syncsort

DMX-h ETL Use Case Accelerator

File Lookup

Table of Contents

1 Introduction

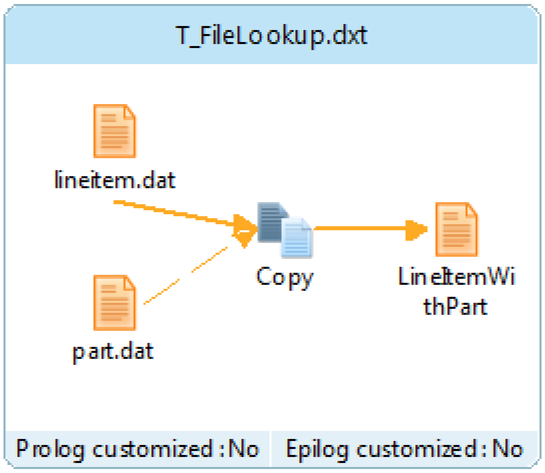
DMX-h ETL can efficiently look up a value in a dimension table in Spark or MapReduce using Intelligent Execution (IX). In this use case accelerator, DMX-h performs a lookup with two TPC Benchmark H (TPC-H) data sets: a line item file and a part file.

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 File Lookup with DMX-h IX

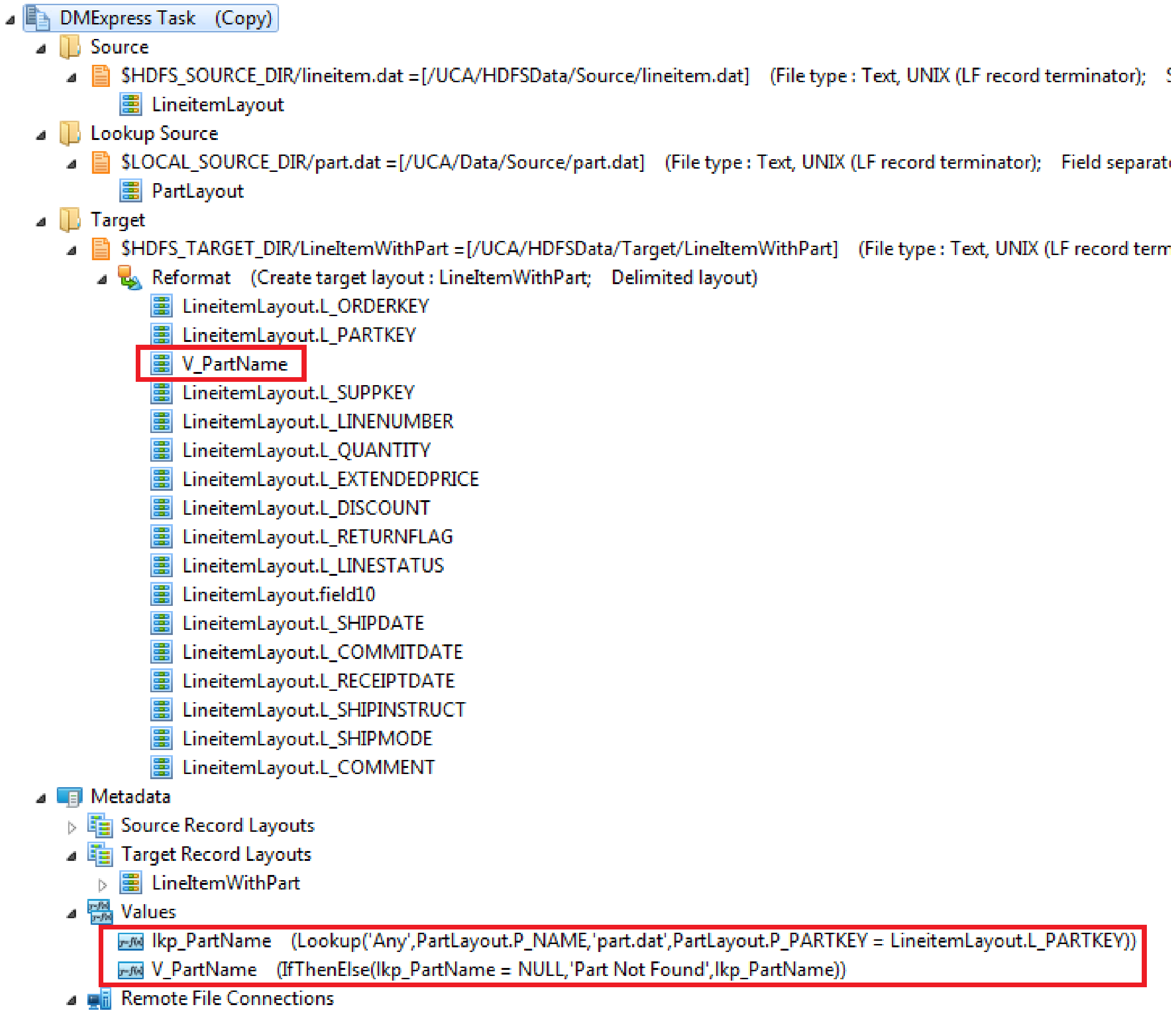
The File Lookup solution in DMX-h ETL consists of a job, J_FileLookup.dxj, with a single copy task that performs a lookup on a part file and outputs a copy of the line item source data matched with the part lookup data.



