INTEGRATING BIG DATA... SMARTER

syncsort

Syncsort DMX-h

Amazon EMR Edition

User Guide

Table of Contents

1 Introduction

DMX-h, Syncsort's Hadoop-enabled version of its DMExpress ETL software, can run on Amazon's Elastic MapReduce (EMR) platform in the cloud.

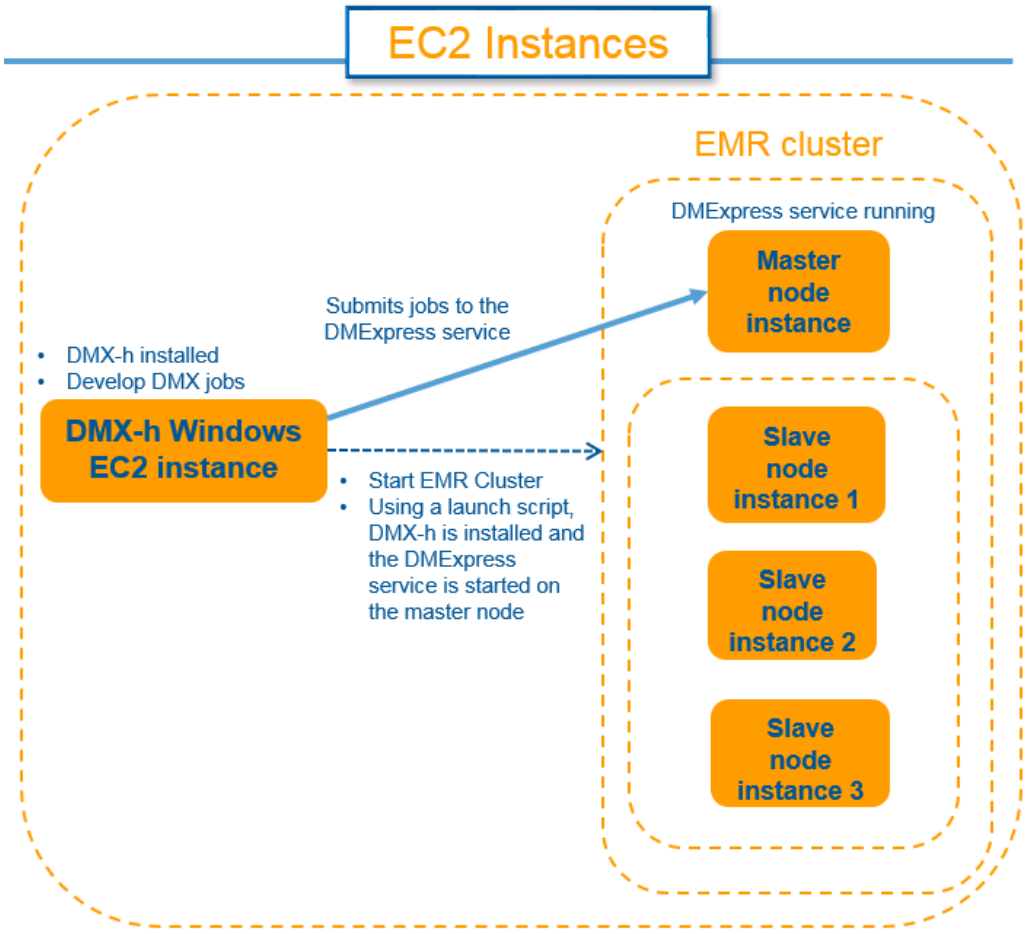
DMX-h enables powerful ETL processing within Hadoop without the need to learn complex MapReduce programming skills or invest in building and maintaining your own Hadoop cluster. With DMX-h, you can:

- develop MapReduce/Spark jobs in an easy-to-use graphical interface without coding
- connect to any data source or target, including mainframe data
- jump-start your Hadoop productivity with included use case accelerators (UCAs).

DMX-h is available in the Amazon Web Services (AWS) Marketplace for up to 5 nodes and is intended for testing, piloting, and proof of concepts. Larger instances can be set up on AWS EMR.

