

Deploy the Informatica[®]Enterprise Data Preparation Solution on the Microsoft Azure Marketplace (10.5.1.1)

Abstract

The automated marketplace solution uses Azure Resource Manager to launch, configure, and run the Azure virtual machine, virtual network, and other services required to deploy a specific workload on Azure. This deployment reference provides step-by-step instructions for deploying Informatica Enterprise Data Preparation on the Microsoft Azure Marketplace.

Supported Versions

• Enterprise Data Preparation 10.5.1.1

Table of Contents

Overview
Apache Log4j RCE Vulnerability
Intended Audience
Costs and Licenses
Architecture
Azure Resources in the Deployment
Informatica Domain
Informatica Clients
Before You Begin.
License Key Prerequisite
Prerequisites
Deploying Enterprise Data Preparation on the Azure Marketplace
Step 1. Begin Provisioning
Step 2. Deploy a Domain and Configure Azure Resources
Step 3. Recycle the Informatica Services
Monitoring Instance Provision and Informatica Domain Creation
Logs
Database Account and User Details
Enterprise Data Preparation FAQ

Overview

Customers of Microsoft Azure and Informatica can execute a Enterprise Data Preparation deployment from the Azure marketplace to create an Informatica domain in the Azure and explore Enterprise Data Preparation functionality.

This deployment reference provides step-by-step instructions for deploying Enterprise Data Preparation on Microsoft Azure. Automated reference deployments use Azure Resource Manager to launch, configure, and run the Azure virtual machine, virtual network, and other services required to deploy a specific workload on Microsoft Azure.

Apache Log4j RCE Vulnerability

Informatica 10.5.1.1 addresses the Log4j RCE vulnerabilities by removing the JndiLookup class from the vulnerable Log4j libraries.

For information about the vulnerabilities, see Informatica Response to Apache Log4j RCE Zero Day Vulnerability.

Intended Audience

This guide is for users who are responsible for deploying the marketplace solution of Enterprise Data Preparation 10.5.1.1 on Microsoft Azure.

As a user with administrator privileges to deploy applications on Microsoft Azure, you must be familiar with Azure platform elements such as Azure Resource Manager, Virtual Machine, Virtual Network, Azure Databricks, Azure Functions, Azure Active Directory, Azure database, and Azure Blob storage. See the Microsoft Azure documentation.

To find Enterprise Data Preparation documentation, see the Informatica documentation portal.

Costs and Licenses

You are responsible for the cost of the Azure services used while running this deployment. There is no additional cost for using this marketplace deployment.

The Azure resource manager template for this deployment includes configuration parameters that you can customize. Some of these settings, such as instance type, will affect the cost of deployment. See the pricing pages for each Azure service that you plan to use for cost estimates.

This deployment requires a license for Informatica Enterprise Data Preparation. To sign up for a license, contact your organization's Informatica sales contact or Informatica Global Customer Support.

Note: You supply the license key value in the Informatica Enterprise Data Preparation License Key parameter when you configure the deployment.

The following table lists the instance types that you can choose based on sizing requirements:

Virtual Machine	Instance Type	Cluster Size
Database	Standard_D3_v2 / Standard_DS4_v2 / Standard_DS11 / Standard_DS11_v2 This includes the Microsoft SQL Server 2019 on Windows Server 2019 Datacenter with pay as you go (PAYG) license model. You will be charged based on the running instances. Note: For information about changing the license mode, see the <u>Microsoft documentation</u> .	Small, Medium, Large
Informatica Domain	Standard_D4_v2 / Standard_DS4_v2 / Standard_DS11 / Standard_DS11_v2 / Standard_D14_v2 / Standard_D5_v2 / Standard_DS5_v2 / This includes the Red Hat Enterprise Linux 7.9 with pay as you go model. You will be charged based on the running model.	Small, Medium, Large

Virtual Machine	Instance Type	Cluster Size
Bastion server	Standard_D2_v2 / Standard_A2_v2 / Standard_DS2_v2 / Standard_B2s / Standard_D14_v2 / Standard_D4ds_v4, Standard_D8ds_v4, Standard_D16s_v4, Standard_D32s_v4	-
	This includes the Microsoft Windows server 2019 Datacenter with pay as you go model. You will be charged based on the running instances.	
Informatica Cluster Service	Standard_D14_v2 / Standard_D5_v2 / Standard_DS5_v2 / Standard_E16_v3 Standard_D4_v2 / Standard_DS4_v2 / Standard_E16_v3	Small Medium,
	Standard_D14_v2 / Standard_DS5_v2 / Standard_D5_v2 This includes the Red Hat Enterprise Linux 7.9 with pay as you go model. You will be charged based on the running model.	Large

