

Planning Your Streams Messaging Deployment

Date published: 2022-02-24

Date modified: 2022-02-24

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.

Legal Notice

© Cloudera Inc. 2024. 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

| | |
|--|----------|
| Deployment scenarios..... | 4 |
| Data Hub cluster definitions..... | 4 |
| Streams Messaging cluster layout..... | 5 |

Deployment scenarios

Before you get started with a Cloudera Data Flow deployment, it is useful to understand which software version is right for your platform and operational objectives. This helps you additionally understand the documentation you need to review to get started.

| If you want ... | Running with the following additional components | Review this documentation |
|---|---|--|
| Kafka for CDP Public Cloud – Streams Messaging clusters | <ul style="list-style-type: none"> • Schema Registry • Streams Messaging Manager • Streams Replication Manager • Cruise Control • Kafka Connect (technical preview, not supported) | Cloudera Data Flow for Data Hub documentation |
| Kafka for CDP Private Cloud Base | | Streaming libraries in Cloudera Runtime CDP Private Cloud Base |
| CDK powered by Apache Kafka for CDH 5.x | | CDK 4.1 Powered by Apache Kafka documentation |
| | <ul style="list-style-type: none"> • Schema Registry | Cloudera Streams Processing 1.0.0 documentation |
| | <ul style="list-style-type: none"> • Schema Registry • Streams Messaging Manager • Streams Replication Manager | Cloudera Streams Processing 2.0.x documentation |
| Kafka in CDH 6.x | | CDH 6.x Kafka documentation |
| | <ul style="list-style-type: none"> • Schema Registry • Streams Messaging Manager • Streams Replication Manager | Cloudera Streams Processing documentation |
| Kafka for HDF and HDP | | HDF documentation |

Data Hub cluster definitions

The Streams Messaging templates include Kafka, Schema Registry, Streams Messaging Manager, Streams Replication Manager and ZooKeeper. You may choose from the following template options, depending on your operational objectives:

- Streams Messaging Heavy Duty for AWS
- Streams Messaging Heavy Duty for Azure
- Streams Messaging Heavy Duty for GCP
- Streams Messaging Light Duty for AWS
- Streams Messaging Light Duty for Azure
- Streams Messaging Light Duty for GCP

Streams Messaging provides advanced messaging and real-time processing on streaming data using Apache Kafka, centralized schema management using Schema Registry, management and monitoring capabilities powered by Streams Messaging Manager, as well as cross-cluster Kafka topic replication using Streams Replication Manager and Kafka partition rebalancing with Cruise Control.

These templates set up fault-tolerant standalone deployments of Apache Kafka and supporting Cloudera components (Schema Registry, Streams Messaging Manager, Streams Replication Manager and Cruise Control), which can be used for Kafka workloads in the cloud or as a disaster recovery instance for on-premises Kafka clusters.

