HITACHI
Inspire the Next

Pentaho and Microsoft Azure

HITACHI
Inspire the Next

Change log (if you want to use it):

Date	Version	Author	Changes

Contents

Overview

This document covers some best practices on installing Pentaho's server and client products on Microsoft Azure. You will learn about the server, network, and storage architecture recommended to run Pentaho in this way.

Our intended audience is Pentaho administrators or anyone with a background in infrastructure architecture or Azure who is interested in installing Pentaho on Azure.

The intention of this document is to speak about topics generally; however, these are the specific versions covered here:

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

Before beginning, use the following information to prepare for the procedures described in the main section of the document.

Terms You Should Know

Here are some terms you should be familiar with:

- **Pentaho Server**: Installed on server equipment and runs unattended, typically 24x7. It executes Pentaho jobs and stores shared definitions of data stores, reports, and so on. Users typically do not log into the server directly.
- **Pentaho Repository**: Database used by Pentaho Server to store definitions and contexts/preserve states necessary to run the server.
- **Pentaho client tools**: Installed on the client equipment like desktops or laptops, where users can interact directly with the tools, allowing users to create reports, define scheduled jobs, and perform other actions. The client tools can also interact with the server to do tasks like store report definitions or schedule background jobs.

Other Prerequisites

The primary focus of this document is the infrastructure architecture to support deployment of Pentaho's server and client products on Azure. The document does not focus on integrating Pentaho with Azure services such as Azure Active Directory, HDInsight, Azure Machine Learning, or other similar services.

Best Practices for Pentaho and Azure

The Pentaho Server, Pentaho Repository, and client tools can be deployed to Microsoft Azure. To do this in the best possible way, we recommend the following actions and settings.

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

- Pentaho Client Tools in Azure
- Virtual Network
- Virtual Machine and Storage
- Pentaho Maintenance

Pentaho Client Tools in Azure

Although there are a few different methods to set up Pentaho products with Azure, following these recommendations will ensure maximum usability and efficiency.

Table 1: Pentaho Client Tools Recommendations

Recommendation	Details
Run Pentaho Server and Pentaho Repository in Azure	If you are already using Azure, and your data sources that Pentaho will access are located in Azure: • Deploy the Pentaho Server to Azure in a virtual machine (VM) • Deploy the associated Pentaho Repository database to an Azure VM
Do *not* run Pentaho client tools in Azure	Users interact directly with the Pentaho client tools, usually with a graphical user interface. Without local installation of the Pentaho client, the user may experience poor performance caused by network latency. When the user's workstation is not in Azure (which is typically the case), then the client tools should be installed either on the user's workstation itself, or on a server that has minimal network latency to the user's workstation.

Virtual Network

We recommend you have a network architect design the logical network within which Pentaho will be hosting, spanning Azure and your own locations, accounting for scale and security to comply with your unique technology standards and business requirements.

With that said, we do have the following recommendations for your virtual network (VNet):

- Run Pentaho Products in Separate VNet
- Extend Corporate WAN to Include Azure
- Network Security Groups and Subnets

Run Pentaho Products in Separate VNet

Use a separate VNet to host the Pentaho products, one per development lifecycle stage. For example:

- **VNet1**: Pentaho Production
- **VNet2**: Pentaho Development
- **VNet3**: Your Applications

This delivers and supports:

- Separation of concerns
- Separation of duties
- Simplified firewall design

Whether you use peering or VNet to VNet connections will depend on whether the VNets in question are in the same geographical region:

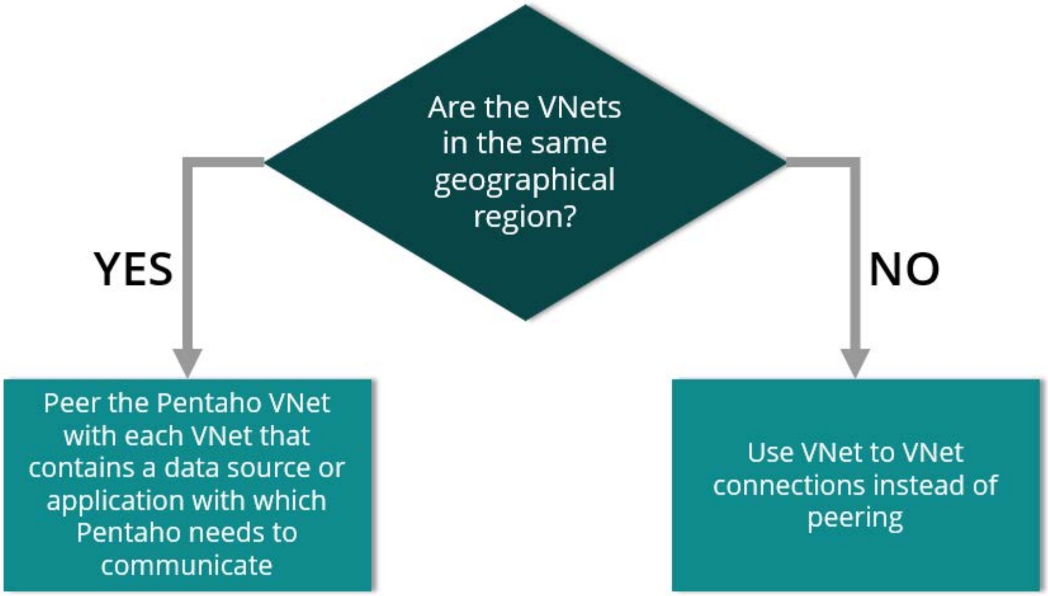
Are the VNets in the same geographical region?
YES
NO
Peer the Pentaho VNet with each VNet that contains a data source or application with which Pentaho needs to communicate
Use VNet to VNet connections instead of peering

Figure 1: Peering or VNet to VNet Connections?

Extend Corporate WAN to Include Azure

Although Pentaho traffic *can* be plainly routed over the public internet, this has the following disadvantages:

Table 2: Disadvantages to Using Public Internet

Feature	Disadvantage(s) to Using Public Internet
Service Level Agreements (SLAs)	There is no guarantee of bandwidth availability, low latency, or connectivity.
Security	Not all protocols used during data transfer are encrypted.
Complexity	All customer sites and all applicable endpoint devices must be configured to talk over the internet to Pentaho in Azure. The work involved for this configuration includes routing and firewall configuration and maintenance.
Throughput	Bandwidth and latency are limited by the internet.

Instead, if your Pentaho installation needs access to any resources residing on your organization's wide area network (WAN), we recommend creating a virtual network between the WAN and Azure.

In order to connect your organization's WAN to Azure, use one of the following two methods:

1. **Site to site virtual private network (VPN) connection**: This method addresses the **security** concern raised previously by ensuring that all traffic to and from Pentaho and your WAN is IPSec-encrypted.
2. **Express route**: This method addresses the **SLA**, **security**, and **throughput** concerns. If the express route is properly designed, it also addresses **complexity** concerns.

Network Security Groups and Subnets

For management simplicity and security transparency, use subnet Network Security Groups (NSGs) instead of VM access control lists (ACLs) to control traffic to and from the Pentaho server(s). This is also a general Azure best practice.

1. Place the Pentaho server (or servers if deployed `active/active`) in a separate NSG and subnet from the Pentaho repository.
2. Open only the minimum number of ports, depending on protocols used to communicate with your applications and data sources and from the Pentaho server to the repository database.

Which specific port numbers to open for your installation will depend on your database product.