For a background on AWS and the component services applicable to DMX-h, see Appendix A.

2 DMX-h Architecture



DMX-h is comprised of the following main components:

DMX-h Windows EC2 Instance

The DMX-h Windows EC2 instance is based on an Amazon machine image (AMI) that includes the following:

- Installed trail version of DMX-h
- Installed AWS command line interface (CLI)
- Standalone JSON processor, jq, which parses the JSON responses that the AWS CLI returns
- Launch script, **launchEMR.ps1**, which uses the AWS CLI to launch and prepare the EMR cluster

This Windows EC2 instance serves as the DMX-h job development machine, on which DMX-h is installed: The root volume is Amazon Elastic Block Store (EBS)-backed:

- You launch this instance through either the **1-Click Launch** tab or the **Manual Launch** tab on the Launch on EC2 page.
- Through this instance, you submit DMX-h jobs to the DMExpress service (**dmxd**) on the master node of the EMR cluster.

EMR Cluster

The EMR cluster is comprised of the EC2 node instances in which you run DMX-h UCA jobs on either a MapReduce framework or a Spark on YARN framework.

You launch the EMR cluster via the **launchEMR.ps1** script, which is located on the DMX-h Windows EC2 instance:

- The EMR cluster is based on EMR release 4.7.0.
- At the end of cluster creation,
    - DMX-h is installed on all cluster nodes.
    - Master node is configured:
        - **dmxd** is running on the master node and master node port 32636 is open.
        - DMX-h UCAs are installed in the home directory of the master node: `/home/dmxuser/UCAs`
        - UNIX user account, `dmxuser`/`dmxuser`, is created on the master-node.
          To ensure security, you may want to change the password for `dmxuser`.
        - Password authentication is enabled on the master node, which enables user `dmxuser` on the DMX-h Windows EC2 instance to connect via a DMX-h secure file transfer protocol (SFTP) file browsing connection to the secure shell (SSH) server on the master node and to browse to the UCAs.
        - Password login is enabled on the master node, which enables user `dmxuser` on the DMX-h Windows EC2 instance to connect via the DMExpress Server Connection dialog to **dmxd** on the master node and to submit jobs to the cluster.
        - AWS creates user `hadoop` with sudo privileges on the master-node.

          To change any setting, `ssh` to the master node, for example:

          ```
          ssh -o ServerAliveInterval=10 -o
          StrictHostKeyChecking=no -i
          D:/Work/EMR/<file_name>.pem hadoop@ec2-54-87-50-
          54.compute-1.amazonaws.com
          ```

3 Support and Charges

3.1 Support

Support is available as follows:

- Register for 30 days of free phone and email support at www.syncsort.com/emrsupport.
- Online support is always available via the Syncsort Online Community at http://community.syncsort.com.

3.2 Charges

There are no DMX-h usage charges for the 5-node edition. For larger clusters, contact Syncsort to obtain a product license.

The following Amazon services will accrue charges from the time each instance is launched until it is terminated:

- DMX-h Windows EC2 instance
  - EBS-backed root volume (defaults to deletion upon termination)
  - DMX-h Windows EC2 instance while running
- EMR cluster

  EMR instance charges per node

Any additional EBS volumes you add/attach to your instances, as well as any Amazon S3 storage you use, will incur additional charges.

4 Launching DMX-h

4.1 Review Product Details

The main point of entry for DMX-h is in the AWS Marketplace, here. You can also just go to aws.amazon.com/marketplace, search for “DMX-h”, and click on the link. On the product page, do the following:

1. Review the details on the DMX-h product offering.
2. Click on the **Continue** button at the top of the page to get to the **Launch on EC2** page.
3. If you’re not already signed in, you will be prompted to sign in or create an AWS account; follow the instructions on screen.

4.2 Launch the DMX-h Windows EC2 Instance

The DMX-h Windows EC2 instance is launched from the **Launch on EC2** page in either of the following ways:

- Via the **1-Click Launch** tab, which is quick and easy, and is recommended for running with a standard configuration to try the UCAs.
- Via the **Manual Launch** tab, which allows for more instance customization and enhanced security, and is recommended when developing and running your own jobs.

