Installation 3

Planning your deployment

Date of Publish: 2018-08-13

http://docs.hortonworks.com

Contents

Deployment Scenarios

Your deployment scenario for installing, configuring, or upgrading your Hortonworks DataFlow (HDF) components depends on your particular use case.

Table 1: Installation Scenarios

Scenario	Installation Scenario	Steps
Installing an HDF Cluster	This scenario applies if you want to install the entire HDF platform, consisting of all flow management and stream processing components on a new cluster. The stream processing components include the new Streaming Analytics Manager (SAM) modules that are in GA (General Availability). This includes the SAM Stream Builder and Stream Operations modules but does not include installing the technical preview version of SAM Stream Insight, which is powered by Druid and Superset. This scenario requires that you install an HDF cluster.	1. Install Ambari. 2. Install databases. 3. Install the HDF management pack. 4. Install an HDF cluster using Ambari.
Installing HDF Services on a New HDP Cluster	This scenario applies to you if you are both an Hortonworks Data Platform (HDP) and HDF customer and you want to install a fresh cluster of HDP and add HDF services. The stream processing components include the new (SAM) and all of its modules. This includes installing the technical preview version of the SAM Stream Insight module, which is powered by Druid and Apache Superset. This scenario requires that you install both an HDF cluster and an HDP cluster.	1. Install Ambari. 2. Install databases. 3. Install an HDP cluster using Ambari. 4. Install the HDF management pack. 5. Update the HDF base URL. 6. Add HDF services to an HDP cluster
Installing HDF Services on an Existing HDP Cluster	You have an existing HDP cluster with Apache Storm and or Apache Kafka services and want to install Apache NiFi or NiFi Registry modules on that cluster. This requires that you upgrade to the latest version of Apache Ambari and HDP, and then use Ambari to add HDF services to the upgraded HDP cluster.	1. Upgrade Ambari 2. Upgrade HDP 3. Install Databases 4. Install HDF Management Pack 5. Update HDF Base URL 6. Add HDF Services to HDP cluster

Table 2: Upgrade Scenarios

Scenario	Upgrade Scenario	Documentation
Upgrading an HDF-only cluster	You have an existing Ambari-managed HDF cluster and want to upgrade it using an Express upgrade.	Resources for each upgrade scenario are available in the Ambari-Managed HDF Upgrade Guide.
Upgrading HDF 3.1.x services on an HDP cluster.	You have an existing Ambari-managed HDP cluster with HDF 3.1.x services installed. In the HDF 3.1.x release, NiFi and NiFi Registry were available to install on HDP. You want to upgrade your HDP cluster to HDP 3.0.0 and you want to upgrade your HDF services to the services available with HDF 3.2.0.	
Upgrading an HDP cluster with HDF 3.0.x	You have an existing Ambari-managed HDP cluster with HDF 3.0.x services installed. In the HDF 3.0.x release, NiFi, SAM, and Schema Registry were available to install on HDP.	

HDF Cluster Types and Recommendations

Cluster Type	Description	Number of VMs or Nodes	Node Specification	Network
Single VM HDF Sandbox	Evaluate HDF on local machine. Not recommended to deploy anything but simple applications.	1 VM	At least 4 GB RAM	
Evaluation Cluster	Evaluate HDF in a clustered environment. Used to evaluate HDF for simple data flows and streaming applications.	3 VMs or no des	• 16 GB of RAM • 8 cores/vCores	
Small Development Cluster	Use this cluster in development environments.	6 VM s/No des	• 16 GB of RAM • 8 cores or vCores	
Medium QE Cluster	Use this cluster in QE environments.	8 VMs/Nodes	• 32 GB of RAM • 8 to16 cores or vCores	
Small Production Cluster	Use this cluster in small production environments.	15 VMs/Nodes	• 64 - 128 GB of RAM • 8 - 16 cores of RAM	1 GB Bonded Nic
Medium Production Cluster	Use this cluster in a medium production environment.	24 VMs/Nodes	• 64 - 128 GB of RAM • 8 - 16 cores of RAM	10 GB bonded network interface card (NIC)
Large Production Cluster	Use this cluster in a large production environment.	32 VMs/Nodes	• 64 - 128 GB of RAM • 16 cores of RAM	10 GB Bonded NIC

