The Data Warehouse Lives On

An Information Island within a Data Lake

JULY 2019

A White Paper by
Dr. Barry Devlin, 9sight Consulting
barry@9sight.com

Finally, we can move beyond the ancient conflict between data warehouse and data lake! It's no longer one vs. the other but, rather, how these two concepts can now work together as a modern data warehouse… for the benefit of both business and IT. This white paper shows how.

After a brief review of how data dominates modern business, the meaning and common usage of the two terms—*data warehouse* and *data lake*—are explained. Then follows a simple architectural model showing how the two are related to and complement one another in today's data-rich environment. This clearly demonstrates the power of envisaging a collaborative union of traditional and new approaches.

We then explore eight areas where the appropriate combination and placement of functionality and data across the two environments optimally supports multiple business and technical needs.

Rounding out the paper is a brief description of Cloudera Data Warehouse and how it subsumes both traditional data warehousing and the data lake, to provide a new hybrid on-premises and multi-cloud solution for modern digital business.

Contents

Copyright © 2019, 9sight Consulting, all rights reserved

SPONSORED BY CLOUDERA
www.cloudera.com

Like two tired prize-fighters, the data warehouse and data lake have been slugging it out for nearly a decade now. In the first bout, around 2010, the data warehouse triumphed: the data lake was too poorly defined, the tools too immature. By the second, mid-decade contest, the data lake was a clear winner: improved technology, combined with burgeoning data volumes and varieties, left the warehouse sprawled on the canvas.

Now, the third fight is at hand. Who will prevail? More than a few data lakes have turned into swamps and some have been drained. New implementation approaches in the Cloud and on premises have emerged. The old distinction between operational and informational computing is being dismantled by real-time business needs.

But the old questions remain. Should we replace our data warehouse with a data lake? Can a data lake offer a more cost-effective solution to business intelligence (BI) and analytics than an enterprise data warehouse (EDW)? Could we transform our proprietary data warehouse into an open source data lake? Should we, and at what price?

We're missing a fundamental point. Framed in either/or terms, we imply that one construct can displace another. Such thinking is flawed, driven by outdated marketing messaging from the early part of this century. Our prize-fighters must call off the fight—now.

Data warehouses and lakes are complementary concepts that emerge from different business needs and technological possibilities. Seen in this manner, three startling conclusions emerge. First, we can—and should—have both. Second, function can be distributed and redistributed between the two environments based on best fit. Third, the warehouse, lake, and other data management constructs can—and should—be fully integrated into a common environment to deliver a better, faster, more agile and cost-effective information preparation and processing to meet rapidly growing business needs.

A data warehouse and data lake complement one another in a modern business.

How can this be? The secret lies in understanding the differences between a data warehouse and a data lake—first, in terms of business requirements and technical possibilities, and then through a simple architectural picture that repositions both concepts in the broader picture of digital transformation.

Not your father's data, nor your mother's insights

Data isn't what it used to be. Business insights take on new meaning as digital transformation demands maximum value from today's big data. But the old need for legitimate decision making still exists. We must support both the modern digital business and the old-fashioned requirement to run and manage it well.

In the old days of your father's data and mother's insights, decision making and reporting was based on data from your own operational systems. Data was managed—from preparation and reconciliation, through use and maintenance, to archival and disposal—in the data warehouse where IT vouched for its (relative) quality. It might have been expensive, but it was doable with the data volumes of the era, and anyway, there wasn't much choice. It may have been slow, but it was fast enough for most business purposes.

Copyright © 2019, 9sight Consulting, all rights reserved

But the world has changed, and business is faster, broader, more future-oriented. With digital transformation, business has moved to real time. Predictive insights into future behavior have supplanted week-old sales reports. A whole new system of insights depends on customer Web activity—likes, clicks, dropped carts, relationships, cross-sells, and more. Analytics has shifted the focus from rearview reports and accurate financial statements to probabilistic assessments of who might do what next.

Big data—from social media, clickstreams, and the Internet of Things (IoT)—has become the foundation. The three "Vs"—volume, velocity, and variety—upended the cost equations for a traditional data warehouse, leading to the open source Hadoop explosion and, in parallel, the drive to take advantage of the cost benefits and elasticity of the Cloud. Another "V"—the veracity of these new sources—is often poor. Together with volume and velocity, this necessitates advanced statistical techniques, machine learning approaches, and multi-function analytics. Such tools operate on real-time data in data lakes and their outcomes directly integrate with and drive operational processes.

