Deploying Enterprise Data Catalog on Microsoft Azure Marketplace
Abstract

This deployment template provides step-by-step instructions for deploying Informatica Enterprise Data Catalog on Microsoft Azure. You can access the deployment template for Informatica Enterprise Data Catalog on Microsoft Azure from Microsoft Azure Marketplace. You can follow the instructions to automate deployments to launch, configure, and run Enterprise Data Catalog with the required compute, network, storage, and other services on a virtual machine.

Supported Versions

- Enterprise Data Catalog 10.2.2 HF1

Table of Contents

Abstract .................................................................................................................................................2
Supported Versions .............................................................................................................................2
Table of Contents ...............................................................................................................................2
Overview ..............................................................................................................................................2
Assumptions .......................................................................................................................................3
Architecture .........................................................................................................................................3
Prerequisites .......................................................................................................................................6
Costs and Licenses ..............................................................................................................................6
Deployment Steps ...............................................................................................................................7
  Accessing the Microsoft Azure Marketplace ..................................................................................7
  Step 1. Configure Basic Settings to Deploy Enterprise Data Catalog ............................................8
  Step 2. Configure Advanced Settings to Deploy Enterprise Data Catalog ....................................10
  Step 3. Configure Infrastructure Settings .......................................................................................12
  Step 4: Summary of Deployment .....................................................................................................14
Logging in to the Informatica Domain Machine .............................................................................16
Accessing Enterprise Data Catalog on Microsoft Azure Marketplace ...........................................17
Verifying and Adding Ports in the Network Security Group ..............................................................19
Tagging a Resource .............................................................................................................................21
Troubleshooting .................................................................................................................................23
Additional Information .......................................................................................................................24

Overview

Informatica Enterprise Data Catalog brings together all data assets in an enterprise and presents a comprehensive view of the data assets and data asset relationships. A data asset is a type of data object, such as a physical data source, Hadoop Distributed File System (HDFS), or big data repository. The data assets in the enterprise might exist in relational databases, purpose-built applications, reporting tools, HDFS, and other big data repositories.
Informatica Enterprise Data Catalog captures the physical and business metadata for data assets that you use to determine the effectiveness of enterprise data. Metadata is data about data. Metadata contains details about the structure of data sources. Metadata also includes information, such as data patterns, data types, relationships between columns, and relationships between multiple data sources.

Informatica Enterprise Data Catalog gathers information related to metadata across the enterprise. The metadata includes column data statistics, data domains, data object relationships, and data lineage information. A comprehensive view of enterprise metadata can help you make critical decisions on data integration, data quality, and data governance in the enterprise.

Assumptions

This document assumes that you are familiar with Microsoft Azure Marketplace, cloud-based application deployments and technologies, concepts and creation of virtual machines, and Apache Hadoop clusters.

Architecture

You can deploy Enterprise Data Catalog on a Hortonworks cluster based on the volume of data. You can select the cluster size based on your requirements.

The following are the cluster sizes you can select to deploy Enterprise Data Catalog:

- **Small.** Includes one cluster node.
- **Medium.** Includes three cluster nodes.
- **Large.** Includes six cluster nodes.

The following image shows the components and architecture for a small-sized Hortonworks cluster deployment:
The following are the components shown in the architecture:

- **Hortonworks cluster:** A single-node or multi-node, embedded cluster based on the cluster size you configure. The Informatica domain runs on a single node with all the associated services. The cluster runs on a separate node or nodes based on your selection. The application services such as the Model Repository Service and the Data Integration Service run on the Informatica domain node.

- **Database Server:** The Microsoft SQL Server database where the repositories for the Informatica domain, the Model Repository Service, the Content Management Service, and the Data Integration Service are created. Enterprise Data Catalog stores profiling data in the profiling warehouse created in the Microsoft SQL Server database.

- **Informatica Server Virtual Machine:** The virtual machine where Informatica domain and Informatica application services are deployed. An Informatica domain is a collection of nodes and services. A node is the logical representation of a machine in a domain. Services for the domain include the Service Manager that manages all domain operations and a set of Informatica application services that represent server-based functionality.

