Informatica
Big Data Management
on the AWS Cloud

Quick Start Reference Deployment

November 2016

Andrew McIntyre, Informatica Big Data Management Team
Santiago Cardenas, AWS Quick Start Reference Team

Contents
Overview..........................................................................................................................2
Costs and Licenses..............................................................................................................3
Architecture.......................................................................................................................3
Informatica Services on AWS.........................................................................................4
Leveraging Amazon EMR for Informatica Big Data Management .............................6
Deploying Big Data Management on AWS.................................................................9
Deployment Steps .........................................................................................................10
  Step 1. Prepare an AWS Account................................................................................10
  Step 2. Launch the Quick Start..................................................................................11
  Step 3. Monitor Instance Provision and Informatica Domain Creation ..............17
  Step 4. Download and Install Informatica Developer ...........................................20
Troubleshooting............................................................................................................20
Additional Resources....................................................................................................21
Send Us Feedback.........................................................................................................21

This Quick Start deployment guide was created by Amazon Web Services (AWS) in partnership with Informatica.
Overview

This Quick Start reference deployment guide provides step-by-step instructions for deploying Informatica Big Data Management on the Amazon Web Services (AWS) Cloud. Quick Starts are automated reference deployments that use AWS CloudFormation templates to launch, configure, and run the AWS compute, network, storage, and other services required to deploy a specific workload on AWS.

Informatica Big Data Management enables your organization to process large, diverse, and fast changing datasets so you can get insights into your data. Use Big Data Management to perform big data integration and transformation without writing or maintaining Apache Hadoop code.

Use Big Data Management to collect diverse data faster, build business logic in a visual environment, and eliminate hand-coding to get insights on your data. Consider implementing a big data project in the following situations:

- The volume of the data that you want to process is greater than 10 terabytes.
- You need to analyze or capture data changes in microseconds.
- The data sources are varied and range from unstructured text to social media data.

You can identify big data sources and perform profiling to determine the quality of the data. You can build the business logic for the data and push this logic to the Hadoop cluster for faster and more efficient processing.

You can view the status of the big data processing jobs and view how the big data queries are performing. You can use multiple product tools and clients such as Informatica Developer (the Developer tool) and Informatica Administrator (the Administrator tool) to access big data functionality. Big Data Management connects to third-party applications such as the Hadoop Distributed File System (HDFS) and NoSQL databases such as HBase on a Hadoop cluster on different Hadoop distributions.

The Developer tool includes the native and Hadoop run-time environments for optimal processing. Use the native run-time environment to process data that is less than 10 terabytes. In the native environment, the Data Integration Service processes the data. The Hadoop run-time environment can optimize mapping performance and can process data that is greater than 10 terabytes. In the Hadoop environment, the Data Integration Service pushes the processing to nodes in a Hadoop cluster.
This Quick Start is for users who deploy and develop Big Data Management solutions on the AWS Cloud.

Costs and Licenses

You are responsible for the cost of the AWS services used while running this Quick Start reference deployment. There is no additional cost for using the Quick Start.

The AWS CloudFormation template for this Quick Start 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 AWS service you will be using or the AWS Simple Monthly Calculator for cost estimates.

This Quick Start requires a license for Informatica Big Data Management, as described in the Prerequisites section. To sign up for a demo license, please contact Informatica.

Architecture

Figure 1 shows how Big Data Management fits in an overall big data solution.

![Figure 1: Big Data Management in a big data solution](image-url)
Big Data Management delivers a comprehensive big data solution to natively ingest, integrate, clean, govern, and secure big data workloads in Hadoop. Big Data Management is the process of integrating, governing and securing data assets on a big data platform. It has three components:

- **Big data integration** is the collection of data from various disparate data sources, such as Salesforce, Marketo, Adobe, and Enterprise Data Warehouse, which is ingested, transformed, parsed, and stored on a Hadoop cluster to provide a unified view of the data. Big data integration plays an important role as a big data management core capability as it allows data engineers to extract data from various marketing SaaS applications, apply business logic as defined by data analysts, and load the data to a big data store such as Hadoop.

