

# Syncsort DMX-h ETL Edition

A Smarter Approach to Hadoop ETL

Everything you need to turn Hadoop into a robust ETL solution

## Big Data Is Breaking Traditional ETL

For many years, organizations have struggled to implement and scale conventional ETL solutions. Today, Big Data is prompting them to look at Hadoop to process more data in less time and for less money. However, Hadoop is not a complete ETL solution. While it offers powerful utilities and massive horizontal scalability, it does not provide the set of functionality users need to deliver enterprise ETL capabilities. Hadoop promises incredible opportunities for organizations that know how to leverage its power. But, many early adopters are frustrated by the complexity and barriers that Hadoop presents. Enterprises are demanding a smarter ETL solution.

## Delivering Smarter ETL through Hadoop

Syncsort DMX-h ETL Edition is high-performance ETL software that turns Hadoop into a more robust and featurerich ETL solution, enabling users to maximize the benefits of MapReduce without compromising on capabilities, ease of use, and typical use cases of conventional data integration tools.

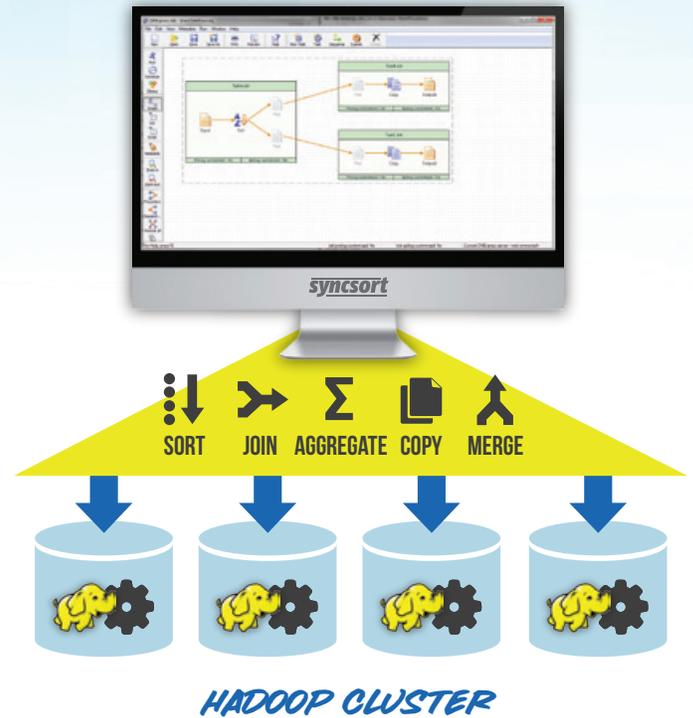
Accelerate your data integration initiatives and unleash Hadoop's potential with the only architecture that runs ETL processes natively within Hadoop.

- ➔ Fast-track your Hadoop productivity with the Use Case Accelerators, a library of pre-built development templates for common ETL use cases
- ➔ Eliminate the need for manual coding in Java, Pig, or HiveQL
- ➔ Develop and test graphically in Windows, prior to deploying in Hadoop
- ➔ Get smarter connectivity to all your data – including mainframe
- ➔ Improve Hadoop's scalability by maximizing the power and efficiency of each node in your cluster

All of this, plus enterprise-grade reliability and support at a price comparable only to open source solutions.

## Smarter Architecture - Get Faster Performance per Node!

DMX-h is the only ETL tool that executes within the Hadoop MapReduce framework via a pluggable sort enhancement, JIRA MAPREDUCE-2454, contributed by Syncsort and now part of Apache Hadoop. Other tools generate code (i.e. Java, Pig, HiveQL) that adds performance overhead and can become a nightmare to maintain and tune. DMX-h is not a



### EVERYTHING YOU NEED TO TURN HADOOP INTO A ROBUST ETL SOLUTION

- ➔ **Smarter Architecture** - No code generation. ETL engine runs natively within MapReduce
- ➔ **Smarter Connectivity** - One tool to connect all your data, even mainframe
- ➔ **Smarter Development** - Hadoop ETL without coding
- ➔ **Smarter Productivity** - Use Case Accelerators for common ETL tasks
- ➔ **Smarter Security** - Enterprise-grade security

### PLUS SMARTER HADOOP...

- ➔ Faster throughput per node
- ➔ **Smart contributions** to open source community

