How-To Library

Informatica

Deploy the Informatica®10.5.3 Data Engineering Streaming Solution on the Microsoft Azure Marketplace

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 Data Engineering Streaming on the Microsoft Azure Marketplace.

Supported Versions

- Data Engineering Streaming 10.5.3

Table of Contents

Overview

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

This deployment reference provides step-by-step instructions for deploying Data Engineering Streaming 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.

Intended Audience

This guide is for users who are responsible for deploying the marketplace solution of Data Engineering Streaming 10.5.3 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 Data Engineering Streaming 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 Data Engineering Streaming. 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 Data Engineering Streaming 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
Database	Standard_D4s_v3 / Standard_D8s_v3 / Standard_D16s_v3 / Standard_E16_v3 / Standard_D4as_v4 / Standard_D16as_v4 / Standard_D16ds_v4 / Standard_D8s_v4 / Standard_D32as_v4 / Standard_E16as_v4 / Standard_E16ds_v4 / Standard_D16ds_v5 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 Microsoft documentation.
Informatica domain	Standard_D4s_v3 / Standard_D8s_v3 / Standard_D16s_v3 / Standard_E16_v3 / Standard_D4as_v4 / Standard_D16as_v4 / Standard_D16ds_v4 / Standard_D8s_v4 / Standard_D32as_v4 / Standard_E16as_v4 / Standard_E16ds_v4 / Standard_D16ds_v5 This includes Red Hat Enterprise Linux 8.3 with pay as you go model. You will be charged based on the running model.
Bastion server Optional	Standard_D4s_v3 / Standard_D8s_v3 / Standard_D16s_v3 / Standard_E16_v3 / Standard_D4as_v4 / Standard_D16as_v4 / Standard_D16ds_v4 / Standard_D8s_v4 / Standard_D32as_v4 / Standard_E16as_v4 / Standard_E16ds_v4 / Standard_D16ds_v5 This includes the Microsoft Windows Server 2019 Datacenter with pay as you go model. You will be charged based on the running instances.

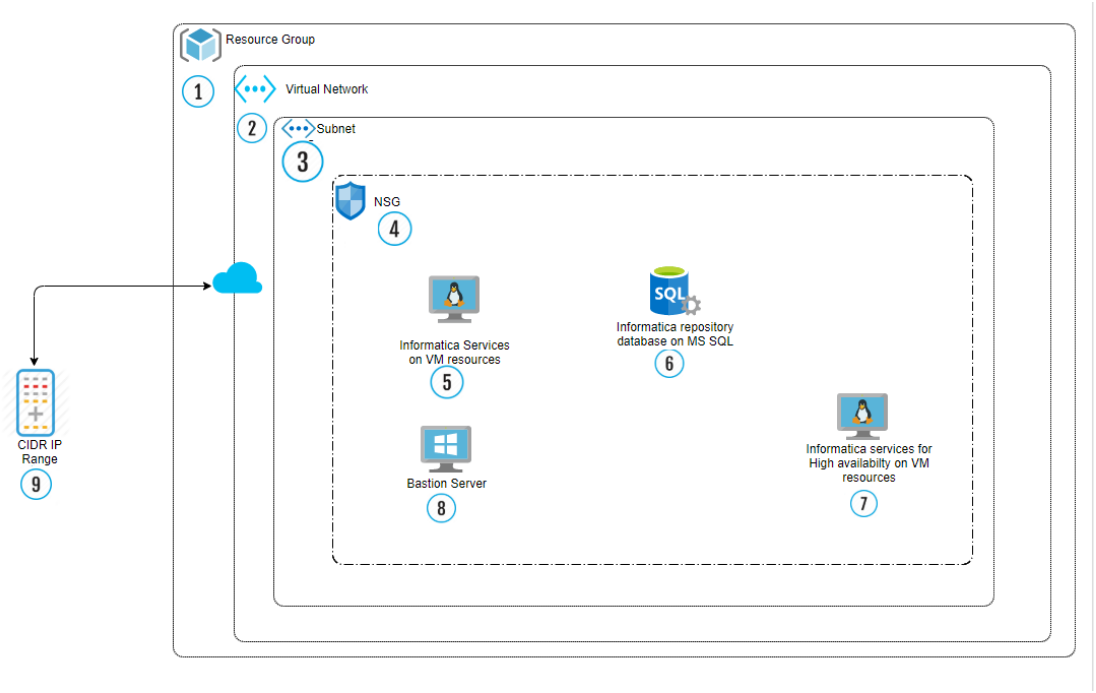
Architecture

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.
- Databricks cluster or Databricks workspace.

The following image shows the architecture of the Data Engineering Streaming on Microsoft Azure:



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

1. A resource group on the Azure platform.
2. A virtual network that includes a subnet.
3. A subnet to contain specific elements of the deployment.
4. A network security group that includes the Data Engineering Streaming deployment.
5. The Informatica services on Azure Virtual Machine.
6. Microsoft SQL Server database instance to act as Informatica domain repositories:
   - Domain configuration repository
   - Model repository
7. Informatica services for high availability on Azure Virtual Machine.
8. Bastion server, if you choose to deploy one.
9. CIDR IP address range that you use to access the Informatica services URL and virtual machines.

After the completion of the automated deployment, you can create a mapping that connects to data sources and targets in existing Azure storage. To do this, manually create the necessary connections to Azure resources. Then you can create mappings and run them on the Databricks runtime engine. For more information about creating connections and configuring access to Azure storage resources, see the Connections Reference.

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	Three virtual machines: - One for Informatica domain. - One for high availability. - One for bastion server.

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 virtual machines: - One for Informatica domain. - One for high availability. - One for bastion server.

