

Cloudera DataFlow for Data Hub 7.2.17

Planning your Streaming Analytics deployment

Date published: 2019-12-16

Date modified: 2023-06-27

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. 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

Streaming Analytics Data Hub cluster definitons.....	4
Streaming Analytics cluster layout.....	4

Streaming Analytics Data Hub cluster definitions

There are four Streaming Analytics cluster definitions available to deploy Streaming Analytics in CDP Public Cloud. You can choose from Light and Heavy duty options, and you can further select the cluster definitions depending on your cloud provider.

You can choose from the following template options based on your operational objectives:

- Streaming Analytics Light Duty for AWS
- Streaming Analytics Light Duty for Azure
- Streaming Analytics Light Duty for GCP
- Streaming Analytics Heavy Duty for AWS
- Streaming Analytics Heavy Duty for Azure
- Streaming Analytics Heavy Duty for GCP

Streaming Analytics offers real-time stream processing and stream analytics with low-latency and high scaling capabilities powered by Apache Flink.

Streaming Analytics templates include Apache Flink that works out of the box in stateless or heavy state environments. Beside Flink, the template includes its supporting services namely YARN, Zookeeper and HDFS. The Heavy Duty template comes preconfigured with RocksDB as state backend, while Light Duty clusters use the default Heap state backend. You can create your streaming application by choosing between Kafka, Kudu, and HBase as datastream connectors.

You can also use SQL to query real-time data with SQL Stream Builder (SSB) in the Streaming Analytics template. By supporting the SSB service in CDP Public Cloud, you can simply and easily declare expressions that filter, aggregate, route, and otherwise mutate streams of data. SSB is a job management interface that you can use to compose and run SQL on streams, as well as to create durable data APIs for the results.

In CDP Public Cloud, you can use the predefined cluster definitions within your environment to connect Flink and SSB with the supported service connectors.

Table 1: Connector support with Cluster Definitions

Cluster Definition	SSB Connector ¹	Flink Connector
Streams Messaging	Kafka, Schema Registry	Kafka, Schema Registry
Real-Time Data Mart	Kudu	Kudu
Data Engineering	Hive ²	Hive
Operational Database	-	HBase

Streaming Analytics cluster layout

The Light and Heavy Duty layouts differ from each other based on the volume of storage. You need to choose between the two options based on your operational objectives and application specifications.

² You can also use the available Hive service from the Data Lake of the environment.

Streaming Analytics: Light Duty cluster layout

You can use a Streaming Analytics: Light Duty cluster definition in development and testing scenarios. The Light Duty cluster definition can also be used in production for stateless Flink jobs or for Flink jobs with minimal state. The Light Duty cluster has the following specifications:

- Flink, SQL Stream Builder, HDFS, YARN, Zookeeper and Kafka are co-located on all instances
- For each node hosting Flink, SQL Stream Builder, HDFS, YARN, Zookeeper and Kafka
 - AWS: m5.2xlarge
 - Azure: Standard_D8_v3
 - GCP: e2-standard-8

