CITO Research
Advancing the craft of technology leadership

APRIL 2014

Putting the Data Lake to Work
A Guide to Best Practices

SPONSORED BY

TERADATA Hortonworks

CONTENTS

Introduction

The concept of a data lake is emerging as a popular way to organize and build the next generation of systems to master new big data challenges. It is not Apache™ Hadoop® but the power of data that is expanding our view of analytical ecosystems to integrate existing and new data into what Gartner calls a logical data warehouse. As an important component of this logical data warehouse, companies are seeking to create data lakes because they manage and use data with increased volume, variety, and a velocity rarely seen in the past.

But what is a data lake? How does it help with the challenges posed by big data? How is it related to the current enterprise data warehouse? How will the data lake and the enterprise data warehouse be used together? How can you get started on the journey of incorporating a data lake into your architecture?

This paper answers these questions and in doing so strikes a blow for clear thinking. Our goal is to share best practices so you can understand how designing a data lake strategy can enhance and amplify existing investments and create new forms of business value.

What Is a Data Lake and Why Has It Become Popular?

The concept of a data lake is closely tied to Apache Hadoop and its ecosystem of open source projects. All discussions of the data lake quickly lead to a description of how to build a data lake using the power of the Apache Hadoop ecosystem. It's become popular because it provides a cost-effective and technologically feasible way to meet big data challenges. Organizations are discovering the data lake as an evolution from their existing data architecture.

The Initial Capabilities of a Data Lake

The data lake arose because new types of data needed to be captured and exploited by the enterprise.[1] As this data became increasingly available, early adopters discovered that they could extract insight through new applications built to serve the business. The data lake supports the following capabilities:

- To capture and store raw data at scale for a low cost
- To store many types of data in the same repository
- To perform transformations on the data
- To define the structure of the data at the time it is used, referred to as schema on read

[1] For more information, see "How to Stop Small Thinking from Preventing Big Data Victories"

- To perform new types of data processing
- To perform single subject analytics based on very specific use cases

The first examples of data lake implementations were created to handle web data at organizations like Google, Yahoo, and other web-scale companies. Then many other sorts of big data followed suit:

- Clickstream data
- Server logs
- Social media
- Geolocation coordinates
- Machine and sensor data

For each of these data types, the data lake created a value chain through which new types of business value emerged:

- Using data lakes for web data increased the speed and quality of web search
- Using data lakes for clickstream data supported more effective methods of web advertising
- Using data lakes for cross-channel analysis of customer interactions and behaviors provided a more complete view of the customer

One drawback with early data lakes was their limited capabilities. They were batch-oriented, and only offered a single means for user interaction with the data. Interacting with these early data lakes meant that you needed expertise with MapReduce and other scripting and query capabilities such as Pig™ and Hive™.

Hadoop 2 paved the way for capabilities that enabled a more flexible data lake. YARN (Yet Another Resource Negotiator) in particular added a pluggable framework that enabled new data access patterns in addition to MapReduce. This meant that data could be accessed in multiple ways, including: interactive, online and streaming. Familiar languages like SQL could be used in addition to MapReduce, and new programming constructs such as Cascading offered an efficient alternative to MapReduce for developers.

Hadoop 2 enabled multiple workloads on the same cluster and gave users from different business units the ability to refine, explore, and enrich data. Enterprise Hadoop has evolved into a full-fledged data lake, with new capabilities being added all the time.

The Data Lake Meets the Enterprise Data Warehouse

The first companies that created data lakes were web-scale companies focused on big data. The challenge was to handle the scale of that data and to perform new types of transformations and analytics on the data to support key applications such as indexing the Web or enabling ad targeting.

But as the wave of big data kept coming, companies that had invested years in creating enterprise data warehouses began creating data lakes to complement their enterprise data warehouses. The data lake and the enterprise data warehouse must both do what they do best and work together as components of a logical data warehouse.