Architecture

The Microsoft Azure marketplace solution, when you deploy on a virtual network, creates and connects the following resources in the network:

- Informatica domain server on a virtual machine, with one additional high availability virtual machine.
- Informatica clients on a bastion server.
- Microsoft SQL Server for the repositories in the Informatica domain.
- Informatica cluster service on a virtual machine.
- An HDInsight cluster.
- A MariaDB database for the Data Preparation repository.



The following image shows the architecture of the Enterprise Data Preparation on Microsoft Azure:

The numbers in the architecture diagram correspond to items in the following list:

- 1. CIDR IP address range that you use to access the Informatica services URL and virtual machines.
- 2. A resource group on the Azure platform.
- 3. The Azure cloud.
- 4. A virtual network that includes a subnet.
- 5. A subnet to contain specific elements of the deployment.
- 6. A network security group that includes the Enterprise Data Preparation deployment.
- 7. The Informatica services on an Azure Virtual Machine.
- 8. An Informatica cluster service running on HDInsight that Enterprise Data Catalog uses to run metadata processing and profiling jobs.
- 9. A MariaDB database for the Data Preparation repository.
- 10. A compute cluster running on HDInsight that Enterprise Data Preparation uses to publish prepared data to the data lake.
- 11. A Microsoft SQL Server database to contain the following Informatica repositories:
 - Domain configuration repository.
 - Model repository.
- 12. A bastion server, if you choose to deploy one.
- 13. Informatica services for high availability on an Azure Virtual Machine.

Azure Resources in the Deployment

The deployment process creates or includes the components listed in this section.

Components in a Deployment in a New Virtual Network

The following components are created when you deploy the software:

Component	Number of Components Created
Virtual Network	One virtual network.
Network security group	One network security group.
Subnet	One subnet.
Microsoft SQL Server	One Microsoft SQL Server instance on virtual machine.
Virtual machines	 Up to nine virtual machines with the following assignments: One for Informatica domain. One for high availability. One for bastion server. One, three, or six for the Informatica cluster service.
HDInsight cluster nodes	 Two for head node. One, three, or six worker nodes.
MariaDB database	One MariaDB database.
Azure Data Lake Storage Gen2 account	One ADLS Gen2 account.
User-assigned managed identity	One managed identity.

Components in a Deployment on an Existing Virtual Network

The following components are created when you deploy the software:

Component	Number of Components Created
Network security group	One network security group.
Microsoft SQL Server	One Microsoft SQL Server instance on virtual machine.
Virtual machines	 Three or up to seven: One for Informatica domain One for high availability One for bastion server Up to six for Informatica depending on whether high availability is enabled One for bastion server
MariaDB database	One MariaDB database.

Bastion Server

You can optionally deploy a bastion server as an element in the solution. The bastion server acts as a firewall between the internet and the cloud platform network where the solution deploys. It can also act as a remote server which you can log in to run Informatica clients.

Informatica Domain

The Informatica domain is a server component that hosts application services, such as the Model Repository Service, cluster configuration for an existing configuration, cloud provisioning configuration for an autodeployed configuration, and Data Integration Service. These services, together with domain clients, enable you to create and run mappings and other objects to extract, transform, and write data.

Application Services

The Informatica domain includes the following application services:

Model Repository Service

Manages the Model repository. The Model repository stores metadata created by Informatica products in a relational database to enable collaboration among the products. Informatica Developer, the Data Integration Service, and the Administrator tool store metadata in the Model repository.

Data Integration Service

Performs data integration tasks for the Developer tool and for external clients.

Metadata Access Service

Allows the Developer tool to access cluster connection information to import and preview metadata. The Metadata Access Service contains information about the Service Principal Name (SPN) and keytab information if the Hadoop cluster uses Kerberos authentication.