The following are the Informatica application services shown in the image:

- **MRS-Model Repository Service.** The Model Repository Service is an application service that manages the Model repository. The Model repository stores metadata created by Informatica clients and application services in a relational database to enable collaboration among the clients and services.
CMS - Content Management Service. The Content Management Service is an application service that manages reference data. A reference data object contains a set of data values that Enterprise Data Catalog searches while performing data domain discovery on source data.

DIS - Data Integration Service. The Data Integration Service is an application service that runs profiles in the Informatica domain. The Data Integration Service generates profile results for resources that you have set up to fetch profile metadata and then writes the profile results to the profiling warehouse.

AS - Analyst Service. The Analyst Service is an application service that runs the Informatica Analyst application in the Informatica domain. The Analyst Service manages the connections between service components and the users that log in to Informatica Analyst. The Analyst Service connects to a Data Integration Service, a Model Repository Service, and a Search Service.

Profiling - Helps you find the content, quality, and structure of data sources of an application, schema, or enterprise. A profile is a repository object that finds and analyzes all data irregularities across data sources in the enterprise and hidden data problems that put data projects at risk. The profiling results include unique values, null values, data domains, and data patterns.

Catalog Service. The Catalog Service is an application service that runs Enterprise Data Catalog in the Informatica domain. The Catalog Service manages the connections between service components and the users that have access to Enterprise Data Catalog search interface and Catalog Administrator.

Informatica Cluster Service. Runs and manages all the Hadoop services, Apache Ambari server, and Apache Ambari agents on an internal Hadoop cluster.

- Network Security Group: The resource where rules to access the VPN are configured.

- VPN: The virtual private network that contains the Hortonworks cluster and Informatica Server virtual machine. You can access the VPN resources using rules in the Network Security Group.

- Data sources: Represents the source databases or metadata sources that Enterprise Data Catalog scans to extract relevant metadata for further use. For on-premised data sources, Enterprise Data Catalog extracts the data through a VPN tunnel.
Prerequisites

• Make sure that you have a valid account with Microsoft Azure Marketplace. See the Microsoft Azure documentation for steps to create a Microsoft Azure Marketplace account.

• Verify that you have a valid subscription in Microsoft Azure Marketplace with the required CPU cores and memory to deploy the virtual machine.

• Verify that you have a valid license key for Informatica Enterprise Data Catalog.

• Verify that you deploy Enterprise Data Catalog in a supported region. See Step 1. Configure Basic Settings to Deploy Enterprise Data Catalog for more information about the supported regions.

• You can add more cores to support the volume of data required after you deploy Enterprise Data Catalog. See the Microsoft Azure documentation for the process and steps to follow to increase the number of cores.

Costs and Licenses

You are responsible for the cost of the Microsoft Azure services used while running Enterprise Data Catalog.

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

<table>
<thead>
<tr>
<th>Virtual Machine</th>
<th>Instance Type</th>
<th>Cluster Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database</td>
<td>Standard_D3_v2</td>
<td>Small, Medium, Large.</td>
</tr>
<tr>
<td>Informatica Domain</td>
<td>Standard_D5_v2</td>
<td>Small, Medium, Large.</td>
</tr>
<tr>
<td>Informatica Cluster Service</td>
<td>Standard_D5_v2</td>
<td>Small.</td>
</tr>
<tr>
<td></td>
<td>Standard_D4_v2</td>
<td>Medium, Large.</td>
</tr>
</tbody>
</table>

This deployment requires a license for Informatica Enterprise Data Catalog. To obtain a license, contact Informatica Global Customer Support.
Deployment Steps

Perform the following steps to deploy Enterprise Data Catalog on Microsoft Azure Marketplace:

**Accessing the Microsoft Azure Marketplace**

1. Go to [https://portal.azure.com](https://portal.azure.com) and log in with your Microsoft Azure credentials. The Microsoft Azure Dashboard appears as shown in the following image:

![Microsoft Azure Dashboard](image)

2. Click on Create a resource, type Informatica Enterprise DataCatalog 10.2.2 HF1 BYOL in search bar.

3. Click Create from the Enterprise Data Catalog 10.2.2 HF1 BYOL pane. Page appears.
Step 1. Configure Basic Settings to Deploy Enterprise Data Catalog