4.2.1 1-Click Launch

On the **1-Click Launch** tab of the **Launch on EC2** page, review the default settings and click on the arrow to expand each section to make any necessary changes as follows:

1. Select the desired **Version** of DMX-h.
2. Select the **Region** in which your AWS resources are located to reduce latency and minimize costs. See Understanding Amazon EMR region limitations.
3. Select the **EC2 Instance Type** that meets your hardware requirements. Review the updated **Monthly Estimate** based on the selected instance type.
4. Select the VPC Settings that meet your security requirements.

   EC2 Classic exposes security risks and is not recommended except for trial runs of the UCAs. For enhanced security, select a previously created VPC or create a new one and select it, then select a subnet. The 1-Click Launch does not provide a public IP to connect to the launched instance, so you will need to attach an Elastic IP to your instance after launching or consider using the EC2 Console Launch for improved VPC configuration.
5. Select **Create new based on seller settings** to allow AWS to auto-generate a new **Security Group** that includes the required port for DMX-h (3389 for the Remote Desktop Protocol (RDP) used to connect from your desktop to the DMX-h Windows EC2 instance) or choose a previously defined **Security Group** that includes this port.

6. Select an existing **Key Pair** if you already created one and have the private key file on your machine. Otherwise, create a new one as described here, then come back to this screen and select the newly created one. You will need to supply the private key file when attempting to log into the instance, so be sure to have it accessible on your local machine.
7. Once all settings are correct and you've read and agreed to the terms as outlined and linked to on this page, click on the **Accept Terms & Launch with 1-Click** button at the top of the page. Note that once subscribed to a given AMI, you will not need to accept terms on subsequent launches of that AMI.
8. When first subscribing to an AMI, a pop-up will be displayed indicating that charges will begin to accrue once the instance is deployed upon subscription completion; on subsequent launches, it indicates that the instance is deploying. Instance details are shown, along with a link to the AWS Management Console, where you can start managing your DMX-h Windows EC2 instance.
9. If you plan to develop and run your own jobs in an instance launched via 1-Click, you will either need to store your job and task files in Amazon S3 or you will need to create and/or attach an (existing) EBS volume to your instance to store your job and task files. Be sure that this EBS volume is not set to be deleted on instance termination, and that it is in the same Availability Zone as your instance.

4.2.2 Manual Launch

On the **Manual** L**aunch** tab of the **Launch on EC2** page, do the following:

1. Read and agree to the terms as outlined and linked to on this page, then click on the **Accept Terms** button at the top right if not already subscribed to this AMI.
2. Select a **Version**, then click on the **Launch with EC2 Console** button corresponding to the **Region** in which your AWS resources are located. See Understanding Amazon EMR region limitations.

This will take you to the EC2 Console, where you can configure the instance on each successive screen by clicking on the **Next** button at the bottom of the page as follows:

1. On the **Choose Instance Type** page, select the instance type that meets your hardware requirements by clicking on the categories in the left pane and then selecting the desired instance type in the table.
2. On the **Configure Instance Details** page, do the following:
    a. For the **Network** setting, we recommend selecting an existing VPC or creating a new one by clicking on **Create new VPC** and following the on-screen instructions:
        i. Ideally this would be done by a network administrator aware of the needs for your environment, but at a minimum, you can accept the defaults, noting that the number of IP addresses in the **Public subnet** will need to be big enough to hold the DMX-h Windows EC2 instance and EMR cluster node instances and that the **Availability Zone** must match that of any EBS volumes you will use.

ii. Once the VPC is created, click on the refresh button next to the **Network** dropdown to repopulate it and select the newly created VPC.

iii. Select the desired **Subnet** if more than one was added to the VPC.

iv. Select **Enable** from the **Auto-assign Public IP** dropdown so that you can access the instance from your local machine.

b. If you choose **Launch into EC2-Classic** for the **Network** setting, set the **Availability Zone** in which the instance will be launched to match the Availability Zone of any existing EBS volume that you will attach to the instance. Make any other appropriate selections to meet your requirements.

