Configuring Big Data Management in the Amazon EMR Cloud Environment

Abstract

You can enable Informatica Big Data Management for Amazon EMR in the Amazon cloud environment. When you create an implementation of Big Data Management in the Amazon cloud, you bring online virtual machines where you install and run Big Data Management. Then you use Informatica Developer (the Developer tool) to design and implement mappings for big data solutions in the Amazon cloud. This article shows you how to provision Amazon resources and create an instance of Big Data Management in the Amazon cloud environment, then download the Developer tool.

Supported Versions

- Informatica Big Data Management 10.1

Table of Contents

Overview 2
Prerequisites. 2
Provision Amazon EMR Cluster Resources and the Informatica Domain. 3
Monitor Instance Provision and Informatica Domain Creation. 7
Download and Install Informatica Developer. 8
Next Steps. 9

Overview

You can choose to enable Informatica Big Data Management for Amazon EMR in the Amazon cloud environment. When you create an implementation of Big Data Management in the Amazon cloud, you bring online virtual machines where you install and run Big Data Management.

First, verify prerequisites. Then perform the following steps to create an implementation of Big Data Management in the Amazon cloud:

1. Provision Amazon EMR cluster resources and the Informatica domain.
2. Monitor creation of cluster resources and the Informatica domain.
3. Download and install Informatica Developer.

After you perform these steps, you are ready to use Big Data Management.

Prerequisites

Before you configure Big Data Management in the Amazon EMR cloud environment, verify the following prerequisites:

- You have an account with Amazon Web Services (AWS), with the account login information available.
- You have purchased a license for Informatica Big Data Management and have uploaded the Big Data Management license file to an Amazon S3 bucket.
  The license file has a name like `BDMLicense.key`.
- You have configured an Amazon PEM key to use for authentication during setup.
  When you log in to any EC2 system or EMR cluster, you use a password file for authentication. The file is called a PEM key and has a file suffix of .pem.

If you do not have an existing PEM key to use, perform these steps to get this file:

1. Log in to the EC2 system.
2. In the EC2 dashboard, select **Network & Security** > **Key Pairs**.
3. Click **Create Key Pair** to create a .pem file, or select any of the key pairs in the account.

   Note: Your administrator might ask you to use a particular existing key pair.

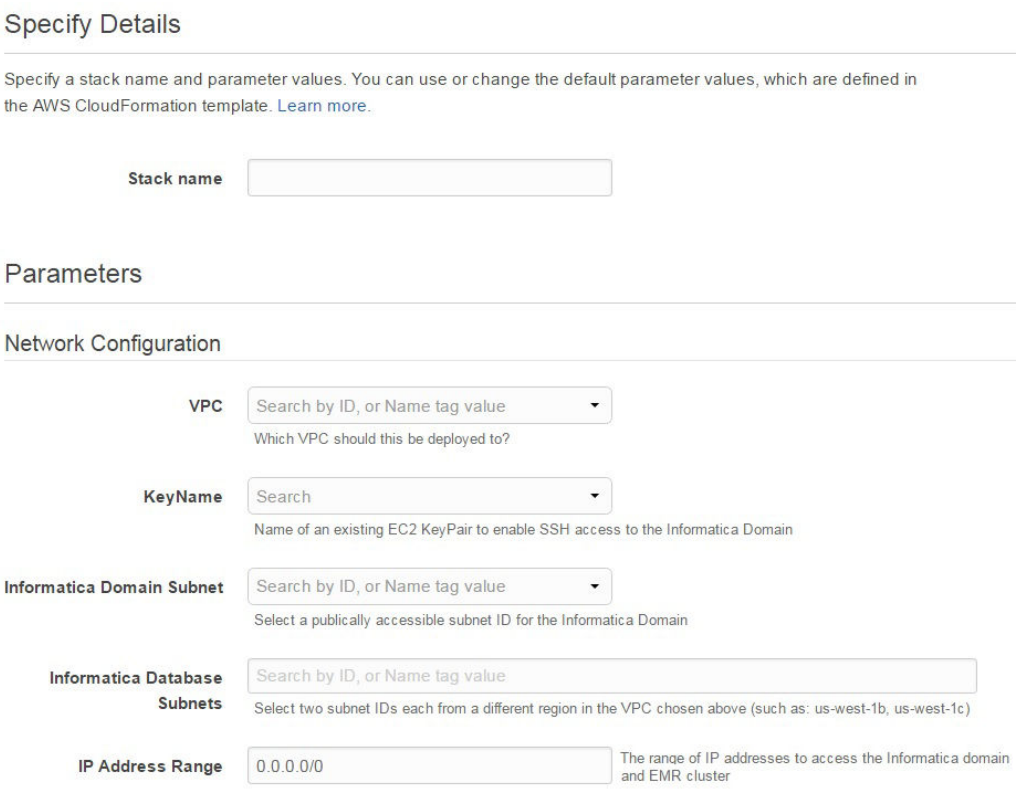
   When you create a key pair, you save the .pem file to your desktop system. Simultaneously, the EC2 system saves the key pair to your account. Make a note of the key pair that you want to use for the Big Data Management instance, so that you can provide the key pair name during network configuration.

