Solution Brief

Dell | Cloudera | Syncsort Data Warehouse Optimization – ETL Offload Reference Architecture

Accelerate your journey to advanced data analytics with a Hadoop-based solution

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.

New challenges for the enterprise data warehouse

Many large organizations have invested heavily in the development and growth of enterprise data warehouses. 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.

Diverse data formats—including new unstructured and semi-structured data types—are adding ETL complexity and creating heavy processing burdens. Data integration and transformation workloads can now consume as much as 80 percent of EDW capacity, according to Gartner. It's no wonder that 70 percent of today's data warehouses are performance- and capacity-constrained, according to the firm.[1] In some cases, just a few heavy jobs can bog down an EDW, threatening service level agreements with business units.

These backend challenges driven by an overloaded EDW can impact the frontlines of the business, because more processing means less query capacity. For example, business segments might not be able to run crucial reports in time to make critical business decisions, and business analysts might not be able to query data for ad hoc analysis, resulting in slower decision processes. This is a problem that isn't going to go away. It's only going to grow worse as the universe of big data grows in size and complexity.

So how do you overcome these challenges? It might seem that the logical path forward is to scale out your EDW, but that can be a costly proposition in terms of software licensing and consulting fees. And even at that, many EDWs are not equipped to handle the diverse variety of today's data—from social media feeds to machine-generated data streams.

Instead, these challenges drive the need for solutions that offload the heavy lifting of ETL processing from the data warehouse to a complementary, lower-cost processing environment built for the diversity of today's data. Additional benefits include higher performance in the EDW for business reporting and queries, resulting in increased velocity for business insights. This is where the Apache™ Hadoop™ platform enters the picture.

1 Gartner. "The State of Data Warehousing in 2014." June 19, 2014.

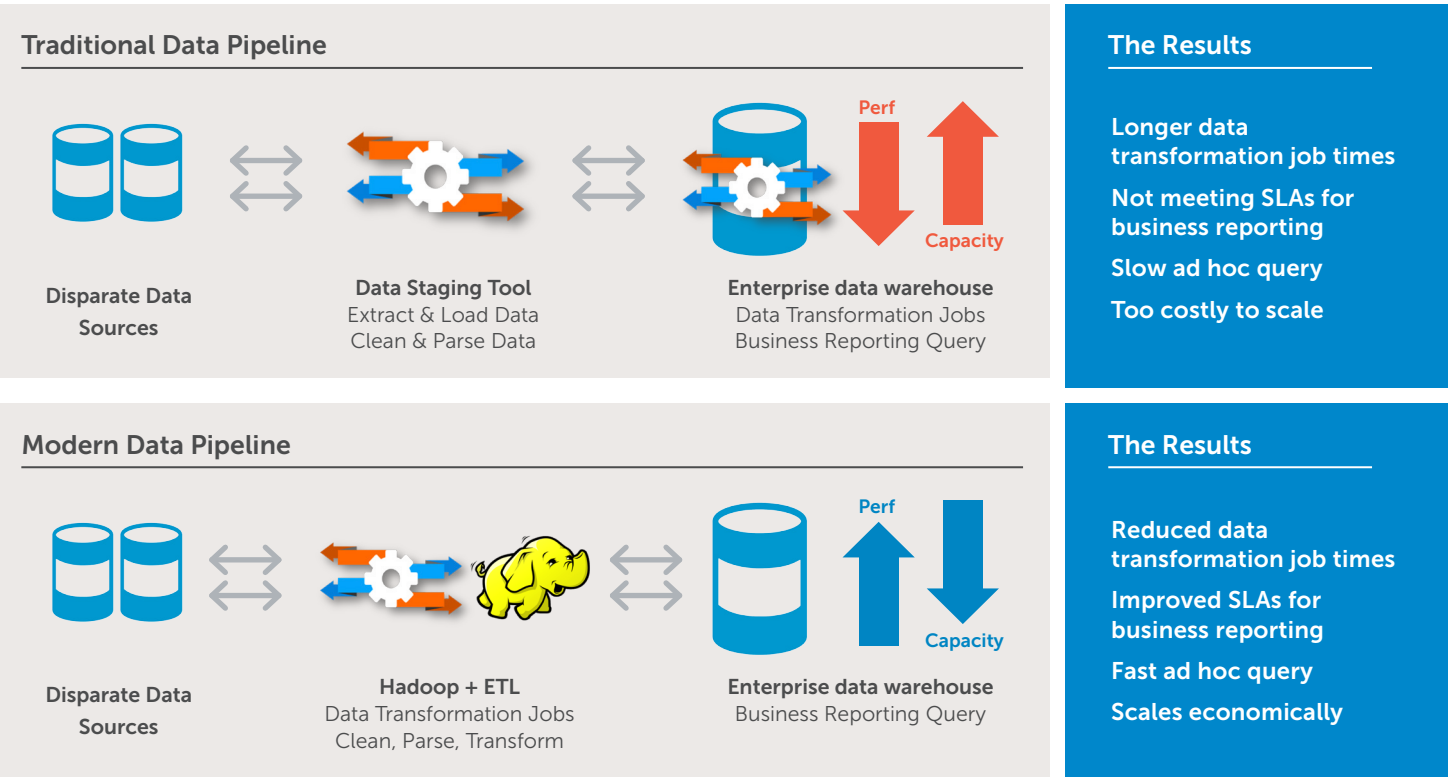

Figure 1: Modernize the Data Pipeline

The power of the Hadoop platform

As their EDWs hit their limits, many forward-looking organizations are deploying the open source 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 new data sources and formats, and make better use of existing EDW investments.