- **Big data governance** is the strategy of managing and controlling data, while quality is the ability to provide cleansed and trusted datasets that can be consumed or analyzed by an intelligent data application or a big data analytics tools. Big data governance is steadily emerging as a need for Big Data Management as data stewards must manage and maintain the ever-increasing variety of data, specifically customer data, by breaking down the taxonomy of each marketing data type at a granular level.

- **Big data security** is concerned with securing sensitive data across the big data solution. As part of the big data strategy, the team must discover, identify, and ensure any customer data stored in weblogs, CRM applications, internal databases, and third-party applications are secure based on defined data security policies and practices. The team needs full control and visibility into any data in the data lake and monitor for unusual behavior or non-compliance. Additionally, sensitive customer data must be masked to prevent unauthorized users from accessing sensitive information.

**Informatica Services on AWS**

Deploying this Quick Start with default parameters builds the Informatica Big Data Management environment illustrated in Figure 2 in the AWS Cloud. The Quick Start deployment automatically creates the following Informatica elements:

- Domain
- Model Repository Service
- Data Integration Service

The deployment then assigns the connection to the Amazon EMR cluster for the Hadoop Distributed File System (HDFS) and Hive.
The Informatica domain and repository database are hosted on Amazon Relational Database System (Amazon RDS) using Microsoft SQL Server, which handles management tasks such as backups, patch management, and replication.

To access Informatica Services on the AWS Cloud, you can install the Informatica client on a Microsoft Windows machine.

The following figure shows Informatica Services running on AWS:

![Informatica Big Data Management on AWS](image)

**Figure 2. Informatica Big Data Management on AWS**

The Quick Start configures the following information services during the one-click deployment:

- **Informatica domain** is the fundamental administrative unit of the Informatica Platform. The Informatica Platform has a service-oriented architecture that provides the ability to scale services and share resources across multiple machines. High availability helps minimize service downtime due to unexpected failures or scheduled maintenance in the Informatica environment.

- **Informatica node** is a logical representation of a physical or virtual machine in the Informatica domain. Each node in the Informatica domain runs application services, such as the Model Repository and the Data Integration Service.

- **Informatica Model Repository Service** manages the Model Repository, which is a relational database that stores all the metadata for projects created using Informatica
Client tools. The Model Repository also stores run-time and configuration information for applications that are deployed to a Data Integration Service.

- **Informatica Data Integration Service** is a compute component within the Informatica domain that manages requests to submit big data integration, big data quality, and profiling jobs to the Hadoop cluster for processing.

- **Informatica Client**, specifically the Informatica Developer tool, allows data engineers to design and implement big data integration, big data quality, and a profiling solution that execute on the Hadoop cluster.

The Informatica domain and the Informatica Model Repository databases are configured on Amazon RDS using Microsoft SQL Server.

**Leveraging Amazon EMR for Informatica Big Data Management**

**Common Amazon EMR Architecture Patterns for Informatica Big Data Management**

Amazon EMR provides two methods to configure a cluster: transient and persistent. Transient clusters are shut down when the jobs are complete. For example, if a batch-processing job pulls web logs from Amazon Simple Storage Service (Amazon S3) and processes the data once a day, it is more cost-effective to use transient clusters to process web log data and shut down the nodes when the processing is complete. Persistent clusters continue to run after data processing is complete. An infrastructure architect needs to consider which configuration method works best for the organization’s use case as both have advantages and disadvantages. For more information, please refer to the [Amazon EMR best practices whitepaper](#). Informatica Big Data Management supports both cluster types.

Informatica Big Data Management supports two patterns to move data to an Amazon EMR cluster for processing.

