

Streams Replication Manager for HDF and HDP 1.0.0

Understanding Replication Flows

Date published: 2019-09-20

Date modified: 2019-09-20

The Cloudera logo is displayed in a bold, orange, sans-serif font. The word "CLOUDERA" is written in all caps, with a stylized 'E' that has a horizontal bar extending to the right.

<https://docs.cloudera.com/>

Legal Notice

© Cloudera Inc. 2022. All rights reserved.

The documentation is and contains Cloudera proprietary information protected by copyright and other intellectual property rights. No license under copyright or any other intellectual property right is granted herein.

Unless otherwise noted, scripts and sample code are licensed under the Apache License, Version 2.0.

Copyright information for Cloudera software may be found within the documentation accompanying each component in a particular release.

Cloudera software includes software from various open source or other third party projects, and may be released under the Apache Software License 2.0 (“ASLv2”), the Affero General Public License version 3 (AGPLv3), or other license terms. Other software included may be released under the terms of alternative open source licenses. Please review the license and notice files accompanying the software for additional licensing information.

Please visit the Cloudera software product page for more information on Cloudera software. For more information on Cloudera support services, please visit either the Support or Sales page. Feel free to contact us directly to discuss your specific needs.

Cloudera reserves the right to change any products at any time, and without notice. Cloudera assumes no responsibility nor liability arising from the use of products, except as expressly agreed to in writing by Cloudera.

Cloudera, Cloudera Altus, HUE, Impala, Cloudera Impala, and other Cloudera marks are registered or unregistered trademarks in the United States and other countries. All other trademarks are the property of their respective owners.

Disclaimer: EXCEPT AS EXPRESSLY PROVIDED IN A WRITTEN AGREEMENT WITH CLOUDERA, CLOUDERA DOES NOT MAKE NOR GIVE ANY REPRESENTATION, WARRANTY, NOR COVENANT OF ANY KIND, WHETHER EXPRESS OR IMPLIED, IN CONNECTION WITH CLOUDERA TECHNOLOGY OR RELATED SUPPORT PROVIDED IN CONNECTION THEREWITH. CLOUDERA DOES NOT WARRANT THAT CLOUDERA PRODUCTS NOR SOFTWARE WILL OPERATE UNINTERRUPTED NOR THAT IT WILL BE FREE FROM DEFECTS NOR ERRORS, THAT IT WILL PROTECT YOUR DATA FROM LOSS, CORRUPTION NOR UNAVAILABILITY, NOR THAT IT WILL MEET ALL OF CUSTOMER’S BUSINESS REQUIREMENTS. WITHOUT LIMITING THE FOREGOING, AND TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW, CLOUDERA EXPRESSLY DISCLAIMS ANY AND ALL IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO IMPLIED WARRANTIES OF MERCHANTABILITY, QUALITY, NON-INFRINGEMENT, TITLE, AND FITNESS FOR A PARTICULAR PURPOSE AND ANY REPRESENTATION, WARRANTY, OR COVENANT BASED ON COURSE OF DEALING OR USAGE IN TRADE.

Contents

Replication Flows Overview.....	4
Remote Topics.....	4
Bi-directional Replication Flows.....	5
Fan-in and Fan-out Replication Flows.....	5

Replication Flows Overview

Get familiar with the concept of replications and replication flows.

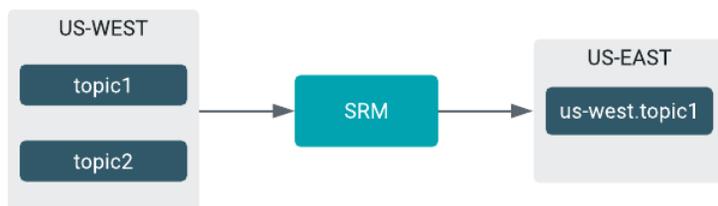
Replication involves sending records from a source cluster to a target cluster. In SRM a replication refers to a source and target cluster pair, the direction in which data is flowing and the topics that are being replicated. Source target cluster pairs can be specified in the SRM configuration file; they are notated `source->target`. Initially, when `source->target` pairs are set up they are considered inactive, as no data is being replicated between them. To start replication users need to specify which topics to replicate with the `srms-control` command line tool.

It is important to understand that replication in SRM is configured independently for each `source->target` cluster pair. Moreover, configuration is done on a per topic basis. This means that each topic in a source cluster can have a different direction or target that it is being replicated to. A set of topics in the source cluster can be replicated to multiple target clusters while others are being replicated to only one target cluster. This allows users to set up powerful, topic specific replication flows.

The term replication flow is used to specify all replications set up in a system. This document uses the term when referring to the visual representation of SRM replications.

A basic example of a replication flow is when topics are being sent from one cluster to another cluster in a different geographical location. Note that in this example there is only one replication or `source->target` pair. Moreover, only one of the two topics on the source cluster are being replicated to the target cluster.

