HITACHI
Inspire the Next

Installing Pentaho Server on Hadoop Edge Nodes

HITACHI
Inspire the Next

Change log (if you want to use it):

Date	Version	Author	Changes

Contents

This page intentionally left blank.

Overview

This document discusses the pros and cons of installing Pentaho Server software on Hadoop edge nodes, as well as considerations for splitting the web server tier from the web application tier (Tomcat container).

While we publish minimum and recommended requirements for server hardware suitable for hosting Pentaho software, much of that is based on environments where data is stored in traditional datastores. In these cases, the decision of where to install Pentaho Server software is usually straightforward. These requirements do not discuss the function (dedicated purpose) or location of the server within the corporate network with respect to where data is stored. This is especially relevant for environments where Pentaho is used to ingest, store, process, and analyze large amounts of data into and out of Hadoop clusters.

This document covers these Pentaho versions:

Software	Version(s)
Pentaho	7.x, 8.x

The Components Reference in Pentaho Documentation has a complete list of supported software and hardware.

Before You Begin

This document assumes that you are familiar with and have installed Pentaho Server in your environment. It also assumes that you are familiar with general networking concepts and that you have an overview of Apache Tomcat web server and Apache web server technologies.

If Pentaho is only deployed in the edge node of Hadoop, but not storing and processing data in Hadoop, a Hadoop add-on license is not required. However, if Pentaho is accessing Hadoop data services, or using Hadoop to execute PMR or Spark, then Hadoop node licenses will be required.

Implement the recommendations in this document considering the specific characteristics of each Pentaho Server and Hadoop cluster environment. Create a security and implementation plan with your Hadoop administrators and other stakeholders.

Use Case: Moving Pentaho Off Edge Server or Not?

Amy's company is currently running Pentaho Server on an edge node within their Cloudera Hadoop cluster. The installation uses Apache Web Server to route network traffic to two Tomcat web application servers running a clustered Pentaho Server instance. Amy's company has recently tightened security around the Hadoop cluster and would like to move part or all of the Pentaho Server application off the edge node onto a non-Hadoop network server. Amy needs to know what to consider in making the decision to move part of all of Pentaho off the edge server.

Introduction to Hadoop and Edge Nodes

A typical Hadoop cluster comprises three different server node types: master nodes, worker nodes, and edge or gateway nodes. Edge nodes are named so because they act as a connection or interface between the secured Hadoop cluster and the general corporate network.

You can find more information on these topics in the following sections:

- Hadoop Server Node Types
- Purpose of Edge Nodes

Hadoop Server Node Types

A typical Hadoop cluster comprises the following server node types: master nodes, worker (or slave) nodes, and edge nodes.

A **master node** is a node that runs the services that manage the worker nodes. Examples include the `Namenode` service that manages the storing of data in the Hadoop Distributed Filesystem (HDFS): the Resource Manager service runs parallel data processing computations on that data using MapReduce and/or YARN applications, and the Hive `Metastore` service which uses a relational database to store metadata for Hive tables and partitions, while using its `metastore` service API to let clients (including Hive) access the information.

A **worker node** is a node that runs services that process data. Worker nodes make up most servers (~90+%) in a Hadoop cluster, and they perform the jobs of storing data locally and running computations on that data. Worker nodes run a `DataNode` service (that communicates with, and is subordinate to, the `Namenode` service on the master node), and a `NodeManager` service (that communicates with, and is subordinate to, the Resource Manager service on the master node).

An **edge node** is a node within a Hadoop cluster that is neither a master node nor a worker node. Rather, it is a node that is configured like any other Hadoop node in terms of binaries (libraries) and Hadoop configuration settings, but that does not run any of the master or worker nodes services. It is typical for a Hadoop cluster to have at least one edge node based on performance needs.

Purpose of Edge Nodes

Edge nodes act as an interface between a Hadoop cluster and the rest of the corporate or cloud network. Edge nodes are typically multihomed, having two or more bonded 10GbE network connectors, with one connected to the private subnet of the firewall-protected Hadoop cluster and the other connected to the corporate network. This separation maintains the security of the Hadoop cluster while allowing the edge node to serve as a controlled access point into the cluster. This limitation improves reliability and security of the cluster.