**Pattern 1: Using Amazon S3**

In this first pattern, data is loaded to Amazon S3 using Informatica Big Data Management and PowerExchange for Amazon S3 connectivity. For data processing, Informatica Big Data Management mapping logic pulls data from Amazon S3 and sends it for processing to Amazon EMR. Amazon EMR does not copy the data to the local disk or HDFS. Instead, the mappers open multithreaded HTTP connections to Amazon S3, pull data to the Amazon EMR cluster, and process data in streams, as illustrated in Figure 3.
Pattern 2: Using HDFS and Amazon S3 as Backup Storage

In this pattern, Informatica Big Data Management writes data directly to HDFS and leverages the Amazon EMR task nodes to process the data and periodically copy data to Amazon S3 as the backup storage. The advantage of this pattern is the ability to process data without copying it to Amazon EMR. Even though this method may improve performance, the disadvantage is durability. Since Amazon EMR uses ephemeral disk to store data, data could be lost if the Amazon EMR EC2 instance fails. HDFS replicates data within the Amazon EMR cluster and can usually recover from node failures. However, data loss could still occur if the number of lost nodes is greater than your replication factor. We recommend that you back up HDFS data to Amazon S3 periodically.
Process Flow

Figure 5 shows the process flow for using Informatica Big Data Management on Amazon EMR as it relates to this reference architecture. It illustrates the data flow process using Informatica Big Data Management and Amazon EMR, Amazon S3, and Amazon Redshift. The process follows Pattern 1, discussed in the earlier section. The numbers in the diagram identify six fundamental processes that an enterprise architect must consider when architecting a Big Data Management solution.

The numbers in Figure 5 refer to the following steps:

1. Offload infrequently used data and batch load raw data onto a defined landing zone in an Amazon S3 bucket. Marketing data stored on a data warehouse application server can be offloaded to a dedicated area, freeing space in the current Enterprise Data Warehouse. Instead of feeding data from source systems into the warehouse, raw transactional and multi-structured data is loaded directly onto Amazon S3, further reducing impact on the warehouse.

2. Collect and stream real-time machine and sensor data. Data generated by machines and sensors, including application and web log files, can be collected in real time and streamed directly into Amazon S3 instead of being staged in a temporary file system or the data warehouse.
3. Discover and profile data stored on Amazon S3. Data can be profiled to better understand its structure and context, and requirements can be added for enterprise accountability, control, and governance for compliance with corporate and governmental regulations and business SLAs.

4. Parse and prepare data from web logs, application server logs, or sensor data. Typically, these data types are either in multi-structured or unstructured formats that can be parsed to extract features and entities, and data quality techniques can be applied. Prebuilt transformations and data quality and matching rules can be executed natively in Amazon EMR, preparing data for analysis.

5. After data has been cleansed and transformed using Amazon EMR, high-value curated data can be moved from Amazon EMR to an Amazon S3 output bucket or to Amazon Redshift, where data is directly accessible by BI reports, applications, and users.

6. Additionally, high-value curated data stored in Amazon S3 can be migrated to the on-premises enterprise data warehouse where existing enterprise BI reports, applications, and users can access the curated data.

Deploying Big Data Management on AWS

The Quick Start provides two deployment options:

- **Deployment of Big Data Management into a new VPC** (end-to-end deployment) builds a new virtual private cloud (VPC) with public and private subnets, and then deploys Informatica Big Data Management into that infrastructure.

- **Deployment of Big Data Management into an existing VPC** provisions Big Data Management components into your existing infrastructure.

Prerequisites

Before you deploy this Quick Start, we recommend that you become familiar with the following AWS services:

- [Amazon VPC](#)
- [Amazon EC2](#)
- [Amazon EMR](#)

If you are new to AWS, see [Getting Started with AWS](#).

Before you configure Big Data Management in the Amazon EMR cloud environment, verify the following prerequisites:
- You have an account with 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. (To sign up for a demo license, please contact Informatica.)

  The license file has a name like BDMLicense.key.
- You have configured an Amazon private key (.pem file) to use for authentication during setup.

Deployment Steps

Step 1. Prepare an AWS Account

1. If you don’t already have an AWS account, create one at http://aws.amazon.com by following the on-screen instructions.

2. Use the region selector in the navigation bar to choose the AWS Region where you want to deploy Informatica Big Data Management on AWS.

