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

Table of Contents

Abstract .................................................................................................................................................. 2
Supported Versions .............................................................................................................................. 2
Table of Contents ............................................................................................................................... 2
Overview ............................................................................................................................................... 2
Assumptions .......................................................................................................................................... 3
Architecture .......................................................................................................................................... 3
Prerequisites .......................................................................................................................................... 6
Costs and Licenses ............................................................................................................................... 7
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 .................................................................................................... 13
Accessing Enterprise Data Catalog on Microsoft Azure Marketplace .................................................. 14
Troubleshooting ..................................................................................................................................... 16
Additional Information ........................................................................................................................ 16

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, Azure HDInsight, cloud-based application deployments and technologies, concepts and creation of virtual machines, and Apache Hadoop clusters.

**Architecture**

You can deploy Enterprise Data Catalog on an Azure HDInsight cluster based on the volume of data. You can select the Azure HDInsight cluster size based on your requirements.

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

- Minimum. Includes two cluster head nodes and one cluster worker node.
- Small. Includes two cluster head nodes and three cluster worker nodes.
- Medium. Includes two cluster head nodes and five cluster worker nodes.
- Large. Includes two cluster head nodes and seven cluster worker nodes.

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

- **Azure Storage WASB**: The Azure cloud storage known as WASB storage or catalog where the metadata extracted by the resources are stored in HDFS. The catalog is an indexed inventory of all the data assets in the enterprise. The catalog stores all the metadata extracted from data sources. Enterprise Data catalog uses scanners to extract metadata from data sources. A scanner is a pluggable component of Enterprise Data Catalog that extracts metadata from a specific data source. For example, to extract data from an Oracle data source, Enterprise Data catalog uses an Oracle scanner. A resource is a catalog object that represents a data source from where scanners extract metadata. You configure resources in Catalog Administrator. The metadata stored in the catalog is encrypted using 256-bit Advanced Encryption Standard (AES) encryption. The catalog includes the metadata persistence store, search index, and graph database.

- **Azure HDInsight 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. The cluster includes the following components:
- Head node virtual machines and worker node virtual machines for load balancing.
- Internet gateway head node that can be accessed using a public IP address.

- SQL Server Database Service: The SQL Server as service where the repositories for the Informatica domain, the Model Repository Service, the Content Management Service, and the Data Integration Service are created. The access to the SQL Server Database service is through the firewall rules set at the time of creation of the service. Enterprise Data Catalog stores profiling data in the profiling warehouse created in the 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.
o 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.

o 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.

o 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 Azure HDInsight 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-premise 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 for the deployment. 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 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 Microsoft Azure service you will be using for cost estimates.

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 a new virtual machine:

**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 **Virtual machines** from the options on the left panel.
The Virtual machines page appears as shown in the following image:

3. Click Add.

The Compute pane appears as shown in the following image:

5. Click Create from the Enterprise Data Catalog pane. The Create Enterprise Data Catalog page appears.

Note: Select Resource Manager from the Select a deployment model drop-down list to create a new virtual machine. If you want to connect to an existing virtual network, select the required virtual network from the drop-down list.

Step 1. Configure Basic Settings to Deploy Enterprise Data Catalog

Perform the following steps on the Create Enterprise Data Catalog 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:

- Canada Central
- Central US
- East Asia
- East US
- East US2
- Japan East
- Korea South
- North Central US
4. Click OK. The Informatica Enterprise Data Catalog Configure Settings page appears.

**Step 2. Configure Advanced Settings to Deploy Enterprise Data Catalog**

Perform the following steps on the Informatica Enterprise Data Catalog Configure Settings page to configure the advanced settings to deploy Enterprise Data Catalog:

1. Specify the password for Informatica Administrator in the Informatica Administrator Password text box.

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

3. Select one of the following options from the Deployment Size drop-down list to specify the size of the cluster required based on the volume of the data:
   - **Minimum**. Includes two cluster head nodes and one cluster worker node.
   - **Small**. Includes two cluster head nodes and three cluster worker nodes.
   - **Medium**. Includes two cluster head nodes and five cluster worker nodes.
   - **Large**. Includes two cluster head nodes and seven cluster worker nodes.
Based on the deployment size you select, the deployment sizing parameters change.

