Hortonworks DataFlow

Planning Your Deployment

(August 9, 2017)

docs.hortonworks.com

Hortonworks DataFlow: Planning Your Deployment

Table of Contents

1. Deployment Scenarios

1.1. Identify your Deployment Scenarios

Depending on your use case, your deployment scenario for installing and configuring HDF components is different. These scenarios are covered in the following table.

Scenario	Deployment Scenario	Scenario Steps
Installing HDF Services on a New HDP Cluster	This scenario applies to you if you are both an 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 Streaming Anlatyics Manager (SAM) and **all** of its modules. This includes installing the technical preview version of SAM's Stream Insight module which is powered by Druid and SuperSet. This requires that you install both an HDF and an HDP cluster.	1. Install Ambari 2. Install Databases 3. Install HDP Cluster using Ambari 4. Install HDF Management Pack 5. Update HDF Base URL 6. Add HDF Services to HDP cluster
Installing an HDF Cluster	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 Anlatyics Manager (SAM) modules that are **GA**. This includes the SAM's Stream Builder and Stream Operations modules but **does not** include installing the technical preview version of SAM's Stream Insight module which is powered by Druid and SuperSet. This requires that you install an HDF cluster.	1. Install Ambari 2. Install Databases 3. Install HDF Management Pack 4. Install HDF cluster using Ambari
Installing HDF Services on an Existing HDP Cluster	You have an **existing** HDP cluster with Storm and or Kafka services and want to install NiFi or SAM's modules on that cluster. This requires that you upgrade to the latest version of Ambari and HDP, and 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
Performing any of the above deployment scenarios using a local repository. See *Using Local Repositories* in the instructions appropriate for your scenario.	Local repositories are frequently used in enterprise clusters that have limited outbound internet access. In these scenarios, having packages available locally provides more governance, and better installation performance. This requires that you perform several steps to create a local repository and prepare the Ambari repository configuration file.	1. Obtain the Public Repositories 2. Set Up the Local Repository 3. Prepare the Ambari Repository Configuration File

1.2. HDF Cluster Types and Recommendations

Cluster Type	Description	Number of 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/Nodes	• 16 GB of RAM • 8 cores/vCores	
Small Development Cluster	Use this cluster in development environments.	6 VMs/Nodes	• 16 GB of RAM • 8 cores/vCores	
Medium QE Cluster	Use this cluster in QE environments.	8 VMs/Nodes	• 32 GB of RAM • 8 - 16 cores/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 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

