INTEGRATING BIG DATA... SMARTER

syncsort

Syncsort Ironcluster™ ETL Docker Edition

User Guide

© Syncsort® Incorporated, 2014

All rights reserved. This document contains proprietary and confidential material, and is only for use by licensees of DMExpress. This publication may not be reproduced in whole or in part, in any form, except with written permission from Syncsort Incorporated. Syncsort is a registered trademark and DMExpress is a trademark of Syncsort, Incorporated. All other company and product names used herein may be the trademarks of their respective owners.

The accompanying DMExpress program and the related media, documentation, and materials ("Software") are protected by copyright law and international treaties. Unauthorized reproduction or distribution of the Software, or any portion of it, may result in severe civil and criminal penalties, and will be prosecuted to the maximum extent possible under the law.

The Software is a proprietary product of Syncsort Incorporated, but incorporates certain third-party components that are each subject to separate licenses and notice requirements. Note, however, that while these separate licenses cover the respective third-party components, they do not modify or form any part of Syncsort's SLA. Refer to the "Third-party license agreements" topic in the online help for copies of respective third-party license agreements referenced herein.

Table of Contents

1 Introduction 1
1.1 What's Included 1
1.2 Support 1
2 Installation and Setup 2
2.1 Installing Docker 2
2.2 Launching an Ironcluster ETL Container 2
2.3 Licensing Ironcluster ETL in the Docker Container 3
2.4 Installing the Ironcluster ETL Windows Workstation 3
3 Use Case Accelerators 5
3.1 Running the Use Case Accelerators 5
4 Container Management 6
4.1 Naming the Ironcluster ETL Container 6
4.2 Stopping the Ironcluster ETL Container 6
4.3 Starting the Ironcluster ETL Container 6
4.4 Removing the Ironcluster ETL Container 6

1 Introduction

Syncsort Ironcluster ETL is a powerful, low-cost, and easy-to-use data integration tool that includes the following:

- A graphical user interface for developing and maintaining ETL jobs
- A library of Use Case Accelerators to quickly deploy typical ETL applications such as web log processing, CDC, multi-level aggregations, data cleansing, and more
- A self-tuning ETL engine for maximum scalability at the lowest price
- Ability to access and transform virtually any data from relational to social, structured to unstructured, and even mainframe and Hadoop

Docker is an open-source environment for creating lightweight, portable application containers that can be run anywhere.

Syncsort Ironcluster ETL Docker Edition makes Ironcluster ETL even easier to deploy and scale, with the following benefits:

- Deploy in just minutes, with a small footprint
- Run anywhere – physical or virtual, cloud or on premises
- Facilitate private & public cloud deployments
- Move seamlessly from development to testing and into production
- Easily scale with your data and workloads, with no tuning or configuration changes
- Eliminate third-party dependencies

Take a test drive of the Ironcluster ETL Docker container, available in DockerHub, along with the Ironcluster ETL Workstation and license key downloadable from Syncsort upon registration.

1.1 What's Included

The Ironcluster ETL Docker image includes the following:

- Ironcluster ETL
- Use Case Accelerators – common ETL use case examples

In addition to launching the container image, you will need to go to the Syncsort website to register. Upon registration, you will receive an email with a 30-day trial license key along with a link to download the Windows Workstation software. The key is good for both the container image and the Windows software.

1.2 Support

Support is available for Ironcluster ETL via the Syncsort Online Community.

Also, check out our resources page, where you will find helpful videos and links to the community, free online training, and this document.

2 Installation and Setup

2.1 Installing Docker

The Docker framework can be downloaded and installed on various platforms from the Docker site. Follow the instructions there for your chosen platform.

The machine (or VM) on which you install Docker is referred to as the Docker host. Once installed, you can run the `docker` client to get/list images and list/run/stop/start containers.

- **Images** are the read-only templates that define what a container built from that image would contain.
- **Containers** are running instances of the image from which they were launched.

For additional details on Docker, see the following pages on their website:

- Understanding Docker
- Command Line Reference

2.2 Launching an Ironcluster ETL Container

To launch Ironcluster ETL in a Docker container, do the following:

1. Go to the Syncsort Docker ETL page to register; you will receive an email containing a trial license key and a link to the Windows workstation installation file.
   a. Save the license key file on the Windows machine where you will install the Ironcluster ETL Windows Workstation. You will use this same key to license Ironcluster in the container.
2. On your Linux machine, do the following:
   a. Set an environment variable to represent the port on the Linux host that will redirect to port 22 on the Ironcluster ETL Docker container for SFTP connectivity to the container. The usual port is 22, but if your Linux machine is already using this port for its own SSH/SFTP connectivity, the container can't share it. A recommended port is 2222 as long as it's not already in use. For example:

```
export DMX_SFTP_PORT=2222
```

   b. Run the following command to instantiate the Ironcluster ETL Docker image as a running container; it will download the image if not already available on your machine. Port 32636 is used for the `dmxd` service, which allows jobs to be initiated from the Windows Workstation and run on the Ironcluster ETL server in the container.

```
sudo docker run -d -p $DMX_SFTP_PORT:22 -p 32636:32636 \
   syncsort/ironcluster_etl
```

