9sight Consulting

The Data Warehouse Lives On

New Memes, Old Themes

April 2022

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, integrated construct… 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 usage of the terms *data warehouse* and *data lake* are explained. Then follows a simple architectural model showing how the two are related and complement one another in today's environment. This clearly demonstrates the power of envisaging a collaborative union of traditional and new approaches.

Three new architectural patterns—data fabric, data lakehouse, and data mesh are also discussed and related to this model.

We then explore ten 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 the Cloudera Data Warehouse and how it subsumes both traditional data warehousing and the data lake, to provide a new hybrid on-premises, multi-cloud solution for modern digital business.

Contents

Sponsored by CLOUDERA

www.cloudera.com

In 2010, the data management world awoke to a new meme. James Dixon's blog post[1], introduced us to the data lake: "If you think of a data mart as a store of bottled water—cleansed and packaged and structured for easy consumption—the data lake is a large body of water in a more natural state. The … various users of the lake can come to examine, dive in, or take samples." Like all good memes, the image is simple and memorable.

Leading-edge enterprises rapidly took to the water, driven by technical needs to use big data pouring in from the internet. Promised cost savings from open-source software and commodity hardware also inspired. Elsewhere, the motive was political—an effort to distance IT from a struggling data warehouse project. Some claimed to be replacing data warehouses with new data lakes. Throughout the decade, the debate between lakers and warehousers escalated. Vendors and consultants chose sides as new implementation approaches in the cloud came to dominate the market. Architecture and technology were incorrectly conflated. Vague high-level patterns drove often shaky implementations. More than a few data lakes turned into swamps; some have been drained.

Confusion reigned, which the 2019 version of this white paper[2], attempted to dispel, declaring that these concepts are different, and both are required: "Data warehouses and lakes are complementary concepts that emerge from different business needs and technological possibilities." The warehouse focuses on getting correct results, such as regulatory reporting and management decision making; the lake provides for exploration and innovation, such as data science and machine learning.

Since then, a consensus has emerged. Lakes and warehouses do indeed complement one another, and data, as well as management processes, should be shared between them. Advanced and hybrid cloud technologies have enabled this conclusion: most lakes and warehouses have a major cloud component. Lake and warehouse terminology and meaning have become intertwined. According to the 2021 Forrester Wave™ Cloud Data Warehouse[3], the most common use cases are analytical in nature and would have been previously most often delivered via data lakes.

A data warehouse and data lake complement one another in a modern digital business and should be built on a common platform.

However, continuing implementation challenges have led to the emergence of three new memes (or, at a more detailed level, architectural patterns): data fabric, data mesh, and data lakehouse. Despite claims by their proponents that each offers the ultimate solution to data management issues, each has distinct strengths and weaknesses. Furthermore, inconsistent terminology, definitions, and competing claims continue to sow new seeds of confusion in a market that often still blurs basic data management concepts.

So, to questions about the meaning of *lake* and *warehouse* are added those about *mesh* and *fabric*. And additional queries arise. Is a fully decentralized approach to data management now mandatory? Can artificial intelligence solve the old problems of metadata? Is a single technology base possible or even desirable?

This paper offers possible answers these and other complex questions. And provides a firm foundation for thinking about them. Because, in fact, there are multiple answers and options, depending on your business needs and current solutions.

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 were based on data from your own operational systems. Data was managed—from preparation and reconciliation, through use and maintenance, to archival and disposal—in a centralized data warehouse where IT vouched for its (relatively high) quality. It might have been expensive, but it was doable, and anyway, there wasn't much choice. It may have been slow, but it was fast enough for most business purposes.

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[*]—social media, clickstreams, and the Internet of Things (IoT)—has become the foundation. Its three "Vs"—volume, velocity, and variety—upended the cost equations for traditional data warehousing, driving enterprises to seek the cost benefits and elasticity of open-source and commodity solutions. Advanced statistical techniques, machine learning, and multi-function analytics, operating on real-time data in data lakes must directly integrate with and drive operational processes—requirements seldom seen in data warehouses. Increasingly, cloud is supplementing or replacing on-premises solutions. But, contrary to some trendy views, the need for old-fashioned BI and reporting never disappeared, especially in strategic management, financial reporting, and regulation. Today's data and insights must live beside those of your father and mother. The founding principles of data warehousing—combining and reconciling quality data from disparate sources—are as necessary as ever and apply to a growing percentage of lake data.

The mixed characteristics of big data, together with open-source software and now, cloud technology, are at the heart of the data lake.

The challenge now is to integrate these disparate needs while adopting a highly distributed hybrid cloud environment. It will be a complex and expensive task to move business-critical, on-premises warehouse and operational systems to cloud-centric environments. Similarly, for on-premises data lakes. Current cloud-based lakes and/or warehouses are often poorly integrated; consolidation is not straightforward.