   **Note**  The Quick Start isn’t currently supported in the following regions: US West (N. California), US East (Ohio), Asia Pacific (Singapore), Asia Pacific (Mumbai), EU (Frankfurt).

3. Create a key pair in your preferred region.

   When you log in to any Amazon EC2 system or Amazon EMR cluster, you use a password file for authentication. The file is called a private key file and has a file name extension of .pem.

   If you do not have an existing .pem key to use, follow the instructions in the AWS documentation to create a key pair.

   **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, AWS 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.

4. If necessary, request a service limit increase for the Amazon EC2 M3 and M4 instance types. You might need to do this if you already have an existing deployment that uses
this instance type, and you think you might exceed the default limit with this reference deployment.

**Step 2. Launch the Quick Start**

1. Choose one of the following options to deploy the AWS CloudFormation template into your AWS account.

   ![Launch Quick Start (for new VPC)](image1)
   ![Launch Quick Start (for existing VPC)](image2)

   The templates are launched in the US East (N. Virginia) region by default. You can change the region by using the region selector in the navigation bar (note the restrictions listed in step 1).

   Each stack takes approximately two hours to create.

   **Note:** You are responsible for the cost of the AWS services used while running this Quick Start reference deployment. There is no additional cost for using this Quick Start. See the pricing pages for each AWS service you will be using or the AWS Simple Monthly Calculator for full details.

2. On the **Select Template** page, keep the default setting for the template URL, and then choose **Next**.

3. On the **Specify Details** page, review the parameters for the template. Enter values for the parameters that require your input. For all other parameters, you can customize the default settings provided by the template.

   In the following tables, parameters are listed and described separately for deploying Big Data Management into a **new VPC** or an **existing VPC**.
Note  The templates for the two scenarios share most, but not all, of the same parameters. For example, the template for an existing VPC prompts you for the VPC and subnet IDs in your existing VPC environment. You can also download the templates and edit them to create your own parameters based on your specific deployment scenario.

Parameters for deployment into a new VPC:

View template

Network Configuration:

<table>
<thead>
<tr>
<th>Parameter (name)</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Availability Zones (AvailabilityZones)</td>
<td>Requires input</td>
<td>Choose two Availability Zones that will be used to deploy the Informatica Big Data Management components. The Quick Start preserves the logical order you specify.</td>
</tr>
<tr>
<td>VPC CIDR (VPCCIDR)</td>
<td>10.0.0.0/16</td>
<td>CIDR block for the VPC.</td>
</tr>
<tr>
<td>Private Subnet 1 CIDR (PrivateSubnet1CIDR)</td>
<td>10.0.0.0/19</td>
<td>CIDR block for the private subnet located in Availability Zone 1.</td>
</tr>
<tr>
<td>Private Subnet 2 CIDR (PrivateSubnet2CIDR)</td>
<td>10.0.32.0/19</td>
<td>CIDR block for the private subnet located in Availability Zone 2.</td>
</tr>
<tr>
<td>Public Subnet 1 CIDR (PublicSubnet1CIDR)</td>
<td>10.0.128.0/20</td>
<td>CIDR block for the public (DMZ) subnet located in Availability Zone 1.</td>
</tr>
<tr>
<td>Public Subnet 2 CIDR (PublicSubnet2CIDR)</td>
<td>10.0.144.0/20</td>
<td>CIDR block for the public (DMZ) subnet located in Availability Zone 2.</td>
</tr>
<tr>
<td>IP Address Range (RemoteAccessCIDR)</td>
<td>Requires input</td>
<td>The CIDR IP range that is permitted to access the Informatica domain and the Amazon EMR cluster. We recommend that you use a constrained CIDR range to reduce the potential of inbound attacks from unknown IP addresses. For example, to specify the range of 10.20.30.40 to 10.20.30.49, enter 10.20.30.40/49.</td>
</tr>
</tbody>
</table>