2.3 Licensing Ironcluster ETL in the Docker Container

The trial license key file needs to be applied to Ironcluster ETL in the Docker container as follows:

1. Find the IP address of the Docker host by running `ifconfig` and finding the `inet addr` in the `eth0` interface for your Linux machine. For example:

```
eth0      Link encap:Ethernet  HWaddr 00:0C:29:15:72:D0
          inet addr:192.168.244.130  Bcast:192.168.244.255
             Mask:255.255.255.0
          inet6 addr: fe80::20c:29ff:fe15:72d0/64 Scope:Link
          ...
```

2. Use a utility such as `scp` to copy the license key file to the `/usr/dmexpress` directory of the container as the `syncsort` user. For example, with `scp`, the command would be as follows, where *`host_IP`* is the IP address of the Docker host:

```
scp -P 2222 DMExpressLicense.txt \
syncsort@host_IP:/usr/dmexpress
```

3. Use a utility such as `ssh` to connect to the container on the Docker host as the `syncsort` user. For example, with `ssh`, the command would be as follows, where *`host_IP`* is the IP address of the Docker host:

```
ssh syncsort@host_IP -p 2222
```

4. You will be prompted to apply the license key; enter `Y` to have the applykey application run automatically; enter `N` to run it later as indicated.
5. When prompted by the `applykey` application, enter the location of the `DMExpressLicense.txt` file that you just copied, i.e. `/usr/dmexpress`.
6. Review and accept the license agreement.

2.4 Installing the Ironcluster ETL Windows Workstation

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

1. Follow the link sent to you via email to download and extract the Ironcluster ETL Windows Workstation.
2. Double-click on the `dmexpress_`*`version`*`_windows_x86.exe` file.
3. Follow the on-screen instructions, and be prepared to point to the Windows location where you saved the trial license key.
4. After installation is complete, the DBMS and SAP verification screens can be skipped.

5. Set an environment variable for the container port on the Docker host to the same value that was specified on the Linux machine as follows:

   a. Open the Windows **Control Panel**, select **User Accounts**, then **Change my environment variables**.

   b. Click on **New…** under the User variables section of the **Environment Variables** dialog, and specify the following:

      Variable name: `DMX_SFTP_PORT`
      Variable value: <enter the port you set for this variable on the Linux machine, e.g. `2222`>

   c. Click **OK** successively to dismiss the dialogs.

3 Use Case Accelerators

Syncsort Ironcluster ETL comes with a built-in set of use case accelerators to quickly and easily demonstrate both the development and running of Ironcluster ETL jobs covering a variety of common ETL use cases.

The following use case accelerators are included in the Ironcluster ETL container at `/usr/dmexpress/examples/UseCaseAccelerators/`*`UCA name`*`/`. See "Use Case Accelerators" in the DMExpress Help for details on each application and how to run it, noting that the ones that interface with Excel do not run on the Linux server.

- **BatchChangeDataCapture** – Compare two versions of the same data to identify added, deleted, or updated records
- **DataCleansing** – Standardize and merge data files with different layouts
- **MultiLevelAggregation** – Aggregate data across multiple dimensions
- **WebLogProcessing** – Analyze web traffic data

3.1 Running the Use Case Accelerators

1. Start the DMExpress Job Editor from the Windows start menu by selecting **All Programs->DMExpress->DMExpress Job Editor**.
2. Select **File->Open Job…**, select the **Remote Servers** tab and double-click on **New file browsing connection**. Specify the connection information as follows, then click on **Verify connection**, then **OK**, and then **OK** again.

Server:	*`Docker host IP address`*
Connection type:	Secure FTP
Authentication:	Password
User name:	`syncsort`
Password:	`syncsort`

3. Double-click on the newly created file browsing connection, browse to the location of the job you want to run, and double click on the `.dxj` file to open up the job in the DMExpress Job Editor.
4. Set the Server to the Ironcluster ETL container as follows:
   a. Click on the **Status** button in the Job Editor toolbar.
   b. 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 Docker host IP in **Connect to server**, enter the **User name** and **Password** as indicated in step 2 above, and click **OK**, then **Close**.
   c. Click on the **Run** button in the Job Editor toolbar, then click **OK** in the **Run Job** dialog.
   d. 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 and job statistics.
5. Once you've successfully run some of the use case accelerators, you are ready to start creating your own jobs. See the DMExpress Help for more information.

4 Container Management

4.1 Naming the Ironcluster ETL Container

To make it easier to identify the container, you can create it with a name by specifying the `--name="my_name"` option to the `docker run` command. Then, in any command where you would specify the container ID, you can specify the container name instead.

4.2 Stopping the Ironcluster ETL Container

To stop the Ironcluster ETL container, run the following commands, where the first one lists the running containers and the second one stops the specified container ID:

```
sudo docker ps
sudo docker stop container_ID     // or container_name if created
```

4.3 Starting the Ironcluster ETL Container

To start a previously stopped container, run the following command:

```
sudo docker start container_ID    // or container_name if created
```

4.4 Removing the Ironcluster ETL Container

To remove the Ironcluster ETL container, run the following command:

```
sudo docker rm container_ID // or container_name if created
```

This will permanently remove the container and any files stored directly in it.

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