Perform the following steps on the Create Enterprise Data Catalog 10.2.2. HF1 BYOL page to configure the basic settings to deploy Enterprise Data Catalog:

1. Select the subscription associated with your account from the Subscription drop-down list.

2. Select any of the following options from the Resource group section:
   - **Create new.** Select this option to create a new resource group to which you want to associate the Enterprise Data Catalog deployment. Make sure that you do not provide an empty value when you create a new resource group.
   - **Use existing.** Select this option to associate the deployment with an existing resource group. You can select an existing resource group from the drop-down list.

   A resource group represents a group of resources that share the same lifecycle, policies, and permissions.

3. Select the required regional location where you want to deploy Enterprise Data Catalog, from the Location drop-down list.

   The following are the list of regions supported by Enterprise Data Catalog on Azure Marketplace:
   - Australia East
   - Australia Southeast
   - Brazil South
   - Canada Central
   - Canada East
- Central India
- Central US
- East Asia
- East US
- East US 2
- France Central
- Japan East
- Japan West
- Korea Central
- Korea South
- North Central US
- North Europe
- South Africa North
- South Central US
- South India
- Southeast Asia
- UK South
- UK West
- West Central US
- West Europe
- West India
- West US
- West US 2

4. Click OK. The Informatica Enterprise Data Catalog Configure Settings page appears.
1. Specify the name of the virtual machine where you want to install the Informatica domain, the Model Repository Service, and the Data Integration Service in the **Informatica Domain Server VM** box.

   **Note:** To select an Informatica domain server with a higher configuration, click **Change size** under **Informatica Domain Server Machine Size**. For more information about the sizes of the domain server, see the **Costs and Licenses** section.
2. Specify the name of the virtual machine where you want to deploy the Hortonworks cluster in the Informatica Cluster Service VM box.

3. Select one of the following options from the Enterprise Data Catalog Deployment Size drop-down list to specify the size of the cluster required based on the volume of the data:
   - Small. Includes one cluster node.
   - Medium. Includes three cluster nodes.
   - Large. Includes six cluster nodes.
   
   **Note:** To select an Informatica Cluster Service machine with a higher configuration, click Change size under Informatica Cluster Service Machine Size. For more information about the sizes of the Informatica Cluster Service machine, see the Costs and Licenses section.

4. Specify the username that you want to use to access the virtual machines that host the Informatica domain and the Hortonworks cluster in the VM Username box.

5. Specify the username to access Informatica Administrator in the Informatica Domain Username box.

6. Specify the password for the username in the Password box.

7. Confirm the password in the Confirm Password box.
   
   **Note:** Enterprise Data Catalog uses the same password to access the virtual machines that host the Informatica domain, the Hortonworks cluster, and to log in to Informatica Administrator.

8. Click in the License Key section to browse and select a valid license key for Enterprise Data Catalog.

9. Click OK. The Infrastructure Settings page appears.
**Step 3. Configure Infrastructure Settings**

Perform the following steps on the **Infrastructure Settings** page to configure the infrastructure settings to deploy Enterprise Data Catalog:

<table>
<thead>
<tr>
<th>1</th>
<th>Basics</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Informatica Enterprise Data</td>
</tr>
<tr>
<td>3</td>
<td>Infrastructure Settings</td>
</tr>
<tr>
<td></td>
<td>Configure Database, Storage, and...</td>
</tr>
<tr>
<td>4</td>
<td>Summary</td>
</tr>
<tr>
<td></td>
<td>Informatica Enterprise Data Catalog</td>
</tr>
<tr>
<td>5</td>
<td>Buy</td>
</tr>
</tbody>
</table>

### Infrastructure Settings

- **Database Machine Name**
- **Database Machine size**
  - 1x Standard D3 v2
  - 4 vcpus, 14 GB memory
  - [Change size](#)
- **Informatica Database Instance Username**
  - `dbuser`
- **Informatica Database Instance Password**
- **Confirm Informatica Database Instance Password**
- **Virtual network**
  - (new) InfraVNET
  - [Configure subnets](#)
  - [CIDR IP Address Range](#)

### Steps

1. Provide the name of the machine on which Microsoft SQL Server database runs in the **Database Machine Name** box.  
   **Note:** To select database machine with a higher configuration, click **Change size** under **Database Machine size**. For more information about the sizes of the database machine, see the [Costs and Licenses](#) section.
2. Provide the database user name to be created for the Informatica domain and services in the **Informatica Database Instance Username** text box.

