Four Key Steps in Implementing the Data Warehouse of the Future

Digital transformation is creating new data requirements that traditional data warehouses can no longer address cost-effectively, but immediately switching to a truly modern approach is simply not an option for enterprises with existing data warehouse investments. This road map explains how to start leveraging the benefits of Hadoop today — solving real, current challenges with immediate return on investment and minimal disruption — to scale over time to the data warehouse of the future.
Today, most business leaders recognize that “data is the new oil.” Tapping into their data is helping organizations optimize their processes and improve customer interactions, as well as powering new business models. The companies that best capitalize on the opportunities afforded by digital transformation will be those that do the best job of converting their data into actionable insights.

To that end, 85% of the senior business and technology leaders surveyed for the NewVantage Partners Big Data Executive Survey 2017 said that their firms aspire to be data driven. However, so far only 37% of the companies said they have been successful in those efforts.

What is holding them back?

In many cases, legacy technology investments are interfering with efforts to become more data driven. Historically, data warehouses have served as the “single source of truth” that businesses use for analytics and reporting. But traditional data warehouses are no longer meeting enterprises’ data needs.

Many existing data warehouses are simply too slow to keep up with current demands for real-time insights based on huge volumes of data. Making matters worse, they require proprietary technology that can be very expensive to purchase and run.

To overcome these challenges, organizations have begun investigating Hadoop, which offers the performance and lower costs they need. However, most organizations already have existing investments in data warehouse technology. They can’t just “rip and replace” their data warehouse technology; that’s neither practical nor cost-effective.

Instead, enterprises need a practical road map that shows them how to get from their current data warehouse environment to a future, Hadoop-based environment that has the speed, scalability, and flexibility they need.

Problems with Today’s Data Warehouses

In the past, business analysts were satisfied with analyzing relatively small sets of data and providing reports on a weekly, monthly, or even quarterly basis. But enterprises using analytics today have much broader and deeper needs. They want to run more analytics on more data.

While older data warehouses contained primarily, or even exclusively, internal business data, organizations now want to bring in less transactional, semistructured data, such as support logs or customer trip reports, as well as data from external sources such as social media feeds, market research, and the internet. Much of that data doesn’t reside in a traditional database; it’s unstructured data, such as documents, email messages, images, videos, and audio files. While this data is rich with potential insights, traditional data warehouses are not designed to analyze it.

Today, businesses want to analyze data in real time so they can respond to emerging opportunities and threats. But they also want to be able to access more granular historical data to provide context and help them spot trends. Traditionally, data warehouses have stored raw data for three years, but that time frame is no longer sufficient to meet evolving business needs.

While 85% of firms aspire to be data driven, only 37% have been successful.
Analysts and business users also need greater flexibility than data warehouses have been able to provide. They want to combine data from disparate sources in novel ways so they can tease out subtle relationships and discover insights that otherwise would remain hidden.

Many organizations have attempted to meet these changing business requirements with their current data warehouse technologies. That requires them to spend premium resources on lower-value workloads and storing noncritical data. As a result, many businesses cannot afford to scale their traditional data warehouses in pace with the needs of their business users. Plus, those current data warehouses don’t offer the necessary performance rates, which means that, in many cases, critical reports are not ready on time.

The Data Warehouse of the Future
To provide the necessary performance at an affordable price, enterprises need a new kind of data warehouse. Ideally, they should move toward lambda architecture based on Hadoop and in-memory database technology for their data processing needs.

Lambda architecture is a data-processing architecture designed to handle massive quantities of data by taking advantage of both batch- and stream-processing methods. In this arrangement, Hadoop data stores supplement traditional data warehouses, offloading lower-value tasks and enabling new capabilities, while in-memory database technology improves performance and enables real-time analytics.

This data warehouse of the future addresses the challenges organizations are experiencing with their current data warehouses. For example, Hadoop’s distributed processing capabilities are highly scalable, making it possible to store vast quantities of data on inexpensive hardware.

In addition, Hadoop is particularly well suited to storing and analyzing unstructured data, which is often challenging for traditional data warehouses. And because Hadoop is open source software that runs on clusters of industry-standard hardware, it is much more affordable.

Also, unlike a traditional data warehouse, lambda architecture is schema-less. It doesn’t need the structured table system of a relational database or data warehouse. That means data architects don’t need to plan out and then rigidly maintain how the data storage will be organized, so users have much greater flexibility in data storage and reporting, helping them adapt to changing business conditions.

Lambda architecture’s in-memory technology can process data very quickly. It can even analyze large quantities of real-time streaming data in great depth — all in a fraction of the time required by traditional data warehouses.

In-memory technology can be pricey, but the cost is offset by the savings from using Hadoop, which is much less expensive than traditional data warehouses.