Edge nodes also act as a staging area for landing, processing, and loading into and out of Hadoop. They are particularly well-suited for loading data into Hadoop because they have high-speed connections to the cluster. Being part of the Hadoop cluster's network, edge nodes can initiate direct connections to worker nodes. To this end, Hadoop client and management command line and web-based tools that facilitate loading and processing of data are installed on edge nodes, including:

- Hive client/Hive Server 2 Server
- Sqoop
- Flume
- HttpFS
- Oozie
- Pig
- Beeline
- Ambari and Hue (management tools)

Edge nodes also contribute to enhanced performance by enabling uniform data and work distribution across a Hadoop cluster. They do this by dynamically accessing a wide and varying number of worker nodes when submitting jobs, as opposed to repeatedly using a small set of worker nodes, resulting in data skew and possibly performance issues in overworked nodes.

The distinction of Hadoop cluster nodes by type does not follow hard-and-fast rules, but rather follows best practices recommendations on how to efficiently use server resources within a cluster and optimize cluster performance. Hadoop administrators can assign Hadoop services to any node type they choose.

Pentaho on Edge Nodes

In this section, we discuss specifics of running Pentaho on edge nodes, including faster data transfer, more even resource allocation, and lower network management costs. You will also find information about security constraints for your edge node configuration.

You can find details on these topics in the following sections:

- Advantages of Running Pentaho on Edge Nodes
- Addressing Security Concerns When Running on Edge Nodes

Advantages of Running Pentaho on Edge Nodes

Install Pentaho Server on edge nodes, instead of on a regular network server, in environments where Pentaho software will be interacting with one or more Hadoop clusters.

Benefits of running Pentaho this way include:

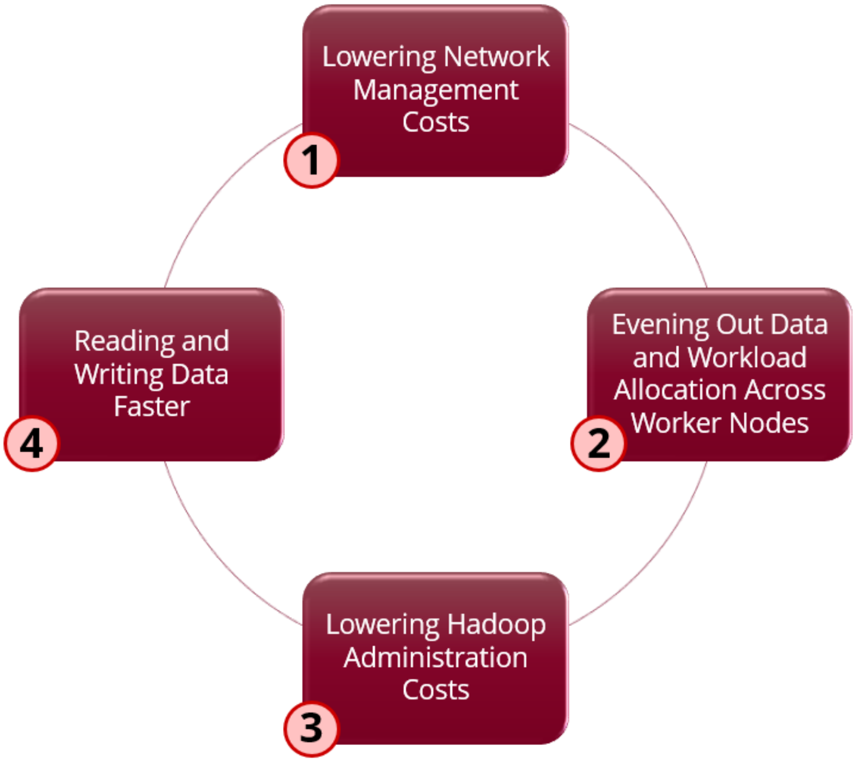
Lowering Network Management Costs
1
Evening Out Data and Workload Allocation Across Worker Nodes
2
Lowering Hadoop Administration Costs
3
Reading and Writing Data Faster
4

Figure 1: Benefits of Running Pentaho on Edge Nodes

1 You can lower your network management costs by making use of the **edge node that is already configured to access the cluster in a secured and restricted manner**.

Benefit: From the edge node, Pentaho jobs and transformations can readily access all necessary (Hadoop/Spark) services and nodes within the cluster, without needing any additional network configuration or long-term management. Edge nodes allow application users to contact worker nodes when necessary, providing a network interface for the cluster without leaving the entire cluster open to communication. As part of the Hadoop cluster, the edge node natively enforces the cluster's security settings and automatically allows access to all master and worker nodes as the cluster grows. For example, updates to Kerberos authentication settings or Hadoop configuration files (site XML files) are pushed out to all nodes including the edge nodes, reducing the overhead that would otherwise be associated with running Pentaho outside the cluster.