Amazon EC2 Configuration:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key Pair Name (KeyPairName)</td>
<td>Requires input</td>
<td>Public/private key pair, which allows you to connect securely to your instance after it launches. When you created an AWS account, this is the key pair you created in your preferred region.</td>
</tr>
<tr>
<td>EMR Master Instance Type (EMRMasterInstanceType)</td>
<td>m3.large</td>
<td>The EC2 instance type for the Amazon EMR master node.</td>
</tr>
<tr>
<td>EMR Core Instance Type (EMRCoreInstanceType)</td>
<td>m3.large</td>
<td>The EC2 instance type for the Amazon EMR core nodes.</td>
</tr>
<tr>
<td>Parameter</td>
<td>Default</td>
<td>Description</td>
</tr>
<tr>
<td>------------------------------------------------</td>
<td>-------------</td>
<td>--------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Informatica Domain Instance Type</strong> (InformaticaServerInstanceType)</td>
<td>m4.large</td>
<td>The EC2 instance type for the instance that hosts the Informatica domain.</td>
</tr>
</tbody>
</table>

**Amazon EMR Configuration:**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMR Cluster Name (EMRClusterName)</td>
<td>Requires input</td>
<td>The name of the Amazon EMR cluster where the Big Data Management instance will be deployed.</td>
</tr>
<tr>
<td>EMR Core Nodes (EMRCoreNodes)</td>
<td>Requires input</td>
<td>The number of core nodes. Enter a value between 1 and 500.</td>
</tr>
<tr>
<td>EMR Logs Bucket Name (EMRLogBucket)</td>
<td>Requires input</td>
<td>The S3 bucket where the Amazon EMR logs will be stored.</td>
</tr>
</tbody>
</table>

**Amazon RDS Configuration:**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informatica Database Username (DBUser)</td>
<td>Requires input</td>
<td>User name for the Informatica domain and the Model Repository database instance. You can specify any string.</td>
</tr>
<tr>
<td>Informatica Database Password (DBPassword)</td>
<td>Requires input</td>
<td>Password for the Informatica domain and the Model Repository database instance. You can specify any string.</td>
</tr>
</tbody>
</table>

**Informatica Big Data Management Configuration:**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informatica Administrator Username (InformaticaAdminUser)</td>
<td>Requires input</td>
<td>Administrator user name for accessing Big Data Management. You can specify any string. 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.</td>
</tr>
<tr>
<td>Informatica Administrator Password (InformaticaAdminPassword)</td>
<td>Requires input</td>
<td>Administrator password for accessing Big Data Management. You can specify any string. 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.</td>
</tr>
<tr>
<td>BDM License Key Location (InformaticaBDMKeyS3Bucket)</td>
<td>Requires input</td>
<td>The S3 bucket in your account that contains your Big Data Management license key file.</td>
</tr>
<tr>
<td>BDM License Key Name (InformaticaBDMKeyName)</td>
<td>Requires input</td>
<td>The path and file name for the Big Data Management license key file. The path must include the subdirectories under the S3 bucket name. For example,</td>
</tr>
</tbody>
</table>
### AWS Quick Start Configuration:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quick Start S3 Bucket Name</td>
<td>aws-quickstart</td>
<td>S3 bucket name for the Quick Start assets. This bucket name can include numbers, lowercase letters, uppercase letters, and hyphens (-), but should not start or end with a hyphen. You can specify your own bucket if you copy all of the assets and submodules into it, if you want to override the Quick Start behavior for your specific implementation.</td>
</tr>
<tr>
<td>Quick Start S3 Key Prefix</td>
<td>quickstart-informatica-bdm/</td>
<td>S3 key prefix for the Quick Start assets. This prefix can include numbers, lowercase letters, uppercase letters, hyphens (-), and forward slashes (/), but should not start or end with a forward slash (which is automatically added). This parameter enables you to override the Quick Start behavior for your specific implementation.</td>
</tr>
</tbody>
</table>