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.

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.

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 Model Repository Service documentation.

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 "Database Account and User Details" on page 12.

Informatica Clients

You can use several different clients with Data Engineering Streaming:

Administrator tool

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

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 Data Engineering Streaming.

You can upload the license key file from your local system to the container in the Microsoft Azure storage account. When you configure the Data Engineering Streaming 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 Data Engineering Streaming 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
  - Databricks environment:
    - Create a Databricks workspace. To create a workspace, follow the guidelines in the Azure documentation.
    - Generate a Databricks token ID for authentication, or use an existing token.
- You have a Contributor or higher role.
- You have sufficient number of CPU cores based on the instance types in the region where you plan to deploy the Data Engineering Streaming solution.

The following table lists supported regions for Data Engineering Streaming with Databricks:

Americas	Europe	Asia and Oceania
- Brazil South - Canada central - Canada East - Central US - East US - East US 2 - North Central US - South Central US - West Central US - West US - West US 2	- North Europe - UK South - UK West - West Europe	- Australia East - Australia Southeast - Australia Central 2 - Central India - East Asia - Japan East - Japan West - Korea central - Korea South - South India - Southeast Asia

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.

Deploying Data Engineering Streaming on the Azure Marketplace

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

When you provision the Data Engineering Streaming 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 Data Engineering Streaming deployment.

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

1. Search for and select the Data Engineering Streaming solution.
   a. Log in to the Azure marketplace website. Use the search bar to search for Informatica Data Engineering Streaming.
   b. Select **Informatica Data Engineering Streaming 10.5.3.**

      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 Data Engineering Streaming 10.5.3 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.

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 Data Engineering Streaming.
Region	Required. Azure location where you deploy Data Engineering Streaming.

Informatica Data Engineering Streaming

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 Data Engineering Streaming deployment. Default is Disabled. For information about high availability for the Informatica domain, see the High Availability documentation.
Informatica Server	Required. Indicates the virtual machine size of the Informatica domain. Default is Standard_D16ds_v4.
Database Server	Required. Indicates the virtual machine size of the database server. Default is Standard_D8s_v3.
Password	Indicates the password for SSH, RDP, database, and database users.
Confirm Password	Confirms the password that you entered.

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_D4s_v3.

Databricks Configuration

Enter values for the following parameters:

Parameter	Description
Options	Choose how to integrate the solution. Choose from the following options: - Existing - Autodeploy - Skip Default is Skip.

Depending on your choice of an existing or autodeployed Databricks cluster, enter values for an existing or autodeploy configuration.

Existing Configuration

To use an existing Databricks configuration, enter values for the following parameters:

Parameter	Description
Cluster ID	Required. Cluster ID of the existing Databricks cluster.
Token ID	Required. Databricks token ID. Required for authentication.
Domain URL	Required. URL for Databricks workspace access. **Note:** Use domain URL without `https://`.

Autodeploy Configuration

You can use a cluster workflow to create an ephemeral cluster to run jobs. An ephemeral cluster is also called an auto-deployed cluster. For more information about cluster workflows, see the Informatica documentation.
To use an autodeployed Databricks cluster, enter values for the following parameters:

Parameter	Description
Token ID	Required. Databricks token ID. Required for authentication.
Domain URL	Required. URL for Databricks workspace access. **Note:** Use domain URL without `https://`.

You can use cluster workflows with Data Engineering Streaming 10.4.0 solution on Azure marketplace. For more information, see Cluster Workflows.

Network Configuration

Enter values for the following parameters:

Parameter	Description
CIDR IP Address Range	Required. The CIDR public IP range of clients that are permitted to access the Informatica Data Engineering Streaming. Format is x.x.x.x/x.
Assign Public IP	Assigns a public IP address to the network interface that is attached to the virtual machine. Default is Yes.
Virtual Network	Required. The identifier for the Azure virtual network where you want to deploy Data Engineering Streaming. **Note:** The deployment supports new and existing virtual networks. The Azure location must be same for the virtual network resource group and the deployment resource group.
Subnet	Required. The identifier for the subnet within the virtual network where Data Engineering Streaming is deployed.

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 Data Engineering Streaming and creates resources in the environment that you configured.

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

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 Data Engineering Streaming deployment.
   The dashboard displays the **Overview** view of the resource group, with resource deployment status as a clickable link in the upper right corner.
2. 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.

Logs

After the completion of the Data Engineering Streaming 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: The directory path `/var/lib/waagent/custom-script/download/0` contains the `stdout` and `stderr` logs. The directory also contains the file `convert.sh`, which contains the script that was executed to install Azure resources and services.

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.

The following table describes the database and user account information:

Database Name	User	Applicable For
domaindb	domainuser	Informatica Domain
mrsdb	mrsuser	Model Repository Service

Database Name	User	Applicable For
pwhdb	pwduser	Data Integration Service as profiling warehouse connection
monitordb	monitoruser	Monitoring Model repository
wfhdb	wfhuser	Data Integration Service as workflow connection

Data Engineering Streaming FAQ

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

A. You can ignore the error message and unsuccessful deployment status. Data Engineering Streaming is successfully deployed and you can start to use Data Engineering Streaming. 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 Data Engineering Streaming 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 database and domain.
443	Storage	Required to download the product license.
443	`13.91.47.76/32`	Required for the Azure RHUI content delivery servers. For more information, see: https://docs.microsoft.com/en-us/azure/virtual-machines/workloads/redhat/redhat-rhui
443	`40.85.190.91/32`	
443	`52.187.75.218/32`	
443	`52.174.163.213/32`	
443	`52.237.203.198/32`	

Author

Ashwin G