At most companies, the enterprise data warehouse was created to consolidate information from many different sources so that reporting and analytics could serve everyone. The enterprise data warehouse was designed to create a single version of the truth that could be used over and over again. Here's a snapshot of the enterprise data warehouse in the past:

Like a data lake:

- The enterprise data warehouse supported batch workloads

Unlike a data lake:

- The enterprise data warehouse also supported simultaneous use by hundreds to thousands of concurrent users who were performing reporting or analytics tasks.

In addition, many of these users accessed the data through tools that were powered by SQL, the query language used by enterprise data warehouses.

Another major difference is that the enterprise data warehouse is a highly designed system. Most of the time the data repository is carefully designed before the data is stored. This model, **schema on write,** exists to support many different types of activities in which a canonical form of data is required to create a single, shared version of the truth that can be used by hundreds to thousands of people and applications. One drawback to this method is that careful design and modeling can be time consuming and reduce flexibility.

So if we look at the important dimensions comparing the enterprise data warehouse and the data lake, we can start to see the sweet spot of each.

Table 1. Comparing the Enterprise Data Warehouse and the Data Lake

Dimension	Enterprise Data Warehouse	Data Lake
Workload	Hundreds to thousands of concurrent users performing interactive analytics using advanced workload management capabilities to enhance query performance. Batch processing	Batch processing of data at scale. Currently improving its capabilities to support more interactive users
Schema	Typically schema is defined before data is stored. **Schema on write** means required data is identified and modeled in advance. Requires work at the beginning of the process, but offers performance, security, and integration. Works well for data types where data value is known	Typically schema is defined after data is stored. **Schema on read** means data must be captured in code for each program accessing the data. Offers extreme agility and ease of data capture, but requires work at the end of the process. Works well for data types where data value is not known
Scale	Can scale to large data volumes at moderate cost	Can scale to extreme data volumes at low cost
Access methods	Data accessed through standard SQL and standardized BI tools, which are supported by many different systems for reporting and analytics	Data accessed through programs created by developers, SQL-like systems, and other methods
Benefits	Very fast response times Consistent performance High concurrency Easy to consume data Rationalization of data from multiple sources into a single enterprise view Clean, safe, secure data Cross-functional analysis Transform once, use many	Executes on tens to thousands of servers with superb scalability Parallelization of traditional programming languages (Java, C++, Python, Perl, etc.) Supports higher level programming frameworks such as Pig and HiveQL Radically changes the economic model for storing high volumes of data
SQL	ANSI SQL, ACID compliant	Flexible programming, evolving SQL
Data	Cleansed	Raw
Access	Seeks	Scans
Complexity	Complex joins	Complex processing
Cost/Efficiency	Efficient use of CPU/IO	Low cost of storage and processing

The emergence of the data lake in companies that have enterprise data warehouses has led to some interesting changes. The change comes from the data lake's role in a large ecosystem of data management and analysis.

A Very Visible Data Lake Impact: ETL Migration

The ability of the data lake to store and process data at low cost and to use many different methods for transforming and distilling data has expanded the role of the data lake as a location for "extract-transform-load" or ETL, the process of preparing data for analysis in a data warehouse. Data lakes are a natural fit for ETL on big data. This sort of "scale-out ETL" allows big data to be distilled into a form that is loaded into a data warehouse for wider use.

Data lakes are also a good fit for migration of ETL processes that take up processing cycles of enterprise data warehouses which could be used for analytic and operational applications. Data can be migrated from source systems into the data lake and ETL can take place there. This migration has the advantage of allowing the ETL process to run against data from enterprise applications and from big data sources at the same time. There is a huge development effort underway by almost every ETL vendor to port their technology to Hadoop.

While there is no reason to immediately port all ETL to Hadoop, the advantages of Hadoop for this purpose make it likely that more ETL workloads will find a home there.