### Parameters for deployment into an existing VPC:

**View template**

**Network Configuration:**

<table>
<thead>
<tr>
<th>Parameter (name)</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPC ID (VPCID)</td>
<td>Requires input</td>
<td>ID of your existing VPC where you’d like to deploy Big Data Management (e.g., vpc-0343606e). The VPC must meet the following requirements:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• It must be set up with public access through the Internet via an attached Internet gateway.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• The DNS Resolution property of the VPC must be set to Yes.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• The Edit DNS Hostnames property of the VPC must be set to Yes.</td>
</tr>
<tr>
<td>Informatica Domain Subnet</td>
<td>Requires input</td>
<td>Publicly accessible subnet ID for the Informatica domain.</td>
</tr>
<tr>
<td>(PublicSubnetID)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Informatica Database Subnets</td>
<td>Requires input</td>
<td>IDs of two private subnets in the selected VPC. These must be in different Availability Zones in the selected VPC (e.g., us-west-1b, us-west-1c).</td>
</tr>
<tr>
<td>(DBSubnetIDs)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Amazon EC2 Configuration:

<table>
<thead>
<tr>
<th>Parameter (name)</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key Pair Name (KeyPairName)</td>
<td>Requires input</td>
<td>Public/private key pair, which allows you to connect securely to your instance after it launches. When you created an AWS account, this is the key pair you created in your preferred region.</td>
</tr>
<tr>
<td>EMR Master Instance Type (EMRMasterInstanceType)</td>
<td>m3.large</td>
<td>The EC2 instance type for the Amazon EMR master node.</td>
</tr>
<tr>
<td>EMR Core Instance Type (EMRCoreInstanceType)</td>
<td>m3.large</td>
<td>The EC2 instance type for the Amazon EMR core nodes.</td>
</tr>
<tr>
<td>Informatica Domain Instance Type (InformaticaServerInstanceType)</td>
<td>m4.large</td>
<td>The EC2 instance type for the instance that hosts the Informatica domain.</td>
</tr>
</tbody>
</table>

Amazon EMR Configuration:

<table>
<thead>
<tr>
<th>Parameter (name)</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMR Cluster Name (EMRClusterName)</td>
<td>Requires input</td>
<td>The name of the Amazon EMR cluster where the Big Data Management instance will be deployed.</td>
</tr>
<tr>
<td>EMR Core Nodes (EMRCoreNodes)</td>
<td>Requires input</td>
<td>The number of core nodes. Enter a value between 1 and 500.</td>
</tr>
<tr>
<td>EMR Logs Bucket Name (EMRLogBucket)</td>
<td>Requires input</td>
<td>The S3 bucket where the Amazon EMR logs will be stored.</td>
</tr>
</tbody>
</table>
Amazon RDS Configuration:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informatica Database Username (DBUser)</td>
<td>Requires input</td>
<td>User name for the Informatica domain and the Model Repository database instance. You can specify any string.</td>
</tr>
<tr>
<td>Informatica Database Password (DBPassword)</td>
<td>Requires input</td>
<td>Password for the Informatica domain and the Model Repository database instance. You can specify any string.</td>
</tr>
</tbody>
</table>

Informatica Big Data Management Configuration:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informatica Administrator Username (InformaticaAdminUser)</td>
<td>Requires input</td>
<td>Administrator user name for accessing Big Data Management. You can specify any string. 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.</td>
</tr>
<tr>
<td>Informatica Administrator Password (InformaticaAdminPassword)</td>
<td>Requires input</td>
<td>Administrator password for accessing Big Data Management. You can specify any string. 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.</td>
</tr>
<tr>
<td>BDM License Key Location (InformaticaBDMKeyS3Bucket)</td>
<td>Requires input</td>
<td>The S3 bucket in your account that contains your Big Data Management license key file.</td>
</tr>
<tr>
<td>BDM License Key Name (InformaticaBDMKeyName)</td>
<td>Requires input</td>
<td>The path and file name for the Big Data Management license key file. The path must include the subdirectories under the S3 bucket name. For example, if the S3 bucket location is myBucketName/SubDir1/SubDir2/BDMLicense.key, type in SubDir1/SubDir2/BDMLicense.key.</td>
</tr>
</tbody>
</table>

When you finish reviewing and customizing the parameters, click Next.