Content Management Service

Manages reference data and is responsible for compiling rule specifications into mapplets. The Content Management Service provides reference data information to the Data Integration Service and to the Developer tool and Analyst tool. The Content Management Service stores reference data in a database that you specify.

Analyst Service

Manages the connection between the service components and the users who log in to Analyst tool. You can perform column and rule profiling, manage scorecards, and manage bad records and duplicate records in the Analyst tool. The Analyst Service stores profiling, scorecarding, and bad and duplicate record data in databases that you specify.

Informatica Cluster Service

Runs and manages the nodes and services associated with Enterprise Data Catalog.

Catalog Service

Runs Enterprise Data Catalog and manages connections between service components and external applications.

Enterprise Data Preparation Service

Runs the Enterprise Data Preparation application.

Interactive Data Preparation Service

Manages data preparation within the Enterprise Data Preparation application.

Domain Repositories

Informatica repositories, hosted on Microsoft SQL Server databases, store metadata about domain objects. The Informatica domain includes the following repositories:

Domain configuration repository

The domain configuration repository stores configuration metadata about the Informatica domain. It also stores user privileges and permissions.

Model repository

Stores metadata for projects and folders and their contents, including all repository objects such as mappings and workflows. For more information, see the <u>Model Repository Service</u> documentation.

Data Preparation repository

Stores worksheet metadata created when you use the Enterprise Data Preparation application to prepare data.

In addition to these domain repositories, the solution also requires a repository for Hive metadata. This repository is hosted on an SQL database. It stores Hive table metadata to enable Hadoop operations.

Application Service Databases

The Informatica domain and the application services use a series of databases to store information. You must set up the databases with the database names and user names that the domain and the services expect.

To read a list of the database names and user names that you must apply to the databases, see <u>"Database Account</u> and User Details" on page 16.

Clusters

Informatica uses the following HDInsight clusters.

Cluster Service

An Informatica cluster service that the Catalog Service uses to run metadata processing and profiling jobs.

Informatica compute cluster

A compute cluster deployed on HDInsight that Enterprise Data Preparation uses to publish prepared data to the data lake.

Informatica Clients

You can use several different clients with Enterprise Data Preparation:

Administrator tool

The Administrator tool enables you to create and administer services, connections, and other domain objects.

Analyst tool

The Analyst tool is a web-based application that enables you to analyze, profile, and create scorecards for data. You can also create and manage reference tables in the Analyst tool.

Developer tool

The Developer tool enables you to create and run mappings and other objects that enable you to access, transform, and write data to targets.

Command line interface

The command line interface offers hundreds of commands to assist in administering the Informatica domain, creating and running repository objects, administering security features, and maintaining domain repositories.

Bastion server

You can optionally deploy a bastion server as an element in the solution. The bastion server is a Windows instance installed with the Developer tool and the command line interface clients. The bastion server acts as a firewall for access to the network. It can also act as a remote Windows server that you can log in to run Informatica clients.

Before You Begin

Before you launch the automated deployment on Microsoft Azure, verify the prerequisites and make the choices described in this section.

License Key Prerequisite

Verify that you have a license to deploy Enterprise Data Preparation.

You can upload the license key file from your local system to the container in the Microsoft Azure storage account. When you configure the Enterprise Data Preparation deployment, you then need to add the license key from your Microsoft Azure storage account.

Microsoft Azure Storage Account

You can use your existing Microsoft Azure storage account or create a new account.

To create a new Microsoft Azure storage account, complete the following the steps:

- 1. Log in to the Microsoft Azure portal.
- 2. Click Create a resource.
- 3. Search for the storage account.
- 4. Click Create and enter the required parameters.
- 5. Click **Review + create**.
- 6. Navigate to Storage resource > Containers > + Container.
- 7. Enter a container name.
- 8. Select Private (no anonymous access level) for the public access level.
- 9. Click to open the created container.

Prerequisites

Before you deploy Enterprise Data Preparation on Microsoft Azure, verify the prerequisites.

- You must have a Microsoft Azure subscription with owner role.
- You must have access and permissions to create the following resources on the Azure platform:
 - Virtual network
 - Network security group
 - Virtual machines
- You have sufficient number of CPU cores based on the instance types in the region where you plan to deploy the Enterprise Data Preparation solution.