# Syncsort DMX-h ETL Edition

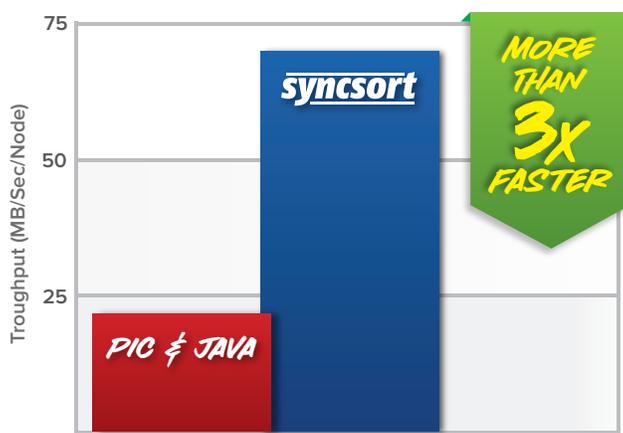
A Smarter Approach to Hadoop ETL

**syncsort**

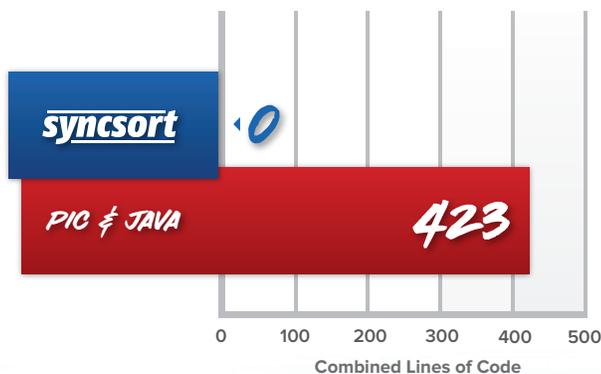
code generator. Instead, Hadoop MapReduce automatically invokes the highly-efficient DMX-h engine at runtime, which executes natively on all nodes as an integral part of the Hadoop framework. Once deployed, DMX-h automatically optimizes resource utilization – CPU, memory and I/O - on each node to deliver the highest levels of performance, with no tuning required. Simply stated, higher performance and efficiency per node means you can process more data in less time, with fewer servers.

## CDC BENCHMARK

**NO CODING, NO SCRIPTING, JUST FASTER!**



Total data per side: 100 GB (200 GB total)



## Smarter Development - Experience Hadoop ETL without Coding

A typical ETL deployment in Hadoop requires organizations to acquire a completely new set of advanced programming skills that are expensive and difficult to find. DMX-h enables people with a much broader range of skills - not just MapReduce programmers - to create ETL tasks

that execute within the MapReduce framework, replacing complex Java, Pig, or HiveQL code with a powerful, easy-to-use graphical development environment. DMX-h makes it easier to develop, maintain, and re-use applications running on Hadoop via:

- ➔ A Windows-based, graphical development environment
- ➔ Comprehensive built-in transformations
- ➔ Built-in metadata capabilities, for greater reusability, impact analysis, and data lineage

## Smarter Productivity – Fast-track Your Way to Successful Hadoop ETL

Leveraging Hadoop for ETL requires overcoming a steep learning curve. Moreover, with many potential use cases, it's difficult to know where to start. Therefore, most organizations struggle to make their staff productive and deliver impactful results when they begin testing Hadoop. DMX-h ETL Edition helps you get started and fully productive with Hadoop quickly by providing a library of Use Case Accelerators to implement common ETL tasks such as joins, change data capture (CDC), web log aggregations, mainframe data access, and more.

## Smarter Connectivity - Connect to All Your Data, Including Mainframe, with One Tool

Big Data comes from all data sources across an organization. In most cases, this leads to data ingestion challenges – requiring programmers to manually write custom scripts to parse, transform and then load data into HDFS. DMX-h ETL Edition alleviates these hurdles with connectivity capabilities critical to successful Hadoop ETL deployments.

With DMX-h ETL Edition, you need only one tool to connect all sources and targets to Hadoop, including relational databases, appliances, files, XML, cloud, and even mainframe. No coding or scripting needed. DMX-h ETL Edition can also be used to pre-process data - cleanse, sort, partition, and compress - prior to loading it into Hadoop, resulting in enhanced performance and significant storage savings.

