# DELL EMC CLOUDERA SYNCSORT ETL OFFLOAD HADOOP SOLUTION

## Primary use case

Accelerate extract, transform and load (ETL) on existing enterprise data warehouses

## Solution benefits

- Integrates easily with Hadoop®
- No coding necessary for easy deployment
- No need for expertise on Apache Pig™, Hive™, and Sqoop™
- Closes the skills gap using Syncsort SILQ™

## Differentiation

- Reduces EDW admin costs up to 76 percent.[1]
- Transforms data 60 percent faster for analysis.[2]
- Designs transformation jobs up to 54 percent faster.[3]

[1] Cost advantages report
[2] Performance advantages report
[3] Design advantages report

## DRIVE OPERATIONAL EFFICIENCY

Many large organizations have invested heavily in the development and growth of enterprise data warehouses (EDWs). Today, the EDW is often the central data store for business reporting, extract/transform/load (ETL) processes, and data ingestion from diverse sources, both inside and outside the enterprise.

While the EDW plays an all-important role in the effort to leverage big data to drive business value, it is not without its challenges. As the volume, velocity and variety of data increases, many EDWs are being pushed to their limits. This overload stems in part from the many processes, such as transform jobs, that have been moved into the EDW because traditional ETL tools could not deal with the size of the data volume.

As their EDWs hit their limits, many forward-looking organizations are deploying the open source Apache™ Hadoop® platform as a complement to their existing EDWs, and then offloading data warehouse processing functions to Hadoop. With a robust offload solution, an organization can accelerate ETL processing, work easily with a wide range of new data sources and formats, and make better use of existing EDW investments.

That's the idea behind the Dell EMC™ Cloudera™ Syncsort™ ETL Offload Hadoop Solution for augmenting existing EDWs. This groundbreaking solution allows you to capitalize on the unique technical and cost advantages of the Hadoop platform while protecting your investments in your EDW.

Since 2011, Dell EMC, in partnership with Cloudera and Intel, has helped organizations solve the skills gap by providing expert guidance and knowledge to streamline the architecture, design, planning and configuration of Hadoop environments. Today, this expert guidance and knowledge is embodied in Dell EMC Cloudera Hadoop Solutions, including the ETL offload solution with Cloudera and Syncsort.

Cloudera™ Enterprise

Syncsort™ DMX-h™ and SILQ

Dell OpenManage™ / iDRAC with Lifecycle Controller

**Pod Network**
2x Dell EMC Networking S4048 10GbE Pod Switches
1x S3124 iDRAC Switch

**Cluster Network**
2x Dell EMC Networking S6000 40GbE Cluster

**Infrastructure Nodes**
1x Dell EMC PowerEdge™ R630 Admin Node
3x PowerEdge R730xd Name Nodes
1x PowerEdge R730xd Edge Node or
1x PowerEdge FC630 Name Nodes Admin Node
3x PowerEdge FC630 Name Nodes
1x PowerEdge FC630 Edge Node

**Data Nodes**
10x Dell EMC PowerEdge R730xd with 3.5 Drives – 48 TB or
10x PowerEdge R730xd with 2.5" Drives – 24TB or 20x
PowerEdge FC630 / FD332 – 32 TB

## DELL EMC CLOUDERA SYNCSORT ETL OFFLOAD HADOOP SOLUTION

The Dell EMC Cloudera Syncsort ETL Offload Hadoop Solution incorporating version 5.9 of the Cloudera distribution of Hadoop (CDH 5.9) is the 19th iteration of our proven architecture based on expert engineering and full validation. Dell EMC engineers have validated and certified CDH 5.9 on our Intel-based Dell EMC PowerEdge™ R730xd servers and PowerEdge™ FX2 FC630 server nodes with FD332 storage blocks and Syncsort DMX-h 9.1 with Dell EMC Networking switches S4048ON and S3048ON.

The result is a solution that allows your organization to build an ETL offload Hadoop cluster without the guesswork. You can leverage the Dell EMC solution to streamline the entire process, from bare-metal server configuration to network setup to running CDH 5.9 on a certified architecture.