The following table lists CPU core requirements for Enterprise Data Preparation:

Deployment Type	Minimum Virtual CPU Cores*	Maximum Virtual CPU Cores**
Small	60	78
Medium	84	102
Large	132	150

*Minimum Virtual CPU Cores: Informatica High Availability and Bastion server are disabled. **Maximum Virtual CPU Cores: Informatica High Availability and Bastion server are enabled.

Note: Not all Azure resources are supported in all regions. See the Azure documentation to verify that the resources for your solution are supported in your desired region.

In addition to geographical regions, the solution supports government cloud regions. Contact Informatica Global Customer Support to check support for your desired region.

Existing HDInsight Cluster Prerequisites

Before you deploy Enterprise Data Preparation on an existing HDInsight cluster, verify the following prerequisites.

- Your password is the same for the cluster login and SSH credentials.
- The HDInsight cluster is within your deployment region.
- Use the Vnet associated with the existing HDInsight cluster.
- Verify that you have fulfilled the prerequisites for launching an HDInsight cluster. See the Informatica documentation.

Deploying Enterprise Data Preparation on the Azure Marketplace

The automated deployment of Enterprise Data Preparation on the Azure marketplace uses the Azure Resource Manager template to guide your choices and launch the solution deployment.

When you provision the Enterprise Data Preparation solution on the Azure marketplace, launch the wizard and configure the basic properties. Later, configure the solution.

Step 1. Begin Provisioning

Use the Azure Marketplace website to provision Azure cluster resources including a Enterprise Data Preparation deployment.

When you implement the Enterprise Data Preparation solution on Azure marketplace, you launch the wizard, configure basic properties.

- 1. Search for and select the Enterprise Data Preparation solution.
 - a. Log in to the <u>Azure marketplace</u> website. Use the search bar to search for Informatica Enterprise Data Preparation.
 - b. Select Informatica Enterprise Data Preparation 10.5.1.1.

Click Get it now to launch the solution wizard.

c. Read the details of the terms of use and click Continue.

The wizard redirects the browser window to the Enterprise Data Preparation 10.5.1.1 solution on the Azure portal.

d. Click Create.

A series of panels opens to enable you to configure the solution on the Azure platform.

2. Enter the information in the Basics panel, and click **OK**.

Step 2. Deploy a Domain and Configure Azure Resources

Create an Informatica domain and configure new or existing Azure resources to use with it.

You must download the latest mysql.jar, and then upload the archive as part of the deployment process. The file name must be mysql-connector-java-<versiondetails>.jar.

Basics

Enter values for the following parameters:

Parameter	Description
Subscription	Required. Azure subscription you use to manage the deployment.
Resource Group	Required. The Azure resource group containing the Virtual Network where you deploy Enterprise Data Preparation.
Region	Required. Azure location where you deploy Enterprise Data Preparation.

Informatica Enterprise Data Preparation

Enter values for the following parameters:

Parameter	Description
Informatica License Key	Required. Redirects you to the list of storage account under your subscription. Select the container that has the license file for upload.
Informatica High Availability	Indicates whether you want to enable high availability for the Enterprise Data Preparation deployment.
	Default is Disabled.
	For information about high availability for the Informatica domain, see the <u>High Availability</u> documentation.
Informatica Server	Required. Indicates the virtual machine size of the Informatica server. Default is Standard_D5_v2.
Database Server	Required. Indicates the virtual machine size of the database server.
	Default is Standard_D3_v2.
Cluster Service Virtual	Size of the virtual machine that hosts the Informatica cluster service.
Machine Size	Default is Standard_D5_v2 / Standard_D4_v2.
Password	Indicates the password for SSH, RDP, database, and database users.
Confirm Password	Confirms the password that you entered.

Existing HDInsight Cluster Settings

Enter values for the following parameters to deploy Enterprise Data Preparation on an existing HDInsight cluster:

Parameter	Description
Microsoft HDInsight Cluster URL	Required. URL for the existing HDInsight cluster. Specify the full URL including https://.
	To get the URL, complete the following steps in the Ambari Web UI: 1. Select All Resources . 2. Filter the list based on HDInsight cluster.
	3. Click on the HDInsight cluster, and then select the URL.
Azure Data Lake Storage Gen2 Account Name	Required. Azure Data Lake Storage Gen2 account to use as the primary storage type for the HDInsight cluster.
Azure Data Lake Storage Gen2 Account Resource Group	Required. Resource group to which the Gen2 storage account belongs.
Cluster login user name	Required. User name to use to log in to the HDInsight Cluster Web UI.
Password	Required. Password to use to log in to the HDInsight Cluster Web UI and SSH.
Cluster secure shell (SSH) user name	Required. User name to use to log in via SSH.