3. Confirm the password in the **Confirm Informatica Database Instance Password** box.

4. Provide the password for the database user name in the **Informatica Database Instance Password** text box.
   
   **Note:** To select an Informatica Cluster Service machine with a higher configuration, click **Change size** under **Database Machine size**. For more information about the sizes of the domain server, see the **Costs and Licenses** section.

5. Click **Virtual network**. The **Choose virtual network** pane appears with the list of virtual networks associated with the subscription and location you specified.

6. Select the required virtual network or click **Create new** to create a new virtual network.
   
   Configure the following parameters if you want to create a new virtual network:
   
   • **Name**. Specify a unique name for the new virtual network.
   
   • **Address space**. Specify the range of IP addresses for the virtual network.

7. Click **OK**.

8. Click **Subnets** to specify the subnets that must include all resources such as cluster components and virtual machines. The **Subnets** pane appears.

9. Configure the following properties for the Subnets:
   
   • **Subnet name**. Specify a unique name for the subnet.
   
   • **Subnet address prefix**. Specify a range of IP addresses for the subnets.
   
   • Click **OK**.

10. Specify the Classless Inter-Domain Routing (CIDR) IP address range for which you want to permit access to the Informatica Server, in the **CIDR IP Address Range** text box. You can specify a range of IP addresses as shown in the following sample if you want to specify permissions to IP addresses within the range 10.0.0.0 through 10.0.0.255: 10.0.0.0/24.

11. Select **Yes** from the **Import Sample Content** drop-down list if you want to import the sample content bundled with Enterprise Data Catalog. You can use the sample content to get started with Enterprise Data Catalog.

12. Click **OK**. The **Summary** page appears.
Step 4: Summary of Deployment

The Summary page displays a summary of your selection as shown in the following image:

The template validates the properties and values that you provided for correctness. If any of the properties is incorrect and the validation fails, you cannot proceed further without correcting the property values.

Click **OK** to deploy Enterprise Data Catalog on Microsoft Azure Marketplace. The deployment progress appears on the top right corner of your screen.

The template deploys the components in the following order:

1. VPN, network security group, Microsoft SQL Server database.
2. Enterprise Data Catalog on an embedded Hortonworks cluster.
3. Informatica Server.
Logging in to the Informatica Domain Machine

Perform the following steps if you want to log in to virtual machine that hosts Informatica domain:

1. Login to Microsoft Azure Marketplace with valid credentials.
2. Enter the name of the resource group in the search bar and press Enter as shown in the following image:

   ![Resource Group Search](image)

3. Click the resource group that you created. The list of resources in the resource group appears as shown in the following image:

   ![Resource Group List](image)

4. Click the name of the virtual machine that you provided for the Informatica domain when you created the resource as shown in the following image:

   ![Virtual Machine Details](image)
5. Click Connect from the menu. The Connect to virtual machine dialog box appears as shown in the following image:

6. Select Public IP address from the IP address drop-down list on the Connect to virtual machine dialog box.

7. Copy the content listed after ssh in the Login using VM local account box. For example, if the box lists the content as ssh sshuser@<Public IP Address of Informatica Domain>, copy sshuser@<Public IP Address of Informatica Domain>.

8. Install and run PuTTY or any terminal to connect to a Linux virtual machine.

9. Provide the copied content as the hostname to connect to the Linux virtual machine. The Linux virtual machine login interface appears.

