The EDW Lives On

The Beating Heart of the Data Lake

April 2017, last updated May 2018

A White Paper by

Dr. Barry Devlin, 9sight Consulting
barry@9sight.com

Finally, we can move beyond the conflict between data warehouse and data lake! It's no longer one vs. the other but, rather, how these two concepts can complement one another for the benefit of both business and IT.

First, we explore how to optimally support different business and technical requirements by the appropriate placement of functionality such as data preparation, archival and business access across the two environments. A simple architectural model defines what warehouses and lakes actually are and how they complement one another. This clearly demonstrates the power of such collaborative thinking between traditional and new approaches.

A brief description of Hortonworks' Enterprise Data Warehouse Optimization Solution rounds out the paper.

Contents

Sponsored by:

HORTONWORKS®

Like two tired prize-fighters, the data warehouse and data lake have been slugging it out for half a decade now. Should we replace our data warehouse with a data lake? Can a data lake offer a more cost-effective solution to business intelligence (BI) 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?

These and related questions miss a fundamental point. Framed in either/or terms, the implication is that one construct can displace another, that one approach can be chosen to the exclusion of all others. If fact, such thinking is flawed, driven by outdated marketing messaging from the early part of this decade and, indeed, earlier.

The reality is that data warehouses and lakes are largely complementary concepts that emerge from different business needs and technological possibilities. Seen in this manner, two startling possibilities emerge. First, we can—and should—have both. Second, function can be distributed and redistributed between the two environments based on best fit. Together, they promise performance enhancements and cost reductions—a better, faster, more agile and cost-effective BI to meet rapidly growing business needs.

A data warehouse and data lake are complementary to 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.

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

In the good old days, all decision support and reporting was based on data that came from operational systems you owned or developed. This 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 volumes of the era, and anyway, there wasn't much choice. It may have been a bit slow, but it was fast enough for most business purposes.

Then the world changed. With the Internet and e-commerce, business moved to real time. Week-old sales reports were superseded by predictive insights into future customer behavior. This whole new system of insights depended on customer behavior on the Web—likes, clicks, comments, dropped carts, relationships, upsells and cross-sells, etc. BI became analytics: the focus shifted from backward-looking reports and accurate financial statements to probabilistic assessments of who might do what next.

The business need had evolved and expanded: faster, broader, more future-oriented. Big data, first from social media and clickstreams, and more recently, from the Internet of Things, became the foundation. The now famous three "Vs"—volume, velocity, and variety—upended the cost equations for a traditional data warehouse, leading to the open source Hadoop explosion. Note here, however, that the quality and reliability of these new sources was and is often poor. And contrary to the views of some trend setters, the need for old-fashioned BI data, reporting and analysis did not disappear. Today's data and insights must live beside those of your father and mother.

This is the challenge of today's digitalized business. We must have the urgent new insights based on modern and mostly external data sources. But we also need to run the business in compliance with legal and accounting imperatives, based on the operational and data warehouse systems developed over the past thirty years. It would be a complex and expensive task to re-engineer or replace these legacy systems. Indeed, why would you? Today's data warehouses are more powerful and sophisticated than ever before. Years of investment in these platforms and systems by vendors and internal IT have delivered functional and business-critical applications.

But, can the new technology help to enhance or simplify the legacy environment?

Old wine in new bottles

On the foundation of new data and novel insights together, a new data management and delivery ecosystem based on Hadoop has emerged over the past decade: a data lake (whose definition we'll return to later). Now, as this ecosystem matures, the opportunity arises to use it to provide better and/or cheaper solutions to some of the more intractable problems of traditional data warehousing:

A data lake offers the possibility of better and/or cheaper solutions to old data warehouse problems.

1. *Preparation and enrichment:* getting data ready for the warehouse has long been the most complex and computationally expensive component of a data warehouse. Traditionally labelled extract, transform, and load (ETL), such processing is performed either in a dedicated ETL server, within the relational database of the warehouse (often called ELT—extract, load and transform), or in a combination of both. In many cases, these systems are based on proprietary software, leading to high licensing cost. Furthermore, when performed in the warehouse, such processing is costly and can interfere with business-critical BI or analytic tasks.