New HDInsight Cluster Settings

Enter values for the following parameters to deploy Enterprise Data Preparation on a new HDInsight cluster:

Parameter	Description
HDInsight Virtual Machine Size	Required. Size of the nodes in the HDInsight cluster. Default is Standard_D13_v2.

Bastion Server

Enter values for the following parameters:

Parameter	Description
Deploy Bastion Server	Deploys a bastion server to access other resources in the virtual network. Default is No.
Bastion server size	Virtual machine size. Applicable only when you choose to deploy the bastion server. Default is Standard_D2_v2.

Infrastructure Settings

Enter values for the following parameters:

Parameter	Description
CIDR IP Address Range	Required. Classless Inter-Domain Routing IP address range permitted to access Enterprise Data Preparation. Format the string as x.x.x/x.
Assign Public IP Address	Required. Assigns a public IP address to the network interface that is attached to the virtual machine.
Virtual Network	Required. Select to create a Virtual Network where you deploy Enterprise Data Preparation.
Subnet	Required. The identifier for the subnet within the Virtual Network where you deploy Enterprise Data Preparation.
Upload MySQL JAR	Required. Path to the mysql.jar for the MySQL database.
	The file name must be mysql-connector-java- <versiondetails>.jar.</versiondetails>
	The .jar version must be equal to or higher than 5.1.28.
	For more information about MySQL database requirements, see the <u>Informatica documentation</u> .

After you configure the parameters, verify the choices in Review + create, read the terms of use, and click Create.

When you click **Create**, Azure deploys the Enterprise Data Preparation and creates resources in the environment that you configured.

Step 3. Recycle the Informatica Services

After the deployment completes, use the Administrator tool to recycle the Informatica application services.

- 1. Log in to the Administrator tool.
- 2. Click the Services and Nodes tab.
- 3. In the Domain Navigator, select the Model Repository Service.
- 4. From the Actions menu, select Search Index > Re-index.
- 5. From the **Search Index > Re-index** menu, select **Recycle Service**.
- 6. Select the Data Integration Service, and then select **Actions > Recycle Service**.
- 7. Select the Content Management Service, and then select Actions > Recycle Service.
- 8. Select the Interactive Data Preparation Service, and then select Actions > Recycle Service.
- 9. Select the Enterprise Data Preparation Service, and then select Actions > Recycle Service.
- 10. Log in to the Enterprise Data Preparation application.
- 11. Log out of the Enterprise Data Preparation application, and then log in again.

Monitoring Instance Provision and Informatica Domain Creation

You can use cloud platform dashboards, logs, or other artifacts to see whether cluster creation succeeded and how to locate and identify the Informatica domain on the cloud platform.

During Deployment

After you finish configuring the solution and start the deployment process, the Azure dashboard indicates deployment status in the top right corner.

To view the detailed status of the deployment job, including resources, click Deployment in progress...

When Deployment is Complete

The automated deployment includes the following resources:

- Virtual network
- Network security group
- Microsoft SQL Server database
- Informatica domain
- Informatica cluster service
- HDInsight cluster
- MariaDB database

Perform the following steps to use your Azure dashboard to verify the status of resource deployment:

1. Use the dashboard search bar to search for the resource group that contains the Enterprise Data Preparation deployment.

The dashboard displays the **Overview** view of the resource group, with resource deployment status as a clickable link in the upper right corner.

- Click the resource deployment status link. When you click the deployment status link, a detail window opens listing the failed and successful deployments.
- 3. Click Error details for information about failed resource deployments.
- 4. Click **Overview** to see a list of the resources in a resource group.
- 5. You can click column headings in the display to sort by name, type, or location of the resource.

When the deployment is complete, you can open the Informatica Administrator tool in a browser. The Administrator tool URL has the following format:

https://<Public IP address_or_DNS name_or_Private IP address>:8443

User name: infauser

You can read the values from the properties of the Informatica services virtual machine in your resource group.