Provision Amazon EMR Cluster Resources and the Informatica Domain

You can use the AWS marketplace to provision cluster resources and install Big Data Management in the cluster.

1. Go to the Amazon AWS marketplace (https://aws.amazon.com/marketplace).
2. Search for and select **Informatica Big Data Management 10.1**, and then click **Continue**.

   The **Create Stack** screen opens. The following image shows part of the **Create Stack** screen:



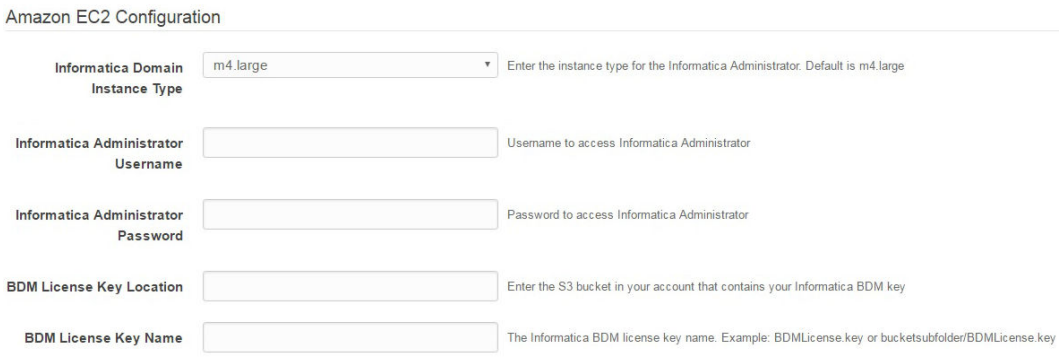
3. In the **Stack name** field, type the name of the stack instance to create.

4. In the **Parameters** section, enter the following information in the **Network Configuration** area:

Property	Description
VPC	Select the Virtual Private Cloud (VPC) location to install Big Data Management. The VPC is a provisioned computing instance on Amazon's AWS cloud. Amazon AWS provides one or more VPC with each account. Each VPC has a range of IP addresses. The VPC must meet the following requirements: - Set up with public access through the internet via an attached internet gateway. - The DNS Resolution property of the VPC must be set to Yes. - The Edit DNS Hostnames property of the VPC must be set to Yes.
KeyName	Select an existing EC2 KeyPair name to enable SSH access for Informatica services to the EC2 instance. This might be the key pair that you created in the Prerequisites section.
Informatica Domain Subnet	Select a publically accessible subnet for the Informatica domain.
Informatica Database Subnets	Specify the IDs of three different subnets. One of these subnets must meet the following requirements: - Must be a member of the VPC. - Must have access to the internet via gateway. The Informatica domain must use this subnet. The other two subnets must meet the following requirements: - Must be members of the VPC. - Must be in two separate regions, to enable database failover. - May be private or public. It is not necessary to choose subnets in the domain subnet.
IP Address Range	IP address range to use to limit SSH access from the Informatica domain to the EC2 instance. For example, to specify the range of 10.20.30.40 to 10.20.30.49, enter the following string: `10.20.30.40/49`

5. In this step and the following steps, you enter information that the Informatica installer needs to create the Informatica domain and the domain repository database.

