syncsort

DMX-h Quick Start for your Hadoop Cluster

Table of Contents

1 Introduction

DMX-h ETL is Syncsort's high-performance ETL software for Hadoop. It combines powerful ETL data processing capabilities with the enhanced performance and scalability of Hadoop, without the need to learn complex Spark/MapReduce programming skills.

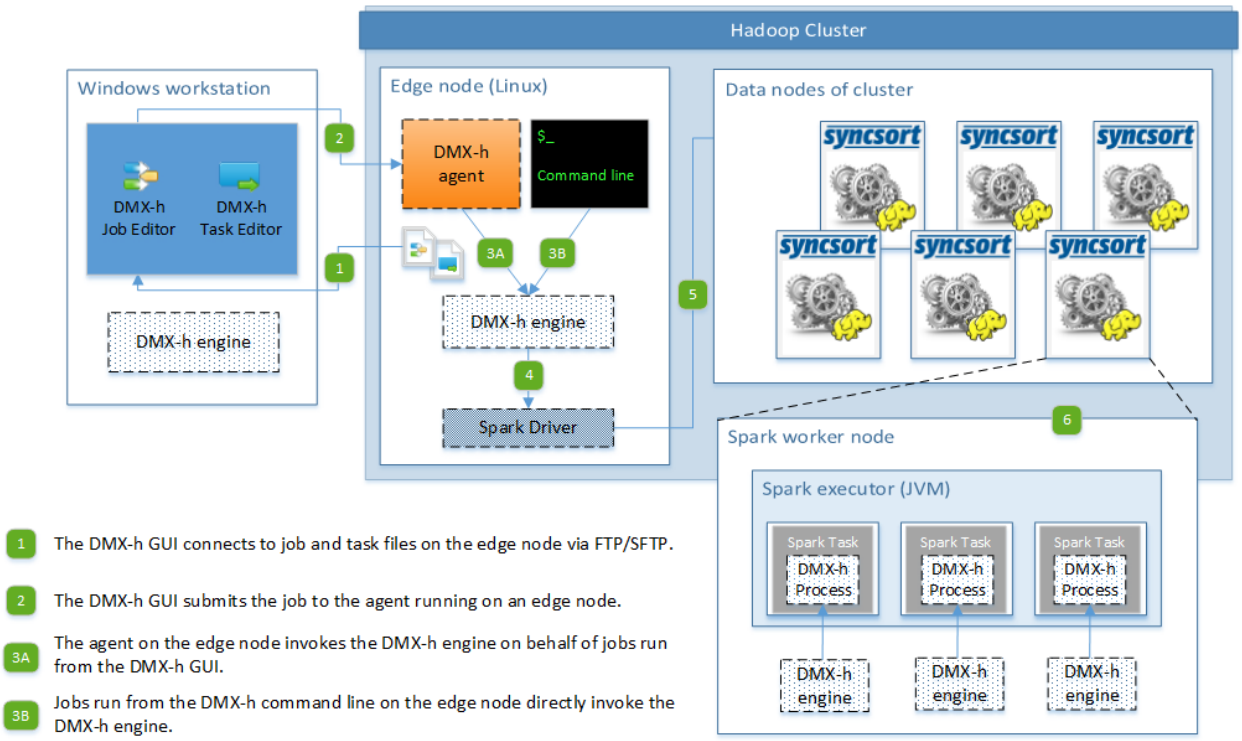
A downloadable package of use case accelerators demonstrates how common ETL applications, easily developed in DMExpress, can be run in the Hadoop environment.

Installing the DMX-h software and setting up the use case accelerators in your Hadoop cluster is fast and easy. Just follow the instructions in this document, and try out DMX-h for yourself.

1.1 DMX-h ETL Architecture

DMX-h ETL jobs and tasks are created in the DMX-h GUI (Job and Task Editors) on the Windows workstation. Jobs can then be submitted to the Spark/MapReduce cluster from either the Windows GUI or from the Linux command line on the edge node.

1.1.1 DMX-h Spark Execution Architecture



1 The DMX-h GUI connects to job and task files on the edge node via FTP/SFTP.

2 The DMX-h GUI submits the job to the agent running on an edge node.

3A The agent on the edge node invokes the DMX-h engine on behalf of jobs run from the DMX-h GUI.