**Note:** For medium and large deployment sizes, you might need to increase the number of cores. See the Microsoft Azure documentation for the process and steps to follow to increase the number of cores.

4. Click **Informatica Server Instance Type**.

The Choose a Size pane appears as shown in the following image with the Recommended tab open by default:

![Choose a Size pane](image)

5. Select the required size for the Informatica Server instance and click **Select**.

Note: The **Recommended** tab displays the recommended sizes for the Informatica Server instance based on your selections for the previous properties. You can click **View all** to view all the available options from which you can select the required size for Informatica Server instance.

6. Click **Cluster Head Node Instance Type**. The Choose a Size pane appears.

7. Select the required cluster gateway node instance type and click **Select**. The cluster master services such as the YARN resource manager and the HDFS name node run on the cluster gateway node.

Note: The **Recommended** tab displays the recommended sizes for the cluster gateway node instance based on your selections for the previous properties. You can click **View all** to view all the available options from which you can select the required size for the cluster head node instance.

8. Click **Cluster Worker Node Instance Type**. The Choose a Size pane appears.

9. Select the required cluster worker node instance type and click **Select**. The cluster slave services such as the YARN node manager and the HDFS data node run on the cluster worker node.

Note: The **Recommended** tab displays the recommended sizes for the cluster worker node instance based on your selections for the previous properties. You can click **View all** to view all the available options from which you can select the required size for the cluster worker node instance.

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.

The default value * allows access to the Informatica Server from all public IP addresses.

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 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:

1. Provide the database user name to be created for the Informatica domain and services in the Informatica Database Instance Username text box.
2. Provide the password for the database user name in the Informatica Database Instance Password text box.
3. Click Storage account. The Choose storage account pane appears with the list of storage accounts associated with the subscription and location you specified.
4. Select the required storage account or click Create new to create a new storage account.
   - Configure the following parameters if you want to create a new storage account:
     - Name. Specify a unique name for the new storage account.
     - Performance. Select Standard or Premium. Standard performance storage accounts provide backup in magnetic drives and are cost
effective. Premium performance storage accounts provide backup in SSD drives and are recommended for I/O-intensive applications and consistent low-latency performance.

- **Replication.** Choose the data replication strategy for data high availability that you want to use for the storage account. Select any of the following options:
  - Locally-redundant storage (LRS).
  - Geo-redundant storage (GRS).
  - Read-access geo-redundant storage (RA-GRS).

See the Microsoft Azure documentation for more information about data replication strategies.

- Click OK.

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.

10. 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. Azure storage WASB, VPN, network security group, SQL Server Service, and databases.
3. Informatica Server.

**Accessing Enterprise Data Catalog on Microsoft Azure Marketplace**

After you complete the deployment of Enterprise Data Catalog on Microsoft Azure, perform the following steps to access Enterprise Data Catalog:

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

   ![Resource Groups page](image)

   1. 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. Click the virtual machine configured for the resource group. The virtual machine page appears as shown in the following image:

4. Copy the DNS name from the page and substitute the <host> parameter in the following URLs with the DNS name:
   - http://<host>:<port>/ldmadmin to access Catalog Administrator.
   - http://<host>:<port>/ldmcatalog to access Enterprise Data Catalog.
   - http://<host>:<port> to access the Informatica Administrator.

   **Note:**
   - <host> represents the host name of the Informatica Enterprise Data Catalog virtual machine that you created.
   - <port> represents the port configured for the Catalog Service. Default is 6705 for both Catalog Administrator and Enterprise Data Catalog.
   - 6007 is the default port number to access Informatica Administrator.
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 Server installation Silent installer log files- /opt/Informatica/10.2.0/installed/Informatica_10.2.0_Services_*.log directory on the virtual machine where you installed Informatica Server.
- Informatica services deployment log files- /var/lib/waagent/custom-script/download/0/installation.log directory in the virtual machine where you installed Informatica Server.

Additional Information

- See Informatica Enterprise Data Catalog Community for more information about Enterprise Data Catalog.
- See the following Enterprise Data Catalog Guides:
  - Enterprise Data Catalog Tutorial
  - Catalog Administrator Guide
  - Enterprise Information Catalog User Guide

Authors

Suraj Jayan
Lead Technical Writer

Acknowledgements

The author would like to acknowledge Prem Krishna, Robin Sharma, and Abdullah Rao for their contributions to this article.