### VALIDATED BENEFITS

Research by Principal Technologies, a technology testing and analysis firm, found that the Dell EMC Cloudera Syncsort ETL Offload Hadoop Solution can help organizations drive operational efficiency by completing Hadoop ETL jobs faster, simplify the Hadoop ETL design process, and save thousands of dollars on Hadoop ETL jobs.

In specific terms, the firm determined:

- A Dell EMC Cloudera Syncsort ETL Offload Hadoop Solution fully implemented by an entry-level employee could reduce administrative costs by 76 percent.[1]
- ETL jobs created by an entry-level technician using the Dell EMC Cloudera Syncsort ETL Offload Hadoop Solution ran up to 60 percent faster than a solution created by a Hadoop expert using open source tools.[2]
- A Dell EMC Cloudera Syncsort ETL Offload Hadoop Solution enables less-experienced users to develop and deploy Hadoop ETL jobs in less than a week.[3]

Dell EMC developed an architecture document that provides the guidance that can help your organization build a Hadoop DAS cluster from bare-metal hardware. With the expertise and experience of Dell EMC to take you through the steps, you can save valuable time and resources — as you accelerate time to value.

### REDUCE THE RISK

According to Gartner, through 2018, 70 percent of Hadoop deployments will fail to meet cost savings and revenue-generation objectives due to skills and integration challenges.[4] The high failure rate has left many organizations wary of adopting Hadoop, yet the demands of the business will continue to require big data technologies. With this knowledge, Dell EMC recognized the need to assist and enable organizations to lower the risk using a tested, validated and certified plan for implementing Hadoop.

Using the many years of experience Dell EMC has in building Hadoop DAS architectures, you can leverage that expertise to fill the skills gap and build an architecture that meets the needs of your business, all while reducing the risks that come with technology projects.

### DELL EMC POWEREDGE SERVERS

Today's Dell EMC Cloudera Hadoop Solutions are validated and certified on two of the industry-leading PowerEdge servers.

You can maximize server-based storage flexibility and performance with the Intel-based Dell EMC PowerEdge R730xd server, part of the new 13th generation of Dell EMC PowerEdge servers. The R730xd server offers an optimal balance of storage utilization, performance and cost with an optional in-server hybrid storage configuration that can support tiering and capacity for up to 28 drives in a 2S/2U system, including up to 18 1.8-inch SATA SSDs.

Additionally, we offer our solutions using the Dell EMC PowerEdge FX2 server, a 2U hybrid rack-based computing platform that combines the density and efficiencies of blades with the simplicity and cost benefits of rack-based systems.

Scale workloads quickly, as needed, adding resources incrementally without the expense and inefficiency of overprovisioning. The groundbreaking FX architecture combines efficient management with flexible data center building blocks to optimize Hadoop workloads.

Combine various modules with the FX2 chassis, a compact 2U rack-based enclosure that shares cooling, power, networking and PCI expansion slots, to quickly deploy the exact combination of compute, storage and networking resources you need to best fulfill your business requirements.

---

[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.

[4] Gartner. "Market Guide for Hadoop Distributions." January 6, 2015.

FX2 solution components include:

- Server blocks at the heart of the FX converged architecture powered by the latest Intel® Xeon® processors
- FC630 server nodes — 2-socket, half-width 1U workhorse server blocks ideal for a wide variety of business applications
- FD332 storage blocks — flexible, high-density, half-width 1U storage modules that enable you to rapidly scale direct attached storage (DAS) in your FX-based infrastructures

## DELL EMC NETWORKING S-SERIES 10GBE SWITCHES

These switches allow you to deploy modern workloads and applications designed for the open networking era with an optimized data center top-of-rack (ToR) networking solution.