The mixed characteristics of big data, together with the emergence of Hadoop, are at the heart of the data lake and the statistical techniques used to analyze it.

Contrary to some trendy views, the need for old-fashioned BI data and reporting never disappeared. Today's data and insights must live beside those of your father and mother.

These are the challenges of today's digital business. We must have urgent new insights based on modern, mostly external data sources. But we also need to run the business in compliance with legal and accounting imperatives, based on the operational and EDW systems developed over thirty years. Today's warehouses are more powerful and sophisticated than ever before. Years of investment in these platforms and operational systems by vendors and internal IT have delivered functional and business-critical applications.

It would be a complex and expensive task to re-engineer or replace these legacy systems. Where replacement is necessary, a transition plan, perhaps spanning several years, will likely be needed. Can new open source technology help to enhance or simplify the legacy environment? Will the Cloud make data management easier? Building on a combination of data warehouse, data lake, and operational systems, can we build a path to take us where we need to go? Answering these questions requires some architectural thinking and a logical picture that positions the various components.

A warehouse on an island in a lake (beside operational systems)

Understanding what the concepts *data warehouse* and *data lake* really mean shows how they complement one another. This allows the creation of a conceptual picture showing how they can—and must—work together to create a realistic approach to managing and utilizing ***all*** data in a digital business.

After more than thirty years, the conceptual definition of a data warehouse is stable, although in functional terms, some differences (such as Kimball's dimensional / star schema data model) exist A high-level overview is shown below, based on my 2013 book *"Business unIntelligence"*[1]. The definition reflects the evolution of the concept in its initial years, with components optimized for specific purposes based on the evolving characteristics of relational databases. The ***Enterprise Data Warehouse (EDW)***, responsible for cleansing and

Copyright © 2019, 9sight Consulting, all rights reserved

reconciling data from many operational sources, is central to differentiating between a data warehouse and a data lake.

The primary purpose of a data warehouse is to provide a set of reliable and consistent data to business users in support of decision making, especially for legally relevant actions, performance tracking and problem determination. This detailed data originates from operational systems and may be subdivided or summarized in appropriately structured *data marts* by the time a businessperson sees it.

A data warehouse provides reliable and consistent data for decision making in critical situations.

In contrast, a data lake, is often defined in terms of characteristic attributes, as seen in this excerpt (lightly edited) from a 2014 blog post[2]: *"A Data Lake is characterized by:*

1. *Collect everything: contains all data, both raw sources over extended time periods and any processed data*
2. *Dive in anywhere: enables users across all business units to refine, explore and enrich data on their terms*
3. *Flexible access: enables multiple data access patterns across a shared infrastructure: batch, interactive, online, search, in-memory, etc."*

The challenge with this definition is that it implies that the data lake contains every imaginable data item, allows all sorts of processing, and can meet every business or technical need, including those covered by pre-existing systems. I propose the more limited and useful definition below, based on the original needs noted by James Dixon[3] and focused on functionality outside the scope of a traditional data warehouse and operational system environment. Although other experts may disagree with this restriction, one clear advantage is that it focuses effort in areas of most benefit to the many enterprises that have previously invested in data warehousing for existing reporting needs and are still happy with their current operational systems environment to run their business.

This division of needs allows the creation of the simple logical architecture as shown in figure 1 that positions the data warehouse and data lake relative to one another and to operational systems in a way that can be understood by the business, as well as IT.

Basic definitions

Data warehouse: a data collection, management and storage environment for decision making support, consisting of:

- ***Enterprise data warehouse (EDW):*** a detailed, cleansed, reconciled and modeled store of cross-functional, historical data, fed mostly from operational systems
- **Data marts:** subsets of decision support data optimized and physically stored for specific uses by businesspeople

Data lake: a multi-structured, often distributed data store built for:

- Ingestion and processing of high-volume, raw data from multiple sources, both external and internal, without prior structuring to a preferred model
- Accessing, formatting, processing, and managing data as required for business or technical purposes, particularly in support of advanced analytics needs

Copyright © 2019, 9sight Consulting, all rights reserved

At the heart of this figure is the data warehouse, but let's start from the data lake and work in. The data lake is as described above, fed from external and internal big data sources, such as clickstream, social media, and the IoT, via the data streams on the left. This raw data is prepared and processed into a variety of stores for use in analytics, machine learning, and a wide range of predictive and prescriptive business applications. These are *illustrative* processes that allow inferences about what is happening and may happen in the "real world." Data *timeliness and rawness* is key to illustrative computing; delays or summarization often degrade analytic value. And while there may not be the time (or need) to fully cleanse and reconcile such data, it does require enough *context-setting information*[*] *(CSI)* to make it meaningful for business and maintainable by IT.