3. On the **Add Storage** page, click on **Add New Volume** to add a new EBS volume to store your job and task files. Be sure the **Delete on Termination** checkbox is unchecked if you want the EBS volume to persist after the instance is terminated. Note that this will result in additional EBS fees from Amazon. If you have an existing EBS volume that you want to attach instead, follow the instructions here after the instance is launched.
4. On the **Add Tags** page, enter a meaningful name in the **Value** field for the **Name** tag to easily identify your DMX-h Windows EC2 instance in the EC2 Console.
5. On the **Configure Security Group** page, do one of the following:

   a. Choose **Select an existing security group** and select a previously created group configured as described below based on your network settings.

   b. Choose **Create a new security group** and do the following:

      i. Enter an appropriate name and description.

      ii. Set the **Source** for the SSH port to **My IP** (which will detect your local IP address automatically) so that you can connect to the DMX-h Windows EC2 instance from your desktop.

      iii. Click on **Add Rule** to add a **RDP** rule entry for TCP protocol, port 3389, with **Source** set to **Anywhere**, but be aware that this open-to-the-world port presents an increased security risk and you will receive a warning upon launching.

6. Click on the **Review and Launch** button at the bottom right. Verify your selections on the **Review Instance Launch** page, and click on **Launch** to launch the instance as configured.

7. You will be prompted to select an existing **Key Pair** if you already created one and have the private key file on your machine. Otherwise, create a new one as described here, then select the newly created one. You will need to supply the private key file when attempting to log into the instance, so be sure to have it accessible on your local machine, and check the box to acknowledge that you do.

8. Click on the **Launch** button. This will launch the DMX-h Windows EC2 instance and charges will begin to accrue.

4.3 Connect to the DMX-h Windows EC2 Instance

The Windows EC2 instance may take some time to come up and pass its status checks (displayed on the **Instances** page of the EC2 console), especially with a micro or small instance type. Once the status checks have passed, you can connect to it as follows:

1. From the EC2 Console, select the DMX-h Windows EC2 instance, and click on the **Connect** button at the top of the page.
2. In the **Connect To Your Instance** screen:
   a. Click on **Get Password** to retrieve the initial Administrator password; note that this can take up to 30 minutes from the time the Windows EC2 instance was started.
   b. When prompted, paste in the full contents of the private key file used to launch the instance, and click on the **Decrypt Password** button to populate the **Password** field. Keep this window open or copy the password, though you can always come back and retrieve it again if the password hasn't been manually changed.
   c. Click on the **Download Remote Desktop File** button to download the RDP file, then open it to connect to the DMX-h Windows EC2 instance. You will get an Unknown publisher warning; click **Connect** to continue.
   d. When prompted, enter the decrypted password to log in as Administrator. You will get a security certificate warning; click **Yes** to continue.

When you see the DMX-h Windows EC2 instance desktop, you are automatically connected.

4.4 Launch the EMR Cluster

Once the DMX-h Windows EC2 instance is up and running, launch the EMR cluster:

Note: Do not copy/paste from this document to invoke the commands.

1. Open a PowerShell session.

2. To launch the cluster, run **launchEMR.ps1** with the applicable parameters. Consider the following:

   - Launch script location

     launchEMR.ps1 is located on the DMX-h Windows EC 2 instance: `C:\DMX-h_LaunchFiles\launchEMR.ps1`

   - `Get-Help` command line option

     For a command line description of parameters and an example, execute the following:

     ```
     Get-Help C:\DMX-h_LaunchFiles\launchEMR.ps1
     ```

   - Launch script usage and results

     Reference the following:

     - Launch script usage
     - Launch script console display

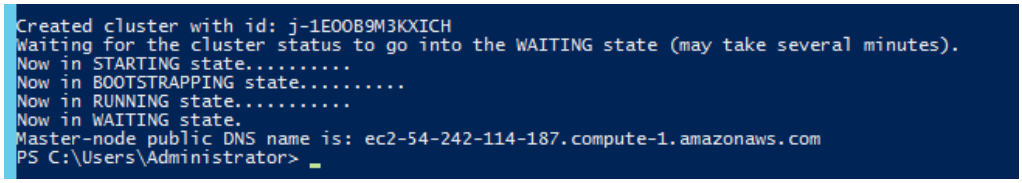
   - Cluster node limit

     Important: The default value for the number of nodes is 2. The maximum number of permissible nodes is 5.