Among other benefits, this switch series:

- Delivers low latency, superb performance and high density with hardware and software redundancy
- Offers Active Fabric designs using S- or Z-Series core switches to create a two-tier, 1/10/40 GbE data center network architecture
- Provides an ideal solution for applications in high-performance data center and computing environments

## DELL EMC ISILON DATA LAKE FOR EDW

Another option for getting started with Hadoop is a Dell EMC Data Lake. At Dell EMC, we continue to keep customer use cases first as we offer robust Data Lake opportunities using Cloudera Enterprise together with Dell EMC scale-out NAS Isilon storage running the OneFS operating system.

Dell EMC Isilon scale out NAS with Hadoop provides customers with a highly efficient, massively scalable, and secure data lake to optimize enterprise data warehouse resources. With this solution, data wrangling and curation of data can be offloaded from your EDW and performed in Cloudera Enterprise on the Isilon Data Lake.

The combination of Cloudera and Isilon shared storage helps organizations speed time to insights, improve storage utilization, eliminate islands or silos of storage, and lower storage management costs of migration, security and protection.

## DELL EMC SERVICES

After your exploration of the technologies, Dell EMC Services makes getting started easy. Options include custom solution design, hardware and software deployment, ongoing support and training. With Dell EMC, you have the assurance that your Cloudera solution is backed by expert hardware and software support that can be tailored to your specific needs.

## CLOUDERA ENTERPRISE

The right technology is key for turning your data into real business value. Powered by Apache Hadoop, Cloudera Enterprise is a fast, easy and secure modern data platform. From analytics to data science, anyone can now get results from any data and across any environment — all within a single, scalable platform.

When you make Cloudera Enterprise the center of your business, you open up limitless possibilities with your data. Whether you're powering data engineering and data science workloads, building an operational or analytic database, or looking to bring them all together in an enterprise data hub, Cloudera has the right platform to fit your needs.

Cloudera Enterprise delivers:

- **Data engineering** — Bring your data engineers and data scientists together to build real-time pipelines, speed data processing, and develop and train data models.
- **Analytic database** — Modernize your IT architecture to enable ELT and high-performance SQL analytics for reporting, exploration and self-service business intelligence.
- **Operational database** — Build data-driven applications that deliver real-time insights for monitoring and detection, as well as streaming applications like Internet of Things and model scoring and serving.

## CLOUDERA CDH 5.9

Cloudera CDH 5.9 is at once:

- **Fast for business** — From analytics to data science and everything in between, Cloudera delivers the performance you need to unlock the potential of unlimited data.
- **Easy to manage** — Focus on results, not fighting fires. Cloudera provides the operations that keep mission-critical applications up and running — especially at scale.
- **Secure without compromise** — Meet your most stringent data security and compliance needs without sacrificing business agility and innovation. Cloudera provides a comprehensive, integrated approach to data security and governance.

## SYNCSORT DMX-H 9.1

Syncsort DMX-h is specifically designed to help you achieve your modern data strategy objectives with a single interface for accessing and integrating all your enterprise data sources — batch and streaming — across Hadoop, Spark, Linux, Unix or Windows, on premises or in the cloud.

Syncsort DMX-h was designed from the ground up to make big data integration simple, combining a long history of innovation with significant contributions Syncsort has made to improve Apache Hadoop.

With DMX-h, you get:

- A single software environment for accessing and integrating all your enterprise data sources — batch and streaming — while managing, governing and securing the entire process
- Software that evolves with the Hadoop ecosystem to keep you current without rewriting jobs or acquiring new skills
- The best mainframe access and integration capabilities in the world
- An easy-to-use graphical interface with the flexibility to quickly extend the software for your unique needs
- Access to expert support and services to help ensure your success

## GAIN MORE VALUE FROM YOUR INVESTMENTS

The Dell EMC Cloudera Syncsort ETL Offload Hadoop Solution gives you everything you need to capitalize on your ETL offload opportunities — including software, hardware, services and a validated reference architecture. With this robust offload solution, your organization can accelerate ETL processing, work easily with a wide range new data sources and formats, and make better use of your existing EDW investments.