The original data lake was defined as a purely informational environment, used solely for BI and analytics, where no new data was created. This has changed dramatically with the increasing focus on prescriptive analytics and machine learning. These processes required feedback loops from the data lake—new data, information and models—into internal sources and operational systems, indicated by the dashed blue arrows in the figure.

Within this data lake lies an *information island*, a foundation for storage and management of well-structured, fully described, and cleansed data—in essence, information. On this island is built the data warehouse, consisting of the classical EDW and data mart constructs and fed via a traditional ETL-based environment (black arrows) from the operational systems on the lakeshore. One difference of note is that while classical data warehouse architecture depicts one-way data flows from operational systems to the EDW and on to the marts, this more modern take shows two-way data movement in some of the feeds. CSI is also prominent and important in the data warehouse and may overlap or be consolidated with the CSI in the lake.

Every data lake needs a central information island where structure and order can be applied in support of data management and governance needs.

The data warehouse and operational systems contain *functional* data and information that is at the heart of running and managing a business according to ethical, legal, and accounting practices. It begins with the collection or creation of legally binding transactions that represent real business activities like creating a customer account or accepting an order. It proceeds through the operational processes that deliver value and ends in the informational processes used to track progress and address problems. Thus, it spans from Cobol programming in the 1950s to "typical" data warehouse and BI tools today. *Accuracy and consistency* of the data used is vital to functional computing: if the data is wrong, the business breaks or the regulator intervenes. Before the Internet age, these transactions were all business had to use and all that IT had to manage.

The separation of concerns underpinning the functional and illustrative concepts keeps data and processes that must be well-managed for business continuity and legality apart from those that require less management but allow more creativity. A data lake supports these latter needs, a warehouse the former. For businesspeople, this separation of storage is hidden and managed by technology as discussed in the next section.

Separately considering functional and illustrative concerns is vital balancing the legal and creative needs of the business.

These functional and illustrative concepts, with their opposing data characteristics and uses define the shores of the data lake, both external and island. They do *not* imply that

[*] Context-setting information is a broader and more meaningful name I use for metadata.

Copyright © 2019, 9sight Consulting, all rights reserved

data should never cross those boundaries. In fact, the opposite is true: the more that data can permeate these boundaries the better it is for improved business value and reduced IT costs—provided that such data movement is well understood and managed.

Further note that this architectural picture does not imply any physical placement of either box on premises, in the cloud—private or public—or any combination of these. In fact, in the emerging cloud environment, the most likely placement is a hybrid approach of on premises and cloud depending on the sources of the main types of data involved.