Migration of Analytic Results to the Data Warehouse

Companies using both an enterprise data warehouse and a data lake are often creating a distributed form of analytics. Data in the data lake is often divided into categories. Some data, such as ***video, audio, images, and other assets***, are stored in a filesystem and then use the power of Hadoop to apply various analytic techniques to extract insights. Other data may include ***unstructured or partially structured text*** that is also stored in the filesystem but that require different forms of analytics. The number of categories and the types of analytics applied to each category vary widely across industries. Call detail records may be the focus in the telecommunication industry while sensor data is especially critical in manufacturing.

In most cases, the results of analytics provide actionable insights or distilled evidence to support other forms of analytics. As data moves through this process, it gets smaller and more structured. In addition, the data becomes more interesting to more people. In other words, the business value density of the data rises.

As the business value density rises, the more natural home for the data is the enterprise data warehouse, where it can be operationalized and reused more effectively. Following this same pattern, distilled insights from data discovery platforms, in-memory databases, graph analytics engines, NoSQL repositories, and third-party services also find their way to the enterprise data warehouse.

Maturation of the Data Lake for Enterprise Use

Apache Hadoop was conceived as a tool for developers, who were more concerned with the power to transform and analyze data and create applications at low cost and high-scale than with anything else. As companies create hybrid systems that combine data lakes and enterprise data warehouses, these questions must be answered:

- Is the data secure?
- Is access controlled?
- Are all compliance regulations taken care of?
- Is activity tracked with an audit trail?
- Is data controlled and managed through a lifecycle?

Expanded Discovery and Exploration

One of the most interesting developments as the data lake and the enterprise data warehouse have become a hybrid, unified system is how users can ask questions that can be answered by more data and more analytics with less effort.

Often, enterprise data warehouse capabilities have been expanded to allow data from a data lake to be incorporated in queries using various methods. Teradata, for example, has created a SQL transparency layer called Teradata SQL-H™ that allows data in Hadoop to be explored through queries issued by the Teradata Enterprise Data Warehouse.

A powerful alternative to doing data discovery with just the data lake and the enterprise data warehouse is incorporating a specialized “discovery platform” to the data lake integration architecture. A discovery platform like the Teradata® Aster® Discovery Platform provides users with powerful, high-impact insights from big data through an integrated solution optimized for multiple analytics on all data with speed and minimal effort.

When it comes to integration between the Teradata Aster Discovery Platform and the data lake, Teradata Aster also supports SQL-H capabilities for query transparency against the data lake. In addition to SQL-H, the Teradata Aster File Store™ (AFS) enables quick and easy ingestion of multi-structured data coming from the data lake in raw format. AFS supports many Hadoop APIs that make this data transfer from the data lake fast and seamless.

For its part, Hortonworks is also connecting Hadoop with any and every system to enable exploration. Through the Hadoop YARN project, different types of workloads like streaming, search, and NoSQL can now connect with Hadoop.

It is clear that in the long term, the location of data and to some extent the analytics to answer questions will be hidden from end users. The job of end users and analysts is to ask questions. The logical data warehouse, made up of an enterprise data warehouse, a data lake, and a discovery platform to facilitate analytics across the architecture, will determine what data and what analytics to use to answer those questions.