   Pumping through the data lake: data preparation and enrichment in the Hadoop environment is maturing for external data sources, starting with batch and moving toward streaming approaches. While some differences in approach remain (incremental loads predominate in data warehouses), data preparation on Hadoop is becoming increasingly attractive and powerful as a way of reducing the cost and impact of ETL processes performed in the data warehouse itself.

2. *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 business users. In addition, users must use different tools to request and/or access historical data, creating an artificial barrier to its daily use.

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

3. *Access:* with increasing quantities of mostly externally sourced data being ingested into the Hadoop environment, business users face challenges in accessing such data. Until recently, much of this access has been through tools that are beyond the experience of business users, or involve programmatic approaches more suited to IT developers. Furthermore, using Hadoop-based data together with data traditionally found in the warehouse or data marts could involve copying and pasting data from one environment to the other, adding cost and effort to business users lives.

   Swimming in the data lake: business use of data has centered around a "rows and columns" paradigm since the earliest days of BI. Whether through spreadsheets, multi-dimension cubes, or SQL queries, offering such access to data in the data lake is vital to its widespread use by "ordinary" business users. With data of business interest now spread over two or more environments, it becomes increasingly important to join data across physically distinct locations—a facility known as data virtualization. Such tools are vital to entice business users to dip their toes into the data lake.

These instances and others emerging—such as streaming data, hybrid transactional / analytical processing (HTAP), and time series—all point to the fact that a data lake has value to offer to a data warehouse and *vice versa*. The complementary nature of the two environments is clear. It's time to look more closely at how they interact.

A warehouse by a lake

Since its inception[1] in the mid-1980s, the data warehouse had become the go-to source for all BI reporting, querying by the beginning of the millennium. Its original, primary driver was to offer a consistent, reconciled data foundation (often called *single version of the truth*) across all business functions and departments. Soon, it was declared that all decision support data should flow through the warehouse for quality control, specialized analytics, (and, also, because there was no other obvious platform for such work). This approach had predictable impacts on performance, agility, costs and so on. These challenges, together with new business needs led to continuous improvements in the underlying software, leading to the modern and powerful environment seen today. However, bear in mind the original driver mentioned above—data reconciliation: it's far more important than the idea that all data should reside there.

The data lake concept was first floated in 2010 by James Dixon[2]. Its initial description was simply a large store of raw data, driven in part by the burgeoning growth of big data, but also in reaction to the often expensive and limiting structuring of data in the warehouse and data marts. Soon, I and others[3] criticized the looseness of the definition, troubled by the possibility of a "data swamp" of poorly managed data. Dixon, among others, expanded the scope of the data lake to include all data, including even that traditionally stored in the data warehouse, an approach that gives rise to addition concerns about the costs and challenges of "ripping and replacing" the data warehouse ecosystem.

Despite such concerns, the popularity of the data lake concept has grown rapidly. An 9sight/EMA survey[4] published in November 2016, shows that fully two-thirds of the respondents reported that they had currently adopted a data lake strategy, up from just over 50 percent in a year and a quarter. Furthermore, almost 15 percent of the 2016 survey respondents said that a data lake had replaced their data warehouse. However, the survey showed considerable confusion about what might actually be the definition of a data lake. Data lake components scoring highly included data warehouse, operational data, departmental and analytical data marts, as well as data mining sandboxes.

So, what is a data lake? And, while we're in definition mode, what is a data warehouse?

Defining terminology: data warehouse and data lake

After more than thirty years, the conceptual definition of a data warehouse is stable, although in logical or functional terms, some differences are evident[*]. A high-level overview is shown in the accompanying box, based on my 2013 book *"Business unIntelligence"*[5]. The definition reflects the evolution of the data warehouse concept in its initial years, with particular components optimized for specific purposes based on the evolving characteristics of relational databases over three decades. The EDW, with its role in cleansing and reconciling data from many sources, is central to understanding the difference between a data warehouse and a data lake.