IX runs this as a map-only job. Each map node processes a subset of the lineitem.dat file and performs a lookup on the full part.dat file, which is a small distributed cache file.

2.1 T_FileLookup Task

This task calls the Lookup() function, which uses PARTKEY to find a matching part name (P_NAME) record in the "part.dat" lookup source file. An IfThenElse() condition is used to change NULL values to "Part Not Found". Finally, the part name from the lookup is added to the target record layout via a reformat, and the resulting records are written to the target.



- DMExpress Task (Copy)
  - Source
    - $HDFS_SOURCE_DIR/lineitem.dat =[/UCA/HDFSData/Source/lineitem.dat] (File type : Text, UNIX (LF record terminator); S
      - LineitemLayout
  - Lookup Source
    - $LOCAL_SOURCE_DIR/part.dat =[/UCA/Data/Source/part.dat] (File type : Text, UNIX (LF record terminator); Field separat
      - PartLayout
  - Target
    - $HDFS_TARGET_DIR/LineItemWithPart =[/UCA/HDFSData/Target/LineItemWithPart] (File type : Text, UNIX (LF record term
      - Reformat (Create target layout : LineItemWithPart; Delimited layout)
        - LineitemLayout.L_ORDERKEY
        - LineitemLayout.L_PARTKEY
        - V_PartName
        - LineitemLayout.L_SUPPKEY
        - LineitemLayout.L_LINENUMBER
        - LineitemLayout.L_QUANTITY
        - LineitemLayout.L_EXTENDEDPRICE
        - LineitemLayout.L_DISCOUNT
        - LineitemLayout.L_RETURNFLAG
        - LineitemLayout.L_LINESTATUS
        - LineitemLayout.field10
        - LineitemLayout.L_SHIPDATE
        - LineitemLayout.L_COMMITDATE
        - LineitemLayout.L_RECEIPTDATE
        - LineitemLayout.L_SHIPINSTRUCT
        - LineitemLayout.L_SHIPMODE
        - LineitemLayout.L_COMMENT
  - Metadata
    - Source Record Layouts
    - Target Record Layouts
      - LineItemWithPart
    - Values
      - lkp_PartName (Lookup('Any',PartLayout.P_NAME,'part.dat',PartLayout.P_PARTKEY = LineitemLayout.L_PARTKEY))
      - V_PartName (IfThenElse(lkp_PartName = NULL,'Part Not Found',lkp_PartName))
    - Remote File Connections

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***