Figure 1: Simple Replication Flow Example

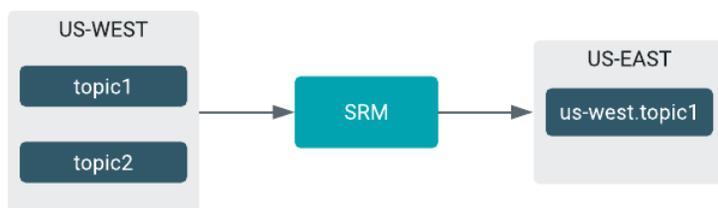


Remote Topics

Learn about SRM's remote topics.

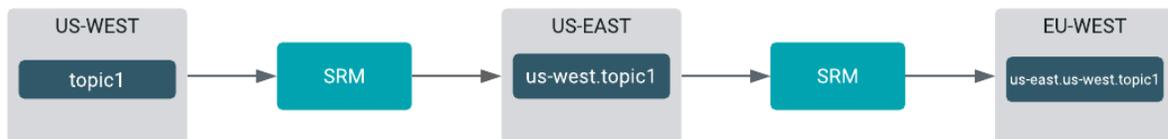
In any replication flow, the selected source topics are replicated to remote topics on the target cluster. Remote topics reference the source cluster via a naming convention. For example, the `topic1` topic from the `us-west` source cluster creates the `us-west.topic1` remote topic on the target cluster.

Figure 2: Simple Replication Flow Example



Remote topics can themselves be replicated. In this case, the remote topic references all source and target clusters. The prefix in the name will start with the cluster closest to the final target cluster. For example, the `topic1` topic replicated from the `us-west` source cluster to the `us-east` cluster and then to the `eu-west` cluster will be named `us-east.us-west.topic1`.

Figure 3: Complex Replication Flow Example



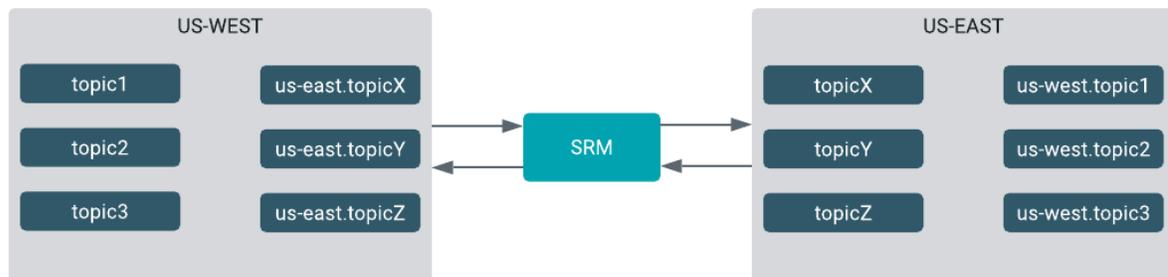
Tip: You might want to have your Kafka consumers read messages from both source and remote topics simultaneously. To achieve this, Kafka consumers should include a wildcard topic name pattern. For example, suppose that you want your consumer to read from `topic1` located in `us-west` and its remote counterpart, `us-west.topic1`, located in `us-east`. In such a case, you can use the `. *topic1` pattern, which matches any topic that ends with `topic1`.

Bi-directional Replication Flows

Learn more about bi-directional replication flows.

SRM understands cycles and will never replicate records in an infinite loop. This enables bi-directional replication flows in which clusters are mutually replicated. In this case, records sent to one cluster will be replicated to the other and the other way around. You can configure any number of clusters in this way.

Figure 4: Bi-directional Replication Flow

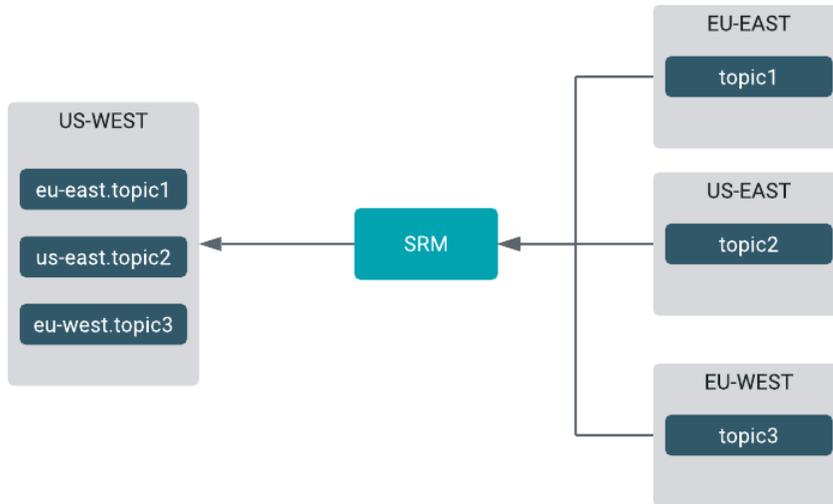


Fan-in and Fan-out Replication Flows

Learn about fan-in and fan-out replication flows.

You can construct fan-in replication flows, where records from multiple source clusters are aggregated in a single target cluster.

Figure 5: Fan-in Replication Flow



Similarly, you can construct fan-out replication flows as well, where a single cluster is replicated to multiple target clusters.

Figure 6: Fan-out Replication Flow

