Data Integration Service logs are located at `<INFA_HOME>/logs/<node name>/services/DataIntegrationService/disLogs/`

To resolve this error, configure the following advanced properties of the Data Integration Service:

- **Maximum Heap Size.** Specify a minimum value of 2048M. Default is 640M.
- **JVM command Line Options.** Specify a minimum value of 1024M for the **XX:MaxMetaspaceSize** attribute. Default is 192M.

The streaming mapping execution fails with the following error in the application logs of the Hadoop cluster:

```
Cleaning up the staging area /tmp/hadoop-yarn/staging/cloudqa/.staging/
job_1475754687186_0406
PrivilegedActionException as:cloudqa (auth:PROXY) via yarn (auth:SIMPLE)
cause:org.apache.hadoop.security.AccessControlException:
Permission denied: user=cloudqa, access=EXECUTE, inode="/tmp/hadoop-yarn/
staging":yarn:supergroup:drwx-----
at
org.apache.hadoop.hdfs.server.namenode.DefaultAuthorizationProvider.checkFsPermission(DefaultAuthorizationProvider.java:281)
at
org.apache.hadoop.hdfs.server.namenode.DefaultAuthorizationProvider.check(DefaultAuthorizationProvider.java:262)
at
org.apache.hadoop.hdfs.server.namenode.DefaultAuthorizationProvider.checkTraverse(DefaultAuthorizationProvider.java:206)
at
org.apache.hadoop.hdfs.server.namenode.DefaultAuthorizationProvider.checkPermission(DefaultAuthorizationProvider.java:158)
at
org.apache.hadoop.hdfs.server.namenode.FSPermissionChecker.checkPermission(FSPermissionChecker.java:152)
at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkPermission(FSNamesystem.java:6621)
at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkPermission(FSNamesystem.java:6603)
at org.apache.hadoop.hdfs.server.namenode.FSNamesystem.checkOwner(FSNamesystem.java:6522)
```

This error occurs when a YARN user, Spark engine user, or mapping impersonation user does not have sufficient permission on the /tmp/hadoop-yarn/staging directory. Assign required permissions and run the mapping again.

When I run a Streaming mapping that contains an HBase data object, I get the following error:

```
HBaseDataAdapter : java.lang.NullPointerException
at
com.informatica.products.extensions.adapter.hadoop.hive.storagehandler.utils.PwxWriter.close(PwxWriter.java:165)
at
com.informatica.products.extensions.adapter.hadoop.hive.storagehandler.PwxHiveRecordWriter.close(PwxHiveRecordWriter.java:119)
at com.informatica.platform.dtm.executor.hive.boot.storagehandler.INFAOutputFormat$INFAHiveRecordWriter.close(INFAOutputFormat.java:145)
at org.apache.spark.sql.hive.SparkHiveWriterContainer.close(hiveWriterContainers.scala:109)
at
org.apache.spark.sql.hive.SparkHiveWriterContainer.writeToFile(hiveWriterContainers.scala:194)
at org.apache.spark.sql.hive.execution.InsertIntoHiveTable$$anonfun$saveAsHiveFile$3.apply(InsertIntoHiveTable.scala:131)
at org.apache.spark.sql.hive.execution.InsertIntoHiveTable$$anonfun$saveAsHiveFile$3.apply(InsertIntoHiveTable.scala:131)
at org.apache.spark.scheduler.ResultTask.runTask(ResultTask.scala:70)
at org.apache.spark.scheduler.Task.run(Task.scala:86)
at org.apache.spark.executor.Executor$TaskRunner.run(Executor.scala:274)
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1142)
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor
```

This error occurs when you try to write a null value to a ROW column of an HBase table.

Ensure that you do not write a null value to a ROW column of an HBase table.

When I test a MapRStreams connection, the Developer tool crashes.

This error occurs if you have not completed the required prerequisites.

Ensure that you copy the conf files to the following directory: <INFA_HOME>\clients\DeveloperClient\hadoop\mapr_5.2.0\conf

For more information about the prerequisite tasks, see the *Informatica Big Data Management Integration Guide*.

When I import a data object with Avro schema in a streaming mapping the mapping fails with the following error:

```
com.informatica.adapter.sdkadapter.exceptions.AdapterSDKException: [SDK_APP_COM_20000]
error [getSchemaConfig():java.io.IOException: Not a data file.]
```

This occurs when the sample schema file is invalid. When you specify Avro format, provide a valid Avro schema in a .avsc file.

When I run a streaming mapping with an Event Hub source or target, the mapping fails with the following error:

```
com.microsoft.azure.eventhubs.AuthorizationFailedException: Put token failed. status-
code: 401, status-description: InvalidSignature: The token has an invalid signature.
```

This occurs when the Shared Access Policy Name or Shared Access Policy Primary Key is configured incorrectly in the Azure EventHub read data operation, Azure EventHub write data operation, or Azure Eventhub connection. Configure valid values for these properties. When you configure the properties in the Azure EventHub connection, ensure that the policy applies to all data objects that are associated with the connection.

A streaming mapping created in 10.2.0 might not run after upgrading to 10.2.1 when you use JSON schema for column projection for the data you are writing

This might occur if the payload has malformed data.

Remove the malformed data run the mapping.

When I run a streaming mapping on a Cloudera CDH cluster that contains a Kafka source or target, the application does not process the data.

This error occurs when the offset commit topic replication factor configured for the Kafka broker does not match the number of nodes running on the Cloudera CDH cluster.

Perform the following steps to resolve this error:

1. Delete all Kafka topics.
2. Stop the Kafka broker and clear the Kafka log directory specified in the `log.dirs` property of the Kafka broker.
3. Stop ZooKeeper and clear the log directory specified in the `dataDir` property of ZooKeeper.
4. Configure the `Offset Commit Topic Replication Factor` property for the Kafka broker. Specify a value of 1 or 2 for `offsets.topic.replication.factor` property depending on the number of nodes in the cluster.
5. Start ZooKeeper and the Kafka broker.
6. Create the Kafka topics and run the mapping again.

When I run a streaming mapping that contains a JMS target, I get the following error:

```
WebSphere MQ call failed with compcode '2' ('MQCC_FAILED') reason  
'2053' ('MQRC_Q_FULL').  
at com.ibm.msg.client.wmq.common.internal.Reason.createException
```

This error occurs when the JMS queue that you are writing to is full.

Increase the queue depth of the JMS server and run the mapping again.

In a streaming mapping, when I perform a self-join on source data that has metadata of complex data type, the mapping validation fails at design time.

This error might occur if you have selected the **Sorted Input** property in the advanced properties of the Joiner transformation.

To resolve this error, deselect the **Sorted Input** property and run the mapping again.

When I run a streaming mapping that contains a JMS source, it fails with an Unexpected JMSMessage payload type error.

This error might occur in the following situations:

- There is a mismatch between the data type that you write to the queue and the data present in the queue. Clear the JMS queue and then run the mapping.
- There is a mismatch in the data type that you configured for the JMS source and the data type of the streamed data. Verify that the data type that you configure for the source is the same as the data type of the streamed data.

When I edit the schema of a data that is of complex data type, the schema type does not change.

This error occurs because the **Project Column as Complex Data Type** option is not selected.

To resolve this error, when you edit the schema, select the **Project Column as Complex Data Type** option in the columns projection properties of the data object read or write operation properties.

When I run a streaming mapping with a Kafka source and a Hive target, the mapping fails with the following error message:

```
java.io.IOException: Mkdirs failed to create file
```

To resolve this error, set 777 permissions on the hive.exec.scratchdir directory for the user on all the nodes of the cluster, and then run the mapping.

CHAPTER 6

Window Transformation

This chapter includes the following topics:

- [Window Transformation Overview, 107](#)
- [Window Transformation Types, 107](#)
- [Window Transformation Port Properties, 109](#)
- [Window Transformation Window Properties, 109](#)
- [Tumbling Window Transformation Example, 110](#)
- [Sliding Window Transformation Example, 110](#)
- [Rules and Guidelines for the Window Transformations, 112](#)

Window Transformation Overview

Use the Window transformation when you want to accumulate streamed data into data groups and then process the data sets. The Window transformation is a passive transformation.

