COST ADVANTAGES OF HADOOP ETL OFFLOAD WITH THE INTEL PROCESSOR-POWERED DELL | CLOUDERA | SYNCSORT SOLUTION

Save thousands on Hadoop® ETL jobs

An entry-level employee using the Dell™ | Cloudera® | Syncsort® solution for Hadoop reduced administrative costs by up to **76.3 percent.**

Save $425,972

compared to contracting a senior engineer to implement an open-source solution*

Save $146,823

compared to paying a senior engineer on staff to implement an open-source solution*

*assuming the design of four ETL jobs every month over three years

Four Dell PowerEdge™ R730xd servers and two Dell PowerEdge R730 servers, powered by the Intel® Xeon® processor E5-2600 v3 product family, in a Hadoop cluster.

Many companies are adopting Hadoop solutions to handle large amounts of data stored across clusters of servers. Hadoop is a distributed, scalable approach to managing Big Data that is very powerful and can bring great value to organizations. Companies use extract, transform, and load (ETL) jobs to bring together data from many different applications or systems on different hardware in order to modify or adjust the data in some way, and then put it into a new format that they can mine for useful information.

Without the right tools, ETL on Hadoop can require highly experienced, expensive, and hard-to-find programmers to create the jobs in order for execution. Dell, Cloudera, and Syncsort offer an integrated Hadoop ETL solution that allows entry-level technicians—after only a few days of training—to perform the same tasks that these Hadoop specialists perform, often even more quickly. In our tests, we found that the unique design of the Dell | Cloudera | Syncsort solution can allow end users with little experience using Hadoop to develop and deploy optimized ETL jobs faster than an expert-driven do-it-yourself (DIY) solution deployed using open-source tools, which may require iterative revisions and tweaks to reach a similar level of performance.

In our testing and assumptions, these advantages mean the three-year administrative costs are up to $425,972 lower using the Dell | Cloudera | Syncsort solution than using senior engineer time with a DIY approach.

JULY 2015

A PRINCIPLED TECHNOLOGIES TEST REPORT

Commissioned by Dell

SAVE MONEY AND TIME WITH THE DELL | CLOUDERA | SYNCSORT SOLUTION

Extract, Transform, and Load

ETL refers to the following process in database usage and data warehousing:

- Extract the data from multiple sources
- Transform the data so it can be stored properly for querying and analysis
- Load the data into the final database, operational data store, data mart, or data warehouse

Hadoop implementations suffer from several barriers to effectiveness. They often have primitive integration with infrastructure, and today we find there is currently a lack of available talent to run Hadoop clusters and perform data ingest and processing tasks using the cluster.[1] The Dell | Cloudera | Syncsort solution offers help with both of these problems.

The Dell | Cloudera | Syncsort solution is a reference architecture that offers a reliable, tested configuration that incorporates Dell hardware on the Cloudera Hadoop platform, with Syncsort's DMX-h ETL software. The Dell | Cloudera | Syncsort reference architecture includes four Dell PowerEdge R730xd servers and two Dell PowerEdge R730 servers, powered by the Intel Xeon processor E5-2600 v3 product family.

For organizations that want to optimize their data warehouse environments, the Dell | Cloudera | Syncsort reference architecture can greatly reduce the time needed to deploy Hadoop when using the included setup and configuration documentation as well as the validated best practices. The Syncsort DMX-h software means Hadoop ETL jobs can be developed using a graphical interface in a matter of hours with minor amounts of training, and with no need to spend days developing code. The Dell | Cloudera | Syncsort solution also offers professional services with Hadoop and ETL experts to help fast track your project to successful completion.[2]

To understand the cost advantages of the Dell | Cloudera | Syncsort solution, we draw on testing performed in the Principled Technologies labs.[3] We had an entry-level technician and a highly experienced Hadoop expert work to create three Hadoop ETL jobs using different approaches to meet the goals of several use cases.

The entry-level worker, who had no familiarity with Hadoop and less than one year of general server experience, used Syncsort DMX-h to carry out these tasks. Our expert had 18 years of experience designing, deploying, administering, and benchmarking enterprise-level relational database management systems (RDBMS). He has deployed, managed, and benchmarked Hadoop clusters, covering several Hadoop distributions and several Big Data strategies. He set up the cluster and designed and created the use cases using only free open-source DIY tools.