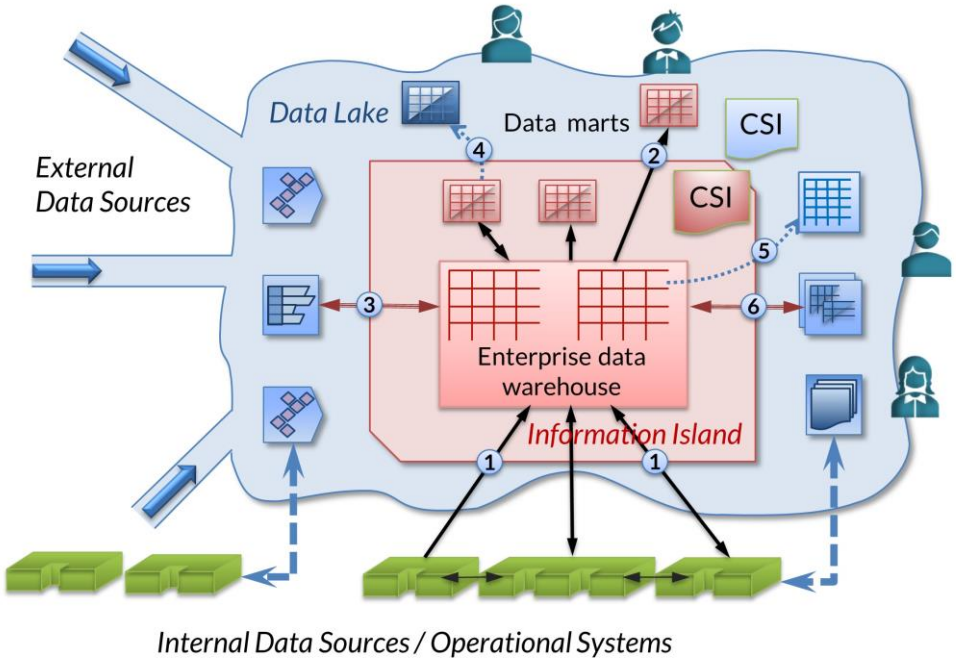


Figure 1: The warehouse on an island in a lake beside operational systems

Reinventing the data warehouse for a new era

With an integrated view of data warehouse and data lake, as well as operational systems, the extended Hadoop ecosystem offers the possibility of better and/or cheaper solutions to old data warehouse problems and opportunities to build novel business solutions on new and evolving platforms.

I hear you say, this image of a warehouse on an island in a lake is very elegant, but… how does it help to understand what to do with existing warehouse and lake systems? The answers lie mainly in the numbered tags in figure 1. They point to opportunities to combine the strengths of the two environments to provide better and/or cheaper solutions to some of the more intractable problems of traditional data warehousing and opportunities to build novel business solutions on new and evolving platforms:

1. *EDW preparation and enrichment:* getting data ready for the warehouse has long been the most complex and costly aspect of data warehousing. Such extract, transform, and load (ETL) processing often occurs in a dedicated server. It may also occur in the EDW relational database (ELT—extract, load and transform), or in a combination of both. In many cases, these systems are based on proprietary software, leading to

Copyright © 2019, 9sight Consulting, all rights reserved

high licensing costs. Furthermore, when performed in the warehouse, such processing can interfere with business-critical BI or analytic tasks. Today, these data flows are often bidirectional, adding to data management challenges.

2. *Data mart preparation:* populating data marts is similarly complex and heavy on computation, using the same ETL and ELT techniques as for EDW population.

3. *EDW population from external data sources:* while most externally sourced data is destined solely for the data lake, there are circumstances when some data must be reconciled with EDW data, for example, when external customer identifiers must be matched with internal customer numbers. Such data flows may also be bidirectional between the EDW and lake.

Optimized data processing: data preparation and enrichment in the extended Hadoop ecosystem [†] is maturing for external sources, both for batch and streaming approaches. While differences in approach remain (incremental loads predominate in data warehouses), data preparation in the data lake—in all three cases mentioned above—is becoming increasingly attractive and powerful as a way of reducing the cost and impact of ETL/ELT performed in the data warehouse environment.

Optimizing preparation and enrichment of data is a key benefit of a conjoined warehouse and lake.

4. *Data mart migration and upgrade:* data marts in a traditional data warehouse are mainly implemented in optimized dimensional relational databases (or spreadsheets). Today's multi-function analytics needs require a wide variety of non-traditional platforms, up to an including machine learning.

5. *EDW migration:* re-platforming an EDW is a significant and risky task and should not be undertaken lightly. Nonetheless, with increasing data volumes, especially of externally sourced data, that needs to be reconciled with traditional operational data, this option is becoming more interesting for many companies.

Warehouse migration to a new platform: In the case of data mart upgrade, with many advances in analytics occurring first or only in the extended Hadoop ecosystem, migrating some traditional data marts to this new environment provides worthwhile opportunities to build new business solutions. For technology-driven migration of both EDW and marts to the extended Hadoop ecosystem on premises or, particularly, in the cloud, the effort and cost of can be justified by lower licensing and operating costs, and increased elasticity compared to the legacy data warehouse.

Modern analytics needs may benefit from early migration of data marts to the extended Hadoop ecosystem.

6. *Archival:* the traditional approach to archival from data warehouses is to magnetic tape storage. While still offering by far the lowest storage cost per terabyte, tape systems often require manual IT intervention or at minimum physical tape mounting delays for retrieval, which significantly slows access for businesspeople. In addition, they must use different tools to request and/or access historical data, creating an artificial barrier to its daily use.

[†] By "extended Hadoop ecosystem", I mean the multitude of components (such as Parquet, Kudu, Druid, etc.) that have been developed around the core, as well as the core components themselves.