The following diagram outlines our recommended high-level network architecture, including the recommendations from this section:

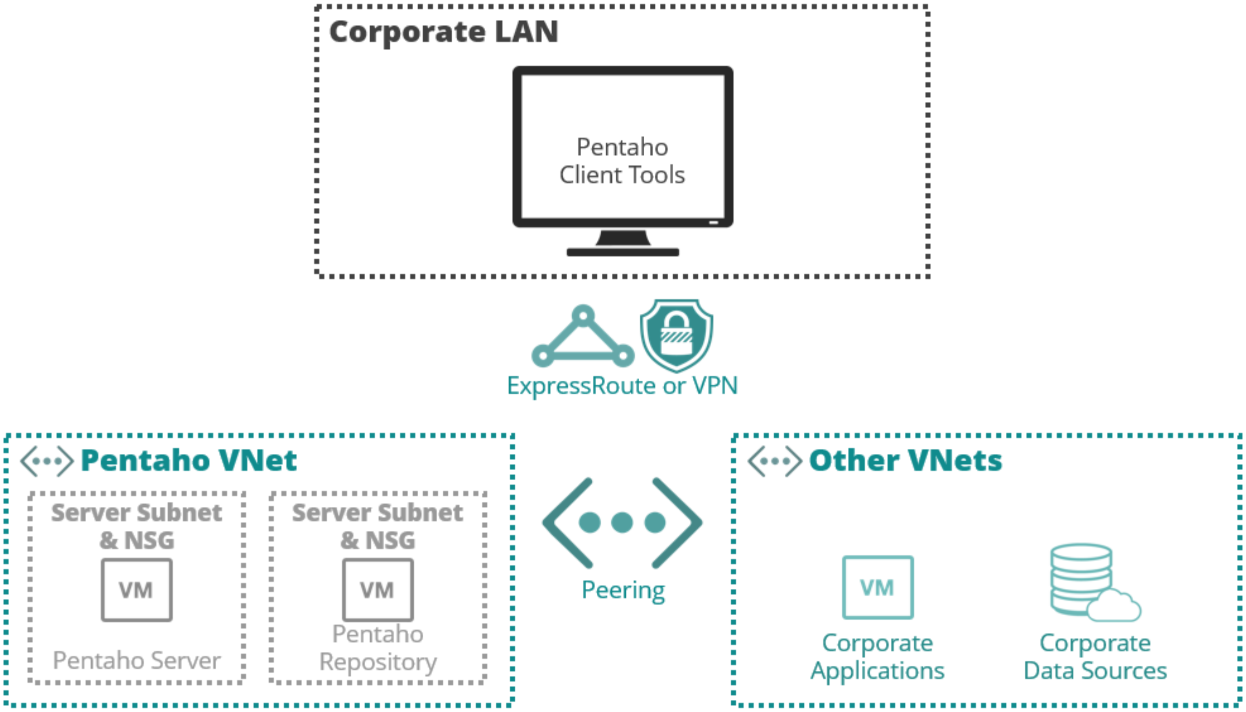

Figure 2: Network Architecture for Pentaho and Azure

Virtual Machine and Storage

We recommend that you commission an infrastructure architect to design the server and storage environments within which Pentaho will be hosted. It is critical the infrastructure be designed for scale and security to comply with your unique organizational standards, capacity, and performance requirements.

This section applies to designing the Azure resources that will host Pentaho. It does not dictate the infrastructure design, but instead highlights specific server and storage configurations and design fragments that should be incorporated into your overall infrastructure design to allow for best hosting of Pentaho.

For detailed instructions on how to install Pentaho products on these servers, and the specific operating system (OS) configurations required, please see Pentaho Installation and the Components Reference.

- Operating Systems
- Pentaho Server Specification
- Pentaho Database Product

Operating Systems

Pentaho Server and Pentaho Repository are compatible with a range of operating systems, listed in the Components Reference.

Because deploying Pentaho in Azure does not affect the validity and applicability of this list of operating systems, we recommend that you select a supported operating system that best fits with your overall infrastructure architecture.

Pentaho Server Specification

This section is applicable to the Azure virtual machines that host the Pentaho Server and the Pentaho Repository database.

Have an infrastructure architect work with Pentaho personnel to appropriately size the VMs hosting Pentaho to provide adequate performance, capacity, and availability.

Azure-specific concepts and restrictions you will need to incorporate during this process are as follows:

Azure VM Sizing

Azure VMs must not only be sized for central processing unit (CPU) capacity, but also for network and disk input/output (I/O). Each Azure VM type has a cap on the total disk and network I/O it can process.