[1] Source: survey of attendees for the 2014 Gartner webinar Hadoop 2.0 Signals Time for Serious Big Data Consideration. www.informationweek.com/big-data/software-platforms/cloudera-trash-talks-with-enterprise-data-hub-release/d/d-id/1113677

[2] Learn more at en.community.dell.com/dell-blogs/dell4enterprise/b/dell4enterprise/archive/2015/06/09/fast-track-data-strategies-etl-offload-hadoop-reference-architecture

[3] Design advantages of Hadoop ETL offload with the Intel processor-powered Dell | Cloudera | Syncsort solution www.principledtechnologies.com/Dell/Dell_Cloudera_Syncsort_design_0715.pdf

The implementation experience of these two workers showed us that using the Dell | Cloudera | Syncsort solution was faster, easier, and—because a less experienced employee could use it to create ETL jobs—far less expensive to implement. In this cost analysis, we apply those findings to a hypothetical large enterprise.

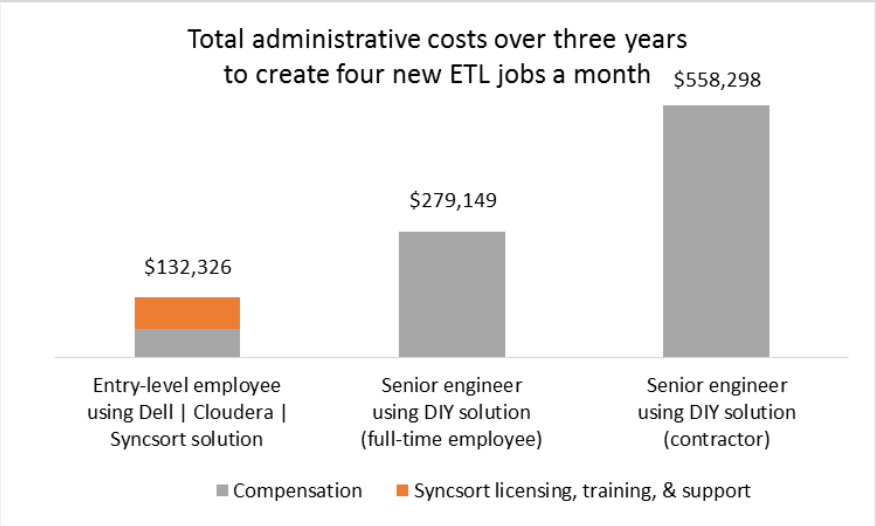

Figure 1: Compensation for workers over three years assuming they created four new ETL jobs every month. (Lower numbers are better.)

As Figure 1 shows, a beginner using the Dell | Cloudera | Syncsort solution can save $146,823 over a senior engineer on staff and $424,972 over a senior engineer working as a contractor (at twice the cost of the on-staff engineer), using our test results and cost assumptions. We define administrative costs as those for software and labor. We present only these costs because the other costs that go into each solution—such as for hardware—would be identical because we performed the tasks using the same platform.

In addition to savings that come from having a less highly compensated employee perform the design work more quickly,[4] the Dell | Cloudera | Syncsort solution offers another avenue to cost-effectiveness: performance. Additional testing in the PT labs[5] revealed that due to their extreme efficiency, the ETL jobs our entry-level workers created using Syncsort DMX-h ran more quickly than those our highly compensated expert created. This can lead to savings in server utilization. While this report focuses on staffing cost savings, the accompanying performance report covers the results of our performance testing. For more information, visit <link here>.

In this paper, we provide a quick overview of the Dell | Cloudera | Syncsort solution and then show how we arrived at the cost savings we present above. Based on our findings, this organization can begin saving on day one by choosing a Dell | Cloudera

[4] Design advantages of Hadoop ETL offload with the Intel processor-powered Dell | Cloudera | Syncsort solution www.principledtechnologies.com/Dell/Dell_Cloudera_Syncsort_design_0715.pdf

[5] Performance advantages of Hadoop ETL offload with the Intel processor-powered Dell | Cloudera | Syncsort solution www.principledtechnologies.com/Dell/Dell_Cloudera_Syncsort_performance_0715.pdf

| Syncsort integrated solution, which will continue to deliver savings throughout the life of the project. The alternative, creating an ad-hoc DIY or "roll-your-own" solution, would take longer, be more expensive to administer, and would likely run less efficiently.