Streams Messaging cluster layout

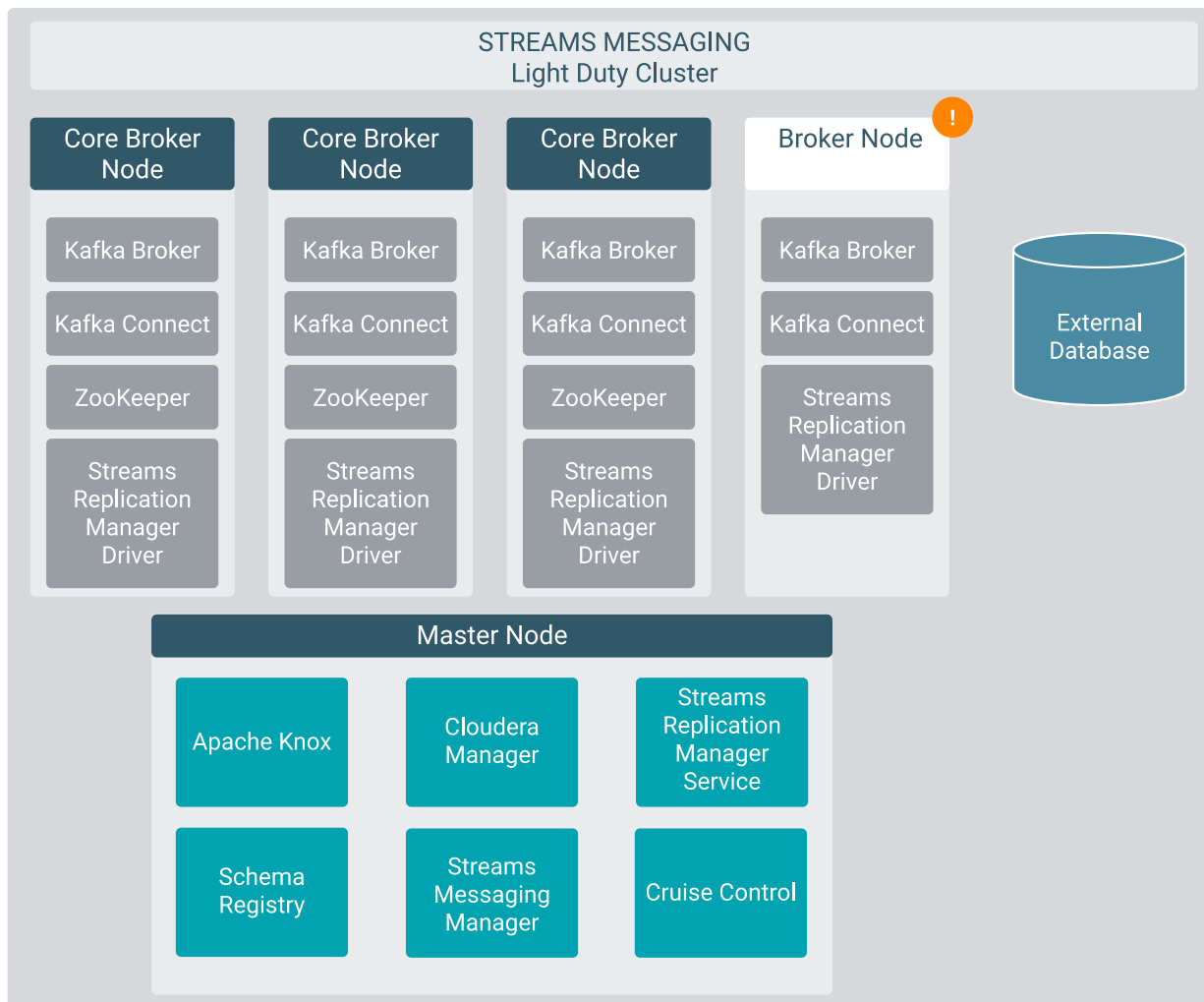
Describes the layout and capacity of the Streams Messaging: Light Duty and Streams Messaging: Heavy Duty cluster definitions

Streams Messaging: Light Duty cluster layout

You can use a Streams Messaging: Light Duty cluster definition in development, testing, or proof of concept scenarios.

- Core Broker and Broker node instance:
 - AWS: m5.2xlarge
 - Azure: D8_v3
 - GCP: e2-standard-8
- Storage configuration per Core Broker and Broker node:
 - AWS: 1 TB Volume EBS ST1
 - Azure: 1 TB Standard Locally-redundant SSD storage
 - GCP: 1 TB Zonal PD-SSD
- Master node instance
 - AWS: m5.2xlarge
 - Azure: Standard_D8_v3
 - GCP: e2-standard-8
- Storage configuration for Master node:
 - AWS: 100 GB Volume EBS Magnetic
 - Azure: 100 GB Standard Locally-redundant SSD storage
 - GCP: 100 GB Zonal PD-Standard

For more information about the cloud provider-specific instance and storage types, see the *Related Information* section.



The Broker node is not provisioned by default. You have the option to manually set how many Broker nodes are created when provisioning the cluster. After the cluster is provisioned, the number of Broker nodes can be changed by scaling your cluster. For more information about scaling, Core Broker and Broker nodes, see *Scaling Streams Messaging Clusters*.



Important: By default, the volume per instance count for Broker and Core Broker nodes is identical. If you customize your cluster during provisioning, Cloudera recommends that Attached Volume per Instances is set to the same value for both node types. Alternatively, if you want to provision a cluster where the number of volumes is not identical, ensure that you complete *Configure data directories for clusters with custom disk configurations* after the cluster is provisioned. Otherwise, Kafka does not utilize all available volumes. Additionally, scaling the cluster might also not be possible.



Note: While Kafka Connect is included and provisioned with this cluster layout, it is considered technical preview and is not supported.