More Information

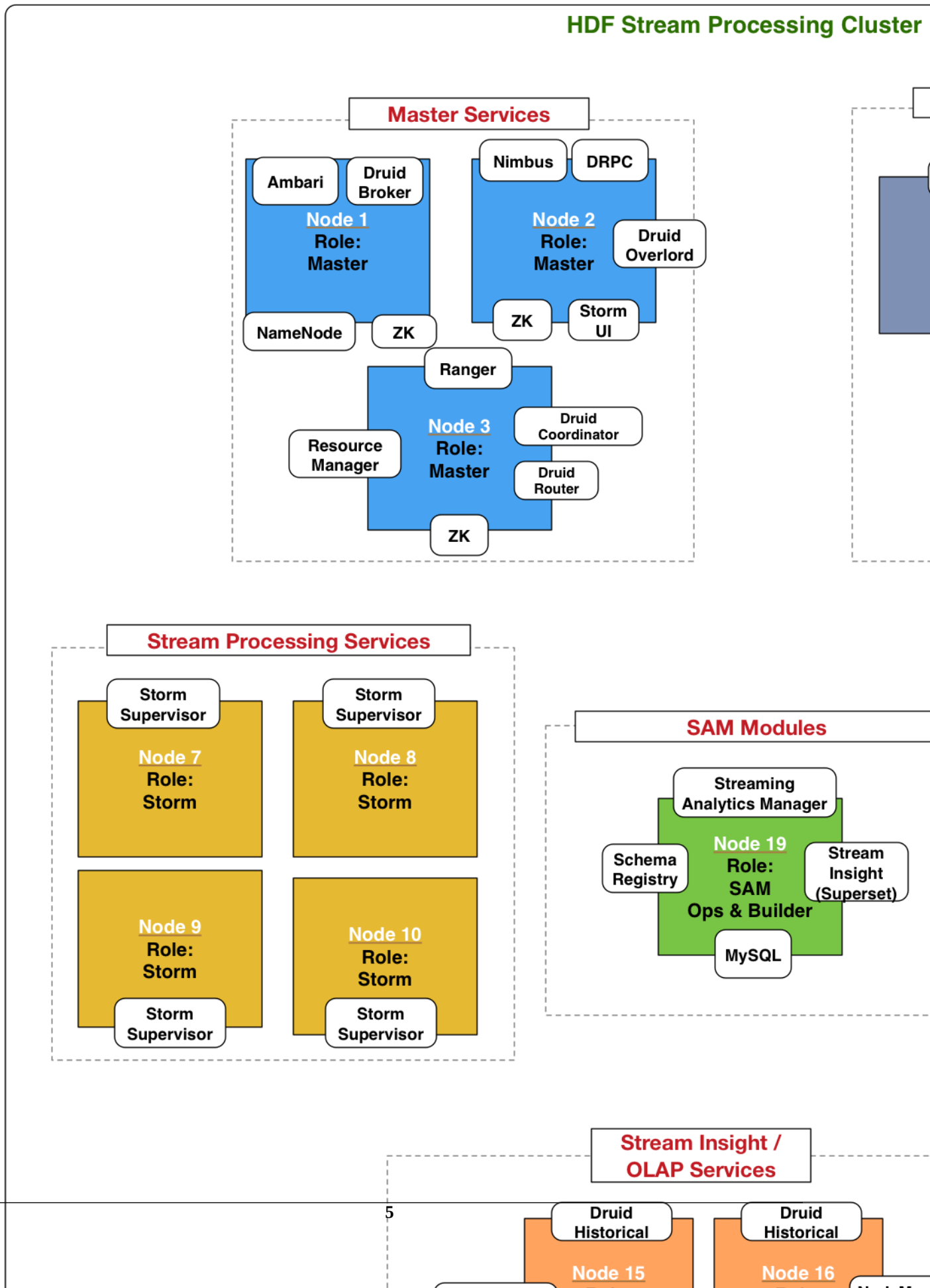
Download the Sandbox

Production Cluster Guidelines

General guidelines for production guidelines for service distribution:

- NiFi, Storm, and Kafka should not be located on the same node or virtual machine.
- NiFi, Storm, and Kafka must have a dedicated ZooKeeper cluster with at least three nodes.
- If the HDF SAM is being used in an HDP cluster, the SAM should not be installed on the same node as the Storm worker node.