ABOUT SYNCSORT DMX-h

Syncsort DMX-h is a high-performance data integration software that runs natively in Hadoop, providing everything needed to collect, prepare, blend, transform, and distribute data. DMX-h, with its Intelligent Execution, allows users to graphically design sophisticated data flows once and deploy on any compute framework (Apache MapReduce, Spark, etc. on premise or in the cloud), future-proofing the applications while eliminating the need for coding.

Using an architecture that runs ETL processing natively in Hadoop, without code generation, Syncsort DMX-h lets users maximize performance without compromising on the capabilities and typical use cases of conventional ETL tools. In addition, the software packages' industrial-grade capabilities to deploy, manage, monitor, and secure your Hadoop environment.

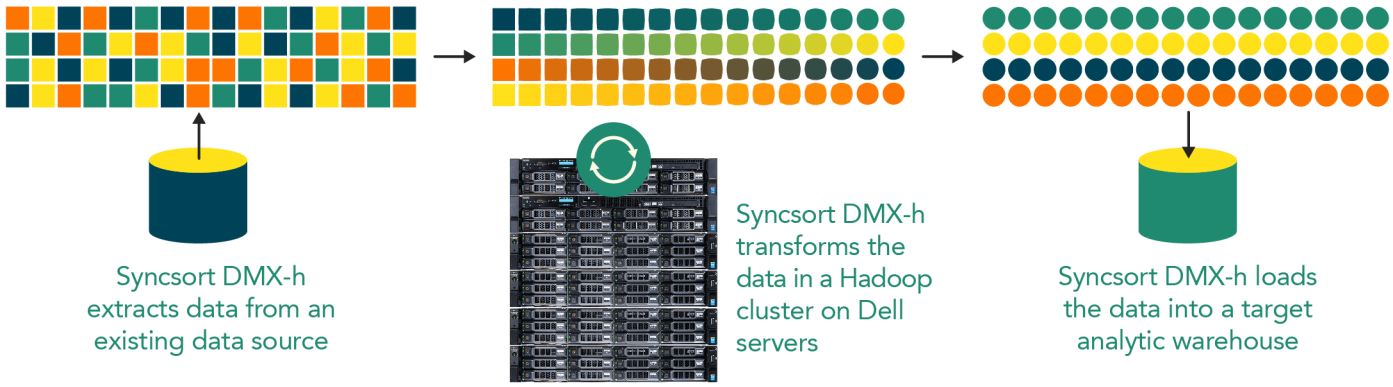

Figure 2: How the Dell | Cloudera | Syncsort solution for Hadoop works.

Syncsort SILQ®

Syncsort SILQ is a technology that pairs well with DMX-h. SILQ is a SQL offload utility designed to help users visualize and move their expensive data warehouse (SQL) data integration workloads into Hadoop. SILQ supports a wide range of SQL flavors and can parse thousands of lines of SQL code in seconds, outputting logic flowcharts, job analysis, and DMX-h jobs. SILQ has the potential to take an overwhelming SQL workload migration process and make it simple and efficient.

LOWER ADMINISTRATIVE COSTS WITH THE DELL | CLOUDERA | SYNCSORT SOLUTION

We base our cost analysis on a hypothetical large enterprise with product, sales, and customer data residing in disparate databases throughout the organization. In our scenario, this enterprise selects the Cloudera distribution of Apache Hadoop. Once they select and assemble a hardware solution to support it, they must select and install front-end tools for data analysis. This organization needs to create four new ETL jobs every month and wants to know if they should hire a senior engineer who specializes in Hadoop, use a contract senior engineer, or invest in a Dell | Cloudera | Syncsort solution. Their primary criteria for the solution are ease of use and cost effectiveness.