Streams Messaging: Heavy Duty cluster layout

You can use the Streams Messaging: Heavy Duty cluster definition in production scenarios. The cluster definition includes:

Azure

- Master Node – Containing Knox, Cloudera Manager, ZooKeeper
 - Instance type – Standard_D8_v3
 - Storage configuration – 100 GB Standard Locally-redundant SSD storage
- Registry Nodes – Containing Schema Registry, ZooKeeper
 - Instance type – Standard_D8_v3
 - Storage configuration – 100 GB Standard Locally-redundant HDD storage
- SMM Nodes – Containing SMM, Schema Registry, ZooKeeper
 - Instance type – Standard_D8_v3
 - Storage configuration – 100 GB Standard Locally-redundant HDD storage
- Core Broker Nodes – Containing a Kafka Broker
 - Instance type – Standard_D8s_v3
 - Storage configuration – 1 TB Premium Locally-redundant SSD storage
- Broker Nodes – Containing a Kafka Broker
 - Instance type – Standard_D8s_v3
 - Storage configuration – 1 TB Premium Locally-redundant SSD storage
- SRM Nodes – Containing the SRM Driver and Service
 - Instance type – Standard_D8_v3
 - Storage configuration – 100 GB Standard Locally-redundant HDD storage
- Connect Nodes – Containing a Kafka Connect role
 - Instance type – Standard_D8_v3
 - Storage configuration – 100 GB Standard Locally-redundant HDD storage