Copyright © 2019, 9sight Consulting, all rights reserved

Optimized cold data storage and retrieval: the extended Hadoop ecosystem is built on commodity hardware and thus offers an attractive archival environment. Although clearly more expensive per terabyte than tape, the added cost may be more than offset by the ease and speed of retrieval of archived data directly by unaided businesspeople. With retrieval in the same language (SQL) as online use, they perceive archived data as equally available (perhaps with a slightly longer access time) as online data, enabling improved use of historical trending data.

The extended Hadoop ecosystem offers new opportunities to improve archival and retrieval of cold data in the warehouse.

7. *Access*[‡]*:* with increasing quantities of mostly externally sourced data being ingested into the data lake, businesspeople face challenges in accessing such data. Until recently, much of this access has been through tools that are beyond their experience or involve programmatic approaches more suited to IT developers. Furthermore, the increased business need to use extended Hadoop-based data together with warehouse or mart data can lead to extensive copying and pasting of data between environments, adding cost, effort and potential error to business insight activities.

An integrated reporting and analytics environment: from simple reports, through spreadsheets, self-service data discovery, and ad hoc queries, to advanced analytics and machine learning, business analysts and data scientists need a powerful and consistent user experience. And with data spread across many environments, seamless access to and joining of data across many physically distinct locations—data virtualization—is vital to encourage extensive uptake by all businesspeople.

Modern, digital business demands a fully integrated reporting and analytics environment spanning a hybrid multi-cloud on-premises implementation.

8. *Hybrid implementation:* as data volumes—particularly on the Internet—have grown, cloud implementations make increasing technical and economic sense. However, for many traditional businesses, significant levels of data and existing processing remain on premises. Hybrid multi-cloud and on-premises implementations of the combined lake and warehouse infrastructure are becoming more prevalent.

Hybrid multi-cloud on-premises computing: a comprehensive data management environment spanning a hybrid implementation of data lakes and warehouses on premises and in the cloud is now needed to ensure that the data swamps that have plagued on premises data lake implementations in the past do not spread to the cloud.

Cloudera Data Warehouse

The Cloudera Data Warehouse brings together functionality from the prior Hortonworks and Cloudera offerings to provide an integrated, comprehensive solution to modern reporting and analytic needs by optimizing and extending existing data warehouse and data lake implementations.

The principle behind figure 1 and, indeed, the entire section "A warehouse on an island in a lake (beside operational systems)," is that the data warehouse, lake and operational systems cannot—and must not—be considered as independent and separate entities today.

‡ Points 7 and 8 are pervasive across figure 1 and no tags are shown there.

Copyright © 2019, 9sight Consulting, all rights reserved

Digital transformation demands that all uses, manipulation and stores of data must be seen as an integrated whole—or, at least, be capable of integration. Cloudera has espoused this principle since its earliest days and has pursued it through its Enterprise Data Hub product.

Cloudera Data Warehouse (CDW) takes this one step further. CDW essentially subsumes both traditional data warehousing and data lakes into a single entity. Recognizing that such a significant conceptual consolidation involves a very broad swathe of business needs, Cloudera identifies three areas of focus:

> Cloudera Data Warehouse subsumes traditional data warehousing and data lakes into a single, easily implemented entity.

1. *Traditional Data Warehouse Optimization:* enables enterprises to modernize and gain more value from existing data warehouse assets by offloading work from often stretched EDWs and migrating data marts to new platforms, in both cases taking advantage of components of the extended Hadoop ecosystem. Regulation, management, and reporting are key business focus areas.

