syncsort

DMX-h Quick Start for MapR VM

Table of Contents

1 Introduction

1.1 Welcome to the DMX-h ETL Test Drive on the MapR VM

The DMX-h ETL Test Drive on the MapR VM is a trial package of Syncsort's Hadoop product offering. It allows you to try DMX-h ETL on your own machine and experience for yourself a smarter approach to Hadoop ETL: powerful data processing capabilities without the need to learn complex MapReduce skills.

The Test Drive provides a ready-to-use virtual machine (VM) pre-installed with the MapR Hadoop distribution, the DMX-h ETL software, and a set of use case accelerators and sample data. We encourage you to try them out and provide feedback directly to our engineers and product managers via the Syncsort User Community.

DMX-h is high-performance ETL software that turns Hadoop into a more robust ETL solution, focused on delivering capabilities and use cases that are standard on traditional data integration platforms. Accelerate your data integration initiatives and unleash Hadoop's potential with the only ETL architecture that runs ETL processes natively within Hadoop.

1.2 What's in the Download

The Test Drive download is a zip file (~3.5 GB zipped) that contains:

- A Test Drive VM (extracts to ~7 GB, details below)
- DMX-h Workstation software (~400 MB)

The Test Drive VM, hostname DMXhTstDrvMapR, includes the following:

- Linux CentOS 6.5, boots to command-line mode
- MapR Sandbox for Hadoop 4.0.2 distribution, MRv2
- DMX-h ETL version 8.1.0 (~320 MB)
- Use case accelerators (~2 MB)
- Sample data for running the use case accelerators (~1.3 GB)

1.3 Getting Started

This document explains how to download and set up the Test Drive, as well as install the DMX-h Workstation software outside of the VM on your Windows system.

Once that is done, you are ready to go for a test drive. You can run the included use case accelerators and then try developing your own solutions.

1.4 Getting Help

For assistance with the DMX-h Test Drive VM, please visit the Syncsort User Community.

2 Setting up the Test Drive

Setting up the Test Drive involves the following steps, each explained in detail in the subsequent sections:

1. Confirm the system requirements.
2. Download and extract the Test Drive zip file.
3. Start the VM.
4. Install the DMX-h Workstation software on Windows.
5. Configure the Windows environment as needed.

2.1 System Requirements for the Test Drive VM

Confirm that your system meets the following requirements for the test drive installation:

- One of the following 64-bit x86 architectures:
  - 1.3 GHz or faster AMD CPU with segment limit support in long mode
  - 1.3 GHz or faster Intel CPU with VT-x support
- Minimum 4 physical cores
- Minimum 8 GB memory
- Minimum 20 GB hard disk space
- 64-bit Windows OS
- VMware Player or VMWare Workstation installed, with the following sub-requirements:
  - The minimum VMWare/Player version must be one of the following:
    - ESXi 5.0
    - Fusion 4.x
    - Workstation 8.x
    - Player 4.x
  - Your host system must meet the hardware and firmware requirements to run 64-bit guest operating systems, as described here.
  - If using an Intel CPU, virtualization technology (VT-x) must be enabled on your VMware host, as described here.
  - The VMware network adapter on the host system must be enabled, and NetBIOS over TCP/IP must be enabled for the adapter. Contact your Network Administrator for assistance.

If these requirements cannot be met, or if you would like to do performance testing, let us know and we can provide you with the DMX-h ETL software to install in your own Hadoop cluster. The VM should only be used for functional testing.

2.2 Download and Extract the Test Drive

Download and extract the test drive as follows:

1. Download the Test Drive zip file to a local folder on your Windows machine.
2. Extract it using a decompression utility such as WinZip or 7zip.

2.3 Start the VM

Bring up the VM and log in as follows:

1. Run VMware Player or Workstation, open the extracted VM (named in the form DMX<version>_MapR<version>_MRv1_VMv<version>.OVF), and then play it. This can take 3-5 minutes or longer to initialize, depending on the speed of your machine.
2. If not already logged in, log in to the VM using the following credentials:

   User Name: dmxdemo
   Password: dmxdemo

   The dmxdemo user has sudo privileges for root access. The 'root' password is 'mapr' in case you need it for other administrative purposes.

2.4 Install the DMX-h Workstation Software on Windows

The DMX-h Workstation software, which includes the Job and Task Editors used to view, create, and run DMX-h ETL applications, must be installed on Windows as follows:

1. Double-click on the dmexpress_<version>_windows_x86.exe file that you extracted from the download.
2. Follow the on-screen instructions.
   a. When prompted, select the option to start a free trial. The trial has a duration of 30 days, starting from the first time you run DMExpress.
   b. When prompted, the DBMS and SAP verification screens can be skipped.

2.5 Configure the Windows Environment

2.5.1 Addressing the VM

The VM external hostname is broadcast to your host Windows system using NetBIOS, so you can reference it when connecting to the VM from the DMX-h ETL Workstation or other tools.

Since the VM is configured to use Network Address Translation (NAT), it is possible to have conflicts if multiple people on the same network are using the Test Drive VM. If this is an issue, you can edit the Windows `hosts` file as an Admin user and add an entry for the IP address of the VM (shown when you connect to the VM) with the hostname `DMXhTstDrvMapR`. For example:

```
192.168.137.128 DMXhTstDrvMapR
```

Since the local hosts file overrides the network broadcast, it will pick up your local VM rather than someone else's on the network.

