syncsort

DMX-h ETL Use Case Accelerators

Mainframe CDC

Table of Contents

1 Introduction

Change data capture (CDC) is a process in which previous and current versions of a large data set are compared to determine the changes between the two versions. This use case accelerator implements CDC in DMX-h ETL where the original source files are located on a remote Mainframe Server and first ingested into HDFS before performing the CDC.

The use case accelerators are developed as standard DMExpress jobs with just minor accommodations that allow them to be run in or outside of Hadoop:

- When specified to run in the Hadoop cluster, DMX-h Intelligent Execution (IX) automatically converts them to be executed in Hadoop using the optimal Hadoop engine, such as MapReduce. In this case, some parts of the job may be run on the Linux edge node if not suitable for Hadoop execution.
- Otherwise, they are run on a Windows workstation or Linux edge node, which is useful for development and testing purposes before running in the cluster.

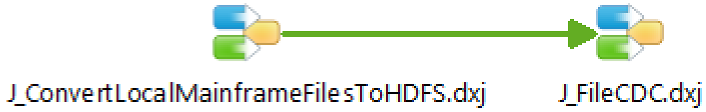
While the method presented in the next section is the simplest and most efficient way to develop this example, the more complex user-defined MapReduce solution is provided as a reference in Appendix A.

For guidance on setting up and running the examples both outside and within Hadoop, see Guide to DMExpress Hadoop Use Case Accelerators.

For details on the requirements for IX jobs, see “Developing Intelligent Execution Jobs” in the DMExpress Help.

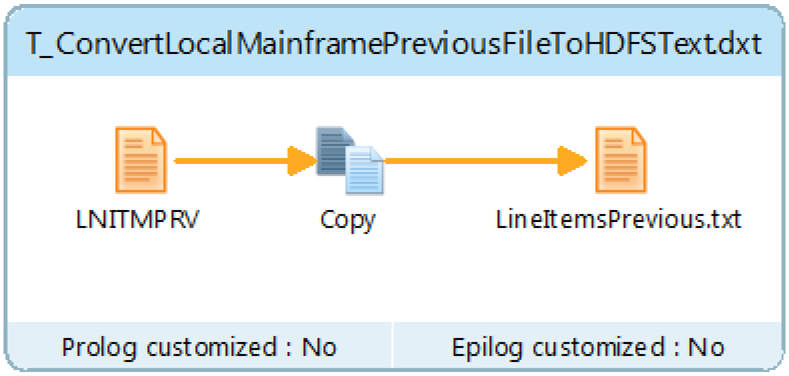
2 Mainframe CDC with DMX-h IX

The Mainframe CDC solution in DMX-h ETL consists of a job, J_MainframeCDCDemo.dxj, containing two subjobs. The first subjob uses a copy task to ingest the Mainframe data into HDFS, and the second subjob performs a CDC on the previous and current data once the HDFS load is completed.



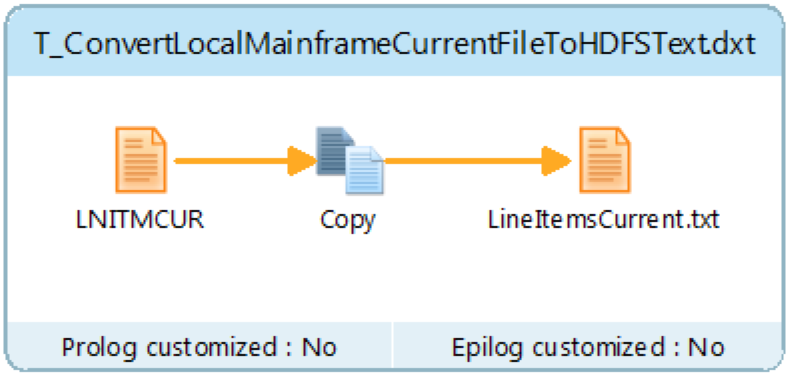
2.1 J_ConvertLocalMainframeFilesToHDFS Subjob

This edge node subjob consists of two independent tasks that concurrently copy the two respective source files from the Mainframe into HDFS, converting the data from EBCDIC fixed length to UTF-8 delimited text.



Prolog customized : No

Epilog customized : No



Prolog customized : No

Epilog customized : No

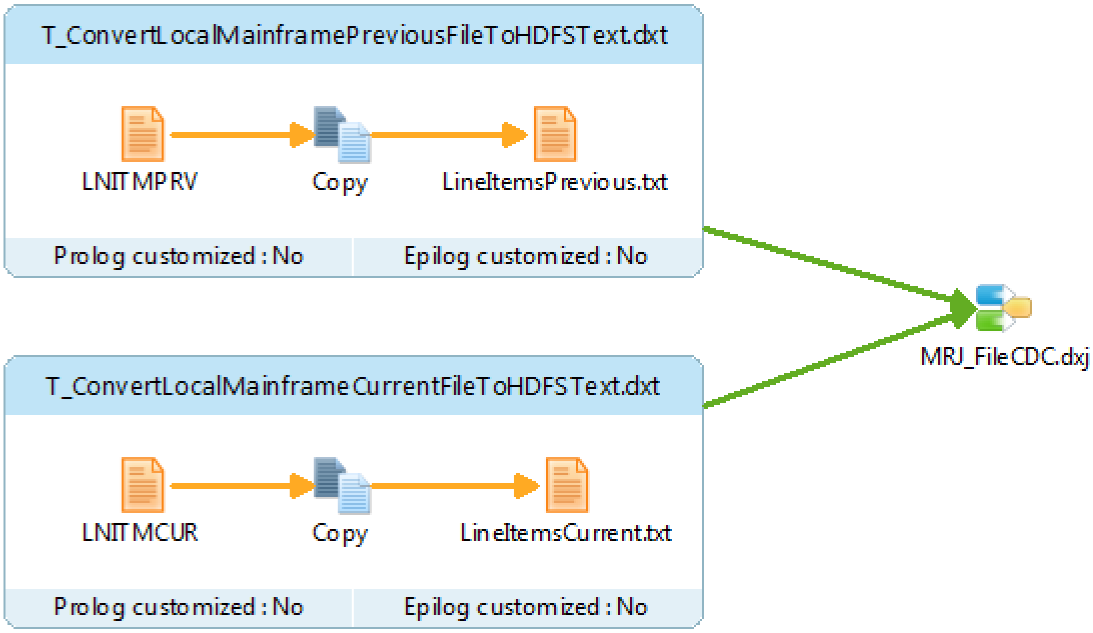
2.2 J_FileCDC Subjob

This MapReduce subjob reads the data that was loaded into HDFS and performs the CDC. It is the same job described in DMX-h Use Case Accelerator: File CDC.

Appendix A Mainframe CDC with DMX-h User-defined MapReduce

This user-defined MapReduce solution is provided as a reference in the event that particular knowledge of your application's data would benefit from manual control of the MapReduce process.

For the top-level job, the only difference between the "standard" version for running with IX and the user-defined MapReduce version is that the two copy tasks don't need to be nested inside of a subjob. However, they still run on the edge node to ingest the mainframe data into HDFS before invoking the FileCDC job as a user-defined MapReduce job.



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***