



Dell | Cloudera | Syncsort

Data Warehouse Optimization

ETL Offload Reference Architecture

Data Sheet

Drive operational efficiency and lower data transformation costs with a Reference Architecture for an end-to-end optimization and offload solution



Augment the Enterprise Data Warehouse (EDW) with Hadoop

The first use case in the big data journey typically begins with a goal to increase operational efficiency. Dell customers understand that they can use Hadoop to cut costs, yet they have asked us to make that process simple. They want defined architectures that provide end-to-end solutions that are validated and engineered to work together.

The Dell | Cloudera | Syncsort Data Warehouse Optimization – ETL Offload Reference Architecture (RA) provides a blueprint to help your organization build an environment to augment your EDW. The RA provides the architecture, beginning from bare-metal hardware, for running ETL jobs in Cloudera Enterprise version 5.5.1 with Syncsort DMX-h 8.5 software.

Many of our customers have a skills-set gap when it comes to utilizing Hadoop for ETL in their environments. They don't have time to build up expertise in Hadoop. The software components of the Reference Architecture help you address this challenge. Dell's solution makes it easy to build and deploy ETL jobs in Hadoop.

Syncsort's high-performance ETL software enables your users to maximize the benefits of MapReduce without compromising on the capabilities and ease of use of conventional ETL tools. Syncsort software enables faster time to value by reducing the need to develop expertise on Pig, Hive and Sqoop, technologies that are essential for creating ETL jobs in MapReduce.

The Dell | Cloudera | Syncsort Data Warehouse Optimization – ETL Offload solution delivers a use-case driven Hadoop Reference Architecture to guide your data warehouse optimization efforts.

The architecture brings together:

- The industry-leading Hadoop distribution version 5.5.1 from Cloudera
- A rich framework and toolset for ETL offload from Syncsort
- Dell™ PowerEdge™ R series servers with Intel® Xeon® processors
- Dell networking components
- Optional consulting and integration services

Traditional Data Pipeline



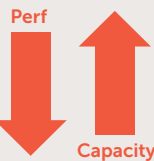
Disparate Data Sources



Data Staging Tool
Extract & Load Data
Clean & Parse Data



Enterprise data warehouse
Data Transformation Jobs
Business Reporting Query



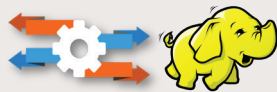
The Results

Longer data transformation job times
Not meeting SLAs for business reporting
Slow ad hoc query
Too costly to scale

Modern Data Pipeline



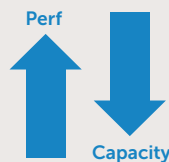
Disparate Data Sources



Hadoop + ETL
Data Transformation Jobs
Clean, Parse, Transform



Enterprise data warehouse
Business Reporting Query



The Results

Reduced data transformation job times
Improved SLAs for business reporting
Fast ad hoc query
Scales economically

Figure 1: Modernize the Data Pipeline

Drive operational efficiency

Third-party research found that the Dell | Cloudera | Syncsort ETL offload solution can help organizations:

- Control costs—Reduce data warehouse administrative costs by up to 76 percent.¹
- Improve productivity—Transform data 60 percent faster for analysis.²
- Simplify ongoing operations—Develop and design complex transformation jobs up to 54 percent faster.³

1 Principled Technologies. "Cost Advantages of Hadoop ETL Offload with the Intel Processor-Powered Dell | Cloudera | Syncsort Solution." July 2015.

2 Principled Technologies. "Performance Advantages of Hadoop ETL Offload with the Intel Processor-Powered Dell | Cloudera | Syncsort Solution." July 2015.

3 Principled Technologies. "Design Advantages of Hadoop ETL Offload with the Intel Processor-Powered Dell | Cloudera | Syncsort Solution." July 2015.

Solution Components

Server	PowerEdge R730XD
Processor	Xeon 2600 Series v3
NIC	Intel 10GB
Controller	PERC H730
O/S	RHEL 6.6
Network Switch	Dell Networking S4048-ON
Management Switch	Dell Networking S3048-ON
Aggregation Switch	Dell Networking S6000
Hadoop Distribution	Cloudera Enterprise Data Hub 5.5.1
ETL Engine	Syncsort DMX-h 8.5



To learn more, visit
Dell.com/Hadoop | Dell.com/BigData