2.5.2 Define the MapR-FS Cluster User

When running the DMExpress GUI and attempting to connect to MapR-FS at design time for file browsing and sampling, DMExpress uses the Windows login user by default for the HDFS connection. However, since the default configuration for MapR-FS requires a cluster user to gain access, the connection will fail unless you define DMX_HDFS_USER as the cluster user (dmxdemo) in the Windows system environment variables before starting the GUI.

3 Start your Test Drive!

3.1 Use Case Accelerators

There are two broad categories of use case accelerators included on the VM in the `/UCA` folder:

- DMExpress Hadoop ETL Jobs
  - Jobs that are eligible for the DMX-h Intelligent Execution Layer (IEL) are created as "standard" DMExpress jobs and are found in a subdirectory named DMXStandardJobs within the example directory structure. When run in Hadoop, they are automatically converted to MapReduce jobs.
  - Jobs that are not currently supported for IEL are created as user-defined MapReduce jobs and are found in a subdirectory named DMXUserDefinedMRJobs within the example directory structure. This folder is also present for IEL-eligible jobs to demonstrate how those jobs would be defined as explicit MapReduce jobs, but the IEL solution is the recommended one to use when available.
- DMExpress HDFS Load/Extract Jobs – these are standard DMExpress jobs that are run on the edge node for extracting and loading HDFS data. They are found in a subdirectory named DMXHDFSJobs within the example directory structure.

A brief description of each use case accelerator is provided below, with links to more detailed descriptions:

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.
Joins and Lookups	Join Large Side | Small Side	Performs an inner join between a small distributed cache file and a large HDFS file.
	Join Large Side | Large Side	Performs a join of two large files stored in HDFS.
	File Lookup	Performs a lookup in a small distributed cache file while processing a large HDFS file.
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 Translation and Connectivity	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.
	Direct Mainframe Redefine Extract & Load	Loads one file residing on a remote mainframe system to HDFS, interpreting REDEFINES clauses and converting to ASCII displayable text.
	Mainframe Redefine File Load	Same as Direct Mainframe Redefine, except that the mainframe file is loaded to HDFS from the 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.

3.2 Running the Use Case Accelerators

Running the use case accelerators is as simple as the following:

1. Log in to the VM as described in section 2.3 and run the prep script to pre-load the sample data to HDFS. The script is located in `$DMXHADOOP_EXAMPLES_DIR/bin`, which is in the path.

   a. This can be done for all use case accelerators using the `ALL` option:

   ```
   prep_dmx_example.sh ALL
   ```

   b. Or it can be done for the specified space-separated list of folder names under `/UCA/Jobs`. For example:

   ```
   prep_dmx_example.sh FileCDC WebLogAggregation
   ```

2. On the Windows Workstation, start the DMExpress Job Editor, and run the desired use case accelerator(s) as follows:

   a. Select **File->Open Job…**, select the **Remote Servers** tab, double click on **New file browsing connection**, specify the connection as follows, and click **OK**:

      i. **Server**: DMXhTstDrvMapR

      ii. **Connection type**: Secure FTP

      iii. **Authentication**: Password

      iv. **User name**: dmxdemo

v. **Password**: dmxdemo

b. Open the desired job as follows:

i. Browse to the location of the job you want to run in one of the following folders as described earlier:

```
/UCA/Jobs/<JobName>/DMXStandardJobs
/UCA/Jobs/<JobName>/DMXUserDefinedMRJobs
/UCA/Jobs/<JobName>/DMXHDFSJobs
```

ii. Select `J_<JobName>.dxj` (or `MRJ_<JobName>.dxj`, as applicable).

iii. Click on **Open**.

c. Click on the **Run** button.

i. Click on **Select Server…**, click on the **UNIX** tab, enter **DMXhTstDrvMapR** for the server, enter the **User name** and **Password** as indicated in step a, and click **OK**.

ii. Select **Run on Hadoop cluster**, and click **OK**.

3. If you want to sample HDFS data, you first need to create an HDFS file browsing connection as follows:

a. Select **File->Open Job…**, select the **Remote Servers** tab, and double click on the **New file browsing connection** entry at the top of the list.

b. Populate the **File Browsing Connection** dialog as follows and click **OK**:

i. Set the **Server** to **DMXhTstDrvMapR**.

ii. Set the **Connection type** to **Hadoop Distributed File System (HDFS)**.

iii. Leave the **Method** for connecting as is – **HTTP (WebHDFS, HttpFS)** – unless you know you are using HTTPS or HFTP.

iv. Set the **Port number** to 14000 for MapR.

c. Click **Cancel** to dismiss the **Browse** dialog without actually selecting a file – it was just opened for the purpose of creating the file browsing connection, which should now be visible in the **Remote Servers** list.

d. You can now sample HDFS data by right clicking on an HDFS source or target in the task tree and selecting **Sample…** from the pop-up menu.

3.3 Additional Information

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

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

About Syncsort

Syncsort provides enterprise software that allows organizations to collect, integrate, sort, and distribute more data in less time, with fewer resources and lower costs. Thousands of customers in more than 85 countries, including 87 of the Fortune 100 companies, use our fast and secure software to optimize and offload data processing workloads. Powering over 50% of the world's mainframes, Syncsort software provides specialized solutions spanning "Big Iron to Big Data", including next gen analytical platforms such as Hadoop, cloud, and Splunk. For more than 40 years, customers have turned to Syncsort's software and expertise to dramatically improve performance of their data processing environments, while reducing hardware and labor costs. Experience Syncsort at www.syncsort.com.

Syncsort Inc. ***50 Tice Boulevard, Suite 250, Woodcliff Lake, NJ 07677*** ***201.930.8200***