When you read from unbounded sources, you might want to accumulate the data into bounded data groups for further processing. To introduce bounded intervals to unbounded data, use a Window transformation.

When you configure a Window transformation, define the type of window and the data boundaries by time. To specify data boundaries, configure the window size and window slide interval. The window size defines the time interval for which data is accumulated as a data group. The slide interval defines the time interval after which the accumulated data group is processed further. The watermark delay defines the threshold time for a delayed event to be accumulated into a data group.

Window Transformation Types

When you create a Window transformation, you can configure sliding or tumbling windows to specify a time interval for selecting a data group from data streams.

Select one of the following window types when you create a Window transformation:

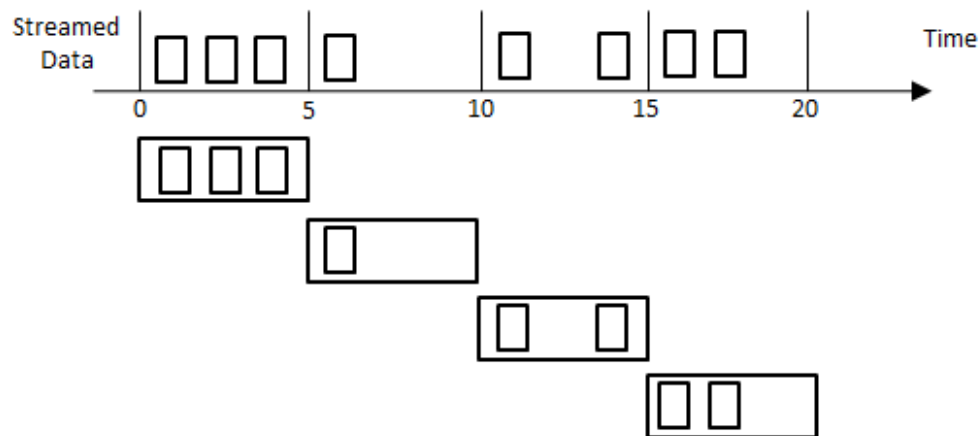
- Tumbling
- Sliding

When you develop a Window transformation, you need to consider factors, such as the type of window, the window size and window slide interval, and watermark delay that you want to use on the data that the source streams.

Tumbling Window

A tumbling window accumulates data and returns a bounded data group. After the output data is sent, the tumbling window is cleared and a new group of data is accumulated for the next output. Tumbling windows do not overlap. In tumbling windows, the window size and slide interval are the same.

The following image shows a sample 5-second tumbling window:

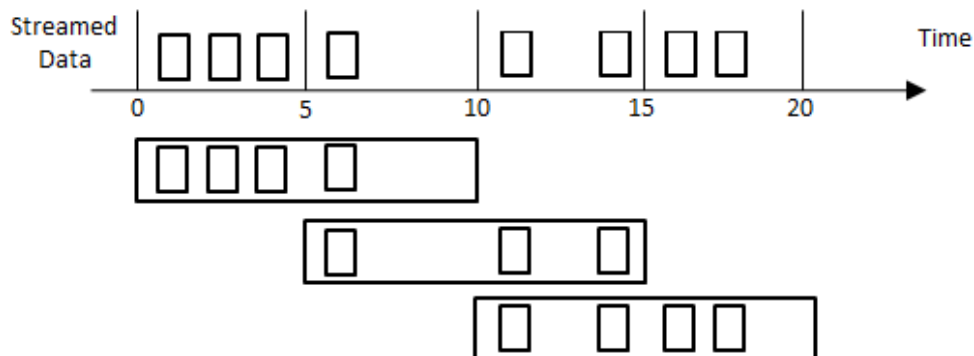


Sliding Window

A sliding window accumulates data and returns a bounded data group. The bounds on the data slide by the time you specify. The data groups that are accumulated can overlap based on the slide interval that you specify.

When you create a sliding Window transformation, the window size must be a multiple of the slide interval.

The following image shows a sample sliding window with a window size of 10 seconds and slide interval of 5 seconds:



Window Transformation Port Properties

The window port specifies the column that contains the timestamp values based on which you can group the events. The Window Port column is applicable for the ports of date/time data type.

The data accumulated contains the timestamp value at which the event was generated rather than the event arrived at the source. When you create a Window transformation, you can determine the column that you want to use to group the data.

Window Transformation Window Properties

A Window transformation has different window types that allow you to accumulate data groups at different time intervals.

Configure the following window properties for a Window transformation:

Window Type

The type of window transformation you want to create. You can choose tumbling or sliding.

Window Size

The window size defines the time interval for which data is accumulated as a data group. The window size should be a multiple of the batch interval. Specify the window size as a value in units of time or as a parameter of type TimeDuration.

Sliding Interval

The slide interval defines the time interval after which the accumulated data group is processed. Specify the slide interval as a value in units of time or as a parameter of type TimeDuration. Specify the sliding interval if you create a sliding window. By default, the window size and sliding interval are same for tumbling windows.

Watermark Delay

The watermark delay defines threshold time for a delayed event to be accumulated into a data group. Watermark delay is a threshold where you can specify the duration at which late arriving data can be grouped and processed.

If an event data arrives within the threshold time, the data is processed, and the data is accumulated into the corresponding data group. You can specify the watermark delay as a value in units of time or as a parameter of type TimeDuration in the window properties.

The following image shows sample window transformation properties:

The screenshot shows the 'Properties' window for a 'Data Viewer' transformation. The 'Window' tab is selected in the left-hand navigation pane. The main area displays the following settings:

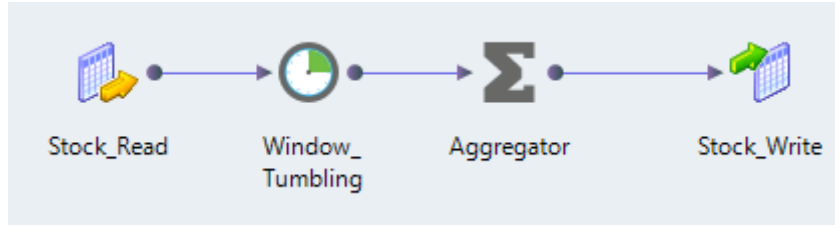
- Window Type:** Sliding (dropdown menu)
- Window Size:** Specify by: Value (dropdown menu), Window Size: 20 (spin box), Seconds (s) (dropdown menu)
- Sliding Interval:** Specify by: Value (dropdown menu), Sliding Interval: 20 (spin box), Seconds (s) (dropdown menu)
- Watermark Delay:** Specify by: Value (dropdown menu), Watermark Delay: 20 (spin box), Seconds (s) (dropdown menu)

Tumbling Window Transformation Example

You want to calculate the maximum value of a stock price every five minutes for stock prices collected over a five-minute time interval. You can use a tumbling Window transformation.

Create a mapping that reads stock prices and calculates the maximum value every five minute.

The following figure shows the example mapping:



You can use the following objects in your mapping:

Kafka Input

The input, Stock_Read, is a Kafka broker.

Window Transformation

The Window transformation, Window_Tumbling, accumulates data and returns a data group every five minute. Configure a window size of 5 minutes. The default slide interval is 5 minutes. The transformation streams data for five minutes and returns a data group every five minutes.

Aggregator

The Aggregator transformation calculates the maximum value of the stock price.

Kafka Output

The output, Stock_Write, is a Kafka broker.