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*

Data marts: *subsets of decision support data optimized and physically stored for specific uses by business people*

The primary purpose of a data warehouse is therefore 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, but may be subdivided or summarized as appropriate by the time a business user sees it.

In contrast, a data lake is often defined in terms of attributes that characterize it, as seen in the following excerpt, lightly edited from Shaun Connolly's 2014 blog post[6]:

"A Data Lake is characterized by three key attributes:

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."*

Data lake: *a data store built for the ingestion and processing of any raw data from multiple sources without prior structuring to a preferred model. In this store, data can be accessed, formatted, processed, and managed as required for business or technical purposes.*

The challenge with this definition is that it implies that the data lake contains every imaginable data item, allows processing however needed, and can basically meet every business or technical need. I propose the more limited and useful definition to the right, based on the original needs noted by James Dixon and focused on functionality outside the scope of a data warehouse. Although other experts may disagree with this latter restriction, one clear advantage is that it focuses effort in areas of most benefit to the majority of enterprises that have previously invested in data warehousing for existing needs.

[*] In particular, Kimball's data warehouse uses a dimensional data model (star schema) as a foundation for "slice-and-dice" analysis. Such a construct appears as a data mart in the above definition.

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

The block labeled ***functional*** 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.

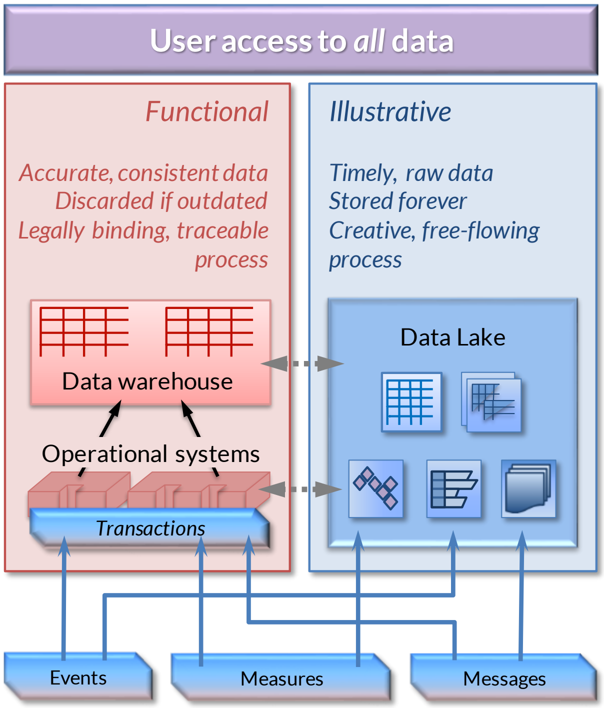

Figure 1: The warehouse beside the lake

E-commerce, social media, and the IoT shows that there is "rawer" data from which transactions arise. This data/information—now all digitized and potentially collected—consists of events (e.g. a click on a website), measures (the speed of your car) and messages (everything from Tweets to videos). Such data supports ***illustrative*** processes that allow inferences about what is happening in the "real world", and are the basis for predictive and prescriptive analytics. Data ***timeliness and rawness*** is key to illustrative computing; delays or summarization often degrade analytic value.

These functional and illustrative purposes, with their opposing data characteristics and uses lead to an architecture that defines the shores of the data lake. Raw data—in the form of events, measures and messages—is ingested into the IT systems of the business. Vast quantities of raw data may be stored in the data lake as the basis for analytics. Traditional operational systems craft raw data into the legally-binding transactions of the business and made it available for decision making through the data warehouse. This separation of concerns 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 business users, this separation of storage is hidden and managed by data virtualization tools and metadata-based approaches. Deep links (the dotted arrows) exist between the two environments for specific business needs such as prescriptive analytic approaches that are becoming more prevalent.

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.