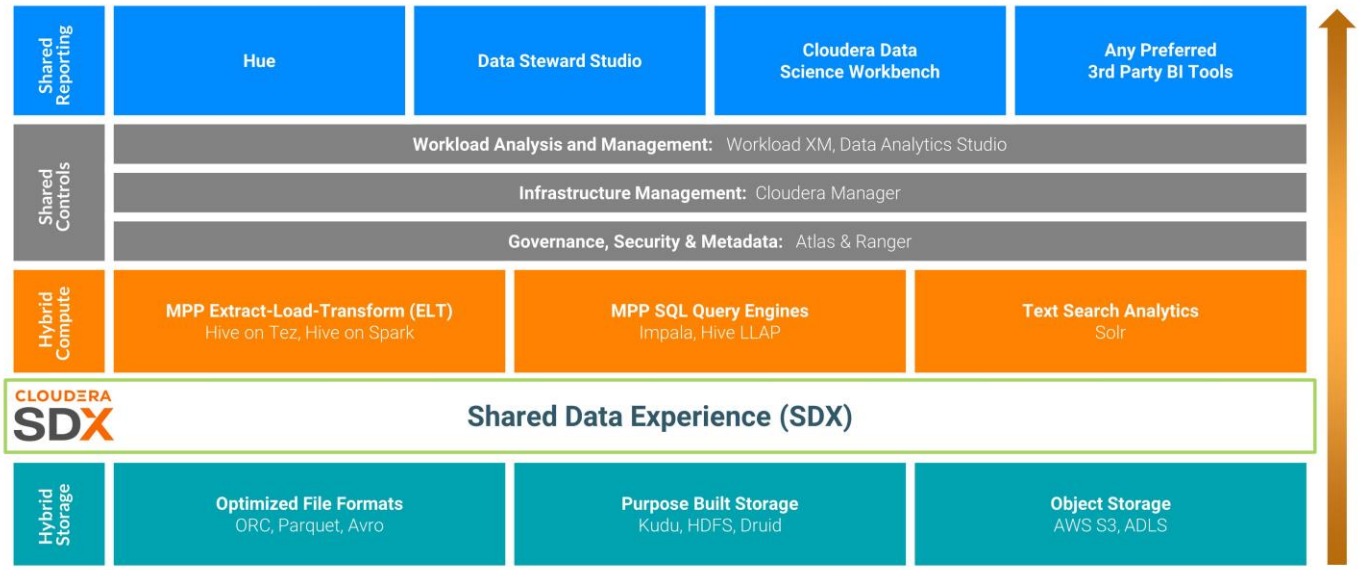
2. *Operations & Events Data Warehouse:* focuses on the growing need for (near) real-time operations, analytics, and data management. Here, the extended Hadoop environment allows tighter integration of traditional operational needs and today's prescriptive analytics.

3. *Research & Discovery Data Warehouse:* brings together all forms of analytics from basic exploratory search and relational query to machine learning-driven predictive and prescriptive analytics, delivering deeper integration between traditional data warehouse and data lake applications.

Following on from its merger with Hortonworks, Cloudera has acquired additional data warehouse optimization tooling in addition to its own extensive stack. Rationalization of these two stacks is ongoing with Cloudera's recent data warehouse announcement which brings together components from Cloudera, Hortonworks, and vendor partners to optimize the data warehouse and provide the functionality in all the areas described in the previous section.

The extensive set of components of the Cloudera Data Warehouse is shown in figure 2.

Figure 2: Cloudera Data Warehouse



Copyright © 2019, 9sight Consulting, all rights reserved

At the base of this stack resides a wide variety of storage formats, both on-premises and in the cloud, ranging from the traditional HDFS data store, through optimized columnar formats, such as ORC and Parquet, to object stores, such as AWS S3.They provide the basic data storage needed when migrating data marts and the EDW to the extended Hadoop ecosystem. Druid, a time series database for real-time analytics at scale and Kudu, a relational database designed for fast analytics.

The Shared Data Experience (SDX), sitting above the file and object stores, provides shared, consistent data context and metadata, including a shared data catalog, reliable, centralized, unified security, consistent governance for safe, compliant self-service access to data, and full data lifecycle management across both on-premises and cloud implementations.

Separate from and independent of the storage layer, the hybrid compute components provide all the data preparation and processing function required to optimize EDW preparation and enrichment, data mart preparation, and EDW population from external sources. Hive on Tez and Hive on Spark are foundational components here and partner tools from Syncsort, Informatica and more offer further levels of support and automation as needed in these areas. Also found in the hybrid compute layer is SQL-based data access and manipulation function such as Hive LLAP and Impala, and text search via Solr.

The layered CDW stack separates storage from computing and uses the Shared Data Experience to provide shared, consistent data context and metadata.

Shared controls provide the systems management function underpinning SDX. Of note here is Workload Experience Manager (Workload XM), which provides enhanced visibility and actionability to reduce migration risk, speed troubleshooting and improve uptime and resource utilization when migrating, optimizing, and scaling workloads.

Finally, at the shared reporting layer, a wide range of open-source and common third-party tools are supported. HUE is a widely used, open source SQL Cloud Editor for browsing, querying and visualizing data. Cloudera Data Science Workbench is a scalable, self-service platform for collaborative data science, using R, Python, or Scala with on-demand and secure access to Spark and Impala processing.