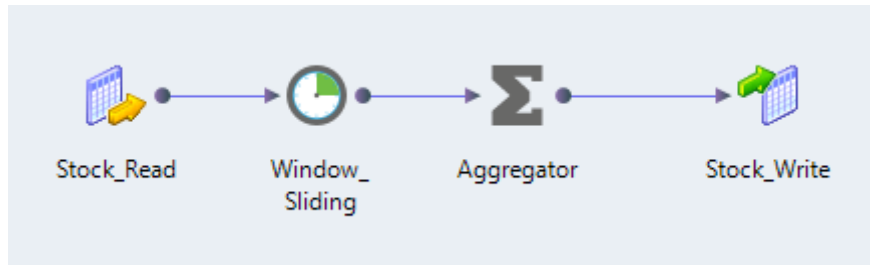
When you run the mapping, the Data Integration Service reads the data from the Kafka broker and passes it to the Window transformation. The window transformation groups the data and provides a data group every five minutes. The Aggregator transformation provides the maximum stock price. The output is written to a Kafka broker.

Sliding Window Transformation Example

You want to calculate the maximum value of a stock price every minute for stock prices collected over a five-minute time interval. You can use a sliding Window transformation.

Create a mapping that reads stock prices and calculates the maximum value every minute.

The following image shows the example mapping:



You can use the following objects in your mapping:

Kafka Input

The input, Stock_Read, is a Kafka broker.

Window Transformation

The Window transformation, Window_Sliding, accumulates data and returns a data group every minute. Configure a window size of 5 minutes and a slide interval of 1 minute. The transformation streams data for five minutes and returns a data group every minute.

Aggregator

The Aggregator transformation calculates the maximum value of the stock price.

Kafka Output

The output, Stock_Write, is a Kafka broker.

When you run the mapping, the Data Integration Service reads the data from the Kafka broker and passes it to the Window transformation. The window transformation groups the data and provides a data group every minute. The Aggregator transformation provides the maximum stock price. The output is written to a Kafka broker.

Rules and Guidelines for the Window Transformations

Certain transformations are valid with restrictions with the Window transformation. The following table describes the rules and guidelines for transformations:

Transformation	Rules and Guidelines
Aggregator	<p>The Aggregator transformation is a multi-group active transformation.</p> <p>The following rules apply to Aggregator transformations:</p> <ul style="list-style-type: none">- You must use a Window transformation directly upstream from an Aggregator transformation in a streaming mapping.- You cannot use multiple Aggregator transformations in the same streaming pipeline.- If you connect the Window port from a Window transformation to an Aggregator transformation, you cannot connect the Window port to any downstream transformation.- If a mapplet contains an Aggregator transformation, you must include a Window transformation directly upstream from the mapplet.- You cannot perform the group by aggregations on date/time data type port marked as a Window port. If you want to perform aggregations on the date/time data type port, you must create a date/time data type port with the timestamp values, and then perform the group by aggregations on the newly created data type port.
Joiner	<p>The Joiner transformation is a multi-group active transformation.</p> <p>The following rules apply to Joiner transformations:</p> <ul style="list-style-type: none">- You must use a Window transformation directly upstream from a Joiner transformation in a streaming mapping.- You cannot use multiple Joiner transformations in the same streaming pipeline.- You must use a Window transformation between the streaming source and any Joiner transformation in a streaming mapping.- The upstream Window transformations in pipelines to a Joiner transformation must have the same slide intervals.- You cannot use an Aggregator transformation anywhere before a Joiner transformation in a streaming mapping.- If you connect the Window port from a Window transformation to a Joiner transformation, you cannot connect the Window port to any downstream transformation.- If a mapplet contains a Joiner transformation, you must include a Window transformation directly upstream from the mapplet.
Lookup	<p>The following rules apply to Lookup transformations:</p> <ul style="list-style-type: none">- A Lookup transformation does not require a Window transformation between a streaming source and itself.- You can include a Lookup transformation only if the mapping has flat file, Hive, HBase, HDFS or JDBC sources.- You cannot use an Aggregator transformation and a passive Lookup transformation that is configured with an inequality lookup condition in the same pipeline.- You cannot use more than one passive Lookup transformation that is configured with an inequality lookup condition in the same pipeline.
Rank	<p>You must use a Window transformation before a Rank transformation in a streaming mapping.</p> <p>Note: Support for the Rank transformation is deferred. Support will be reinstated</p>

Transformation	Rules and Guidelines
Sorter	<p>The Sorter transformation is an active transformation. You must use a Window transformation between the streaming source and the Sorter transformation in a streaming mapping.</p> <p>Note: Support for the Sorter transformation is deferred. Support will be reinstated in a future release.</p>
Union	<p>The following rules apply to Union transformations:</p> <ul style="list-style-type: none"> - A Union transformation does not require a Window transformation between a streaming source and itself. - A Union transformation cannot be used to merge data from streaming and non-streaming pipelines.
Window	<p>The following rules apply to Window transformations:</p> <ul style="list-style-type: none"> - You cannot add a Window transformation to a Logical Data Object mapping or mapplet. - A Window transformation must have at least one upstream streaming source. - All Window transformations must have a slide interval that is a multiple of the mapping batch interval. - A Window transformation that is downstream from another Window transformation must have a slide interval that is a multiple of the slide interval of the upstream Window transformation. - The slide interval of a sliding Window transformation must be less than window size. - The format of the parameter of the window size must have the TimeDuration parameter type. - The window size and the slide interval of a Window transformation must be greater than 0. - The window port from the Window transformation cannot be connected to more than one downstream transformation.

APPENDIX A

Connections

This appendix includes the following topics:

- [Connections Overview, 114](#)
- [Amazon Kinesis Connection, 115](#)
- [Azure EventHub Connection, 118](#)
- [HBase Connection, 120](#)
- [HDFS Connection, 121](#)
- [Hive Connection, 122](#)
- [JMS Connection, 123](#)
- [JDBC Connection Properties, 126](#)
- [Kafka Connection, 128](#)
- [MapR Streams Connection, 130](#)
- [Microsoft Azure Data Lake Store Connection, 132](#)

Connections Overview

Define the connections that you want to use to access data in Kafka brokers, JMS servers, HDFS files, Hive tables, Amazon Kinesis streams, MapR streams or HBase resources. You can create the connections using the Developer tool and `infacmd`.

You can create the following types of connections:

Hadoop

Create a Hadoop connection to run mappings on the Hadoop cluster. Select the Hadoop connection if you select the Hadoop run-time environment. You must also select the Hadoop connection to validate a mapping to run on the Hadoop cluster.

For more information about the Hadoop connection properties, see the *Informatica Big Data Management User Guide*.

HBase

Create an HBase connection to write data to an HBase resource.

HDFS

Create an HDFS connection to write data to an HDFS binary or sequence file.

Hive

Create a Hive connection to write data to Hive tables.

For more information, see the *Informatica Big Data Management Administrator Guide*.

JDBC

Create a JDBC connection when you perform a lookup on a relational database using Sqoop.

For more information about the JDBC connection properties, see the *Informatica Big Data Management User Guide*.

Microsoft Azure Data Lake Store

Create a Microsoft Azure Data Lake Store connection to write to a Microsoft Azure Data Lake Store.

Messaging

Create a Messaging connection to access data as it becomes available, and to run a streaming mapping on a Spark engine. You can create the following types of messaging connections:

- Amazon Kinesis. Create an Amazon Kinesis connection to read from Amazon Kinesis Streams or write to Amazon Kinesis Firehose Delivery Streams.
- Azure Event Hub. Create an Azure Event Hub connection to read from or write to Microsoft Event Hubs.
- JMS. Create a JMS connection to read from or write to a JMS server.
- Kafka. Create a Kafka connection to read from or write to a Kafka broker.
- MapR Streams. Create a MapR Streams connection read from or write to MapR Streams.

Amazon Kinesis Connection

The Amazon Kinesis connection is a Messaging connection. Use the Amazon Kinesis connection to access Amazon Kinesis Data Streams as source or Amazon Kinesis Data Firehose as target. You can create and manage an Amazon Kinesis connection in the Developer tool or through infacmd.

When you configure an Amazon Kinesis connection, you configure the following properties:

- Type of service.
- Access keys (access key ID and secret access key) to interact with Amazon Web Services.
- Region that hosts the service that you are trying to use.
- Credential profile to interact with Amazon Web Services.

