syncsort

DMX-h ETL Use Case Accelerator

Mainframe Variable Processing

Table of Contents

1 Introduction

Mainframe variable length files are EBCDIC-encoded files with varying record lengths. Each record has a 4-byte prefix which provides the length of the record. As is, these files are non-splittable when stored in HDFS, which effectively means you cannot achieve distributed processing when running a Spark/MapReduce job on such a file.

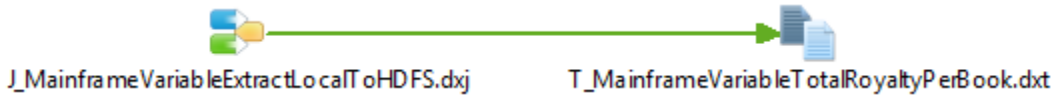
With DMX-h, you can ingest these files into HDFS in a distributable manner by converting them to DMX-h's Mainframe Hadoop Distributable format. You can then run Spark/MapReduce jobs on these EBCDIC-encoded files.

DMX-h ETL use case accelerators are developed as standard ETL jobs and can be run on an edge node, single cluster node, or in the Spark/MapReduce cluster without making any changes to the application.

For guidance on setting up and running this and other use case accelerators, see the Guide to DMX-h ETL Use Case Accelerators.

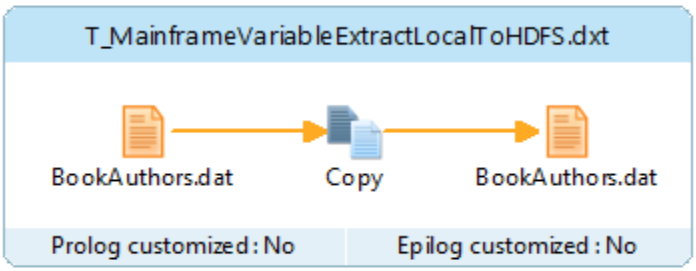
2 Mainframe Variable Processing with DMX-h

The Mainframe variable processing solution consists of a DMX-h job, J_MainframeVariableProcessingTotalRoyaltyPerBook.dxj, with one subjob followed by a copy task.



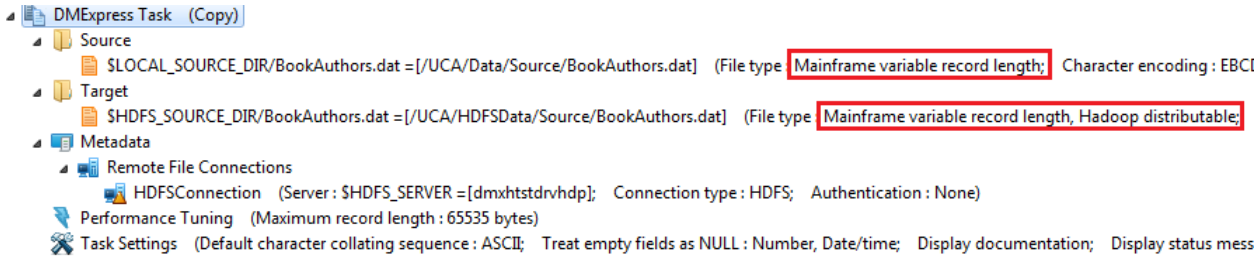
2.1 J_MainframeVariableExtractLocalToHDFS Subjob

This subjob consists of a single copy task that reads the mainframe variable length file and loads it to HDFS. Since it's an ingestion job, DMX-h will run it on the edge node.



2.1.1 T_MainframeVariableExtractLocalToHDFS Task

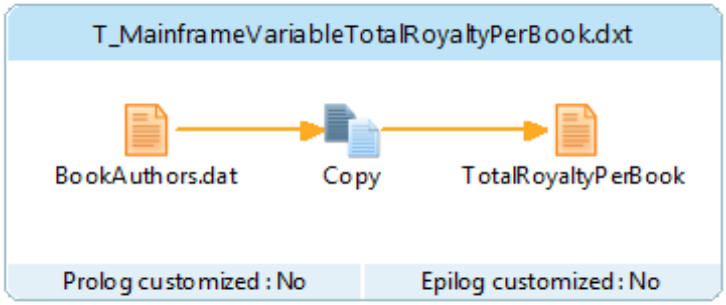
This copy task reads a mainframe variable length file from the local file system and ingests it into HDFS in "Mainframe variable record length, Hadoop distributable" format. You can specify a mainframe server connection in the source file if you are reading directly from the mainframe instead of local disk.