Therefore, the virtual machine size selection must take this into account:

- The Pentaho Server must have an appropriate network total I/O cap to allow communication among the many data sources, applications, and clients.
- The Pentaho Repository server must have an appropriate disk total I/O cap to allow the repository database to be read from and written to.

Availability

Pentaho products can be deployed to provide high availability (HA) and disaster recovery (DR). Because the designs and mechanisms to achieve these are not specific to Azure, you can apply your organization's VM HA and DR standards to Pentaho servers.

We recommend Pentaho Server be deployed in an `active/passive` *(not* `active/active`*) HA configuration. Further details on this are available in Best Practices Library – Pentaho in High Availability (download the file corresponding to the version of Pentaho you are running).*

Pentaho Database Product

Supported database products and versions for the Pentaho Repository database are listed in the Components Reference. Choose a supported database to install on an Azure VM.

Do not use database as a service, including Azure structured query language (SQL). See the known issue later in this document regarding database as a service and Pentaho.

Pentaho Maintenance

These maintenance recommendations will help you keep your Pentaho installation running at peak performance.

Backup

Since the Pentaho products run on standard Azure VMs, those VMs should generally be backed up using the same mechanism as your other Azure VMs.

Similarly, back up the Pentaho database as though it were just another database running on a VM. There are no backup requirements that are specific to Pentaho on Azure.

More information on backing up Pentaho can be found at Tips for Pentaho Backup and Recovery

Monitoring

Like the backups, there are no specific additional monitoring requirements for Pentaho on Azure.

Monitor the Pentaho VMs and repository database just the same way that you monitor other VMs and databases.

Learn more about monitoring Pentaho's processes in any environment, including Azure, at Recommendations - Logging and Monitoring for Pentaho Servers.

Known Issues

The following issues apply to Pentaho products running in Azure. Account for these in your infrastructure and network architecture.

Repository Database Managed Service

Pentaho's repository database is not currently supported on Azure's managed database solution.

Workaround

Install the repository database on an Azure virtual machine conforming to the Pentaho supported configurations listed in the Components Reference.

Azure Load Balancer and Scale Sets

Pentaho Servers can run in active/passive mode, with multiple Tomcat servers running copies of Pentaho Server but only one active at any given time. However, Azure Load Balancer and Scale Sets have not been tested as balancers for this design pattern.

Workaround

If you want to run the Pentaho Server in active/active mode, use Apache Web Server to load balance incoming traffic as described in Best Practices Library – Pentaho in High Availability (download the file corresponding to the version of Pentaho you are running).

Related Information

Here are some links to information that you may find helpful while using this best practices document:

(Include background links or any links you already used in the document. Do not just put the full URL here; embed the link behind the title of whatever the article is named at the link. Generally, do not use Wikipedia, Dummies.com, etc.)

- Microsoft Azure
- Pentaho Components Reference
- Pentaho Installation
- Recommendations – Logging and Monitoring for Pentaho Servers
- Tips for Pentaho Backup and Recovery

Finalization Checklist

This checklist is designed to be added to any implemented project that uses this collection of best practices, to verify that all items have been considered and reviews have been performed.

Name of the Project:____

Date of the Review:____

Name of the Reviewer:____

Item	Response	Comments
Did you deploy the Pentaho server to Azure in a VM?	YES_______ NO_______	
Did you have a network architect design the logical network within which Pentaho will be hosted?	YES_______ NO_______	
Did you use separate VNets to host the Pentaho products?	YES_______ NO_______	
Did you create a virtual network between your corporate WAN and Azure, in case Pentaho needs to access WAN resources?	YES_______ NO_______	
Did you have an infrastructure architect to design the server and storage environments within which Pentaho will be hosted?	YES_______ NO_______	
Did you choose a supported OS that best fits your overall infrastructure architecture?	YES_______ NO_______	
Did an infrastructure architect work with Pentaho personnel to appropriately size the Pentaho-hosting VMs?	YES_______ NO_______	
Did you deploy the Pentaho server in an `active/passive` HA configuration?	YES_______ NO_______	