General Properties

The following table describes the general connection properties for the Amazon Kinesis connection:

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	The 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	The description of the connection. Enter a string that you can use to identify the connection. The description cannot exceed 4,000 characters.
Location	The domain where you want to create the connection. Select the domain name.
Type	The connection type. Select Messaging/Amazon Kinesis.

Connection Properties

The following table describes the connection properties for the Amazon Kinesis connection:

Property	Description
Service	The type of Kinesis Service that the connection is associated with. Select one of the following service types: <ul style="list-style-type: none">- Kinesis Firehose. Select this service to write to Kinesis Firehose Delivery Stream.- Kinesis Streams. Select this service to read from Kinesis Streams
AWS Access Key ID	The access key ID of the Amazon AWS user account.
AWS Secret Access Key	The secret access key for your Amazon AWS user account.

Property	Description
Region	<p>Region where the endpoint for your service is available. You can select one of the following values:</p> <ul style="list-style-type: none"> - us-east-2. Indicates the US East (Ohio) region. - us-east-1. Indicates the US East (N. Virginia) region. - us-west-1. Indicates the US West (N. California) region. - us-west-2. Indicates the US West (Oregon) region. - ap-northeast-1. Indicates the Asia Pacific (Tokyo) region. - ap-northeast-2. Indicates the Asia Pacific (Seoul) region. - ap-northeast-3. Indicates the Asia Pacific (Osaka-Local) region. - ap-south-1. Indicates the Asia Pacific (Mumbai) region. - ap-southeast-1. Indicates the Asia Pacific (Singapore) region. - ap-southeast-2. Indicates the Asia Pacific (Sydney) region. - ca-central-1. Indicates the Canada (Central) region. - cn-north-1. Indicates the China (Beijing) region. - cn-northwest-1. Indicates the China (Ningxia) region. - eu-central-1. Indicates the EU (Frankfurt) region. - eu-west-1. Indicates the EU (Ireland) region. - eu-west-2. Indicates the EU (London) region. - eu-west-3. Indicates the EU (Paris) region. - sa-east-1. Indicates the South America (São Paulo) region.
Connection Timeout (in msec)	Number of milliseconds that the Integration service waits to establish a connection to the Kinesis Stream or Kinesis Firehose after which it times out.
Authentication Type	<p>The type of authentication.</p> <p>Select one of the following values:</p> <ul style="list-style-type: none"> - AWS Credential Profile - Cross-account IAM Role <p>The default value is AWS Credential Profile.</p>
AWS Credential Profile Name	<p>Required if you use the AWS credential profile authentication type. An AWS credential profile defined in the credentials file. A mapping accesses the AWS credentials through the profile name at run time.</p> <p>If you do not provide an AWS credential profile name, the mapping uses the access key ID and secret access key that you specify when you create the connection.</p>
ARN of IAM Role	Required if you use the cross-account IAM role authentication type. The Amazon Resource Name specifying the role of an IAM user.
External ID	Required if you use the cross-account IAM role authentication type and if the external ID is defined by the AWS account. The external ID for an IAM role is an additional restriction that you can use in an IAM role trust policy to designate who can assume the IAM role.

Creating an Amazon Kinesis Connection using Informatica Command Line Program

You can use the `infacmd` command line program to create an Amazon Kinesis connection. Access the command from the `<Informatica installation directory>/clients/DeveloperClient/infacmd` directory.

On UNIX, run the following command:

```
sh infacmd.sh createConnection
```

On Windows, run the following command:

```
infacmd.bat createConnection
```

Enter connection options in the following format:

```
... -o option_name=value option_name='value' ...
```

For example, to create an Amazon Kinesis connection to Kinesis Streams on UNIX using cross-account IAM role, run the following command:

```
infacmd createConnection -dn <domain name> -un <domain user> -pd <domain password> -cn  
<connection name> -cid <connection id> -ct AMAZONKINESIS -o "AWS_ACCESS_KEY_ID=<access key  
id> AWS_SECRET_ACCESS_KEY=<secret access key> ConnectionTimeout=10000 Region=<RegionName>  
ServiceType='Kinesis Streams' RoleArn=<ARN of IAM role> ExternalID=<External ID>  
AuthenticationType='Cross-account IAM Role'"
```

To create an Amazon Kinesis connection to Kinesis Firehose on UNIX using AWS credential profile, run the following command:

```
infacmd createConnection -dn <domain name> -un <domain user> -pd <domain password> -cn  
<connection name> -cid <connection id> -ct AMAZONKINESIS -o "AWS_ACCESS_KEY_ID=<access key  
id> AWS_SECRET_ACCESS_KEY=<secret access key> ConnectionTimeout=10000 Region=<RegionName>  
ServiceType='Kinesis Firehose' Profilename=<AWS credential profile> AuthenticationType='AWS  
Credential Profile'"
```

For more information about the CreateConnection command, see the *Informatica Command Reference Guide*.

Azure EventHub Connection

Use the Azure EventHub connection to access Azure Event Hub as source or target. You can create and manage an Azure EventHub connection in the Developer tool or through infacmd.

General Properties

The following table describes the general connection properties for the AzureEventHub connection:

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	The 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	The description of the connection. Enter a string that you can use to identify the connection. The description cannot exceed 4,000 characters.
Location	The domain where you want to create the connection. Select the domain name.
Type	The connection type. Select Messaging/AzureEventHub

Connection Properties

The following table describes the connection properties for the Amazon Kinesis connection:

Property	Description
Tenant ID	The ID of the tenant that the data belongs to. This ID is the Directory ID of the Azure Active Directory.
Subscription ID	The ID of the Azure subscription.
Resource Group Name	The name of the Azure resource group associated with the event hub namespace.
Client Application ID	The ID of the application created under the Azure Active Directory.
Client Secret Key	The secret key generated for the application.
Event Hub Namespace	The name of the Event Hub Namespace that is associated with the resource group name.
Shared Access Policy Name	The name of the Event Hub Namespace Shared Access Policy. The policy must apply to all data objects that are associated with this connection. To read from Event Hubs, you must have Listen permission. To write to an Event hub, the policy must have Send permission.
Shared Access Policy Primary Key	The primary key of the Event Hub Namespace Shared Access Policy.

Creating an Azure EventHub Connection using Informatica Command Line Program

You can use the `infacmd` command line program to create an Azure EventHub connection. Access the command from the `<Informatica installation directory>/clients/DeveloperClient/infacmd` directory.

On UNIX, run the following command:

```
sh infacmd.sh createConnection
```

On Windows, run the following command:

```
infacmd.bat createConnection
```

Enter connection options in the following format:

```
... -o option_name='value' option_name='value' ...
```

For example, On UNIX you can run the following command to create an Azure EventHub connection:

```
sh infacmd.sh createConnection -dn <domain name> -un <domain user> -pd <domain password> -cn  
<connection name> -cid <connection id> -ct AZUREEVENTHUB -o tenantID='tenant id'  
subscriptionID='subscription id' resourceName='resource name' clientID='client id'  
clientSecretKey='secret key' eventHubNamespace='namespace' sasPolicyName=<policy name>  
sasPolicyPrimaryKey=<policy key>
```

HBase Connection

Create an HBase connection to write data to an HBase table.

You can create and manage an HBase connection in the Developer tool or through infacmd.

General Properties

The following table describes the general connection properties for the HBase connection:

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	The 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	The description of the connection. Enter a string that you can use to identify the connection. The description cannot exceed 4,000 characters.
Location	The domain where you want to create the connection. Select the domain name.
Type	The connection type. Select on the following options: <ul style="list-style-type: none">- HBase- MapR-DB

Connection Properties

The following table describes the connection properties for the HBase connection:

Property	Description
Database Type	The connection type. Select on the following options: <ul style="list-style-type: none">- HBase- MapR-DB
Cluster Configuration	The name of the cluster configuration associated with the Hadoop environment. For more information about cluster configuration, see the <i>Informatica Big Data Management Administrator Guide</i> .
MapR-DB Database Path	Database path that contains the MapR-DB table that you want to connect to. Enter a valid MapR cluster path. When you create an HBase data object for MapR-DB, you can browse only tables that exist in the MapR-DB path that you specify in the Database Path field. You cannot access tables that are available in sub-directories in the specified path. For example, if you specify the path as /user/customers/, you can access the tables in the customers directory. However, if the customers directory contains a sub-directory named regions, you cannot access the tables in the following directory: /user/customers/regions

Creating an HBASE Connection using Informatica Command Line Program

You can use the infacmd command line program to create an HBASE connection. Access the command from the <Informatica installation directory>/clients/DeveloperClient/infacmd directory.