Reasoning: If Pentaho is installed elsewhere, it might have to route communication to *N* number of worker nodes remotely over the corporate network versus using the edge node, creating more potential points of failure or bottlenecks along the way.

② **Make data and workload allocation across worker nodes more equal by executing jobs from the edge node against any or all worker nodes within the cluster.**

Benefit: Running Pentaho jobs and transformations on the edge node allows Pentaho to communicate directly with worker nodes within the cluster without additional configuration. Running Pentaho on a noncluster server would require configuration for all worker nodes in the cluster so that processing is not repeatedly limited to a subset of workers. Using the edge node, Pentaho would then be able to dynamically use random worker nodes and avoid data or workload skewing.

Reasoning: Remember that if Pentaho were running on a non-Hadoop server, unless that server is given access to all worker nodes in the cluster, data and workload skewing would be likely to occur in the long term depending on the configuration of the workloads. For example, Spark workloads that are executed in client mode may only use workers they have access to.

③ Pentaho uses Hadoop configurations and sometimes Hadoop binaries and/or Hadoop command line utilities, so **running on the edge node incurs lower total costs of operation**.

Benefit: No extra effort needs to be spent to update the edge node (binaries or utilities) to run Pentaho jobs. It is a best practice to use certain Hadoop data movement utilities (specifically Sqoop) directly on the edge node, because the edge node has a high data transfer rate to the cluster.

Reasoning: If Pentaho were running on a standalone server, the server would have to be periodically manually refreshed to stay in sync with the Hadoop cluster.

④ **Transformations can read and write large amounts of data to and from Hadoop** if Pentaho is installed on the edge node.

Benefit: Pentaho can use the 10GbE edge node network connector to the cluster.

Reasoning: If Pentaho runs on a server other than an edge node, the network path to access Hadoop master and worker nodes might be subject to the corporate network speeds and to any bottlenecks to access the edge node that serves as a gateway to Hadoop. This could negatively impact the performance of data loads into Hadoop.

Addressing Security Concerns When Running on Edge Nodes

When you run Pentaho directly on an edge node, you may expose your edge node to security vulnerabilities. We have recommendations to alleviate some security concerns:

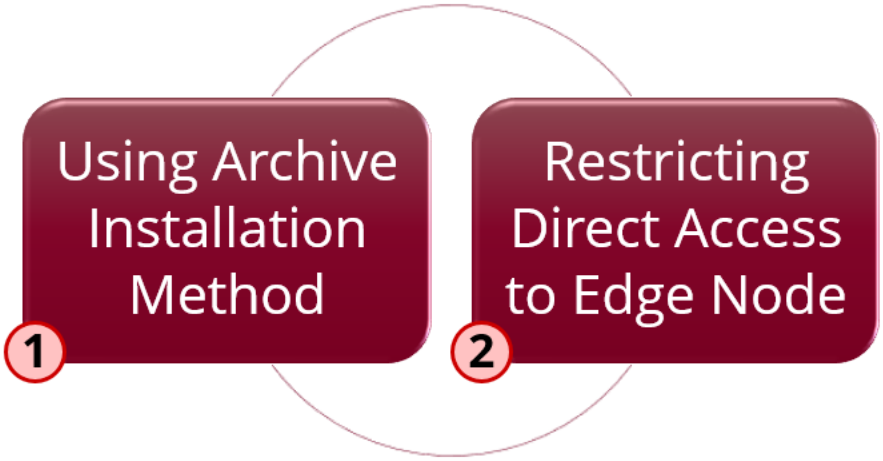
Using Archive Installation Method
1
Restricting Direct Access to Edge Node
2

Figure 2: Addressing Security Concerns for Running Pentaho on Edge Nodes

1 We recommend that you **use the archive installation method** to install Pentaho in production environments.

Benefit: This installation method lets you isolate the Pentaho application inside the preconfigured Tomcat application server, and access the Pentaho repository/PUC web application using the following URLs:

1. For unsecured traffic: `http://<host-server>:<port>/pentaho`
2. For traffic secured using an SSL certificate: `https://<host-server>:<port>/pentaho`

In these URLs, `<host-server>` and `<port>` represent the server name and port through which you access Pentaho.

Reasoning: This isolation keeps the edge node protected. This approach installs the image of an archive file on disk which includes a Pentaho web application already deployed in a Tomcat container, and the Pentaho Data Integration (PDI) engine. While the Pentaho web application serves the Pentaho User Console (PUC) for accessing and managing the Pentaho repository and other web services, the PDI engine is responsible for running PDI jobs and transformations.