With the balanced data warehouse/lake architecture shown here, the data processing, archival and access solutions described in the previous section—as well as other possibilities to use the data lake to enhance, support or extend the data warehouse—can be efficiently delivered on the less costly hardware and software of the data lake.

Hortonworks EDW Optimization Solution

In February 2017, Hortonworks rolled out the first phase of an Enterprise Data Warehouse[†] (EDW) Optimization Solution to facilitate the use of data lake function in support of building, managing, and using EDW data. The approach represents the culmination of an ongoing evolution in thinking about data lakes among the Hadoop community. This is a very welcome development, moving the debate from lake vs. warehouse to a more realistic position of using the strengths of the Hadoop ecosystem to address new business needs, as well as build upon and enhance existing technology investment.

In line with its long-standing strategy of providing fully integrated and tested distributions of a set of Hadoop ecosystem components, Hortonworks now brings together additional components from vendor partners to bridge the gap between the data warehouse and data lake and provide the functionality in the three areas described on pages 3 and 4: preparation and enrichment, archival, and access.

Preparation and enrichment of warehouse data

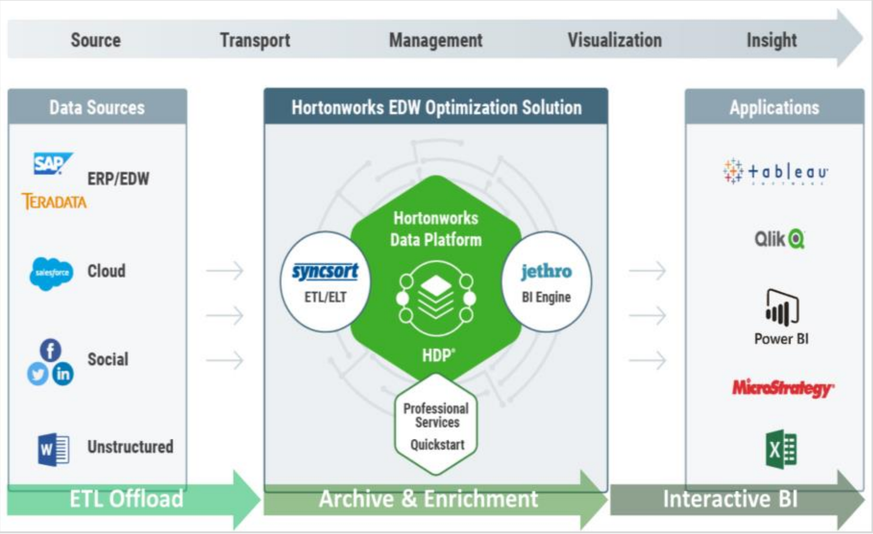

Figure 2: Hortonworks EDW Optimization Solution

The EDW Optimization Solution offers simple drag-and-drop ETL batch and streaming workflows by incorporating DMX-h from Syncsort, which offers access to data from multiple sources, including relational databases, NoSQL stores, mainframe files and databases, etc., and generates highly scalable ETL function in Hive and the Hortonworks Data Platform (HDP).

The Hadoop platform offers great flexibility in terms of what kind of data can be stored there. The traditional EDW is narrowly focused on structured data, with a strict upfront design—known as "schema on write". Hadoop, on the other hand, can store any shape of data, structured, semi-structured, unstructured and will associate to a structure on access— "schema on read". This enables modern data sources that do not easily fit into the EDW—like click streams, web logs, device data, etc. to be more easily prepared and enriched here, as well as storing and supporting previously archived data.

In cases where existing ETL is performed in the data warehouse, this approach can shift over 50% of processing off the data warehouse platform, resulting in significant performance and service level agreement (SLA) improvements. Where the existing ETL is carried out on proprietary systems, these legacy tools can be phased out over time to accrue worthwhile cost savings.

[†] Hortonworks uses the term "Enterprise Data Warehouse" to include both the reconciliation and data mart functions described on page 5 to emphasize the enterprise-wide nature of the overall data warehouse.