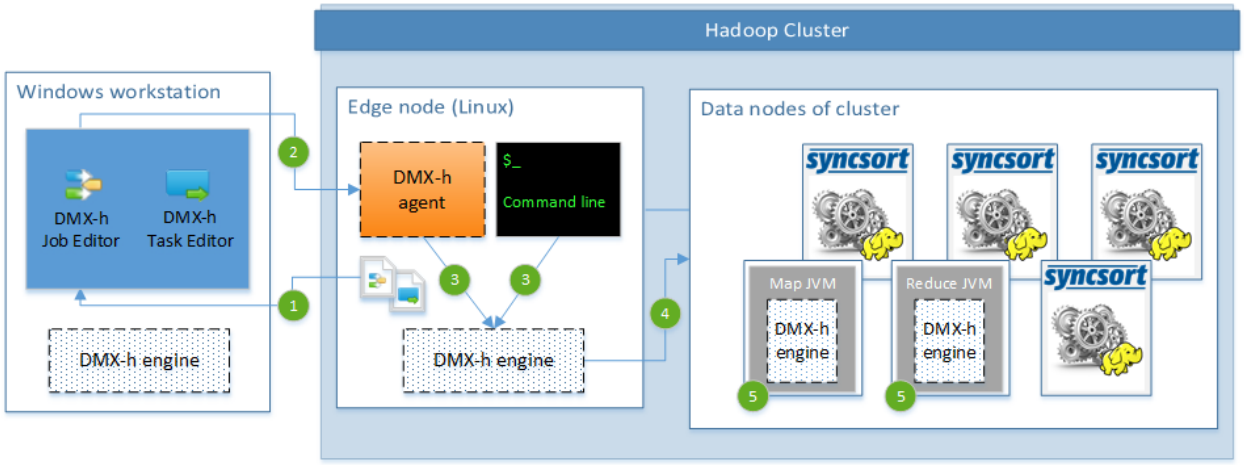
3B Jobs run from the DMX-h command line on the edge node directly invoke the DMX-h engine.

4 DMX-h breaks the job down into one or more Spark jobs and submits them to the Spark master.

5 Based on Spark configuration properties, Spark requests the resource manager (YARN, Mesos, or Spark standalone) to spin up an appropriate number of executors on the Spark worker nodes.

6 Multiple Spark tasks are created within a Spark executor. These tasks invoke the DMX-h engine simultaneously for processing separate blocks of data.

1.1.2 DMX-h MapReduce Execution Architecture



1. DMX-h GUI opens job and task files from the edge node via FTP/SFTP.
2. DMX-h GUI submits a job to the agent running on an edge node.
3. The agent (or a job run from the command line) on the edge node invokes the engine.
4. The engine submits the job, task, and metadata files to distributed cache in the Hadoop cluster.
5. Each map and reduce task executes the job by invoking the DMX-h engine as a child process inside the JVM.

2 Getting Started

2.1 Getting the DMX-h Components

If you haven't already done so, go to the Syncsort Test Drive site, select the desired Hadoop distribution, and submit the subsequent registration form. You will then be able to download a zip file that includes the following:

- DMX-h software for Linux:

  ```
  dmexpress_<DMExpress version>_en_linux_2-6_x86-64_64bit.tar
  ```

- DMX-h software for Windows:

  ```
  dmexpress_<DMExpress version>_windows_x86.exe
  ```

- DMExpress Installation Guide

2.2 Getting the Use Case Accelerators

The use case accelerators, a pre-developed set of DMX-h ETL example applications, along with a set of sample data needed to run them, can be downloaded from Guide to DMX-h ETL Use Case Accelerators.

2.3 Getting Help

For assistance with DMX-h ETL, please visit the DMX-h ETL group in the Syncsort User Community.

3 Installing DMX-h Software

3.1 Installing DMX-h on Linux in your Hadoop Cluster

DMX-h must be installed on all nodes in the Hadoop cluster, as well as on an edge node that has access to the Hadoop cluster. Follow the instructions for Manual/Silent installation in the "Step-by-step Installation on a Hadoop Cluster" section of the Installation Guide delivered with your download.

The DMExpress service, dmxd, only needs to be running on the edge node, from which DMX-h jobs can be run from the command line, or through which they can be run from the Windows GUI. After completing the installations, follow the directions in the Installation Guide for installing the service on the edge node.

Ensure that the user who will be running DMX-h jobs has access to the Hadoop cluster and a home directory in HDFS.

3.2 Installing DMX-h Workstation on Windows

To view the sample DMExpress jobs/tasks and develop your own solutions, you need to install the DMX-h Workstation software on a Windows machine as follows:

1. Double-click on the downloaded DMExpress Windows installation file.
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.

When attempting to browse/sample HDFS files via the DMExpress GUI at design time, DMExpress uses the Windows login user to make the HDFS connection. If your HDFS configuration requires a cluster user to gain access (as is the default in MapR-FS), you need to define DMX_HDFS_USER as the cluster user in the Windows system environment variables before starting the GUI.

3.3 Installing the Use Case Accelerators

1. Using a file transfer utility, connect to the edge node as the user who will run DMX-h jobs, and copy (in binary mode) the downloaded zipped tar files `DMX-h_UCA_Solutions.tar.gz` and `DMX-h_UCA_Data.tar.gz` to the directory in which you want to store the example files.
2. Log into the edge node as the DMX-h user and extract both tar files as follows:

```
tar xvof DMX-h_UCA_Solutions.tar.gz
tar xvof DMX-h_UCA_Data.tar.gz
```

   This will create `Data`, `Jobs`, and `bin` directories under the directory you will later designate as the value of `$DMXHADOOP_EXAMPLES_DIR` as described in the Guide to DMX-h ETL Use Case Accelerators.

4 Using DMX-h

4.1 Use Case Accelerators

Syncsort provides a set of use case accelerators that cover a variety of common ETL use cases to quickly and easily demonstrate both the development and running of DMX-h ETL jobs in Hadoop. A brief description of each one 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.

4.2 Running the Use Case Accelerators

Follow the instructions for running the use case accelerators in your own cluster in the Guide to DMX-h ETL Use Case Accelerators.

4.3 Additional Information

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***