Cloudera DataFlow for Data Hub 7.2.17

# Planning your Streaming Analytics deployment

**Date published: 2019-12-16**
**Date modified: 2023-06-27**

**https://docs.cloudera.com/**

# Legal Notice

# Contents

# Streaming Analytics Data Hub cluster definitons

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

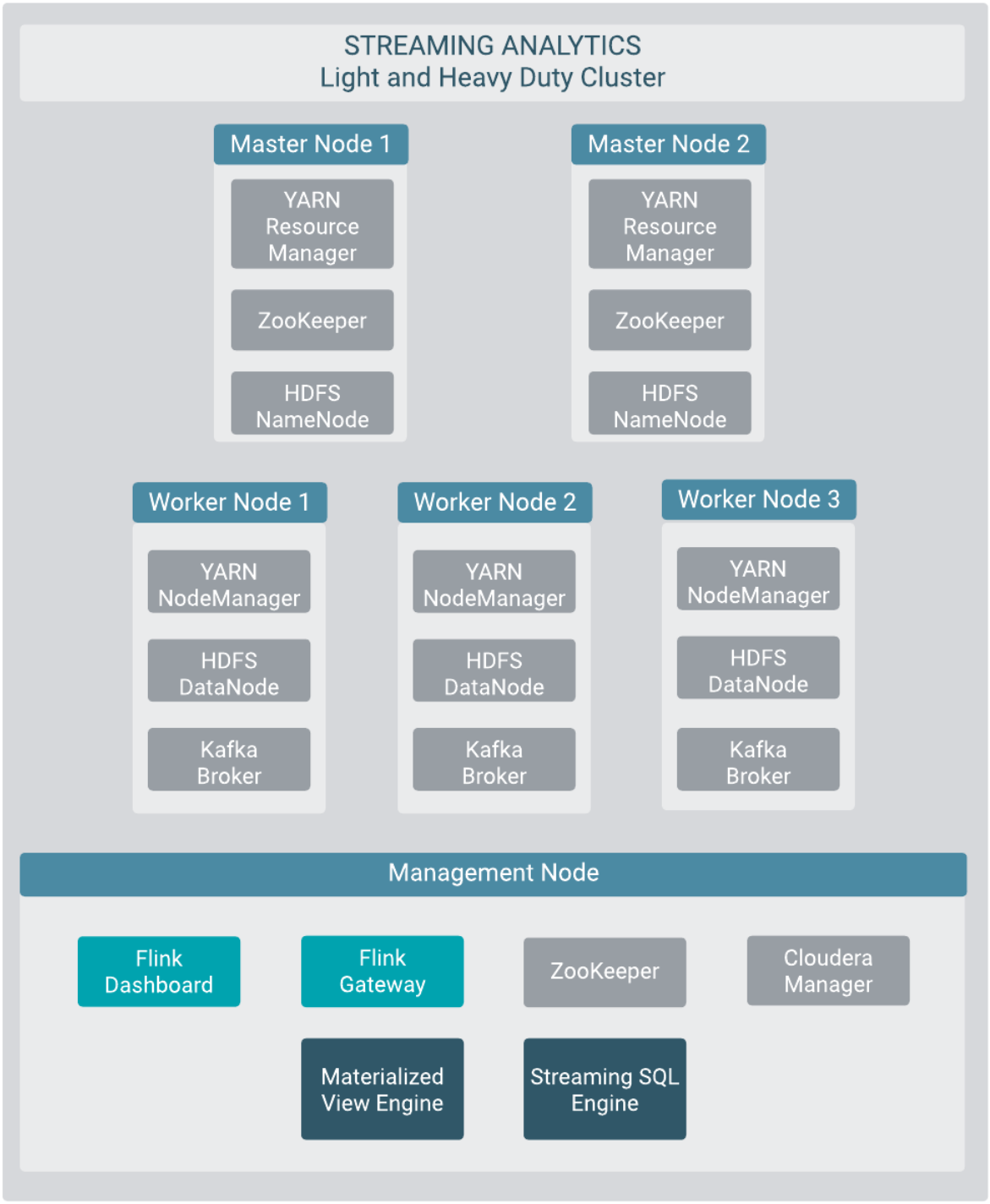
For more information, see your cloud provider-specific information about instance types and storage information.

### 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 ouput 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