Output Tab

If the deployment succeeds, click the deployment named maintemplate>Outputs.

The following table describes the output properties:

Output	Description
username	User name of the Administrator tool user.
password	Password for the Administrator tool user.
databaseUsers	User IDs for the database users of the domain databases created in the deployment.
informaticaAdminConsole_PrivateIP	IP address for the node where the Administrator tool runs.
informaticaAdminConsole_DNS	DNS for the node where the Administrator tool runs.
informaticaPreparation_PrivateIP	IP address for the node where Enterprise Data Preparation runs.

Use the Administrator tool to perform administrator tasks. For a description of domain and Enterprise Data Preparation administrator tasks, see the <u>Informatica Administrator Guide</u> and the <u>Informatica Enterprise Data Preparation Administrator Guide</u>.

Logs

After the completion of the Enterprise Data Preparation deployment, consult logs to see the success or failure of solution element creation.

You can access the following logs on the virtual machine that hosts the Informatica domain:

Azure extension operation logs

Records the installation of Azure resources and services.

You can find the file in the following location:

```
/var/log/azure/custom-script/handler.log
```

Note:

Command execution log

This log records the following events:

- Creation of Informatica connections, cluster configurations, and services.
- Population of the domain and its repositories.

You can find the file in the following location:

/opt/Oneclicksolution.log

Informatica domain and services configuration log

At the top of the log file is a summary section that lists automated tasks and their status. You can view the details about each task under the summary section. If any of the tasks failed complete successfully, you can look at the detailed section for the task to troubleshoot the task.

Database Account and User Details

The Informatica domain and the application services use a series of databases to store information. Verify that database names and user account details on each database match the names and details that the databases expect.

Database Name	User	Applicable For
domaindb	domainuser	Informatica Domain
mrsdb	mrsuser	Model Repository Service
cmsdb	cmsuser	Content Management Service
pwhdb	pwduser	Data Integration Service as profiling warehouse connection
monitordb	monitoruser	Monitoring Model repository
wfhdb	wfhuser	Data Integration Service as workflow connection
analystdb	analystuser	Analyst Service
dpsdb	dpsuser	Enterprise Data Preparation Service
daadb	daauser	Data Asset Analytics

The following table describes the database and user account information:

Enterprise Data Preparation FAQ

Q. The Extension Script time out error message appears and the deployment status of Enterprise Data Preparation is shown as unsuccessful. The Informatica Administrator displays the status of the Enterprise Data Preparation application services as up and running. Is my deployment successful?

A. You can ignore the error message and unsuccessful deployment status. Enterprise Data Preparation is successfully deployed and you can start to use Enterprise Data Preparation. Microsoft Azure Marketplace displays the deployment status as unsuccessful if the deployment time exceeds a specific time limit.

Q. The virtual machine displays the following error message while running the extension script:

```
Enable failed: processing file downloads failed: failed to download file[0]: failed to download the file: http request failed: Get [REDACTED] dial tcp 13.107.246.10:443: i/o timeout
```

The deployment status of Enterprise Data Preparation is shown as unsuccessful. How do I solve this issue?

A. To solve this issue, whitelist the following outbound firewall security rules:

Ports	IP Address/Sites	Description
80 and 443	catalogartifact.azureedge.net	Required to download the extension script from an artefact location.
All	VirtualNetwork	Required to communicate with the Informatica Cluster Service nodes and database to connect to the Domain.
443	Storage	Required to download the product license.

Ports	IP Address/Sites	Description
123	0.rhel.pool.ntp.org	Required for the NTP servers such as
123	1.rhel.pool.ntp.org	Note: The IP address changes dynamically as the NTP servers are connected in the group.
123	2.rhel.pool.ntp.org	
123	3.rhel.pool.ntp.org	
1023	*.pool.ntp.org	Required to communicate with the Informatica Cluster Service nodes and the Catalog Service. Replace * with the name of the NTP Pool server closest to your location. Note: The IP address changes dynamically as the NTP servers are connected in the group.
443	13.91.47.76/32	Required for the Azure RHUI content delivery servers.
443	40.85.190.91/32	For more information, see: https://docs.microsoft.com/en-us/azure/ virtual-machines/
443	52.187.75.218/32	
443	52.174.163.213/32	
443	52.237.203.198/32	

Author

Ashwin G