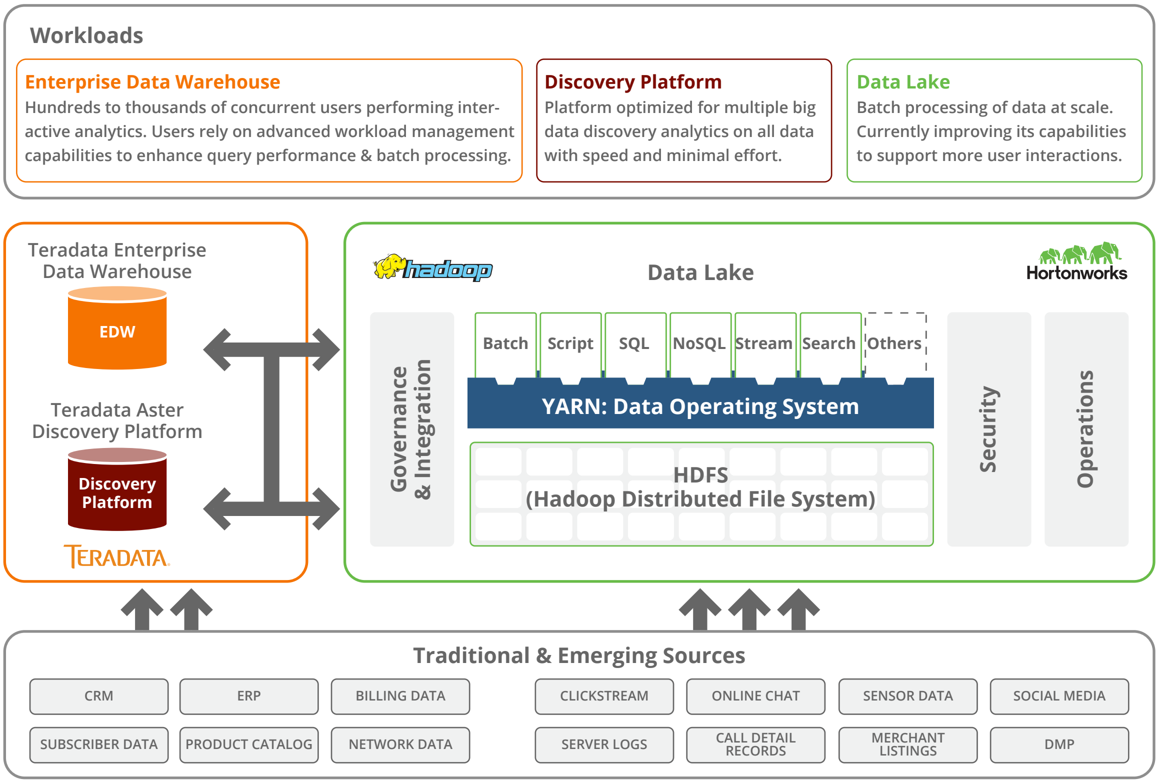
Workloads
Enterprise Data Warehouse
Hundreds to thousands of concurrent users performing interactive analytics. Users rely on advanced workload management capabilities to enhance query performance & batch processing.
Discovery Platform
Platform optimized for multiple big data discovery analytics on all data with speed and minimal effort.
Data Lake
Batch processing of data at scale. Currently improving its capabilities to support more user interactions.
Teradata Enterprise Data Warehouse
EDW
Teradata Aster Discovery Platform
Discovery Platform
TERADATA
hadoop
Data Lake
Hortonworks
Governance & Integration
Batch
Script
SQL
NoSQL
Stream
Search
Others
YARN: Data Operating System
HDFS (Hadoop Distributed File System)
Security
Operations
Traditional & Emerging Sources
CRM
ERP
BILLING DATA
CLICKSTREAM
ONLINE CHAT
SENSOR DATA
SOCIAL MEDIA
SUBSCRIBER DATA
PRODUCT CATALOG
NETWORK DATA
SERVER LOGS
CALL DETAIL RECORDS
MERCHANT LISTINGS
DMP

Figure 1. The logical data warehouse, with enterprise data warehouse, data lake, and discovery platform

Building a Data Lake

Most data lakes have emerged from incremental growth and experimentation. The idea of designing a data lake is something that few people have ever considered. The right approach to creating a data lake is to take the same approach as this white paper: follow the data. Or, perhaps more aptly, follow your data.

The path toward a data lake may differ based on where you start. Is your business entirely focused on big data? Or is big data just starting to come into view? Are you already a company with a data-driven, analytics culture? Or are you building muscle with respect to exploiting data?

The following stages represent the path that many companies take as they start to implement a data lake.