   - Example**: launchEMR.ps1** and parameters

     ```
     C:\PS>C:\DMX-h_LaunchFiles\launchEMR.ps1 -access_key_id
     REIAOI2NUHSNCIH3AQWE -secret_key
     nPU9GDCedgf/EW1IEp3AqUxSjRnP6hB2NmVW70Op -key_name myKey -
     region us-east-1 -tm m3.2xlarge -tc m3.xlarge -n 3 -log_uri
     s3://mybucket/emr
     ```

Starting the EMR cluster node instances may take some time. After successfully launching the EMR cluster, **launchEMR.ps1** displays the EMR cluster id and the master node public DNS name:



3. Make note of the master node public DNS name as you enter this value on the **DMExpress Server Connection** dialog to connect to the master node, from where you browse to and run the UCAs.

4.5 Next Steps

At this point, you can run the UCAs or start developing and running your own jobs.

When you are done running DMX-h jobs, stop or terminate your EC2 instances until you are ready to start or launch them again.

5 Use Case Accelerators

DMX-h comes with a built-in set of use case accelerators (UCAs) to quickly and easily demonstrate both the development and running of DMX-h ETL jobs in the Amazon EMR cluster. These cover a variety of common ETL use cases.

5.1 UCA Overview

The following UCAs are installed on the master node:

Category	Use Case Accelerator	Description
Change Data Capture (CDC)	CDC Single Output	Performs change data capture (CDC) against two large input files, producing a single output file marking records as inserted, deleted, or updated.
	CDC Distributed Output	Same as CDC Single Output, except that it produces three separate output files for the inserted, deleted, and updated records.
	Mainframe Extract + CDC	Same as CDC Single Output, but also converts and loads mainframe data to HDFS before passing the HDFS data to the CDC job.
Join	Join Large Side | Large Side	Performs a join of two large files stored in HDFS.
Aggregations	Web Logs Aggregation	Calculates the total number of visits per site in a set of web logs using aggregate tasks.
	Lookup + Aggregation	Performs a lookup followed by an aggregation.
	Word Count	Performs the standard Hadoop word count example.
Mainframe Access and Integration	Direct Mainframe Extract & Load	Loads two files residing on a remote mainframe system to HDFS, converting to ASCII displayable text.
	Mainframe File Load	Same as Direct Mainframe, except that mainframe files are loaded to HDFS from local file system.
Connectivity	HDFS Extract	Extracts data from HDFS using HDFS connectivity in a DMExpress copy task.
	HDFS Load	Same as HDFS Extract, but loads data to HDFS.
	HDFS Load Parallel	Same as HDFS Load, but splits the data into multiple partitions and loads to HDFS in parallel.

5.2 Running the UCAs

Run the UCAs either on a MapReduce framework or on a Spark on YARN framework in the EMR cluster.

From your desktop, RDP into the Windows EC2 instance as Administrator (if not already logged in there), start the DMExpress Job Editor (**Start-> DMExpress Job Editor**), and run the desired UCA as follows:

1. Click on the **Status** button in the Job Editor toolbar.
    a. In the **DMExpress Server Connection** dialog (automatically raised if **DMExpress server** is empty, otherwise click on **Select Server…**), click on the **UNIX** tab, enter the public DNS name of the master node instance in **Connect to server**, enter the **User name**, dmxuser, and **Password**, dmxuser, and click **OK**.
    b. Select the **Environment Variables** tab, Environment variables" tab and import the variables from the following file: `C:\DMX-h_LaunchFiles\`**`EnvironmentVariablesForUCAs.bat`**
2. Select **File->Open Job…**, click on the **Remote Servers** tab, double click on **New file browsing connection**, specify the connection as follows, click **Verify connection**, then **OK**, and then **OK** again:
    - **Server:** *<public_DNS_name_of_the_master_node>*
    - **Connection type**: Secure FTP
    - **Authentication**: Password
    - **User name**: dmxuser
    - **Password**: dmxuser

    Note: If you updated the dmxuser password, specify the secure password.
3. Double click on the newly created file browsing connection and browse to the location of the job you want to run. Jobs are located in the folder `/home/dmxuser/UCA/Jobs/<JobName>/`. Select `DMXUserDefinedMRJobs/MRJ_<JobName>.dxj` or `J_<JobName>.dxj` (if both are present, select the `J_` version) or select `DMXStandardJobs/<jobname>` and click on **Open**.
4. Click on the **Run** button in the Job Editor toolbar. In the **Run Job** dialog, select Framework and either MapReduce or Spark from the corresponding dropdown. If running on Spark, specify the Spark master URL. Click **OK**.

    This will bring up the **DMExpress Server** dialog, which will show the progress of the running job. Upon completion, select the job and click on **Detail…** to see detailed messages, DMX-h job statistics, and a Hadoop URL to track the cluster job.

    You can view the job output by browsing HDFS from the DMX-h Windows EC2 instance. You can sample the output data on HDFS.
5. To view the Hadoop log files, copy the Hadoop URL from the DMExpress job log and paste it in a browser.

5.3 Additional Information

For details on the UCA directory structure, the automated preparation script, and further instructions on running the jobs, see the Guide to DMX-h ETL Use Case Accelerators.

For additional information on how to develop and run your own Hadoop ETL solutions, see “DMX-h ETL” in the DMExpress Help, accessible via the DMExpress GUI (DMExpress Job Editor or Task Editor) on the Windows EC2 instance.

6 Managing your DMX-h Instances

6.1 Naming your Instances

If you launched via **Manual Launch**, you had an opportunity to create a Name tag in the **Add Tags** step of the instance configuration sequence prior to launching. However, instances launched via **1-Click** or the launch script will not have the Name field populated in the EC2 Console.

To make it easier to keep track of your instances in the EC2 Console, add a name by clicking on the **Name** field for a given instance and typing in a meaningful name. When you select each instance, the instance description in the bottom pane helps you identify the instance as either the DMX-h Windows EC2 instance or one of the EMR cluster node instances (master or slave).

6.2 Stopping and Restarting DMX-h Instances

The DMX-h Windows EC2 instance can be stopped and restarted as needed. Note that EMR cluster node instances cannot be stopped, only terminated.

On the EC2 Console, select the desired instance(s), click on the **Actions** dropdown at the top of the page, and select **Stop** or **Start** to stop/start existing instances.

When an instance is stopped and restarted, it will be assigned a new public DNS name. For the Windows EC2 instance, this will mean that you'll need to download a new RDP connection file, but the previously decrypted password will remain unchanged for the life of the instance.

6.3 Persistent Storage: Source and Target Data

When you run DMX-h jobs in EMR, any target data written to HDFS will be lost when the EMR job flow is terminated. If you don't attach an EBS volume to your instance to store your job and task files, you can use DMX-h to write to Amazon S3 for permanent storage. See "Connecting to Amazon S3 from DMExpress" in the DMExpress Help for details.

6.4 Changing your Cluster Size

You can change the size of a running EMR cluster as described here as long as the cluster size remains within the 5-node limit.

6.5 Terminating your DMX-h Instances

Once all running jobs are completed and any temporary data that you wish to keep has been copied to permanent storage, you can terminate your DMX-h instances as follows:

1. On the EMR Console, select the job id and click on the **Terminate** button at the top to terminate all the node instances in the EMR cluster launched by the script, then click **Yes, Terminate** to confirm. Click on the **Refresh** button and make sure that the **State** shows **TERMINATED**; this may take some time.