On UNIX, run the following command:

```
sh infacmd.sh createConnection
```

On Windows, run the following command:

```
infacmd.bat createConnection
```

Enter connection options in the following format:

```
... -o option_name='value' option_name='value' ...
```

For example, On UNIX you can run the following command to create an HBASE connection:

```
sh infacmd.sh createConnection -dn <domain name> -un <domain user> -pd <domain password> -cn  
<connection name> -cid <connection id> -ct HBase -o DatabaseType=HBase CLUSTERCONFIGID=ConfId
```

For more information about the CreateConnection command and the HBASE connection options, see the *Informatica Command Reference Guide*.

HDFS Connection

Create an HDFS connection to write data to an HDFS binary or sequence file.

You can create and manage an HDFS connection in the Developer tool or through infacmd.

General Properties

The following table describes the general connection properties for the HDFS connection:

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	The 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	The description of the connection. Enter a string that you can use to identify the connection. The description cannot exceed 4,000 characters.
Location	The domain where you want to create the connection. Select the domain name.
Type	The connection type. Select Hadoop File System.

Connection Properties

The following table describes the connection properties for the HDFS connection:

Property	Description
User Name	User name to access HDFS.
NameNode URI	<p>The URI to access HDFS.</p> <p>Use the following format to specify the NameNode URI in Cloudera and Hortonworks distributions:</p> <pre>hdfs://<namenode>:<port></pre> <p>Where</p> <ul style="list-style-type: none">- <namenode> is the host name or IP address of the NameNode.- <port> is the port that the NameNode listens for remote procedure calls (RPC). <p>Use the <code>maprfs:///</code> format to specify the NameNode URI in MapR distribution.</p>
Cluster Configuration	<p>The name of the cluster configuration associated with the Hadoop environment.</p> <p>For more information about cluster configuration, see the <i>Informatica Big Data Management Administrator Guide</i>.</p>

Creating an HDFS Connection using Informatica Command Line Program

You can use the `infacmd` command line program to create an HDFS connection. Access the command from the `<Informatica installation directory>/clients/DeveloperClient/infacmd` directory.

On UNIX, run the following command:

```
sh infacmd.sh createConnection
```

On Windows, run the following command:

```
infacmd.bat createConnection
```

Enter connection options in the following format:

```
... -o option_name='value' option_name='value' ...
```

For example, run the following command to create an HDFS connection on UNIX:

```
sh infacmd.sh createConnection -dn <domain name> -un <domain user> -pd <domain password> -cn  
<connection name> -cid <connection id> -ct HadoopFileSystem -o NameNodeURL='hdfs://  
<namenode>:<port>' USERNAME='username' clusterConfigId='ConfId'
```

For more information about the `CreateConnection` command and HDFS connection options, see the *Informatica Command Reference Guide*.

Hive Connection

Create an Hive connection to write data to a Hive table.

You can create and manage a Hive connection in the Developer tool or through `infacmd`.

Creating a Hive Connection using Informatica Command Line Program

You can use the `infacmd` command line program to create a Hive connection. Access the command from the `<Informatica installation directory>/clients/DeveloperClient/infacmd` directory.

On UNIX, run the following command:

```
sh infacmd.sh createConnection
```

On Windows, run the following command:

```
infacmd.bat createConnection
```

Enter connection options in the following format:

```
... -o option_name='value' option_name='value' ...
```

For example, run the following command to create a Hive connection on UNIX:

```
sh infacmd.sh CreateConnection -dn <domain name> -un <domain user> -pd <domain password> -cn  
<connection name> -cid <connection id> -ct HIVE -o "CLUSTERCONFIGID='ConfId'  
relationalSourceAndTarget='true' pushDownMode='false' metaDataConnString='jdbc:hive2://  
<hostname>:<port>/<db>' enableQuotes='false' HIVEWAREHOUSEDIRECTORYONHDFS='/user/hive/  
warehouse' DATABASENAME='default' BYPASSHIVEJDBCSEVER='false' ConnectString='jdbc:hive2://  
<hostname>:<port>/<db> databaseName=default defaultFSURI= hdfs://<node name>:<port>' "
```

For more information about the `CreateConnection` command and Hive connection options, see the *Informatica Command Reference Guide*.

JMS Connection

The JMS connection is a Messaging connection. Use the JMS connection to read messages from a JMS server. You can create and manage a JMS connection in the Developer tool or through `infacmd`.

Prerequisites to Create a JMS Connection and a JMS Data Object

Before you create a JMS connection or data object, you must include the JMS provider client libraries on the machine running Informatica Big Data Streaming.

To create a JMS connections, copy the following JAR files from the IBM MQ server:

- `com.ibm.mq.allclient.jar`
- `com.ibm.mq.axis2.jar`
- `com.ibm.mq.commonservices.jar`
- `com.ibm.mq.defaultconfig.jar`
- `com.ibm.mq.headers.jar`
- `com.ibm.mq.jar`
- `com.ibm.mq.jmqi.jar`
- `com.ibm.mq.jms.Nojndi.jar`
- `com.ibm.mq.pcf.jar`

- com.ibm.mq.pcf.jar
- fscontext.jar
- jms.jar
- providerutil.jar
- com.ibm.mq.postcard.jar
- com.ibm.mq.soap.jar
- com.ibm.mq.tools.ras.jar
- com.ibm.mq.traceControl.jar
- com.ibm.mqjms.jar
- jndi.jar
- jta.jar
- ldap.jar
- rmm.jar

Place the client JAR files in the following location:

- Developer tool installation directory: <Developer Tool Installation Directory>/clients/DeveloperClient/connectors/thirdparty/infa.jms/common

Prerequisites to Use a JMS Connection and a JMS Data Object

Before you use a JMS connection and JMS data object in your streaming mapping, perform the following tasks:

1. Verify that the IBM MQ client is installed on all the cluster nodes.
2. Copy the JAR files from the IBM MQ server.
3. Place the JAR files in the following directory on the machine where the Data Integration Service runs:
INFA_HOME/services/shared/hadoop/<Hadoop distribution>/extras/spark-auxjars/
For example, /CDH_5.13/extras/spark-auxjars
4. Create the .bindings file using the IBM MQ Server.
The .bindings file must have the details of the MQ Objects that the data objects in the streaming mappings use.
5. Copy the .bindings file to the Developer tool machine.
6. When you create the JMS connection, configure the **Connection URL** property with the location of the .bindings file on the Developer tool machine.
 - On Windows, specify the following format for the location: file:/// <Path to .bindings file>
For example: file:///C:/JMS
 - On Linux, specify the following format for the location: file:/// <Path to .bindings file>
7. Place the .bindings file in the same path on all the cluster nodes.
The default location of the files is /etc
For example, file:///etc
8. Before you run the mapping, update the **Connection URL** property with the location of the .bindings file on the cluster nodes.

General Properties

The following table describes the general connection properties for the JMS connection:

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	The 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	The description of the connection. Enter a string that you can use to identify the connection. The description cannot exceed 4,000 characters.
Location	The domain where you want to create the connection. Select the domain name.
Type	The connection type. Select Messaging/JMS.