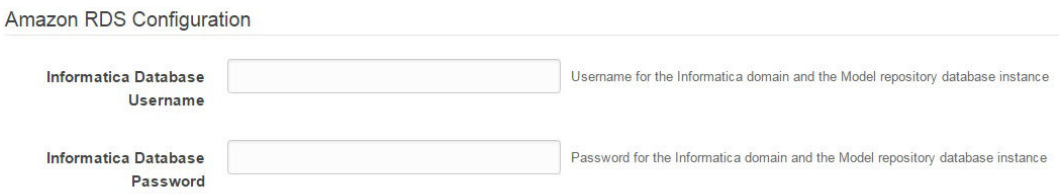
   The following image shows the **Amazon EC2 Configuration** area:



Enter the following information in the **Amazon EC2 Configuration** area:

Property	Description
Informatica Domain Instance Type	Select the type for the instance to host the Informatica domain. Each type corresponds to a different size, in ascending order of size. Default is m4.large. **Note:** When you select an instance type here and in later steps, be aware that Amazon charges more when you select a larger instance type.
Informatica Administrator User Name	Enter the administrator user name for Big Data Management. In this field and the following field, you can specify any user name and password. Make a note of the user name and password, and use it later to log in to the Administrator tool to configure the Informatica domain.
Informatica Administrator Password	Enter the administrator password for Big Data Management.
BDM License Key Location	Enter the location of the Big Data Management license key file. The location is the name of the S3 bucket where the key was saved. For example: `myBucketName`
BDM License Key Name	Enter the path and filename of the BDM license key file in the S3 bucket location. The path must include subdirectories *under* the bucket name. For example, where the entire path including the bucket name is `myBucketName/SubDir1/SubDir2/BDMLicense.key`, type the following: `SubDir1/SubDir2/BDMLicense.key`

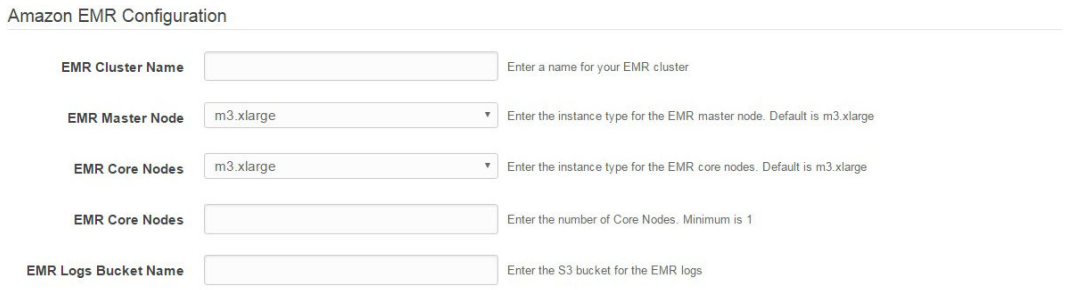
6. The following image shows the **Amazon RDS Configuration** area:



Enter the following information in the **Amazon RDS Configuration** area:

Property	Description
Informatica Database Username	User name for the domain and the Model repository database. In this field and the following field, you can specify any user name and password.
Informatica Database Password	Password for the domain and the Model repository database.

7. The following image shows the **Amazon EMR Configuration** area:



Enter the following information in the **Amazon EMR Configuration** area:

Property	Description
EMR Cluster Name	Enter a name for the Amazon EMR cluster where the BDM instance will be deployed.
EMR Master Node Instance Type	Select the instance type. Each type corresponds to a different size, in ascending order of size. Default is m3.xlarge.
EMR Core Nodes [instance type]	Select the instance type. Each type corresponds to a different size, in ascending order of size. Default is m4.large.
EMR Core Nodes [number of core nodes]	Enter an integer for the number of core nodes to support the BDM deployment. Minimum is 1. Maximum is 500.
EMR Logs Bucket Name	Enter the name of the S3 bucket where EMR stores logs.