2. On the EC2 Console, select the DMX-h Windows EC2 instance, click on the **Actions** dropdown at the top of the page, and select **Terminate** from the bottom of the menu. You will be warned about losing any non-persistent storage; click **Cancel** if you need to go back and save files to your persistent storage, otherwise click **Yes, Terminate** to terminate the instances. Be sure that the **Instance State** shows **terminated**; this may take some time.

3. If you want to re-launch after terminating, click on **Your Account** at the top of the AWS Marketplace page, then click on **Manage your software subscriptions**, and then click on the **Launch more software** button for the desired subscription, which will take you back to the corresponding launch page.

6.6 Understanding Amazon EMR Region Limitations

Consider the following:

- The following regions do not support emr-4.7.0: Ohio, Canada-central, London, and Mumbai.

- Due to an AWS EMR bug, you cannot launch the EMR cluster in the Frankfurt region.

- The default security group setting for the master node does not enable connectivity from the following regions outside the cluster: Singapore, California, Sydney, Ireland, and Sao Paulo.

  To manually update the setting and enable connectivity to the master node from these regions, add the following inbound rules to the master node security group: SSH connection; port 22 and TCP connection; and port 32636 from the DMX-h Windows EC2 instance.

Appendix A Amazon Web Services

Amazon Web Services (AWS) is a cloud computing platform comprised of a set of remote computing services. Following are the primary AWS services/components that pertain to running DMX-h:

- Amazon Elastic Compute Cloud (EC2) provides resizable computing capacity in the AWS cloud, allowing quick and easy launching of any number of virtual servers, or instances, based on your computing needs.
- An Amazon Machine Image (AMI) is a software configuration template consisting of an operating system, applications, libraries, data, and configuration settings. This template is used to launch an EC2 instance.
- An EC2 instance is a virtual server that comprises some combination of CPU, memory, storage, and networking capacity, based on the selected instance type.
  - Each EC2 instance is defined by the instance type, the AMI used to launch it, and the configured storage, security, and network access.
- Amazon Elastic Map Reduce (EMR) is a resizable Hadoop cluster made up of EC2 instances.
- Amazon Elastic Block Store (EBS) provides block level storage volumes that can be attached to EC2 instances for persistent storage.
- Amazon Simple Storage Service (S3) is a scalable, reliable web-based data storage service.
- AWS Identity and Access Management (IAM) is a web service that allows you to manage users and user permissions for your AWS account.
- Amazon Virtual Private Cloud (VPC) is a web service that allows you to create a virtual private network in the Amazon cloud to enhance the security of your cloud computing.
- The AWS Marketplace is an online store where you find, buy, and launch software that runs on the AWS cloud.
- The AWS Management Console is the web interface for managing your AWS services. Clicking on EC2 will take you to the EC2 Console; clicking on Elastic MapReduce will take you to the EMR Console.

Appendix B DMX-h Launch Script Usage

The **launchEMR.ps1**script, which you run from a PowerShell session on the DMX-h Windows EC2 instance, has the following usage:

SYNTAX

```
C:\DMX-h_LaunchFiles\launchEMR.ps1 [-access_key_id] <String> [-
secret_key] <String> [-key_name] <String> [[-tm] <String>] [[-tc]
<String>] [[-n] <Int32>] [-region] <String> [[-log_uri] <String>]
```

Parameters

Parameters	Description
`-access_key_id <String>`	Required. The AWS access key id used to launch the EMR cluster.
`-secret_key <String>`	Required. The AWS secret key used to launch the EMR cluster.
`-key_name <String>`	Required. The name of the AWS key used to launch the EMR cluster.
`-tm <String>`	Optional. The instance type of the master node. If not specified, the parameter defaults to `m3.2xlarge`.
`-tc <String>`	Optional. The instance type of the core nodes. If not specified, the parameter defaults to `m3.xlarge`.
`-n <Int32>`	Optional. Number of core nodes. Value values: `1-5`. If not specified, the parameter defaults to `2`. **Note:** The maximum value is `5`. Do not specify a value greater than `5`.