Based on our experience, the Dell | Cloudera | Syncsort solution can save time and therefore money a number of ways:

- Complete Hadoop DATA ingestion, preparation, and transformation tasks more quickly. Our entry-technician designed and implemented three typical Hadoop ETL jobs in 53.7 percent less time with the Dell | Cloudera | Syncsort solution than our expert doing the same task with the DIY approach.
- Lower-level staffing requirements with Dell | Cloudera | Syncsort. In our tests, entry-level techs were able to run ETL jobs after very little training. Salary for entry-level employees lowers costs compared to the expert required for the DIY solution.
- Avoid expensive recruiting and ramp up efforts.
- Build self-documenting solutions that are much easier to maintain, ramp up on, scale an implementation team for, and transition knowledge of.
- Build portable solutions that will run on new platform and execution frameworks (i.e., MapReduce, Yarn, Spark, etc.) the company might deploy in the future without the need for significant code changes.

We found that due to these factors, the Dell | Cloudera | Syncsort solution would be cheaper to deploy and operate than a DIY solution. You could start realizing savings on the Syncsort DMX-h license costs within the first few weeks based on staff time savings alone, and in even less time when compared to the cost of contract labor.

Figure 3 presents the compensation costs we use in our calculations.

	Annual compensation	Hourly compensation	Notes
Application Systems Analyst I for Dell | Cloudera | Syncsort solution	$81,272	$43	Source: Salary.com
Application Systems Analyst IV for DIY solution	$163,185	$87	Source: Salary.com
Contractor for DIY solution	$326,370	$174	Estimate at double in-house

Figure 3: Cost estimates for compensation of workers. Compensation for full-time employees includes salary and benefits. Compensation for contractor includes only salary.

In our tests, the entry-level technician using Syncsort DMX-h was able to design the ETL jobs for three use cases in 31 hours, while it took the senior engineer using a DIY solution 67 hours (see Figure 4).

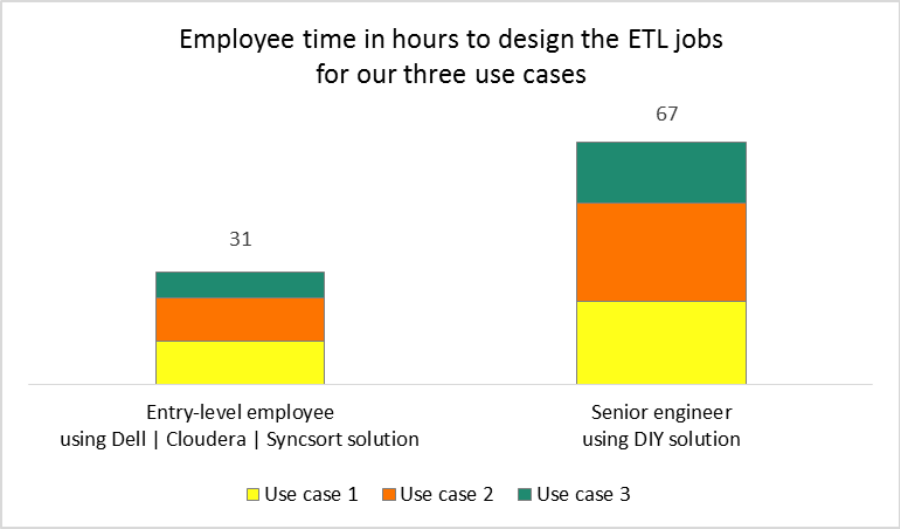

Figure 4: Our entry-level employees were able to complete the three ETL design jobs using Syncsort DMX-h in less than half the time the much more experienced engineer required using other tools. (Lower numbers are better.)

Based on the compensation rates in Figure 3 and the time to complete each task in Figure 4, we calculate the cost of compensation to create the three ETL jobs one time as $1,340 for the entry-level employees. As Figure 5 shows, compensating a senior engineer to do the same work costs 4.3 times as much if he or she is on staff and 8.7 times as much if he or she is a contractor.