Connection Properties

The following table describes the connection properties for the JMS connection:

Property	Description
Connection URL	The location and port of the JMS provider on which to connect. For example: <code>tcp://jndiserverA:61616</code>
User Name	User name to the connection factory.
Password	The password of the user account that you use to connect to the connection factory.
JNDI Context Factory	The JMS provider specific initial JNDI context factory implementation for connecting to the JNDI service. This value is a fully qualified class name of the Initial Context Factory. For example, the class name of the Initial Context Factory for ActiveMQ is <code>org.apache.activemq.jndi.ActiveMQInitialContextFactory</code> For more information, see the documentation of the JMS provider.
JNDI Package Prefixes	A colon-delimited list of package prefixes to use when loading URL context factories. These are the package prefixes for the name of the factory class that will create a URL context factory. For more information about the values, see the documentation of the JMS provider.
JMS Connection Factory	The name of the object in the JNDI server that enables the JMS Client to create JMS connections. For example, <code>jms/QCF</code> or <code>jmsSalesSystem</code> .

Creating a JMS Connection using Informatica Command Line Program

You can use the `infacmd` command line program to create a JMS connection. Access the command from the `<Informatica installation directory>/clients/DeveloperClient/infacmd` directory.

On UNIX, run the following command:

```
sh infacmd.sh createConnection
```

On Windows, run the following command:

```
infacmd.bat createConnection
```

Enter connection options in the following format:

```
... -o option_name='value' option_name='value' ...
```

For example, run the following command to create a JMS connection on UNIX:

```
sh infacmd.sh createConnection -dn <domain name> -un <domain name> -pd <domain password> -cn  
<connection name> -cid <connection id> -ct Jms -o url='<url_val>' username=<u_val>  
password=<pass_val> contextFactory='<cf_val>' packagePrefixes='<pp_val>'<br>jmsConnectionFactory='<jcf_val>'
```

For more information about the `CreateConnection` command, see the *Informatica Command Reference Guide*.

JDBC Connection Properties

You can use a JDBC connection to access tables in a database. You can create and manage a JDBC connection in the Administrator tool, the Developer tool, or the Analyst tool.

Note: The order of the connection properties might vary depending on the tool where you view them.

The following table describes JDBC connection properties:

Property	Description
Database Type	The database type.
Name	Name of the connection. The name is not case sensitive and must be unique within the domain. 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	The description of the connection. The description cannot exceed 765 characters.
User Name	The database user name.
Password	The password for the database user name.

Property	Description
JDBC Driver Class Name	<p>Name of the JDBC driver class.</p> <p>The following list provides the driver class name that you can enter for the applicable database type:</p> <ul style="list-style-type: none"> - DataDirect JDBC driver class name for Oracle: com.informatica.jdbc.oracle.OracleDriver - DataDirect JDBC driver class name for IBM DB2: com.informatica.jdbc.db2.DB2Driver - DataDirect JDBC driver class name for Microsoft SQL Server: com.informatica.jdbc.sqlserver.SQLServerDriver - DataDirect JDBC driver class name for Sybase ASE: com.informatica.jdbc.sybase.SybaseDriver - DataDirect JDBC driver class name for Informix: com.informatica.jdbc.informix.InformixDriver - DataDirect JDBC driver class name for MySQL: com.informatica.jdbc.mysql.MySQLDriver <p>For more information about which driver class to use with specific databases, see the vendor documentation.</p>
Connection String	<p>Connection string to connect to the database. Use the following connection string:</p> <pre>jdbc:<subprotocol>:<subname></pre> <p>The following list provides sample connection strings that you can enter for the applicable database type:</p> <ul style="list-style-type: none"> - Connection string for DataDirect Oracle JDBC driver: jdbc:informatica:oracle://<host>:<port>;SID=<value> - Connection string for Oracle JDBC driver: jdbc:oracle:thin:@//<host>:<port>:<SID> - Connection string for DataDirect IBM DB2 JDBC driver: jdbc:informatica:db2://<host>:<port>;DatabaseName=<value> - Connection string for IBM DB2 JDBC driver: jdbc:db2://<host>:<port>/<database_name> - Connection string for DataDirect Microsoft SQL Server JDBC driver: jdbc:informatica:sqlserver://<host>;DatabaseName=<value> - Connection string for Microsoft SQL Server JDBC driver: jdbc:sqlserver://<host>;DatabaseName=<value> - Connection string for Netezza JDBC driver: jdbc:netezza://<host>:<port>/<database_name> - Connection string for Pivotal Greenplum driver: jdbc:pivotal:greenplum://<host>:<port>;/database_name=<value> - Connection string for Postgres Greenplum driver: jdbc:postgresql://<host>:<port>/<database_name> - Connection string for Teradata JDBC driver: jdbc:teradata://<host>/database_name=<value>,tmode=<value>,charset=<value> <p>For more information about the connection string to use with specific drivers, see the vendor documentation.</p>
Environment SQL	<p>Optional. Enter SQL commands to set the database environment when you connect to the database. The Data Integration Service executes the connection environment SQL each time it connects to the database.</p> <p>Note: If you enable Sqoop, Sqoop ignores this property.</p>
Transaction SQL	<p>Optional. Enter SQL commands to set the database environment when you connect to the database. The Data Integration Service executes the transaction environment SQL at the beginning of each transaction.</p> <p>Note: If you enable Sqoop, Sqoop ignores this property.</p>

Property	Description
SQL Identifier Character	Type of character that the database uses to enclose delimited identifiers in SQL queries. The available characters depend on the database type. Select (None) if the database uses regular identifiers. When the Data Integration Service generates SQL queries, the service does not place delimited characters around any identifiers. Select a character if the database uses delimited identifiers. When the Data Integration Service generates SQL queries, the service encloses delimited identifiers within this character. Note: If you enable Sqoop, Sqoop ignores this property.
Support Mixed-case Identifiers	Enable if the database uses case-sensitive identifiers. When enabled, the Data Integration Service encloses all identifiers within the character selected for the SQL Identifier Character property. When the SQL Identifier Character property is set to none, the Support Mixed-case Identifiers property is disabled. Note: If you enable Sqoop, Sqoop honors this property when you generate and execute a DDL script to create or replace a target at run time. In all other scenarios, Sqoop ignores this property.

Kafka Connection

The Kafka connection is a Messaging connection. Use the Kafka connection to access an Apache Kafka broker as a source or a target. You can create and manage a Kafka connection in the Developer tool or through infacmd.

When you configure a Kafka connection, you configure the following properties:

- The list of Kafka brokers that the connection reads from or writes to.
- The number of seconds the Integration Service attempts to reconnect to the database if the connection fails.
- Version of the Kafka messaging broker. Configure the Kafka messaging broker version to 0.10.1.x-2.0.0.

General Properties

The following table describes the general connection properties for the Kafka connection:

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	The 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	The description of the connection. Enter a string that you can use to identify the connection. The description cannot exceed 4,000 characters.

Property	Description
Location	The domain where you want to create the connection. Select the domain name.
Type	The connection type. Select Messaging/Kafka.

Kafka Broker Properties

The following table describes the Kafka broker properties for the Kafka connection:

Property	Description
Kafka Broker List	Comma-separated list of Kafka brokers which maintain the configuration of the Kafka messaging broker. To specify a Kafka broker, use the following format: <IP Address>:<port>
ZooKeeper Host Port List	Optional. Comma-separated list of Apache ZooKeeper which maintain the configuration of the Kafka messaging broker. To specify the ZooKeeper, use the following format: <IP Address>:<port>
Retry Timeout	Number of seconds after which the Integration Service attempts to reconnect to the Kafka broker to read or write data. If the source or target is not available for the time you specify, the mapping execution stops to avoid any data loss.
Kafka Broker Version	Configure the Kafka messaging broker version to 0.10.1.x-2.0.0.
Additional Connection Properties	Optional. Comma-separated list of connection properties to connect to the Kafka broker. To reduce the time taken to connect to the Kafka broker, ensure that you set the following properties: <ul style="list-style-type: none"> - request.timeout.ms - session.timeout.ms - fetch.max.wait.ms - heartbeat.interval.ms For example, you can use the following syntax: <pre>request.timeout.ms=<value>,session.timeout.ms=<value>, fetch.max.wait.ms=<value>,heartbeat.interval.ms=<value>, security.protocol=SASL_PLAINTEXT,sasl.kerberos. service.name=<kerberos_name>,sasl.mechanism=GSSAPI, sasl.jaas.config=com.sun.security.auth.module. Krb5Login Modulerequired useKeyTab=true doNotPrompt=true storeKey=true client=true keyTab="<Keytab Location>" principal="<principal>";</pre> <p>Technical Preview: The Additional Connection Properties is available for technical preview. Technical preview functionality is supported but is unwarranted and is not production-ready. Informatica recommends that you use in non-production environments only. For more information about the connection properties, see https://kafka.apache.org/documentation/.</p>

Creating a Kafka Connection using Informatica Command Line Program

You can use the `infacmd` command line program to create a Kafka connection. Access the command from the `<Informatica installation directory>/clients/DeveloperClient/infacmd` directory.