## BIG IRON IS BIG DATA TOO!

A particularly underserved area within Hadoop deployments is mainframe. However, Mainframe data can be the critical reference point for new data sources such as web logs and sensor data. DMX-h offers unique capabilities to read, translate, and distribute mainframe data with Hadoop, opening up a wealth of opportunities by delivering deeper analytics, at lower cost.

# Syncsort DMX-h ETL Edition

A Smarter Approach to Hadoop ETL

**syncsort**

DMX-H ETL EDITION KEY FEATURES

## High-Performance Data Transformations

Includes high-performance sort, joins, aggregations, multi-key lookup, advanced textprocessing, hashing functions, and source/record/field-level operations.

## Rapid Development through Windows-based DMX-h Workstation

Lets you develop and test MapReduce ETL jobs locally in Windows through a graphical user interface, then deploy in Hadoop. Expression Builder helps define data transformations based on business rules.

## Use Case Accelerators

Fast-tracks your Hadoop productivity with a library of fully- functional and re-usable templates – including web log processing, CDC, mainframe connectivity, joins, and more –to design your own data flows.

## Data Source & Target Connectivity

Connects any source and target to Hadoop, including all major database management systems, flat files, XML files, mainframe and others.

## Mainframe data ingestion & translation

Reads files directly from the mainframe, parses and transforms the data – packed decimal, occurs depending on, EBCDIC/ASCII, multi-format records, and more – without installing any software on the mainframe and without writing any code. With over 40 years of experience, no one does this better than Syncsort!

## Dynamic ETL Optimizer

Performs data transformations and functions at maximum speed based on hundreds of proprietary algorithms. ETL optimizer automatically chooses best algorithm to maximize performance of each node in Hadoop and adapts in real-time to system conditions.

## File-based Metadata Capabilities

Provides greater transparency into impact analysis, data lineage, and execution flow without dependencies on third-party systems, such as relational databases.

## Support for All Major Hadoop Distributions

Supports all major Hadoop distributions including Cloudera, Greenplum Pivotal HD, Hortonworks, and Apache. Integrates natively with all Hadoop distributions based on Apache 2.0.3-alpha and later.

## Smarter Security - Be Confident with Enterprisegrade Security for Hadoop ETL

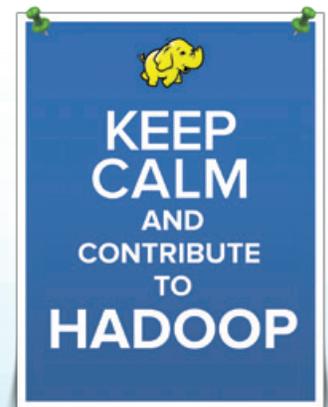
With Big Data comes even bigger responsibility. Therefore, any viable approach to Hadoop ETL must provide ironclad security that meets your organization's and industry's data security requirements. Many of today's Hadoop ETL solutions lack standard security protocols. DMX-h helps you keep all your data secure with market-leading support for common protocols such as LDAP and Kerberos.

## A Smarter Ecosystem for a Smarter Solution!

We're not alone in our quest. We've partner with leading Hadoop players including Cloudera, Greenplum Pivotal HD, Hortonworks, HP Vertica, and more to ensure the best experience for you!

## Better Hadoop for Everyone

Syncsort is particularly focused on lowering the barriers for wider Hadoop adoption and help organizations unleash the complete potential of Hadoop through Smarter ETL. Therefore, we've made important contributions to the open source community, including a patch new feature - JIRA MAPREDUCE-2454 - to allow external implementations of the sort phase in MapReduce and enable more sophisticated use cases.



## ABOUT SYNCSORT

Syncsort provides fast, secure, enterprise-grade software spanning Big Data solutions in Hadoop to Big Iron on mainframes. We help customers around the world to collect, process and distribute more data in less time, with fewer resources and lower costs. 87 of the Fortune 100 companies are Syncsort customers, and Syncsort's products are used in more than 85 countries to offload expensive and inefficient legacy data workloads, speed data warehouse and mainframe processing, and optimize cloud data integration. Experience Syncsort at [www.syncsort.com](http://www.syncsort.com)

**syncsort**

50 Tice Boulevard  
Woodcliff Lake  
NJ 07677  
201.930.8200  
[www.syncsort.com](http://www.syncsort.com)