syncsort

DMX-h Quick Start for Hortonworks Sandbox

Table of Contents

1 Introduction

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

The DMX-h ETL Test Drive on the Hortonworks Sandbox 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 Spark/MapReduce skills.

The Test Drive provides a ready-to-use virtual machine (VM) containing the Hortonworks HDP Hadoop distribution, pre-installed with 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 (~12 GB zipped) that contains:

- A Test Drive VM (unzips to ~19 GB, details below)
- DMX-h Workstation software (~329 MB)

The Test Drive VM, hostname dmxhtstdrvhdp, is based on the Hortonworks Sandbox VM, with additional components pre-installed. It includes the following:

- Linux CentOS 6.7
- Hortonworks 2.4 Hadoop distribution, MRv2
- DMX-h ETL version 9.0 (~215 MB)
- Use case accelerators (~2 MB)
- Sample data for running the use case accelerators (~1.7 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.

2.1 System Requirements for the Test Drive VM

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

- 64-bit Windows OS
- Windows 7/8/10, Windows Server 2008 or higher
- Minimum 30 GB hard disk space
- Memory: 8GB
- VMware Player or VMware Workstation installed, with the following sub-requirements:
  - The minimum VMWare/Player version must be one of the following:
    - 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.
  - Virtualization technology 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, refer to DMX-h Quick Start for your 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>_HDP<version>_VMv<version>.vmx`), 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, login to the VM using the following credentials:

   User Name: `dmxdemo`

   Password: `dmxd3mo`

The `dmxdemo` user has sudo privileges for root access, the password for which is the same. The Ambari Admin login credentials are the same as well.

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.

2.4.1 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 `dmxhtstdrvhdp`. For example:

```
192.168.137.128 dmxhtstdrvhdp
```

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

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 – these are standard DMExpress jobs that can be intelligently executed in Spark or MapReduce, and are found in a subdirectory named DMXStandardJobs within the example directory structure.
- 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.
Message Queues	Fraud Detection with Apache Kafka	Reads from and writes to Apache Kafka message queue topics in a simplified fraud detection example, where non-fraudulent transactions are written to a Hive table.

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.
	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.
	Mainframe Variable Processing	Loads mainframe variable length EBCDIC-encoded files to HDFS, converting to DMX-h's Mainframe Hadoop Distributable format for subsequent distributed processing.
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 running the prep script and then running the jobs.

3.2.1 Run the Prep Script

Login to the VM as described in section 2.3 and run the prep script, located in `$DMXHADOOP_EXAMPLES_DIR/bin` (which is in the path), in one of the following ways to pre-load the sample data to HDFS:

- Run it for all use case accelerators using the `ALL` option:

```
prep_dmx_example.sh ALL
```

- Run it for the specified space-separated list of folder names under `/UCA/Jobs`. For example:

```
prep_dmx_example.sh FileCDC WebLogAggregation
```

3.2.2 Run the Jobs

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

1. Select **File->Open Job…**, click on the **Remote Servers** tab, click on **New file browsing connection**, specify the connection as follows, and click **OK**:
   - **Server**: `dmxhtstdrvhdp`
   - **Connection type**: Secure FTP
   - **Authentication**: Password
   - **User name**: `dmxdemo`
   - **Password**: `dmxd3mo`

2. Open the desired job as follows:

   a. Browse to the location of the job you want to run. Depending on the UCA, the job will be in one of the following folders as described earlier:

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

   Note: In some UCAs, there will also be a subfolder named `DMXUserDefinedMRJobs`, which contains the user-defined MapReduce solution for reference. Details can be found in the corresponding UCA guide.

   b. Select `J_<JobName>.dxj`.

   c. Click on **Open**.

3. Click on the **Run** button.

   a. In the **DMExpress Server Connection** dialog (click on **Select Server…** in the **Run Job** dialog if needed), select the **UNIX** tab, enter `dmxhtstdrvhdp` for the server, enter the **User name** and **Password** as indicated above, and click **OK**.

   b. In the **Runon** section of the **Run Job** dialog, specify the framework in which to run the job:

      i. To run in Spark, select **Spark** for the **Framework**, click on **Define** for the **Spark master URL**, select **yarn** for the **Spark cluster type**, click **OK** to dismiss the **Spark Master URL** dialog, then click **OK** in the **Run job** dialog.

      ii. To run in MapReduce, select **MapReduce** for the **Framework**, and click **OK**.

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.

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***