Conclusions

The long fight between data warehouse and lake has ended—with a win-win outcome. The Cloudera Data Warehouse brings together the functional drives of a warehouse with the illustrative motivation of a lake in a hybrid on-premises, multi-cloud solution based on the extended Hadoop ecosystem.

With thirty years of history, the data warehouse has remained a central component in decision-making support. In the past decade, a new concept—the data lake—has been introduced. At first seen as highly competitive concepts, more evolved thinking makes them equal partners. The purpose of a data warehouse is to provide the reconciled and legally foundational data needed to run and manage the business responsibly. The purpose of a data lake is to offer a place to store raw data and process it in innovative and ever-changing ways. What we call the components matters less than recognizing the different but complementary roles played.

Copyright © 2019, 9sight Consulting, all rights reserved

Drawing a conceptual picture that overlays a data warehouse on a data lake, as well as positioning traditional operational systems, allows us to see that data and functionality can be positioned and moved around within this joint environment to address new business needs and mitigate old data warehouse problems.

Some new business needs—such as multi-function analytics—are best supported by migrating some traditional EDW data or data marts to the data lake's extended Hadoop ecosystem to take advantage of new advances there. Some function—such as data preparation and archiving—can ne moved out of the data warehouse, extending the lifetime of the existing environment, or reducing the operating cost. With the right balance of data and function, a hybrid implementation can be more easily achieved.

Cloudera Data Warehouse is an integrated set of components from the extended Hadoop ecosystem and partner vendors that combines the traditional concepts of data warehouses and data lakes into a single product. It addresses four distinct but interrelated aspects of using data lake componentry to modernize and extend the traditional data warehouse. First, it supports offloading of function, such as data preparation and archive, from the data warehouse or legacy tools to reduce costs and improve performance. Second, it allows migration of selected data and function, such as data marts, to the extended Hadoop ecosystem to take early advantage of developments in these areas. Third, it offers business users the ability to use familiar BI tools to access and use all the data in the data lake, including archived data. Fourth, it offers comprehensive support for a hybrid on-premises and multi-cloud implementation.

Cloudera Data Warehouse is an integrated set of extended Hadoop ecosystem components and partner offerings that combine traditional data warehouses and data lakes into a single product.

This evolution in architecture from warehouse vs. lake to warehouse within lake promises to provide business users with much needed cross-environment illustrative function to explore data creatively, as well as optimizing the warehouse environment to focus on the functional needs of providing correct and consistent data to comply with business, legal, and regulatory needs. Furthermore, the integration and connection of lake and warehouse in the Cloudera Data Warehouse provides the capability to do even more with more data, creating new data driven opportunities for conventional and digitally transformed businesses alike.

Copyright © 2019, 9sight Consulting, all rights reserved

Dr. Barry Devlin is among the foremost authorities on business insight and one of the founders of data warehousing, having published the first architectural paper on the topic in 1988. With over 30 years of IT experience, including 20 years with IBM as a Distinguished Engineer, he is a widely respected analyst, consultant, lecturer and author of the seminal book, "Data Warehouse—from Architecture to Implementation" and numerous White Papers. His book, ***"Business unIntelligence—Insight and Innovation Beyond Analytics and Big Data"*** *(http://bit.ly/BunI-TP2) was published in October 2013.*

Barry is founder and principal of 9sight Consulting. He specializes in the human, organizational and technological implications of deep business insight solutions combining all aspects of internally and externally sourced information, analytics, and artificial intelligence. A regular contributor to TDWI Upside, Twitter (@BarryDevlin), LinkedIn, and more, Barry is based in Bristol. UK and operates worldwide.

Brand and product names mentioned in this paper are trademarks or registered trademarks of Cloudera and other companies.

---

[1] Devlin, B., *"Business unIntelligence"*, (2013), Technics Publications LLC, http://bit.ly/BunI_Book

[2] Connolly, S., *"Enterprise Hadoop and the Journey to a Data Lake"*, (March 2014), https://hortonworks.com/blog/enterprise-hadoop-journey-data-lake/

[3] Dixon, J. "James Dixon's Blog: Pentaho, Hadoop, and Data Lakes", (October 2010), https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/

Copyright © 2017, 9sight Consulting, all rights reserved