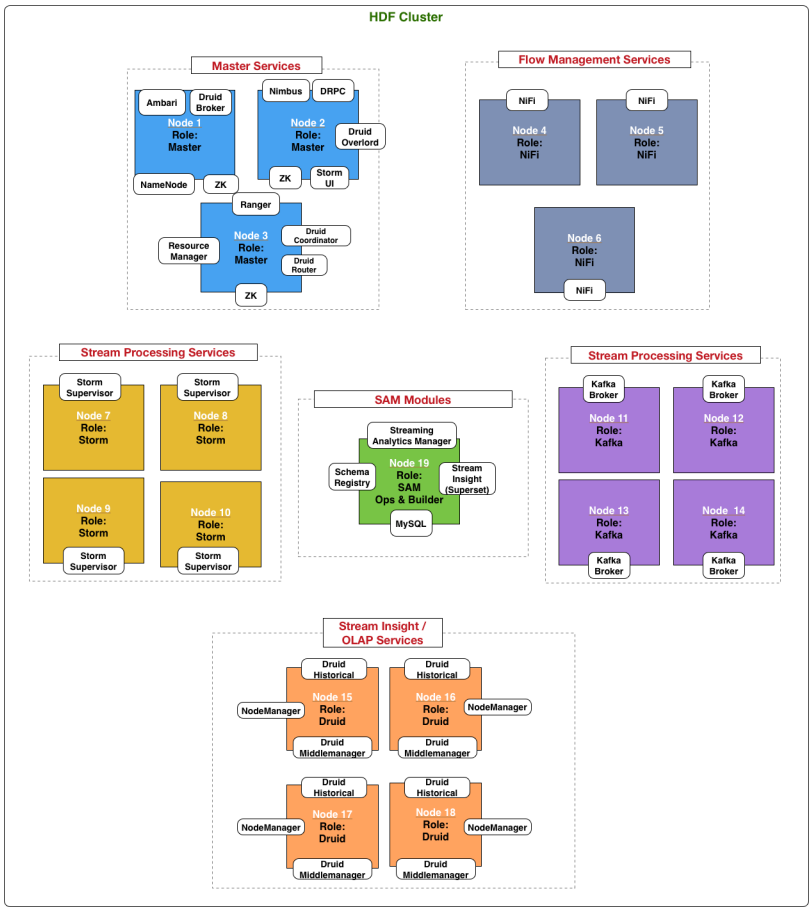
Download the Sandbox

1.3. Production Cluster Guidelines

General guidelines for production guidelines for service distribution:

- NiFi, Storm and Kafka should not be collocated on the same Node/VM.
- NiFi, Storm and Kafka have at least a dedicated 3 Node ZK cluster.
- If HDF's SAM is being used in an HDP cluster, SAM should not installed on the same node as Storm worker node.

The below diagram illustrates how services could be spread out for small production cluster across 15 nodes.



1.4. Hardware Sizing Recommendations

Recommendations for Kafka

- Kafka Broker Node: 8 core, 64-128GB RAM, 2+ 8TB SAS/SSD disk, 10Gige Nic.
- Minimum of 3 Kafka Broker Nodes
- Hardware Profile: More RAM and faster speed disks are better, 10Gige Nic is ideal
- 75 MB/sec per node is a conservative estimate (can go much higher if more RAM and reduced lag between writing/reading and therefore 10GB Nic is required).

With a minimum 3 node cluster, you can expect 225 MB/seccond data transfer.

Further sizing can be done as follows. Formula: num_brokers = desired_throughput(MB/sec) / 75

Recommendations for Storm

- Storm Worker Node: 8 core, 64 GB RAM, 1 Gige Nic
- Minimum of 3 Storm worker nodes
- Nimbus Node: Minimum 2 nimbus nodes, 4 core, 8 GB RAM

- Hardware profile: disk io not that important, more cores are better.
- 50 MB/sec per node with low to moderate complexity topology reading from Kafka and no external lookups. Medium to high complexity topologies may see 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 GB of RAM (though usually not all) on a system

Hence is important that NiFi be running on dedicated nodes. The below are the recommended server and sizing specs 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 ...	Recommended hardware sizing ...
50 MB/second sustained throughput and thousands of events per second	• 1 - 2 nodes • 8 or more cores per node, although more is better • 6 or more disks per node (solid state or spinning) • 2 GB memory per node • 1 GB bonded NICs
100 MB/second sustained throughput and tens of thousands of events per second	• 3 - 4 nodes • 8 or more cores per node, although more is better • 6 or more disks per node (solid state or spinning) • 2 GB of memory per node • 1GB bonded NICs
200 MB/second sustained throughput and hundreds of thousands of events per second	• 5 - 7 nodes • 24 or more cores per node (effective CPUs) • 12 or more disks per node (solid state or spinning)

If you want ...	Recommended hardware sizing ...
	• 4 GB of memory per node • 10 GB bonded NICs
400 - 500 MB/second sustained throughput 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 (solid state or spinning) • 6 GB of memory per node • 10 GB bonded NICs

2. Where to Go Next?

Use this table to help you navigate the HDF documentation library.

If you want to ...	See this document ...
Install or Upgrade an HDF cluster using Ambari	• Release Notes • Support Matrices • Planning Your Deployment • Ambari Upgrade
Manually install or upgrade HDF components. This option is not available for Streaming Analytics Manager and Schema Registry.	• Command Line Installation • MiNiFi Java Agent Quick Start • Manual Upgrade
Get Started with HDF	• Getting Started with Apache NFi • Getting Started with Stream Analytics
Use and administer HDF Flow Management capabilities	• Apache NiFi User Guide • Apache NiFi Administration Guide • Apache NiFi Developer Guide • Apache NiFi Expression Language Guide • MiNiFi Java Agent Administration Guide
Use and administer HDF Stream Analytics capabilities	• Streaming Analytics Manager User Guide • Schema Registry User Guide • Apache Storm Component Guide • Apache Kafka Component Guide