Archival of warehouse data

Storing rarely used—or cold—data in a high-performance data warehouse is poor use of an expensive resource. Traditional archiving moves old, seldom-used data to tape. With a data lake, such data can be moved to Hadoop. In addition, the same approach can be used for any data that is rarely used in the warehouse, including raw data from the Internet of Things, click streams, social media, and other sources.

To Syncsort DMX-h, the data warehouse is just another source, so the advantages listed in the previous section apply equally here. Similarly, all archived data benefits from the access components listed next, making this data easily available at all times for analysis—for the user, it looks as if it never left the data warehouse.

Access to Data Lake Data

Hive 2.0 with LLAP (Live Long and Process) offers as fast as sub-second, scalable SQL analytics and intelligent in-memory caching. The result is a 26x times performance improvement over Hive 1.0, enabling true interactive queries on data stored in HDFS. As a result, traditional data access and query tools from the data warehouse world, such as Qlik, Tableau, and more, can directly use data in the data lake as a basis for queries and analytics.

Together with Jethro Data, the EDW Optimization Solution further accelerates data access through intelligent indexing. This new feature automatically indexes each column, aggregates data for OLAP cubes, and caches highly accessed data, all through self-driving that requires no data engineering by IT or users. Jethro's innovative approach optimizes how data is accessed, delivering performance at scale for thousands of concurrent users with response time in seconds, and scalability to serve billions of rows for query.

Conclusions

Despite three decades of history, the data warehouse remains a central component in any decision making support architecture. In the past five years, a new component—the data lake—has been introduced to the mix. At first seen as highly competitive to the data warehouse, more evolved thinking places it as an equal partner. The data warehouse retains responsibility for reconciled and legally foundational data needed to run and manage the business responsibly. The data lake, on the other hand, offers a place to store raw data and process it in innovative and ever-changing ways.

In addition, the data lake offers an environment to offload from the data warehouse some function that has been problematical in the past. Such function—such as data preparation and archiving—can have performance and SLA impacts on the data warehouse and may more cheaply performed in the data lake. By moving such function out of the data warehouse, the lifetime of the existing environment can be extended or the operating cost reduced.

> *The performance and use of a data warehouse can be optimized by moving some functionality to another appropriate platform such as the data lake.*

Hortonworks EDW Optimization Solution is an integrated set of components from the Hadoop ecosystem and partner software vendors that address three distinct but interrelated aspects of using the data lake to improve and extend the data warehouse. First, it supports offloading data preparation (ETL) from the data warehouse or legacy tools to reduce costs and improve performance. Second, it allows archiving of warehouse data to an online store rather than tape, enabling users faster and simpler access to this historical data. 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. We may envisage that further function will be added in the future.

This evolution in architecture from warehouse vs. lake to warehouse beside 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, this integration and connection of lakes and warehouses provides the capability to do even more with more data, creating new data driven opportunities for traditional and Internet-based businesses alike.

*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 new book, **"Business unIntelligence—Insight and Innovation Beyond Analytics and Big Data"** (http://bit.ly/BunI-Technics) was published in 2013.*

Barry is founder and principal of 9sight Consulting. He specializes in the human, organizational and IT implications of deep business insight solutions that combine operational, informational and collaborative environments. A regular tweeter, @BarryDevlin, and contributor to numerous publications, Barry is based in Cape Town, South Africa and operates worldwide.

Brand and product names mentioned here are trademarks or registered trademarks of the Apache Foundation, Hortonworks and other companies.

[1] Devlin, B. A. and Murphy, P. T., *"An architecture for a business and information system"*, IBM Systems Journal, Volume 27, No. 1, Page 60 (1988) http://bit.ly/EBIS1988

[2] 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/

[3] Stonebraker, M. *"Why the 'Data Lake' is Really a 'Data Swamp'"*, (December 2014), http://cacm.acm.org/blogs/blog-cacm/181547-why-the-data-lake-is-really-a-data-swamp/fulltext

[4] Myers, J., Wise, L. and Devlin, B., *"Charting the Expanding Horizons of Big Data"*, (November 2016), http://bit.ly/BD-survey16

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

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