10. Enter the password that you had provided when you created the resource group.

11. Run the sudo su root command and press Enter.

12. Enter the password that you had provided when you created the resource group.

Accessing Enterprise Data Catalog on Microsoft Azure Marketplace

Before you access Enterprise Data Catalog on Microsoft Azure Marketplace, complete the following step:

- To access the virtual machine instances on Microsoft Azure Marketplace from your Microsoft Windows machine, make sure that you add the following lines in the hosts file:

  - To access the Enterprise Data Catalog applications, enter the following line: <IP address of the Informatica domain machine> <hostname of the machine that hosts the Informatica domain>

  Note: To copy the IP address, perform the following steps:
1. Search for the resource group that you created in Microsoft Azure Marketplace.

2. Select the virtual machine that you configured for the Informatica domain.

3. Copy the IP address that appears.

   - To access the profiling log files, enter the following line: `<IP address of the machine that hosts the Informatica Cluster Service> <hostname of the machine where Informatica Cluster Service runs>

   **Note:** To copy the IP address, perform the following steps:
   
   1. Search for the resource group that you created in Microsoft Azure Marketplace.

   2. Select the virtual machine that you configured for the Informatica Cluster Service.

   3. Copy the IP address that appears.

**Perform the following steps to access Enterprise Data Catalog:**

1. Click **Resource Groups** on the left panel of the **Microsoft Azure Dashboard** page. The **Resource Groups** page appears with the list of resource groups configured as shown in the following image:

![Resource Groups](image-url)

2. Click the resource group you created.

   Alternatively, to search for the required resource group you created, you can type the name of the resource group in the **Filter by name**... text box. The resource
3. From the resource group, click the virtual machine that you configured for the Informatica domain.
   **Note:** 6008 is the default port number to access Informatica Administrator.
5. Login to Informatica Administrator.
6. Perform the following steps to view the URLs to access the Catalog Administrator and the Enterprise Data Catalog Search tool:
   a. After logging in to Informatica Administrator, click the **Services and Nodes** tab. The Informatica domain and the list of application services configured on the domain appear.
   b. Click the Catalog Service to view the URLs to access Enterprise Data Catalog applications.

**Verifying and Adding Ports in the Network Security Group**

Perform the following steps if you want to verify and add ports in the Network Security Group:

1. Login to Microsoft Azure Marketplace with valid credentials.
2. Enter the name of the resource group in the search bar and press Enter.
3. Click the resource group that you created. The list of resources in the resource group appears.
4. Click the Network Security Group from the list of resources for which you want to verify or add ports. The **Network Security Group Overview** page appears as shown in the
5. Select **Inbound security rules** under the **Settings** section on the left panel. The list of configured inbound security rules appear with the configured port, the source, and destination as shown in the following image:
6. If you want to add a port, select a rule for which you want to add a port and click **Add**. The **Add inbound security rule** dialog box appears as shown in the following image:

![Add inbound security rule dialog box](image)

7. Specify the required port numbers and other details in the dialog box and click **Add**.
8. Refresh the **Network Security Group Overview** page to see the changes.

### Tagging a Resource

Tagging a resource with meaningful tags helps in searching for resources using the tags configured.

Perform the following steps if you want to tag a resource in a resource group that you created:

1. Login to Microsoft Azure Marketplace with valid credentials.
2. Enter the name of the resource group in the search bar and press **Enter**.
3. Click the resource group that you created. The list of resources in the resource group appears as shown in the following image:

4. Select the resource that you want to tag as shown in the following image:
5. Click **Tags** from the left panel as shown in the following image:

![Image of Tags panel]

6. Add the required tags and click **Save** as shown in the following image:

![Image of Add Tags panel]

7. Click the **Overview** page for the resource group to see the tags.
Troubleshooting

To troubleshoot issues related to the deployment, you can access the log files from the following locations:

- Informatica Server deployment log files- /var/lib/waagent/custom-script/download/0/ directory on the virtual machine where you installed Informatica Server.
- Informatica application services log files- /home/Informatica/log/node01/service/
- Deployment log file- /home/Informatica/Oneclicksolution_result.log
- Location where the Informatica domain is installed- /home/Informatica
- Location where the installer binaries are located if you want to start a new installation-/opt/Informatica

Additional Information

- See Informatica Enterprise Data Catalog Community for more information about Enterprise Data Catalog.

Authors

Informatica Marketplace R&D Team