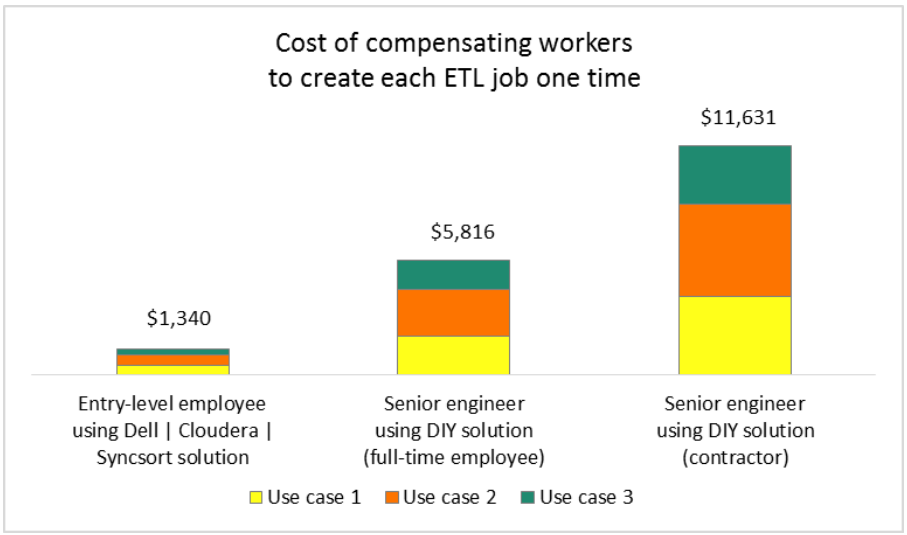

Figure 5: Because of lower compensation rates and shorter time to complete the task, the cost of entry-level workers using Syncsort DMX-h is a fraction of the cost of a more experienced engineer using a DIY approach. (Lower numbers are better.)

Of course, the Dell | Cloudera | Syncsort solution has costs associated with it. Based on prices that Dell provided to us, we calculate the three-year cost of the solution to be $68,000. This price includes licensing for five nodes and the training and professional services that enabled the inexperienced workers in our studey to use the tool to create ETL jobs.

Figure 6 presents the three-year administrative cost to create four ETL jobs every month over three years. As it shows, the enterprise could realize a savings of over $425,972 by selecting the Dell | Cloudera | Syncsort solution over a highly experienced Hadoop contractor using open-source tools.

	Average cost to create a single ETL job	Three-year compensation for creating four ETL jobs monthly	Dell | Cloudera |Syncsort solution licensing, training, & support	Total	Savings with Dell | Cloudera | Syncsort
Entry-level employee using Dell | Cloudera | Syncsort solution	$447	$64,326	$68,000	$132,326	
Senior engineer using DIY solution (full-time employee)	$1,939	$279,149	0	$279,149	$146,823
Senior engineer using DIY solution (contractor)	$3,877	$558,298	0	$558,298	$425,972

Figure 6: Three-year administrative cost to create four ETL jobs every month over three years. (Lower numbers are better.)

CONCLUSION

ETL jobs don't design and run themselves. High-level Hadoop analysis requires custom solutions to deliver the data that you need, and the cost of using senior engineers to create sub-optimal ETL jobs in a DIY hardware and software situation can be staggering.

We found that using the Dell | Cloudera | Syncsort solution could save an organization as much as $425,972 on administrative costs over three years—a savings of 76.3 percent. Because the Dell | Cloudera | Syncsort is an easy-to-use solution, entry-level employees can use it to create optimized ETL jobs after only four days of training. You can spend the money saved to innovate elsewhere in the data center, or to add even more hardware as your business expands.

ABOUT PRINCIPLED TECHNOLOGIES

Principled Technologies, Inc.
1007 Slater Road, Suite 300
Durham, NC, 27703
www.principledtechnologies.com

We provide industry-leading technology assessment and fact-based marketing services. We bring to every assignment extensive experience with and expertise in all aspects of technology testing and analysis, from researching new technologies, to developing new methodologies, to testing with existing and new tools.

When the assessment is complete, we know how to present the results to a broad range of target audiences. We provide our clients with the materials they need, from market-focused data to use in their own collateral to custom sales aids, such as test reports, performance assessments, and white papers. Every document reflects the results of our trusted independent analysis.

We provide customized services that focus on our clients' individual requirements. Whether the technology involves hardware, software, websites, or services, we offer the experience, expertise, and tools to help our clients assess how it will fare against its competition, its performance, its market readiness, and its quality and reliability.

Our founders, Mark L. Van Name and Bill Catchings, have worked together in technology assessment for over 20 years. As journalists, they published over a thousand articles on a wide array of technology subjects. They created and led the Ziff-Davis Benchmark Operation, which developed such industry-standard benchmarks as Ziff Davis Media's Winstone and WebBench. They founded and led eTesting Labs, and after the acquisition of that company by Lionbridge Technologies were the head and CTO of VeriTest.