8. Click **Next**.

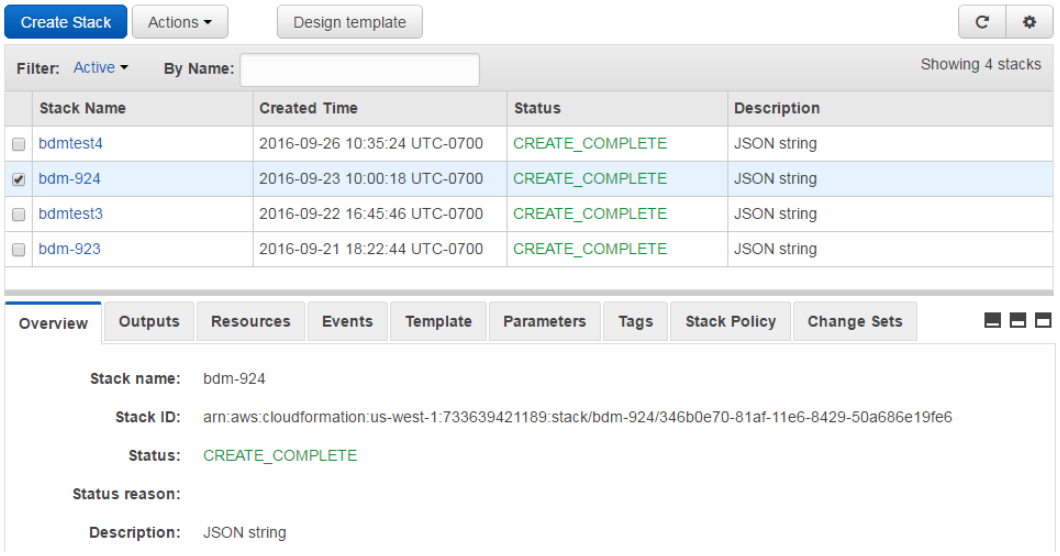
   The **Options** page opens.

9. Optionally enter key labels and values for resources, and then click **Next**.
10. Review the configuration, then click the check box to acknowledge that you are about to create resources.

11. Click **Create**.

    Amazon AWS begins to create the stack. Amazon AWS displays the **Cloud Formation** dashboard.

    The following image shows the **Cloud Formation** dashboard:



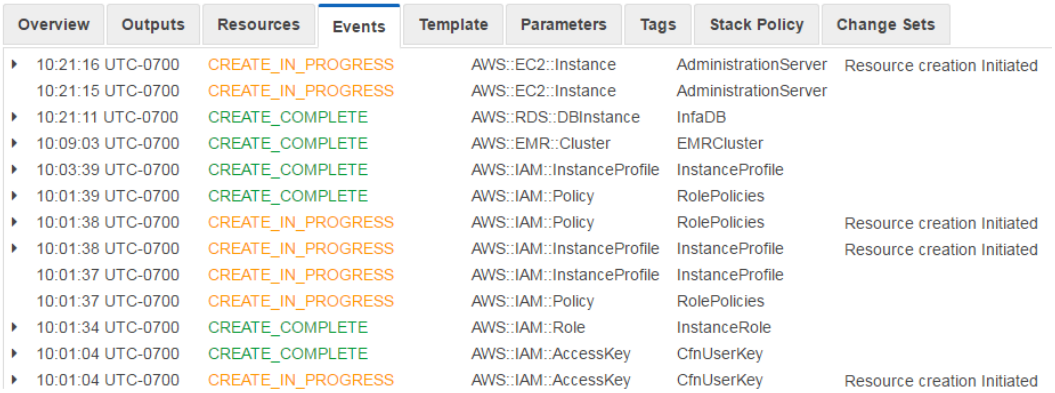
Monitor Instance Provision and Informatica Domain Creation

You can monitor creation of the cluster instance and the Informatica domain, and get more information about system resources.

1. While Amazon AWS creates the stack, you can monitor the process.

   Select the stack that you are creating, then select the **Events** tab to monitor the creation of the stack.

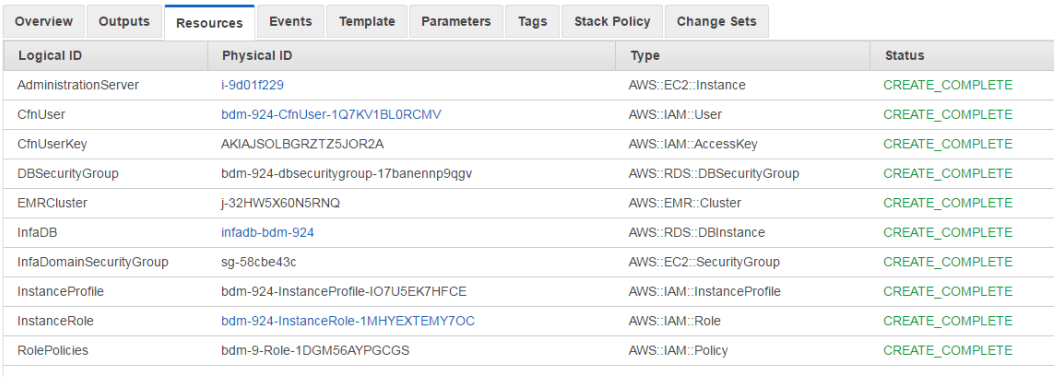
   The following image shows part of the **Events** tab:



2. When stack creation is complete, select the **Resources** tab.

   The **Resources** tab displays information about the stack and Big Data Management instance. You can select the linked Physical ID properties of individual resources to get more information about them.