## TECHNICAL SPECIFICATIONS

| Server Architecture | Infrastructure Nodes 3.5” | General Purpose Data Node 2.5” | Data Node 3.5” | Data Node 2.5” |
|---|---|---|---|---|
| R730XD Server | R730XD | R730XD | R730XD | R730XD |
| Processor | 2 X Intel E5-2650 v4 2.2GHz (12 Core) | 2 X Intel E5-2690 v4 2.6GHz (14 Core) | 2 X Intel E5-2650 v4 2.2GHz (12 Core) | 2 X Intel E5-2690 v4 2.6GHz (14 Core) |
| Memory | 128GB | 128GB | 256GB | 256GB |
| Network Card | Intel X520 Dual Port 10Gbe | Intel X520 Dual Port 10Gbe | Intel X520 Dual Port 10Gbe | Intel X520 Dual Port 10Gbe |
| Hard Drive Controller | H730 | H730 | H730 | H730 |
| Hard Drives | 8 x 1TB 7.2K RPM SAS 12Gbps<br>2 x 600GB 10K RPM SAS 12Gbps (Flex Bay) | 8 x1.2TB 10K RPM SAS 12Gbps<br>2 x 600GB 10K RPM SAS 12Gbps (Flex Bay) | 12 x 4TB 7.2K RPM SAS 6Gbps<br>2 x 600GB 10K RPM SAS 12Gbps | 24 x 1.2TB 10K RPM SAS 12Gbps<br>2 x 600GB 10K RPM SAS 12Gbps |
| RAID Layout | Operating System- 2 HDs, RAID 1<br>Data- 8 HDs, RAID 10 | Operating System- 2 HDs, RAID 1<br>Data- 8 HDs, RAID | Operating System- 2 HDs, RAID 1<br>JBOD- 12HDs, Non-RAID | Operating System- 2 HDs, RAID 1<br>JBOD- 12HDs, Non-RAID |
| **Cluster Network** | **Cluster Data Network** | **BMC Network** | **Edge Network** | |
| Network Switch | S4048ON | S3048ON | S4048ON | |
| Connection | Bonded 10GbE<br>Dual Top of Rack | 1GbE<br>Dedicated Switch Per Rack | Bonded 10GbE, optionally bonded<br>Direct to Edge Network or via aggregation switch | |

| | Infrastructure Nodes | Data Node | |
|---|---|---|---|
| FX2 Server | FC630 | FC630 | |
| Processor | 2 X Intel E5-2650 v4 2.2GHz (12 Core) | 2 X Intel E5-2680 v4 2.4GHz (14 Core) | |
| Memory | 128GB | 256GB | |
| Network Card | Intel X710 Quad Port 10Gbe | Intel X710 Quad Port 10Gbe | |
| Hard Drive Controller | H730 | H730 | |

| | | | |
|---|---|---|---|
| Storage | FD332<br>10 x 1TB 7.2K RPM SAS 12Gbps | On-board Storage<br>2 x 400GB SSD SATA 6Gbps<br>PERC S130<br>FD332<br>16 x 2TB 7.2K RPM SAS 12Gbps | |
| RAID Layout | Operating System- 2 HDs, RAID 1<br>Data- 8 HDs, RAID 10 | Operating System- 2 HDs, RAID 1<br>JBOD- 8 HDs, NON-RAID | |
| **Cluster Network** | **Cluster Data Network** | **BMC Network** | **Edge Network** |
| Network Switch | S4048ON | S3048ON | S4048ON |
| Connection | Bonded 10GbE<br>Dual Top of Rack | 1GbE<br>Dedicated Switch Per Rack | Bonded 10GbE, optionally bonded<br>Direct to Edge Network or via aggregation switch |

| **Cloudera CDH 5.9** | |
|---|---|
| Accumulo | 1.7.2 |
| Impala | 2.7 |
| Kafka | 2.0 |
| Navigator | 2.8 |
| Sentry | Cloudera Manager 5.X & Higher |
| Spark | 1.6 |

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



November 2016 | Rev 1.0