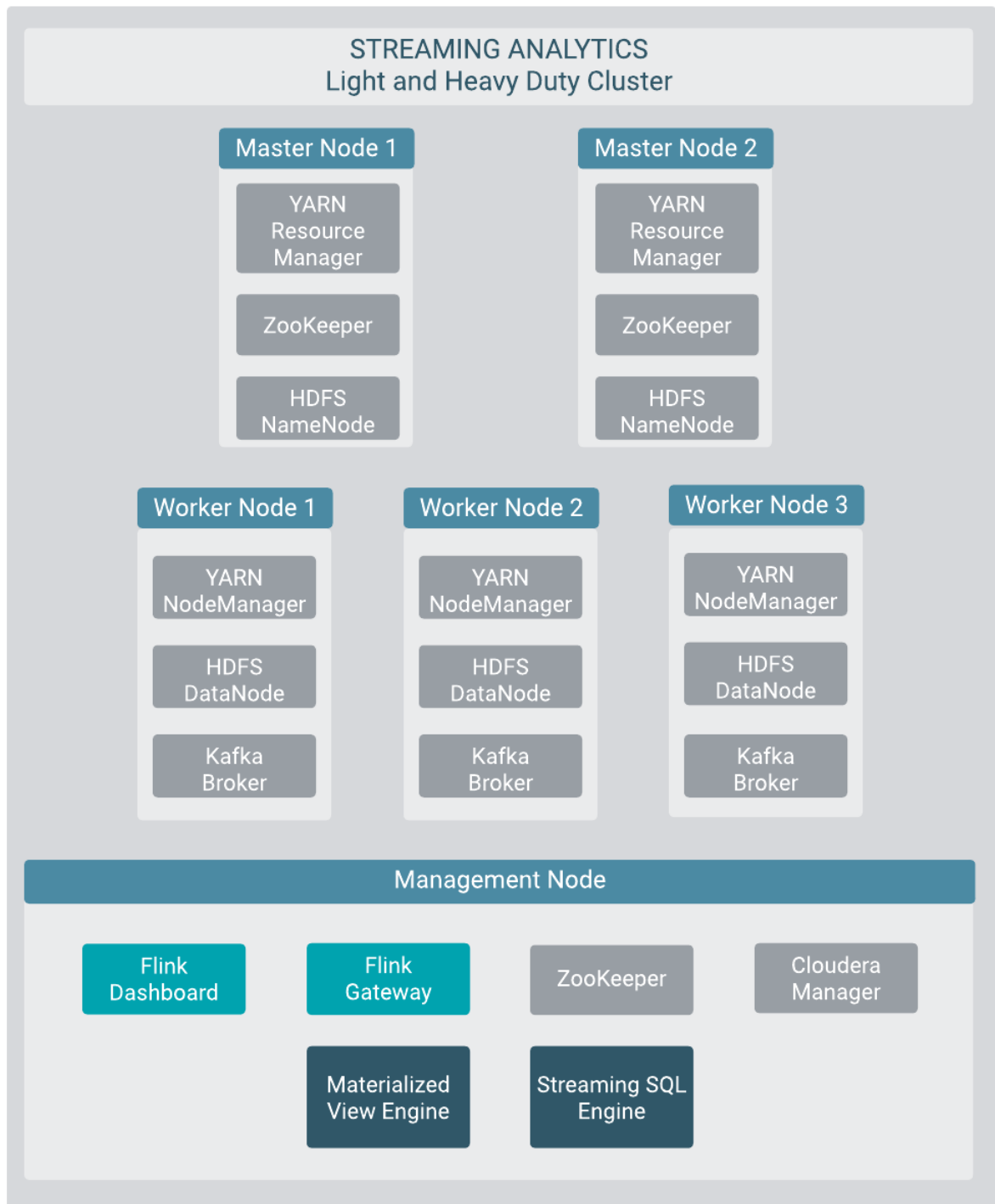
For more information, see your cloud provider-specific information about instance types and storage information.

Streaming Analytics: Heavy Duty cluster layout

You can use a Streaming Analytics: Heavy Duty cluster definition in production for Flink jobs with large state with RockDB as state backend. The Heavy Duty cluster has the following specifications:

- Flink, SQL Stream Builder, HDFS, YARN, Zookeeper and Kafka are co-located on all instances
- For each node hosting Flink, SQL Stream Builder, HDFS, YARN, Zookeeper and Kafka
 - AWS: m5.2xlarge
 - Azure: Standard_D8_v3
 - GCP: e2-standard-8
- For worker nodes:
 - Storage type: SSD
 - Volume size: 1000 GB

For more information, see your cloud provider-specific information about instance types and storage information.



Important: In the Streaming Analytics cluster templates, Kafka service is included by default to serve as a background service only for the websocket output and sampling feature of SQL Stream Builder. The Kafka service in the Streaming Analytics cluster template cannot be used for production, you need to use the Streams Messaging cluster template when Kafka is needed for your deployment.

Related Information

[AWS instance types](#)

[Azure instance types](#)

[GCP instance types](#)

[AWS storage information](#)

[Azure storage information](#)

[GCP storage information](#)