The following image shows the **Resources** tab:



Logical ID	Physical ID	Type	Status
AdministrationServer	i-9d01f229	AWS::EC2::Instance	CREATE_COMPLETE
CfnUser	bdm-924-CfnUser-1Q7KV1BL0RCMV	AWS::IAM::User	CREATE_COMPLETE
CfnUserKey	AKIAJSOLBGRZTZ5JOR2A	AWS::IAM::AccessKey	CREATE_COMPLETE
DBSecurityGroup	bdm-924-dbsecuritygroup-17banennp9qgv	AWS::RDS::DBSecurityGroup	CREATE_COMPLETE
EMRCluster	j-32HW5X60N5RNQ	AWS::EMR::Cluster	CREATE_COMPLETE
InfaDB	infadb-bdm-924	AWS::RDS::DBInstance	CREATE_COMPLETE
InfaDomainSecurityGroup	sg-58cbe43c	AWS::EC2::SecurityGroup	CREATE_COMPLETE
InstanceProfile	bdm-924-InstanceProfile-IO7U5EK7HFCE	AWS::IAM::InstanceProfile	CREATE_COMPLETE
InstanceRole	bdm-924-InstanceRole-1MHYEXTEMY7OC	AWS::IAM::Role	CREATE_COMPLETE
RolePolicies	bdm-9-Role-1DGM56AYPGCGS	AWS::IAM::Policy	CREATE_COMPLETE

3. Click the **Outputs** tab.

   When the Informatica domain setup is complete, the Outputs tab displays the following information:

Property	Description
CloudFormationLogs	Amazon EMS creates the cloud formation logs during the creation of the stack. When stack creation is complete, you can use the logs to verify the successful completion of the installation.
InstanceID	Name of the Informatica domain host.
InformaticaHadoopInstallLogs	Location on the master node of the EMR cluster of the log that records the creation of the Hadoop instance.
InformaticaAdminConsoleURL	URL of the Informatica Administrator. Use the Administrator tool to administer Informatica services.
InformaticaAdminConsoleServerLogs	Location of the Informatica domain installation logs.
EMRMasterNodeHadoopURL	URL of the EMR resource manager and master node.
InformaticaBDMDeveloperClient	Location where you can download the Developer tool client.

   Note: If the **Outputs** tab is not populated with this information, wait for domain setup to be complete.

4. Open the **Resources** tab. Click the **Physical ID** of the AdministrationServer property. The **Physical ID** corresponds to the name of the Informatica domain.

   The **Instance Administration** screen opens.

   You can use the **Instance Administration** screen to launch the Big Data Management instance. You can also get information like the Public DNS and Public IP address.

Download and Install Informatica Developer

Informatica Developer (the Developer tool) is an application that you use to design and implement data integration, data quality, data profiling, data services, and big data solutions. You can use the Developer tool to import metadata,

create connections, and create data objects. You can also use the Developer tool to create and run profiles, mappings, and workflows.

1. On the Amazon AWS console, browse to **Service** > **Cloud Formation**.

   The **Cloud Formation** dashboard opens.
2. Select the **Outputs** tab.
3. Right-click the value of the **InformaticaBDMDeveloperClient** key to download the Developer tool client installer.
4. Uncompress and launch the installer to install the Developer tool on a local drive.

Next Steps

Now that you have provisioned cluster resources and installed the Informatica domain and the Developer tool, read Big Data Management documentation to see how to use Big Data Management.

Access the following documentation on the Informatica Network:

Big Data Management User Guide

Read this guide to see how to use the Developer tool to create mappings that access and process data in your Big Data Management instance.

Big Data Management Security Guide

Read this guide to see how to use security features of Big Data Management.

The Informatica Network also has searchable Knowledge Base articles and other helpful information.

Author

Mark Pritchard
Principal Technical Writer