Four Key Steps in Implementing the Data Warehouse of the Future
It’s easy to see the potential benefit of moving to lambda architecture, but it’s a little more difficult to see the path to getting there. Most organizations have existing technology investments, as well as longstanding processes and procedures that make it difficult to implement lambda architecture all at once.

It is usually simpler for organizations to implement Hadoop and in-memory technology gradually, targeting the use cases with the greatest value potential and scaling out from there. Organizations generally find
that moving through the following steps maximizes the financial benefit while minimizing disruption to the company.

**Step 1: ETL Offload**
Before it can become part of the data warehouse, data must be collected from the applications and databases where it resides and converted into a standardized format. This process, called ETL (short for “extract, transform, load”) often takes a long time and requires a lot of compute resources and storage space. As a result, performing ETL jobs on the proprietary hardware used by most data warehouse systems can be very expensive. Organizations can save time and money by offloading ETL workloads to a Hadoop cluster, which uses inexpensive, industry-standard hardware and processes data much more quickly.

ETL offload is ideal as a first step on the journey to the data warehouse of the future for two reasons. First, it is usually fairly simple to get a Hadoop cluster up and running for this purpose. And second, this project usually yields significant cost savings, which helps build interest in using Hadoop for more purposes.

**Step 2: Active Archive**
As previously mentioned, most data warehouses were designed to hold about three years’ worth of active data, but today’s analytics users often want to access historical data outside that three-year window. Because this older data isn’t accessed very frequently, it isn’t cost-efficient to store it on the proprietary data warehouse infrastructure. However, Hadoop gives organizations the ability to store much larger quantities of data inexpensively. As a result, analysts and business users get access to the data they need at minimal expense.

Once you have a Hadoop implementation up and running for an ETL offload project, it’s easy to expand your use of Hadoop to store this cold data in an “active archive.” The value for the business comes in the form of actionable insights resulting from analyzing a longer range of data, stored in its original rich detail.

**Step 3: Data Repository**
Once organizations have experienced success with Hadoop for their ETL offload and active archive projects, they can build on those efforts by creating a data repository. Like a data warehouse, a data repository provides a location for data stored in other systems throughout the organization. However, while data in the data warehouse has been standardized and organized in a set schema, the data repository stores data in its raw format. It collects structured, unstructured, internal, and external data and stores it in Hadoop clusters for use in analytics.
While creating a data repository from scratch can be a daunting prospect, it’s much more manageable when you already have a Hadoop environment that you are using for ETL offload and an active archive. It’s simply a matter of scaling your use of Hadoop clusters to accommodate the range of data you need to power your analytics.

**Step 4: Lambda Architecture with In-Memory Technology**

The final step in creating the data warehouse of the future is to augment the data repository with real-time capabilities. In this step, the organization gains the full benefit of lambda architecture and is able to analyze both batch and streaming data for immediate insights. The addition of in-memory technology gives the lambda architecture the ability to process data even faster. As a result, the organization gets the timely information it needs to enable sound decision-making.

**Tips for Starting the Journey**

Building the data warehouse of the future doesn’t happen overnight. However, the process is worth it because it pays off in lower costs and the ability to support advanced analytics.

If you’re ready to start the journey, you can begin with a few simple actions:

- **Determine your own direct savings to your current data warehouse operations:** Calculate total cost of ownership for your current data warehouse and a Hadoop-based solution to help you make the business case. Dell EMC sales professionals have a tool that makes it easier to calculate how much you are currently spending and how much you could save by adding Hadoop to your data architecture.
- **Find the teams who have the most to gain from creating new insights from a richer data set.** Talk to people in finance about more accurate forecasts based on richer historical data, and speak to sales and marketing about finding, targeting, and keeping customers by analyzing new types of data.
- **Determine your first step and start small.** Solve a current challenge — whether related to cost of scaling, data timeliness, or analysts’ thirst for more data or desire for innovative analytical approaches. Once you have Hadoop in your environment, you can add new jobs and expand the cluster at your own pace.

Future white papers in this series will explain each of the steps in the journey in more detail, helping you make the transition to the data warehouse of the future as smoothly as possible.

As a member of the Dell Technologies unique family of businesses, Dell EMC serves a key role in providing the essential infrastructure for organizations to build their digital future, transform IT, and protect their most important asset, information. Dell EMC enables our enterprise customers’ IT and digital business transformation through trusted hybrid cloud and big data solutions, built on a modern data center infrastructure that incorporates industry-leading converged infrastructure, servers, storage, and cybersecurity technologies.

Intel® Xeon® processor families deliver the scale-up performance for real-time, high-capacity data analysis that can help businesses derive rapid insights to build innovative services and customer experiences. With high performance, large memory capacity, robust reliability, virtualization, and security features, the Intel® Xeon® processor family is designed for the most mission-critical workloads and the always-on enterprise.

To learn more, visit dell.com/hadoop.