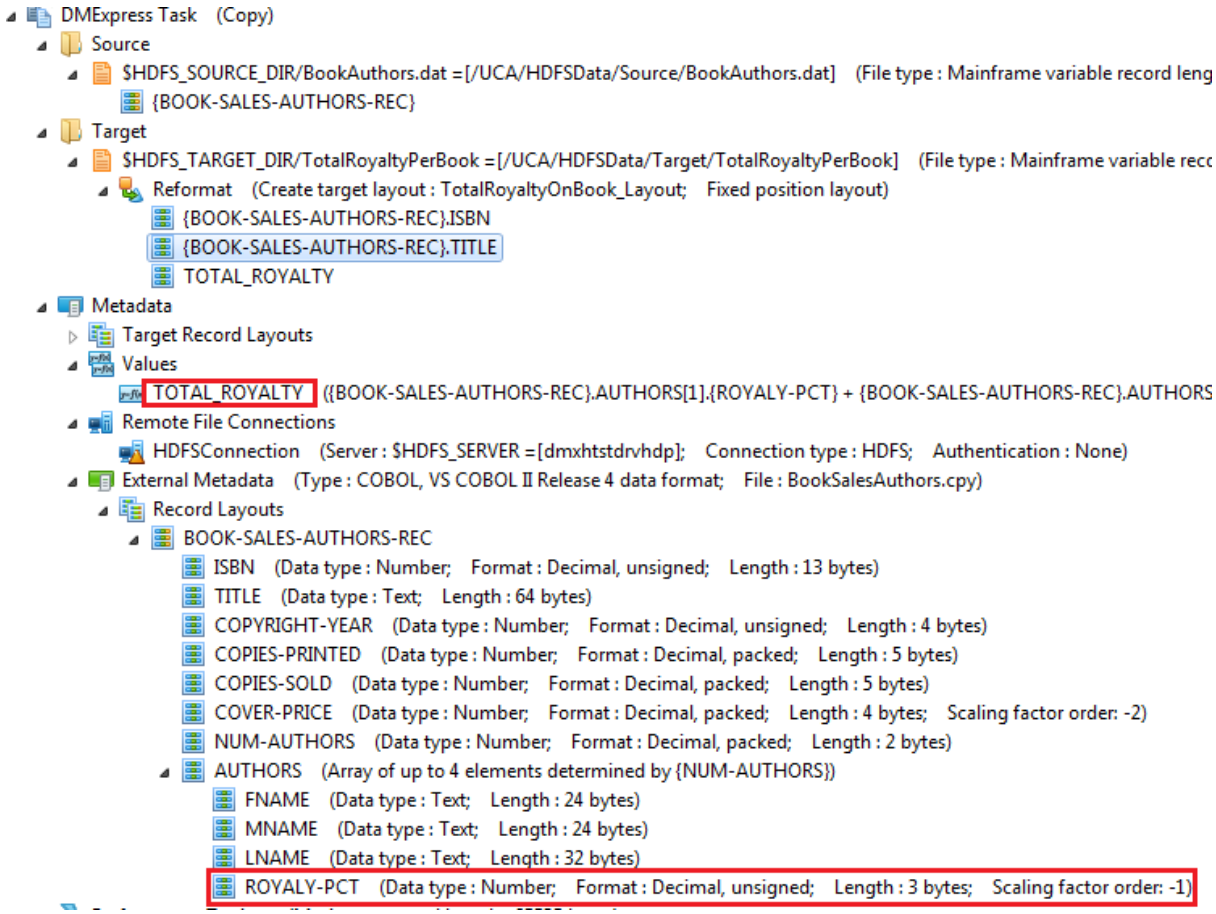
Since this task is ingesting a copy of the mainframe data into Hadoop without interpreting the records, there's no need to provide a copybook to this task.

2.2 T_MainframeVariableTotalRoyaltyPerBook Task

This task, run in the cluster, reads the output of the subjob, which is a Mainframe Hadoop Distributable file. It interprets the EBCDIC data using a copybook, and calculates the total royalty percentage on a book.



Royalty percentage is an element in an OCCURS DEPENDING ON array, the length of which depends on NUM-AUTHORS. We add the royalty percentage for each of the authors in the array to find the total royalty percentage per book. The target of this task is also a mainframe EBCDIC-encoded file, which can later be put back on the mainframe.



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***