On UNIX, run the following command:

```
sh infacmd.sh createConnection
```

On Windows, run the following command:

```
infacmd.bat createConnection
```

Enter connection options in the following format:

```
... -o option_name='value' option_name='value' ...
```

For example, run the following command to create a Kafka connection on UNIX:

```
sh infacmd.sh createConnection -dn <domain name> -un <domain user> -pd <domain password> -cn  
<connection name> -cid <connection id> -ct Kafka -o  
zkHostPortList=<host1:port1>,<host2:port2>,<host3:port3>(optional)  
kfkBrkList=<host1:port1>,<host2:port2>,<host3:port3> kafkabrokerVersion=<version>  
additionalConnectionProperties=<additional properties>
```

For more information about the `CreateConnection` command, see the *Informatica Command Reference Guide*.

MapR Streams Connection

The MapR Streams connection is a Messaging connection. Use the MapR Streams connection to access MapR Streams as source or target. You can create and manage a MapR Streams connection in the Developer tool or through `infacmd`.

When you configure an MapR Streams connection, you configure the path to the MapR Stream and the name of the MapR Stream.

Before you create and use MapR Streams connections in Streaming mappings, complete the required prerequisites. For more information about the prerequisite tasks, see the *Informatica Big Data Management Cluster Integration Guide*.

General Properties

The following table describes the general connection properties for the MapR Streams connection:

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	The 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	The description of the connection. Enter a string that you can use to identify the connection. The description cannot exceed 4,000 characters.
Location	The domain where you want to create the connection. Select the domain name.
Type	The connection type. Select Messaging/MapR Streams.

Connection Properties

The following table describes the connection properties for the MapR Streams connection:

Property	Description
Stream Path and Name	The path that contains the MapR Stream and the name of the MapR Stream. Use the following format: /stream-path:stream-name
Connection Timeout (in msec)	Number of milliseconds that the Integration service waits to establish a connection to MapR Stream after which it times out.

Creating a MapR Streams Connection using Informatica Command Line Program

You can use the `infacmd` command line program to create a MapR Streams connection. Access the command from the `<Informatica installation directory>/clients/DeveloperClient/infacmd` directory.

On UNIX, run the following command:

```
sh infacmd.sh createConnection
```

On Windows, run the following command:

```
infacmd.bat createConnection
```

Enter connection options in the following format:

```
... -o option_name='value' option_name='value' ...
```

For example, run the following command to create a MapR Streams connection on UNIX:

```
infacmd createConnection -dn <domain name> -un <domain user> -pd<password> -cn <connection name> -cid <connection id> -ct MAPRSTREAMS -o connRetryTimeout=10000 streamName=/sample-stream
```

For more information about the CreateConnection command, see the *Informatica Command Reference Guide*.

Microsoft Azure Data Lake Store Connection

Use the Microsoft Azure Data Lake Store Connection to access Microsoft Azure Data Lake Store tables as targets. You can create and manage an Azure EventHub connection in the Developer tool.

You cannot run a mapping with a Microsoft Data Lake Store as target on a MapR distribution.

Microsoft Azure Data Lake Store Connection Properties

Use a Microsoft Azure Data Lake Store connection to access a Microsoft Azure Data Lake Store.

Note: The order of the connection properties might vary depending on the tool where you view them.

You can create and manage a Microsoft Azure SQL Data Warehouse connection in the Administrator tool or the Developer tool. The following table describes the Microsoft Azure Data Lake Store 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	The description of the connection. The description cannot exceed 4,000 characters.
Location	The domain where you want to create the connection.
Type	The connection type. Select Microsoft Azure Data Lake Store.

The following table describes the properties for metadata access:

Property	Description
ADLS Account Name	The name of the Microsoft Azure Data Lake Store.
ClientID	The ID of your application to complete the OAuth Authentication in the Active Directory.
Client Secret	The client secret key to complete the OAuth Authentication in the Active Directory.

Property	Description
Directory	The Microsoft Azure Data Lake Store directory that you use to read data or write data. The default is root directory.
AuthEndpoint	The OAuth 2.0 token endpoint from where access code is generated based on based on the Client ID and Client secret is completed.

For more information about creating a client ID, client secret, and auth end point, contact the Azure administrator or see Microsoft Azure Data Lake Store documentation.

APPENDIX B

Sample Files

This appendix includes the following topic:

- [Sample Files, 134](#)

Sample Files

The data objects in a streaming mapping read and write data in XML, JSON, Avro, CSV format. The following examples contain samples for each schema format.

Sample XSD File

When you configure the data operation properties, specify the format in which the data object reads or writes data. When you specify XML format, provide an XSD.

The following sample XSD describes the elements in an XML file:

```
<?xml version="1.0" encoding="UTF-8" ?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="shiporder">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="orderperson" type="xs:string"/>
        <xs:element name="shipto">
          <xs:complexType>
            <xs:sequence>
              <xs:element name="name" type="xs:string"/>
              <xs:element name="address" type="xs:string"/>
              <xs:element name="city" type="xs:string"/>
              <xs:element name="country" type="xs:string"/>
            </xs:sequence>
          </xs:complexType>
        </xs:element>
        <xs:element name="item" maxOccurs="unbounded">
          <xs:complexType>
            <xs:sequence>
              <xs:element name="title" type="xs:string"/>
              <xs:element name="note" type="xs:string" minOccurs="0"/>
              <xs:element name="quantity" type="xs:positiveInteger"/>
              <xs:element name="price" type="xs:decimal"/>
            </xs:sequence>
          </xs:complexType>
        </xs:element>
      </xs:sequence>
      <xs:attribute name="orderid" type="xs:string" use="required"/>
    </xs:complexType>
  </xs:element>
</xs:schema>
```

Sample JSON Schema

When you configure the data operation properties, specify the format in which the data object reads or writes data. When you specify JSON format, provide a sample JSON file.

The following file is a sample JSON file:

```
{
  "GlossaryCont": {
    "title": {
      "string": "glossaryContTitle0"
    },
    "glossDiv": {
      "title": "glossaryDivTitle0",
      "glossList": {
        "glossEntry": {
          "id":
            "ID0",
          "sortAs": "sortAS0",
          "glossTerm": "GlossTerm0",
          "acronym": "Acronym0",
          "abbrev": "Abbrev0",
          "glossSee": "GlossSee0",
          "glossInteger": 0,
          "glossDouble": 0.234,
          "glossLong": 1000000000000000000,
          "glossDef": {
            "para": "para0",
            "glossSeeAlso": ["glossSeeAlso_0_0"]
          }
        }
      }
    }
  }
}
```

Sample Avro Schema

When you configure the data operation properties, specify the format in which the data object reads or writes data. When you specify Avro format, provide a sample Avro schema in a .avsc file.

The following file is a sample Avro schema:

```
{
  "type" : "record",
  "name" : "Tree",
  "fields" : [
    { "name" : "children", "type" : { "type" : "array", "items": "Tree" } }
  ]
}
```

Sample Flat Format Schema

When you configure the data operation properties, specify the format in which the data object reads or writes data. When you specify Flat format, provide a sample CSV file.

