Cloudera DataFlow for Data Hub 7.3.2

Planning your Cloudera Streaming Analytics deployment

Date published: 2019-12-16
Date modified: 2024-12-11

https://docs.cloudera.com/

Legal Notice

Contents

Cloudera Streaming Analytics Cloudera Data Hub cluster definitons

There are four cluster definitions available to deploy Cloudera Streaming Analytics in Cloudera on 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:

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

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

Cloudera 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 Cloudera SQL Stream Builder in the Cloudera Streaming Analytics template. By supporting the Cloudera SQL Stream Builder service in Cloudera on cloud, you can simply and easily declare expressions that filter, aggregate, route, and otherwise mutate streams of data. Cloudera SQL Stream Builder 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 Cloudera on cloud, you can use the predefined cluster definitions within your environment to connect Flink and Cloudera SQL Stream Builder with the supported service connectors.

Table 1: Connector support with Cluster Definitions

Cluster Definition	Cloudera SQL Stream Builder 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

Cloudera Streaming Analytics cluster layout

Cloudera Streaming Analytics Light Duty and Heavy Duty cluster definitions differ in typical workload, topology, and cost. Heavy Duty adds dedicated worker nodes with SSD-backed storage for large Flink state; Light Duty does not include those workers. Choose the definition that matches your operational goals and application requirements.

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

Cloudera Streaming Analytics: Light Duty and Heavy Duty compared

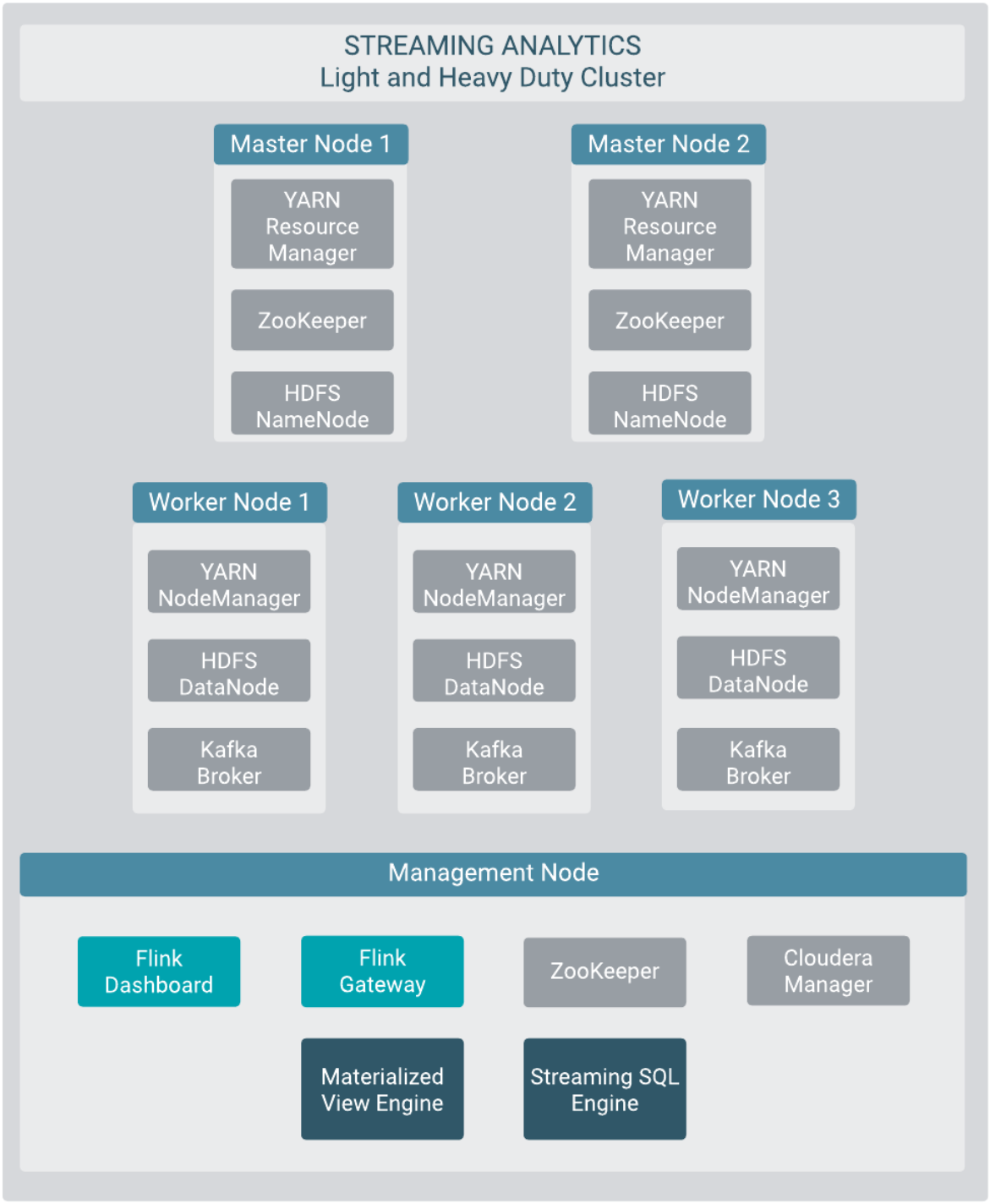
Use the following summary to compare the two cluster definitions at a glance.

Table 2: Cloudera Streaming Analytics Light Duty and Heavy Duty

Consideration	Light Duty	Heavy Duty
Typical use	Development and testing; production for stateless Flink jobs or jobs with minimal state	Production for Flink jobs with large state using RocksDB as the state backend
Topology	Flink, Cloudera SQL Stream Builder, HDFS, YARN, Zookeeper, and Kafka are co-located on all instances. Does not include the separate worker node group described for Heavy Duty.	Same co-located services as Light Duty, plus additional worker nodes with SSD-backed volumes for Flink state and checkpoints (see specifications below).
Cost drivers	Lower baseline footprint: no dedicated worker nodes with large SSD volumes.	Higher cost: worker nodes add instances and 1000 GB SSD storage per worker specification below.

Cluster layout diagram

The following figure illustrates the Cloudera Streaming Analytics cluster templates. The co-located services apply to both Light Duty and Heavy Duty. Only Heavy Duty includes the separate worker nodes with SSD storage; that worker tier is not part of the Light Duty layout.



Important: In the Cloudera 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 Cloudera SQL Stream Builder. The Kafka service in the Cloudera Streaming Analytics cluster template cannot be used for production; you need to use the Cloudera Streams Messaging cluster template when Kafka is needed for your deployment.

Cloudera Streaming Analytics: Light Duty cluster layout

You can use a Cloudera 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 template does not include the Heavy Duty worker nodes with SSD storage; only the co-located node group described in the following specifications applies. The Light Duty cluster has the following specifications:

- Flink, Cloudera SQL Stream Builder, HDFS, YARN, Zookeeper and Kafka are co-located on all instances
- For each node hosting Flink, Cloudera 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.

Cloudera Streaming Analytics: Heavy Duty cluster layout

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

- Flink, Cloudera SQL Stream Builder, HDFS, YARN, Zookeeper and Kafka are co-located on all instances
- For each node hosting Flink, Cloudera 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.

Related Information

AWS instance types
Azure instance types
GCP instance types
AWS storage information
Azure storage information
GCP storage information