The following diagram illustrates how services could be distributed for a small production cluster across 19 nodes:


HDF Stream Processing Cluster
Master Services
Ambari
Druid Broker
Node 1
Role:
Master
NameNode
ZK
Nimbus
DRPC
Node 2
Role:
Master
Druid Overlord
ZK
Storm UI
Ranger
Node 3
Role:
Master
Resource Manager
Druid Coordinator
Druid Router
ZK
Stream Processing Services
Storm Supervisor
Node 7
Role:
Storm
Storm Supervisor
Node 8
Role:
Storm
Node 9
Role:
Storm
Storm Supervisor
Node 10
Role:
Storm
Storm Supervisor
SAM Modules
Streaming Analytics Manager
Schema Registry
Node 19
Role:
SAM
Ops & Builder
Stream Insight (Superset)
MySQL
Stream Insight / OLAP Services
Druid Historical
Node 15
Druid Historical
Node 16

Hardware Sizing Recommendations

Recommendations for Kafka

- Kafka Broker node: eight cores, 64 GB to128 GB of RAM, two or more 8-TB SAS/SSD disks, and a 10- GbE NIC.
- Minimum of three Kafka broker nodes
- Hardware Profile: More RAM and faster speed disks are better; 10 GbE NIC is ideal.
- 75 MB per sec per node is a conservative estimate. You can go much higher if more RAM and reduced lag between writing/reading and therefore 10 GB NIC is required.

With a minimum of 3 nodes in your cluster, you can expect 225 MB/sec data transfer.

You can perform additional further sizing by using the following formula: num_brokers = desired_throughput (MB/sec) / 75

Recommendations for Storm

- Storm Worker Node: 8 core, 64 GB RAM, 1 GbE NIC
- Minimum of 3 Storm worker nodes
- Nimbus Node: Minimum 2 Nimbus nodes , 4 core, 8 GB RAM
- Hardware profile: disk I/O is not that important; more cores are better.
- 50 MB per sec per node with low to moderate complexity topology reading from Kafka and no external lookups . Medium-complexity and high-complexity topologies might have reduced throughput.

With a minimum 2 nimbus, 2 worker cluster, you can expect to run 100 MB/sec of low to medium complexity topology.

Further sizing can be done as follows. Formula: num_worker_nodes = desired_throughput(MB/sec) / 50

Recommendations for NiFi

NiFi is designed to take advantage of:

- all the cores on a machine
- all the network capacity
- all the disk speed
- many gigabytes of RAM (although usually not all) on a system

Hence, it is important that NiFi be running on dedicated nodes. Following are the recommended server and sizing specifications for NiFi:

- Minimum of 3 nodes
- 8+ cores per node (more is better)
- 6+ disks per node (SSD or spinning)
- At least 8 GB

If you want this sustained throughput…	Then provide this minimum hardware ...
50 MB and thousands of events per second	• 1 or 2 nodes • 8 or more cores per node, although more is better • 6 or more disks per node (SSD or spinning) • 2 GB memory per node • 1 GB bonded NICs

If you want this sustained throughput…	Then provide this minimum hardware ...
100 MB and tens of thousands of events per second	• 3 or 4 nodes • 16 or more cores per node, although more is better • 6 or more disks per node (SSD or spinning) • 2 GB of memory per node • 1 GB bonded NICs
200 MB and hundreds of thousands of events per second	• 5 to 7 nodes • 24 or more cores per node (effective CPUs) • 12 or more disks per node (SSD or spinning) • 4 GB of memory per node • 10 GB bonded NICs
400 to 500 MB/sec and hundreds of thousands of events per second	• 7 - 10 nodes • 24 or more cores per node (effective CPUs) • 12 or more disks per node (SSD or spinning) • 6 GB of memory per node • 10 GB bonded NICs