The following file is a sample CSV file:

```
acc_no,acc_type
1,Savings
2,Savings
```


INDEX

A

- Amazon Kinesis
 - configuration [19](#)
 - create AWS credentials [19](#)
- Amazon Kinesis configuration
 - Amazon Kinesis Firehose Targets [20](#)
 - Amazon Kinesis Streams Sources [19](#)
- Amazon Kinesis connection
 - connection properties [116](#)
 - create using infacmd [117](#)
 - general properties [116](#)
 - overview [115](#)
- Amazon Kinesis data object
 - advanced read operation properties [26](#)
 - advanced write operation properties [54](#)
 - column projection read operation properties [26](#)
 - column projection write operation properties [54](#)
 - configuring schema for flat files [27](#)
 - general read operation properties [24](#)
 - general write properties [53](#)
 - overview properties [23](#), [52](#)
 - ports read operation properties [25](#)
 - ports write properties [53](#)
 - run-time read operation properties [25](#)
 - run-time write properties [53](#)
 - source [23](#)
 - sources read operation properties [25](#)
 - target write operation properties [53](#)
- AmazonKinesis data object
 - target [51](#)
- Authentication Systems
 - supported [16](#)
- Azure Data Lake Store data object
 - column projections write operation properties [79](#)
- Azure Event Hub data object
 - general write operation properties [57](#)
- Azure Event Hubs data object
 - advanced read operation properties [30](#)
 - column projection read operation properties [31](#)
 - general read operation properties [29](#)
 - overview properties [28](#), [55](#)
 - ports read operation properties [30](#)
 - run-time read operation properties [30](#)
 - source [28](#)
 - sources read operation properties [30](#)
 - target [55](#)
- Azure Eventhub connection
 - overview [118](#)
- Azure EventHub connection
 - configuration [20](#)
 - connection properties [119](#)
 - create using infacmd [119](#)
 - general properties [118](#)
- Azure EventHub data object
 - advanced write operation properties [58](#)

- Azure EventHub data object (*continued*)
 - column projections write operation properties [58](#)
 - ports write operation properties [57](#)
 - run-time write properties [57](#)
 - target write operation properties [57](#)

B

- Big Data streaming
 - application services [13](#)
 - component architecture [12](#)
 - repository [13](#)
- Big Data Streaming
 - example [14](#)
 - overview [10](#)
 - sources [21](#)
 - streaming process [11](#)
 - targets [49](#)
 - third-party applications [13](#)
- Big Data Streaming Configuration
 - Overview [15](#)
- Big Data Streaming mapping
 - overview [88](#)
- binary data
 - protobuf schema [66](#)
- body fields
 - JMS messages [34](#), [69](#)

C

- complex file data object
 - advanced write operation properties [62](#)
 - column projection write operation properties [63](#)
 - general properties [59](#)
 - general write operation properties [61](#)
 - objects properties [59](#)
 - ports write operation properties [61](#)
 - target [59](#)
 - target write operation properties [61](#)
- complex file write properties
 - execution parameters [63](#)
- complex files
 - compression [60](#)
 - decompression [60](#)
- connections
 - overview [114](#)
- creating
 - data object operation
 - creating [91](#)
 - creating a data object [90](#)

D

- data object
 - column configuration [65](#)
- data object column configuration
 - add columns [65](#)
 - search and add columns [65](#)
- data objects in a streaming [89](#)
- Default Realm
 - configure [17](#)

H

- Hadoop distributions
 - supported [16](#)
- HBase connection
 - connection properties [120](#)
 - general properties [120](#)
 - overview [120](#)
- HBASE connection
 - create using infacmd [121](#)
- HBase data object
 - add columns [65](#)
 - advanced write properties [67](#)
 - get all columns [66](#)
 - overview properties [66](#)
 - search and add columns [65](#)
 - source [32](#)
 - target [65](#)
- HBase data object operation
 - read properties [32](#)
- HDFS connection
 - connection properties [122](#)
 - create using infacmd [122](#)
 - general properties [121](#)
 - overview [121](#)
- header fields
 - JMS messages [34](#), [68](#)
- high availability
 - configuration [102](#)
- Hive connection
 - connection [122](#)
 - create using infacmd [123](#)
- Hive data object
 - data object write operation properties [85](#)
 - general write operation properties [85](#)
 - overview properties [83](#)
 - ports write operation properties [86](#)
 - run-time write operation properties [86](#)
 - write operation properties [84](#)
- Hive write data object
 - advanced write operation properties [87](#)

I

- input properties [77](#)

J

- Java Authorization and Authentication Service
 - configure [17](#)
- JDBC connections
 - properties [126](#)
- JMS connection
 - connection properties [125](#)

- JMS connection (*continued*)

- create using infacmd [126](#)
 - general properties [125](#)
 - prerequisites [123](#)

- JMS data object

- advanced read operation properties [36](#)
 - advanced write operation properties [71](#)
 - column projection read operation properties [37](#)
 - column projection write operation properties [71](#)
 - configuring a schema for flat files [37](#)
 - general read operation properties [35](#)
 - general write operation properties [70](#)
 - message structure [33](#), [68](#)
 - overview properties [34](#), [69](#)
 - ports read operation properties [36](#)
 - ports write operation properties [70](#)
 - run-time read operation properties [36](#)
 - run-time write operation properties [71](#)
 - source [33](#)
 - sources read operation properties [36](#)
 - target [68](#)
 - target write operation properties [71](#)

- JMS messages

- components [33](#), [68](#)

K

- Kafka connection

- create using infacmd [130](#)
 - general properties [128](#)
 - Kafka broker properties [129](#)

- Kafka data object

- advanced read operation properties [42](#)
 - advanced write operation properties [75](#)
 - column projection read operation properties [43](#)
 - column projections write operation properties [76](#)
 - configuring schema for flat files [43](#)
 - general read operation properties [41](#)
 - general write operation properties [74](#)
 - overview properties [39](#), [73](#)
 - ports read operation properties [41](#), [74](#)
 - run-time read operation properties [41](#)
 - run-time write operation properties [75](#)
 - source [38](#)
 - sources read operation properties [42](#)
 - target [72](#)
 - target write operation properties [75](#)

- Kerberized Kafka Clusters

- configuration [17](#)

M

- mapping configurations [97](#)

- MapR Streams connection

- connection properties [131](#)
 - create using infacmd [131](#)
 - general properties [131](#)

- MapR Streams data object

- overview properties [45](#), [80](#)
 - run-time write operation properties [82](#)
 - target write operation properties [82](#)

- MapRStreams connection

- overview [130](#)

- MapRStreams data object

- advanced read operation properties [47](#)
 - column projection read operation properties [47](#)

MapRStreams data object (*continued*)
column projection write operation properties [82](#)
general read operation properties [46](#)
general write operation properties [81](#)
ports read operation properties [46](#)
ports write operation properties [81](#)
run-time read operation properties [47](#)
source [44](#)
sources read operation properties [47](#)
target [79](#)
Messaging connection
JMS connection [123](#)
Kafka connection [128](#)
Microsoft Azure Data Lake Store connection
properties [132](#)
Microsoft Azure Data Lake Store Connection
overview [132](#)
Microsoft Azure Data Lake Store data object
target [76](#)
Microsoft Azure Data Lake Store write operation
properties [77](#)

P

property fields
JMS messages [34](#), [69](#)

R

relational data object
target [83](#)
run configuration [97](#)
run-time properties
HBase data object read operation [32](#)

S

sample file
sample JSON file [135](#)
sample XSD file [134](#)
stateful computing [94](#)
streaming mapping
data object validation [99](#)
environment validation [99](#)
rules [96](#)
run-time validation [101](#)
transformation validation [100](#)
transformations [92](#)
validation types [99](#)
Structured streaming
clients and tools [12](#)
Structured Streaming mapping
monitor jobs [101](#)

T

Troubleshooting
streaming mappings [103](#)

W

Window transformation
overview [107](#)
rules and guidelines for transformations [112](#)
sliding window [108](#)
sliding Window transformation example [110](#)
tumbling window [108](#)
tumbling Window transformation example [110](#)
window properties [109](#)
window types [107](#)