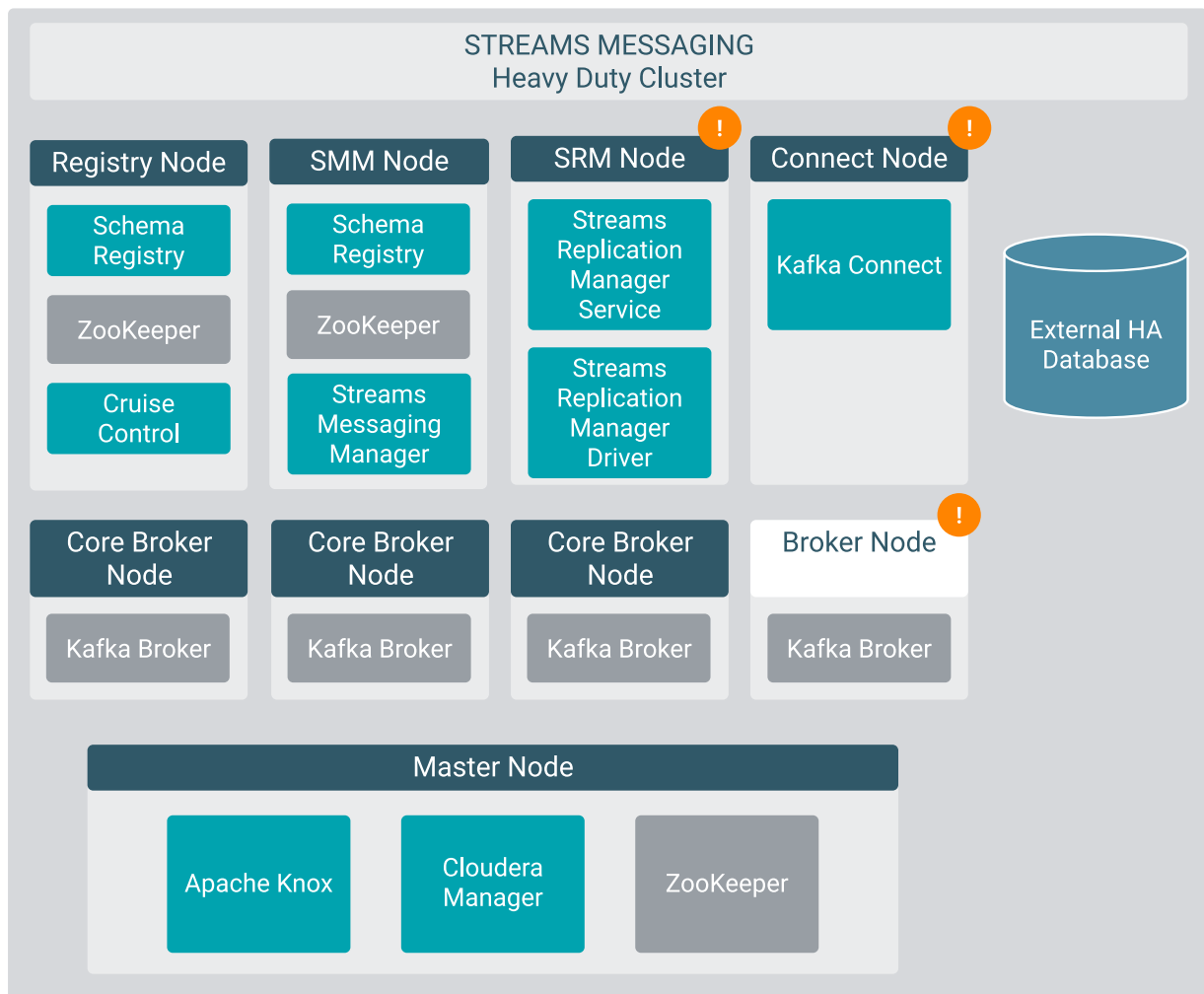
AWS

- Master Node – Containing Knox, Cloudera Manager, ZooKeeper
 - Instance type – m5.2xlarge
 - Storage configuration – 100 GB EBS SC1
- Registry Nodes – Containing Schema Registry, ZooKeeper
 - Instance type – m5.2xlarge
 - Storage configuration – 100 GB EBS SC1
- SMM Nodes – Containing SMM, Schema Registry, ZooKeeper
 - Instance type – m5.2xlarge
 - Storage configuration – 100 GB EBS SC1
- Core Broker Nodes – Containing a Kafka Broker
 - Instance type – m5.2xlarge
 - Storage configuration – 1 TB GP2 SSD
- Broker Nodes – Containing a Kafka Broker
 - Instance type – m5.2xlarge
 - Storage configuration – 1 TB GP2 SSD
- SRM Nodes – Containing the SRM Driver and Service
 - Instance type – m5.2xlarge
 - Storage configuration – 100 GB EBS SC1
- Connect Nodes – Containing a Kafka Connect role
 - Instance type – m5.2xlarge
 - Storage configuration – 100 GB EBS SC1

GCP

- Master Node – Containing Knox, Cloudera Manager, ZooKeeper
 - Instance type – e2-standard-8
 - Storage configuration – 100 GB Zonal PD-Standard
- Registry Nodes – Containing Schema Registry, ZooKeeper
 - Instance type – e2-standard-8
 - Storage configuration – 100 GB Zonal PD-Standard
- SMM Nodes – Containing SMM, Schema Registry, ZooKeeper
 - Instance type – e2-standard-8
 - Storage configuration – 100 GB Zonal PD-Standard
- Core Broker Nodes – Containing a Kafka Broker
 - Instance type – e2-standard-8
 - Storage configuration – 1 TB Premium Locally-redundant SSD storage
- Broker Nodes – Containing a Kafka Broker
 - Instance type – e2-standard-8
 - Storage configuration – 1 TB Premium Locally-redundant SSD storage
- SRM Nodes – Containing the SRM Driver and Service
 - Instance type – e2-standard-8
 - Storage configuration – 100 GB Zonal PD-Standard
- Connect Nodes – Containing a Kafka Connect role
 - Instance type – e2-standard-8
 - Storage configuration – 100 GB Zonal PD-Standard

For more information about the cloud provider-specific instance and storage types, see the *Related Information* section.



The SRM, Broker, and Connect nodes are not provisioned by default. When provisioning a cluster with this definition, you have to manually set the instance count of the appropriate host group to at least 1. Otherwise the host group and its nodes are not provisioned. After a cluster is provisioned, you also have the option to scale these nodes. For more information on scaling, see *Scaling Streams Messaging Clusters*.



Important: By default, the volume per instance count for Broker and Core Broker nodes is identical. If you customize your cluster during provisioning, Cloudera recommends that Attached Volume per Instances is set to the same value for both node types. Alternatively, if you want to provision a cluster where the number of volumes is not identical, ensure that you complete *Configure data directories for clusters with custom disk configurations* after the cluster is provisioned. Otherwise, Kafka does not utilize all available volumes. Additionally, scaling the cluster might also not be possible.



Note: While Kafka Connect is included and can be provisioned with this cluster layout, it is considered technical preview and is not supported.

Related Information

[Scaling Streams Messaging Clusters](#)

[AWS instance types](#)

[Azure instance types](#)

[GCP instance types](#)

[AWS storage information](#)

[Azure storage information](#)

[GCP storage information](#)

[Configure data directories for clusters with custom disk configurations](#)