



DMX-h ETL Use Case Accelerator
File Join Small



© Syncsort® Incorporated, 2015

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
2	File Join Small in DMX-h ETL	2
2.1	Map Step.....	2
2.1.1	MT_Join.dxt.....	3
Appendix A	File Join Small Standard (non-MapReduce) Solution	4

1 Introduction

DMX-h ETL can efficiently join a small data set with a large one in the Hadoop MapReduce framework. In this use case accelerator, DMX-h ETL performs an inner join of two TPC Benchmark H (TPC-H) data sets: a small supplier file and a large line item file.

The use case accelerators can be run outside of Hadoop, either on a Windows workstation or on the Linux edge node, for development and testing before running in a Hadoop cluster. For guidance on setting up and running the examples both outside and within Hadoop, see [Guide to DMX-h ETL Use Case Accelerators](#).

As a point of reference, the standard (non-MapReduce) solution for this use case accelerator is described in Appendix A.

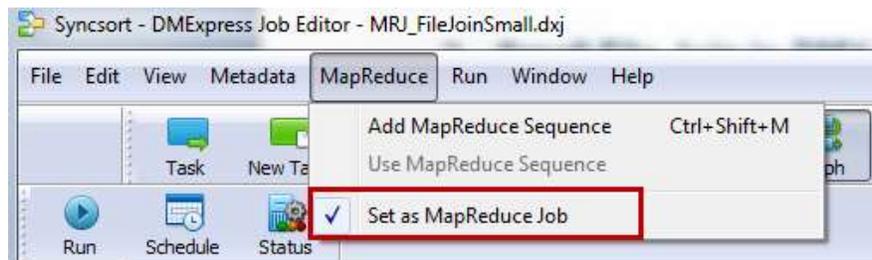
2 File Join Small in DMX-h ETL

The File Join Small job in DMX-h contains only a map step, as follows:



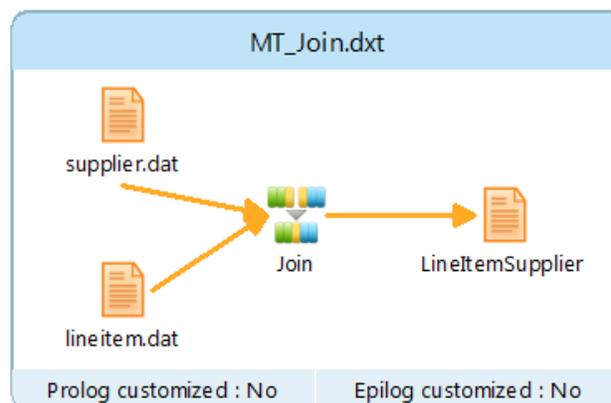
The map step reads two input files and produces a single output stream based on the join key.

The job is defined to be a map-only job by selecting the option **MapReduce->Set as MapReduce Job** as follows:



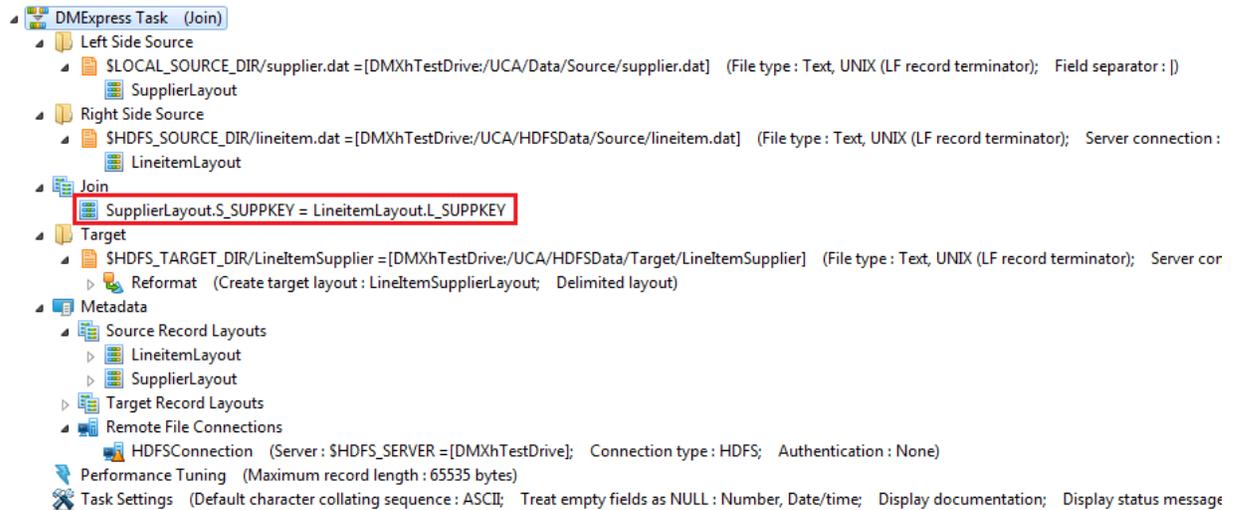
2.1 Map Step

The File Join Small map step in DMX-h consists of one task which reads both input files and joins them based on the join key.



2.1.1 MT_Join.dxt

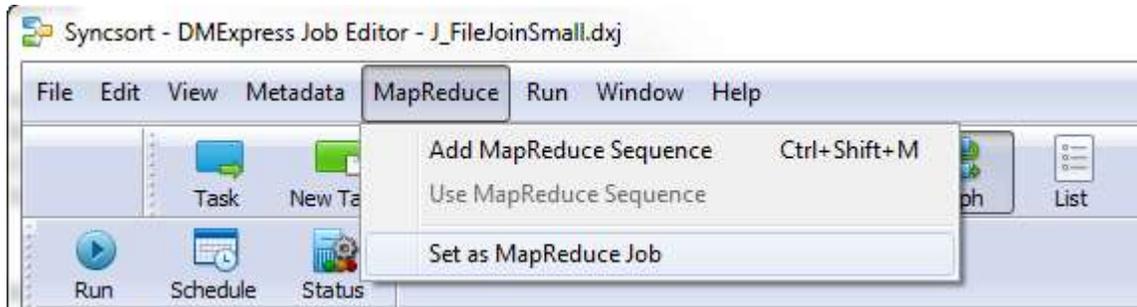
This task reads the small supplier data set directly from a file on the local file system and the large line item data set from a file on the HDFS and joins the two files based on the join key.



Appendix A File Join Small Standard (non-MapReduce) Solution

The standard (non-MapReduce) solution for this use case accelerator is nearly identical to the DMX-h ETL solution provided above, apart from the following:

1. Specify the sources and target files to be local as opposed to HDFS.
2. Uncheck the **MapReduce->Set as MapReduce Job** option as shown below:



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