If Pentaho is installed elsewhere, it might have to route communication to *N* number of worker nodes remotely over the corporate network using the edge node, creating more potential points of failure or bottlenecks along the way.

2 We recommend that you **restrict direct access to the edge node** using methods such as accessing Pentaho through an Apache web server and using load balancing.

Benefit: These methods strengthen your security protocols for your edge node.

Reasoning: Restricting direct access to the edge node puts up further barriers between your edge node and potential vulnerabilities or attacks.

Accessing Pentaho Through an Apache Web Server

Here are the details for accessing Pentaho through a web server for both a single instance of Pentaho and a cluster.

Single Instance of Pentaho

An Apache web server can be installed and configured to serve as the access point to a single instance of Pentaho. In this scenario, you route network traffic from clients through the web server instead of directly through the Tomcat web application server.

The access point for Pentaho would then be: `http(s)://<host-server>pentaho`

It is typical for web servers configured to allow all `https`/`http` traffic to use the default ports of `80` or `443`. However, if your web server is configured on other ports, you would need to include the correct port number in the URLs.

Pentaho Cluster

Try load balancing if you have two or more Pentaho instances configured to operate as a cluster. You can use Apache as a load balancer that receives and dynamically routes traffic to any number of designated Tomcat containers, either on the same physical or virtual server, or on separate servers.

You can do this by configuring the `mod_jk` clustering module in Apache Web Server to communicate over AJP1.3 protocol.

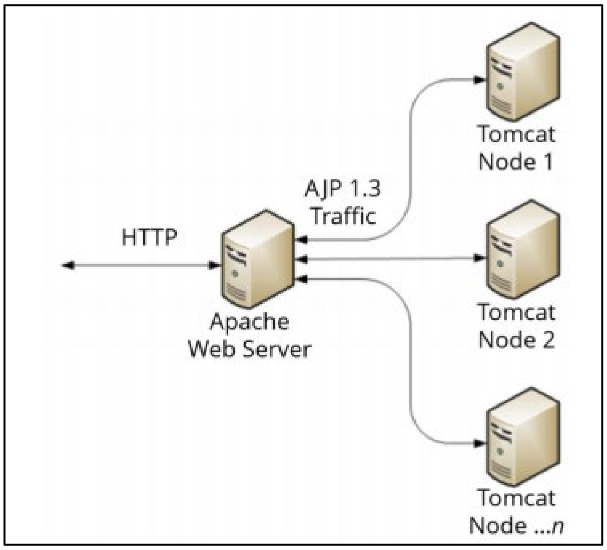

Figure 3: Apache Web Server Configuration

Remediating Security for Running Pentaho on an Edge Node

Considering that the Pentaho web application (deployed in a Tomcat container) and the PDI engine must be installed on the same machine – that is, the archive installation cannot be split across multiple servers – then one way of restricting direct access to the edge node would be to:

1. Install the web server (Apache) on a server in the demilitarized zone (DMZ) and configure it to communicate with one or more Pentaho Server instances running on the edge node.
2. Use a network load balancer as the access point for Pentaho, thereby restricting direct access to the edge node.

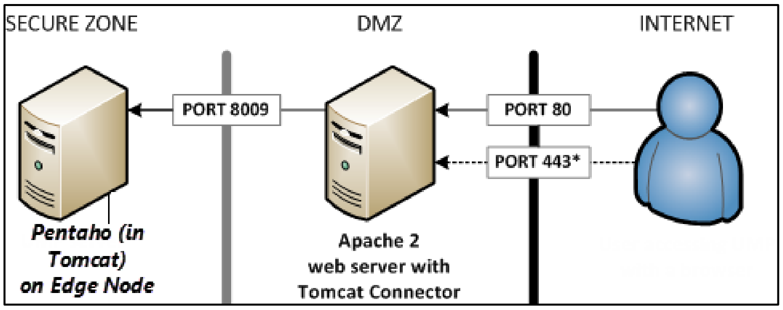
SECURE ZONE
DMZ
INTERNET
PORT 8009
PORT 80
PORT 443*
Pentaho (in Tomcat) on Edge Node
Apache 2 web server with Tomcat Connector

Figure 4: Restricting Direct Access to Edge Node

Related Information

Here are some links to information that you may find helpful:

- Pentaho Components Reference
- Installing Pentaho
- Pentaho in High Availability
- Best Practices for Installation