4. On the Options page, you can specify tags (key-value pairs) for resources in your stack and set advanced options. When you’re done, choose Next.

5. On the Review page, review and confirm the template settings. Under Capabilities, select the check box to acknowledge that the template will create IAM resources.

6. Choose Create to deploy the stack.
Step 3. Monitor Instance Provision and Informatica Domain Creation

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

1. 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:

   ![Events tab](image)

   **Figure 6: Events tab**

   When stack creation is complete, the **Status** field will show **CREATECOMPLETE**, as shown in Figure 7.
2. Select the **Resources** tab.

This tab displays information about the stack and the Big Data Management instance. You can select the linked physical ID properties of individual resources to get more information about them, as shown in Figure 8.
3. Click the **Outputs** tab.

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

<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CloudFormationLogs</td>
<td>Amazon EMS creates the AWS CloudFormation 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.</td>
</tr>
<tr>
<td>InstanceID</td>
<td>Name of the Informatica domain host.</td>
</tr>
<tr>
<td>InformaticaHadoopInstallLogs</td>
<td>Location on the master node of the Amazon EMR cluster of the log that records the creation of the Hadoop instance.</td>
</tr>
<tr>
<td>InformaticaAdminConsoleURL</td>
<td>URL of the Informatica Administrator. Use the Administrator tool to administer Informatica services.</td>
</tr>
<tr>
<td>InformaticaAdminConsoleServerLogs</td>
<td>Location of the Informatica domain installation logs.</td>
</tr>
<tr>
<td>EMRMasterNodeHadoopURL</td>
<td>URL of the Amazon EMR resource manager and master node.</td>
</tr>
<tr>
<td>InformaticaBDMDeveloperClient</td>
<td>Location where you can download the Developer tool client.</td>
</tr>
</tbody>
</table>
Note: If the Outputs tab is not populated with this information, wait for domain setup to be complete.

4. Open the Resources tab. Choose 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 this screen to launch the Big Data Management instance. You can also get additional information about the instance, such as the public DNS and public IP address.

Step 4. 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. Log in to the AWS Management Console and select CloudFormation from the Services menu.
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.

Important If you deploy the Quick Start for a new VPC, Amazon EMR creates security groups that aren’t deleted when you delete the Amazon EMR cluster. To clean up after deployment, you must first delete the Amazon EMR cluster, then delete the Amazon EMR-managed security groups (ElasticMapReduce-master, ElasticMapReduce-slave) by first deleting the circularly dependent rules followed by the security groups themselves, and then delete the AWS CloudFormation stack.

Troubleshooting

If you encounter a CREATE_FAILED error when you launch the Quick Start, we recommend that you relaunch the template with Rollback on failure set to No. (This setting is under Advanced in the AWS CloudFormation console, Options page.) With this setting, the stack’s state will be retained and the instance will be left running, so you can
troubleshoot the issue. Look at the log files in \%ProgramFiles\%Amazon\EC2ConfigService and C:\cfn\log.

**Important:** When you set Rollback on failure to No, you’ll continue to incur AWS charges for this stack. Please make sure to delete the stack when you’ve finished troubleshooting.

For additional information, see Troubleshooting AWS CloudFormation on the AWS website or contact us on the AWS Quick Start Discussion Forum.

**Additional Resources**

**AWS services**
- AWS CloudFormation
  http://aws.amazon.com/documentation/cloudformation/
- Amazon EBS
- Amazon EC2
- Amazon VPC
  http://aws.amazon.com/documentation/vpc/

**Informatica**
- Informatica Network: a source for product documentation, Knowledge Base articles, and other information.
  https://network.informatica.com

**Quick Start reference deployments**
- AWS Quick Start home page
  https://aws.amazon.com/quickstart/
- Community Quick Starts
  https://aws.amazon.com/quickstart/community/

**Send Us Feedback**

We welcome your questions and comments. Please post your feedback on the AWS Quick Start Discussion Forum.
You can visit our GitHub repository to download the templates and scripts for this Quick Start, and to share your customizations with others.