That's the idea behind the Dell | Cloudera | Syncsort Data Warehouse Optimization – ETL Offload Reference Architecture. This groundbreaking solution allows you to capitalize on the unique technical and cost advantages of the Hadoop platform while making better use of your existing EDW investments.

The Hadoop platform was originally developed by the world's largest Internet companies to capture and analyze the massive amounts of data they generate. Unlike earlier platforms, Hadoop can store any kind of data in its native format—structured, unstructured or semi-structured—and be used to perform a wide variety of transformations on that data.

A highly scalable platform, Hadoop allows your organization to store petabytes, and even exabytes, of data cost-effectively. As the amount of data in a cluster grows, you can add new servers with local storage to scale out incrementally and inexpensively.

Hadoop was designed for extreme parallel data processing. When used in conjunction with a tool like Syncsort, Hadoop can help you greatly accelerate ETL processes while reducing data transformation costs in comparison to running ETL jobs in a traditional data warehouse.

With the right tools, Hadoop can serve as the staging area for all your data, allowing you to easily collect, transform and distribute more data in less time and at a significantly lower cost. By offloading ELT workloads into Hadoop, you're poised to reduce batch windows, keep data readily available for as long as you need it, and free up valuable data warehouse capacity for faster analytics and end-user queries.

Validated benefits

Research by Principal Technologies, a technology testing and analysis firm, found that the Dell | Cloudera | Syncsort ETL offload solution can help organizations drive operational efficiency by completing Hadoop ETL jobs faster, simplifying the Hadoop ETL design process, and saving thousands of dollars on Hadoop ETL jobs.

In specific terms, the firm determined:

- A Dell | Cloudera | Syncsort solution for Hadoop fully-implemented by an entry-level employee could reduce administrative costs by 76 percent.[2]

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

A proven approach

Dell has a proven approach to Hadoop solutions. The Dell | Cloudera | Syncsort ETL offload solution is the 17th Dell Hadoop Reference Architecture that has been certified and validated beginning since 2011. These reference architectures are the result of a tested and validated process that Dell has refined over the years to provide the blueprints for an optimal customer experience.

- ETL jobs created by an entry-level technician using the Dell | Cloudera | Syncsort solution for Hadoop ran up to 60 percent faster than a solution created by a Hadoop expert using open source tools.[3]
- A Dell | Cloudera | Syncsort solution for Hadoop enables less-experienced users to develop and deploy Hadoop ETL jobs in less than a week.[4]

A tested and validated reference architecture

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 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

Cloudera components

The ETL offload solution delivers the power of the Hadoop environment via Cloudera Enterprise version 5.7 software. Designed specifically for mission-critical environments, the Cloudera Enterprise offering includes CDH, the world's most popular open source Hadoop-based platform, as well as advanced system and data management tools, open source software components that help with Hadoop usability, and dedicated support from Hadoop experts.

Syncsort components

For sophisticated ETL offload capabilities, the solution incorporates Syncsort DMX-h version 8.5 software. This software suite makes it easy, even for non-data-scientists, to build and deploy ETL jobs in Hadoop. With DMX-h, users can start developing Hadoop ETL jobs within hours, and become fully productive within days, using a drag-and-drop interface and the same ETL skills they already have. There's no need to learn complex technologies like MapReduce, Pig or Hive.

To fast-track your EDW-offload projects, the solution includes Syncsort SILQ, a SQL offload utility designed to help you understand and offload complex SQL data integration workloads from your data warehouse into Hadoop. SILQ takes a SQL script as an input and then provides a detailed flow chart of the SQL logic. Using an intuitive web-based interface, users can easily drill down to get detailed information about each step within the data flow, including tables and data transformations.

Dell and Intel components

At the hardware level, Dell leverages the Dell™ PowerEdge™ R730xd servers for both data and infrastructure nodes. Both of these systems deliver the performance, power efficiency, virtualization and security features of the Intel® Xeon® Processor E5-2600 v4 Product Family, along with large memory capacities and fast storage options.

Professional services and support

Dell Deployment and Consulting can help you quickly realize the full benefit of your data warehouse optimization investments while limiting business disruptions. Dell can provide on-site deployment of the solution hardware, configuration of servers and network switches, and installation of the solution software. Deployment services are performed by trained experts in accordance with Dell, Cloudera and Syncsort best practices. Additional training and consulting services can be added to help your team complete the offloading of ETL workloads into Hadoop.

Support for the overall solution is provided through Dell ProSupport, with collaborative assistance from the Cloudera and Syncsort support teams.

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

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

Key takeaways

Dell has a proven approach to Hadoop solutions. The Dell | Cloudera | Syncsort ETL offload solution is the 17th Dell Hadoop Reference Architecture that has been certified and validated beginning since 2011. These Reference Architectures are the result of a tested and validated process that Dell has refined over the years to provide the blueprints for an optimal customer experience.

The Dell | Cloudera | Syncsort Data Warehouse Optimization – ETL Offload Reference Architecture 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 existing EDW investments.

Among other benefits, the solution helps you:

- Streamline enterprise-class data integration
- Execute ETL processes in less time
- Reduce the costs and resource requirements for ETL processes
- Capitalize on the unique technical and cost advantages of the Hadoop platform
- Make better use of your existing EDW investments

A tested and validated reference architecture

- The industry-leading Hadoop distribution 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

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

May 2016 | Version 5.7