Stage 1: Handling data at scale. The first stage involves getting the plumbing in place and learning to acquire and transform data at scale. In this stage, the analytics may be quite simple, but much is learned about making Hadoop work the way you desire.

> "We have a world-class enterprise data warehouse with Teradata. And we've been leveraging it for years now. Everything we have central to our CRM and BI space is centered on the existence of that centralized data warehouse. It's created a single view of our customers, so we have a true customer profile. We know and understand all the aspects of it. But even with the size of our data warehouse in petabytes, we can't necessarily encompass all of the data, all the different devices, all the different interactions, all the different types of data that are coming at us both internally as well as externally.
>
> Some of the types of data that we're looking at today, the more unstructured and semi-structured types of data, are aspects of data that we haven't been able to retain in the past. Typically that data sits in the warehouse and then falls to tape. Some other data falls to the floor completely. There's no way for retention of that massive amount of data. But what big data is doing is actually allowing us to look at commodity-based hardware for the storage of massive amounts of data, and then taking that and coupling it with the insights that we have from the discovery platform and the combination of the data warehouse."
>
> Rob Smith, Verizon Wireless

Stage 2: Building transformation and analytics muscle. The second stage involves improving the ability to transform and analyze data. In this stage, companies find the tools that are most appropriate to their skillset and start acquiring more data and building applications. Capabilities from the enterprise data warehouse and the data lake are used together.

"Customer goes in, makes a payment; customer goes and calls the call center. We could see specifically what those customers were doing but we didn't understand why. It wasn't until we actually incorporated detailed unstructured data into our big data UDA [Teradata's Unified Data Architecture™] and actually used the discovery platform to garner insights from the interaction of the customer service rep with the customer to find out that the customer knew that they made the payment, but they wanted to verify that their phone service wasn't going to be disconnected or wasn't going to be disrupted.

So it was a timing thing the customer was concerned about. It was a simple insight that allowed us to modify our communications and drive that call-in rate down immediately. A lot of these solutions don't have to be complex; they can be very, very simple, very tactical, but have an end effect in reducing a call-in rate."

Rob Smith, Verizon Wireless

Stage 3: Broad operational impact. The third stage involves getting data and analytics into the hands of as many people as possible. It is in this stage that the data lake and the enterprise data warehouse start to work in unison, each playing its role. One example of the need for this combination is the fact that almost every big data company that started with a data lake eventually added an enterprise data warehouse to operationalize its data. Similarly, companies with enterprise data warehouses are not abandoning them in favor of Hadoop.

"The beauty of Teradata's Unified Data Architecture for us means that we can not only store the data in the right place, but we can minimize or even in a lot of cases eliminate the replication of data. What we want to do is to store it in the right place, but be able to leverage the data where it sits. Aster with the discovery platform actually allows us to have those connections both to Teradata and the data warehouse as well as our connection into the Hadoop distribution so that we can gain insights on the data that sits in each location without having to completely replicate the data as other solutions might require."

Rob Smith, Verizon Wireless

Stage 4: Enterprise capabilities. In this highest stage of the data lake, enterprise capabilities are added to the data lake. Few companies have reached this level of maturity, but many will as the use of big data grows, requiring governance, compliance, security, and auditing.

Conclusion

By following the data we have shown that the emergence of the data lake comes from the need to manage and exploit new forms of data. In essence, the shape of your data lake is determined by what you need to do but cannot with your current data processing architecture. The right data lake can only be created through experimentation.

Together, the data lake and the enterprise data warehouse provide a synergy of capabilities that delivers accelerating returns. Allowing people to do more with data faster and driving business results: That's the ultimate payback from investing in a data lake to complement your enterprise data warehouse. Logical data warehouses deliver the ability to know more about your business and gain more value from all your data.

Learn more about data lake best practices from a Teradata and Hortonworks customer ▸

Find out more about how the Teradata and Hortonworks partnership can help you ▸

This paper was created by CITO Research and sponsored by Teradata and Hortonworks.