`-region <String>`	Required. The region where the cluster is launched.
`-log_uri <String>`	Required. The URI of the logs to which the AWS CLI writes.

Appendix C DMX-h Launch Script Console Display

When the AWS CLI command to create the cluster is invoked from the **launchEMR.ps1** script, the –bootstrap-actions and –steps options are used to install DMX-h on all nodes of the cluster, to install the UCAs on the master node, and to configure the master node to run DMX-h and the UCAs.

Launch Script Display to Console

```
"aws emr --% create-cluster " +
                  "--applications Name=Hadoop Name=Hive Name=Spark
" +
                  "--bootstrap-actions
Path=s3://sserpxemd/9.2/rpminstall.sh " +
                  "--release-label emr-4.7.0 " +
                  "--service-role EMR_DefaultRole " +
                  "--ec2-attributes
KeyName=$key_name,InstanceProfile=EMR_EC2_DefaultRole " +
                  "--instance-groups
InstanceGroupType=MASTER,InstanceCount=1,InstanceType=$tm
InstanceGroupType=CORE,InstanceCount=$n,InstanceType=$tc " +
                  "--steps
Type=CUSTOM JAR,Name=PrepMasterNode,ActionOnFailure=TERMINATE_CLU
STER,Jar=command-runner.jar,Args=[bash,-c,`"aws s3 cp
s3://sserpxemd/9.2/prep_masternode.sh .; chmod +x
./prep_masternode.sh; sudo ./prep_masternode.sh`"] " +

"Type=CUSTOM_JAR,Name=PrepHDFSuserDir,ActionOnFailure=TERMINATE_C
LUSTER,Jar=command-runner.jar,Args=[bash,-c,`"hadoop fs -chmod
a+rwx /user`"] " +

"Type=CUSTOM_JAR,Name=InstallUCAs,ActionOnFailure=TERMINATE_CLUST
ER,Jar=command-runner.jar,Args=[bash,-c,`"aws s3 cp
s3://sserpxemd/9.2/install_UCAs.sh .; chmod a+x
./install_UCAs.sh; sudo -u dmxuser ./install_UCAs.sh`"] "+
      "--output json"
```

–bootstrap-actions Option

The –bootstrap-action option invokes the following script on the all the node of the cluster: `s3://sserpxemd/9.2/`**`rpminstall.sh`**. This script extracts the DMX RPM from its shrink-wrap and installs it.

–steps Option

The following steps are run on the master node:

- `PrepMasterNode` starts dmxd, creates UNIX user account dmxuser/dmxuser, and enables password authentication for SFTP.
- `PrepHDFSuserDir` adds write permission to the Hadoop directory, /user, allowing for the UCAs to write data to this directory.
- `InstallUCAs` accesses UCAs from bucket location `s3://sserpxemd/9.2` and installs them on the cluster.

Appendix D Troubleshooting

Problem	Possible Cause	Action
DMExpress GUI is unable to connect to **dmxd** running on the master node in the cluster	Port where **dmxd** runs is not part of the master node security group	Using the web console, ensure that the following inbound rule exists in the master node security group: TCP, port 32636, source <*DMX-h Windows EC2_instance IP address*>
Unable to establish a SFTP browsing connection from the GUI	Port where the SSH service runs is not part of the master node security group	Using the web console, ensure that the following inbound rule is present in the master node security group: TCP, port 22, source <*DMX-h Windows EC2_instance IP address*>

About Syncsort

Syncsort provides data-intensive organizations across the big data continuum with a smarter way to collect and process the ever expanding data avalanche. With thousands of deployments across all major platforms, including mainframe, Syncsort helps customers around the world overcome the architectural limits of today's data integration and Hadoop environments, empowering their organizations to drive better business outcomes, in less time – with fewer resources and lower total cost of ownership. For more information, please visit www.syncsort.com.

Syncsort Inc. ***2 Blue Hill Plaza #1563, Pearl River, NY 10965*** ***201.930.8200***