A combination of data warehouse, data lake, the emerging data fabric, mesh, and lakehouse, and operational systems will be required. Architectural thinking that positions the various components is needed. Let's paint a new picture for use by both business and IT.

[*] The terms *big data* and the *three "Vs"* have lost their cachet but continue to offer a useful shorthand for the characteristics of much externally sourced data.

A warehouse on an island in a lake

Combining the data warehouse and data lake may seem simple, but they are actually very different concepts. The picture of ***a warehouse on an island in a lake*** offers a straightforward, logical pattern of how they complement one another and how they can—and must—work seamlessly together to manage and utilize ***all*** data in a digital business.

After more than thirty years, the conceptual definition of a data warehouse is largely stable, although in functional terms, some differences in design (such as Kimball's dimensional / star schema data model) still exist. A basic definition below (based on my book *"Business unIntelligence"*[4]) reflects the evolution of the concept, with components optimized for specific purposes driven in part by the evolving characteristics of relational databases. The *Enterprise Data Warehouse (EDW)*, responsible for cleansing and reconciling data from many operational sources, is central to differentiating between a data warehouse and lake.

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

The primary purpose of a data warehouse is to provide a set of reliable and consistent *information* to business users in support of decision making, especially for legally relevant actions, performance tracking and problem determination. The term *information* here is important: it denotes the fact that the warehouse[†] contains more than raw data, but rather real information contextualized and cleansed for its valid and correct use. This detailed information may be further subdivided and/or summarized in appropriately structured *data marts* for performance, ease-of-use, or security by the time any business-person sees or uses it.

The information in a warehouse or marts originates principally from *operational systems*, both traditional on-premises and modern web-based varieties. Other sources may also be defined, provided that the data coming from them is of sufficient quality and can be contextualized into useful and usable information. For example, data in a lake can be ingested into a warehouse via some cleansing and reconciliation process based on agreed data governance rules.

In contrast, a data lake, is often defined in terms of declared attributes: "a data lake is characterized by collect everything, enable anyone to dive in anywhere, and allow flexible access in multiple patterns"—a 2014 definition from Hortonworks' Shaun Connolly, (no longer online). This 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. This, in my view, is too utopian in scope and I prefer to use the more limited and useful basic definition below, based on James Dixon's[Error! Bookmark not defined.] original post, which focuses on function outside the scope of traditional data warehouse and operational environments.

[†] *Data warehouse* should perhaps be called *information warehouse*. Unfortunately, the latter term was trademarked by a major vendor in the early 1990s and thus unavailable for general use.

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

Operational system: a well-managed, process-oriented system, generating the legally binding data to run a business:

- Traditionally, order entry, billing, account management, and similar on-premises systems
- In a digital business, such processes are often replaced or supplemented by self-service, web-based apps generating legally binding data

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

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

These disparate sets of needs and uses has typically led to distinct technology implementations and the creation of silos of disconnected data. However, understanding the differences also allows us to create an integrated, silo-free architectural pattern. Shown in figure 1, this pattern positions the data warehouse and lake relative to one another and to operational systems in a way that can be understood by the business, as well as IT.

At the heart of this figure is the data warehouse, but let's start from the *lake of data* and work into the *island of information*. The lake is, as described above, fed from external 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—but only when needed by data scientists and business analysts (schema on read)—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 metadata or *context-setting information*[‡] *(CSI)* to make it meaningful and maintainable.

The original data lake was defined as a purely informational environment, used solely for analytics and data science, 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 lake of data—new data and models—into

[‡] Context-setting information is a clearer and more meaningful name for traditional metadata plus the broader set of contextual information required in a digital business.

operational systems, both on-premises and web-based, as indicated by the dashed blue arrows in the figure.

Within this lake of data lies an *island of information*, a foundation for storage and management of well-structured, fully described, and cleansed information—fully contextualized data. On this island is built the data warehouse, consisting of the classical EDW and data mart constructs, and fed via traditional ETL-based (extract, transform, and load) tools, shown as black arrows, from the operational systems on the lakeshore. Note 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.

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

CSI is also prominent and important in the data warehouse and may overlap or be consolidated with the CSI in the lake. Metadata/CSI is a key component in this architecture. It was long recognized as central to data warehouses, but early data lake implementations often ignored it and often suffered from severe governance issues as a result. Metadata is now accepted as a core component of data lakes as well as warehouses. In fact, without consolidated and consistent, shared context-setting information, any effort to create a combined warehouse/lake environment will fail.

The data warehouse and operational systems contain *functional* data/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 that business had to use and all that IT had to manage. And today, although

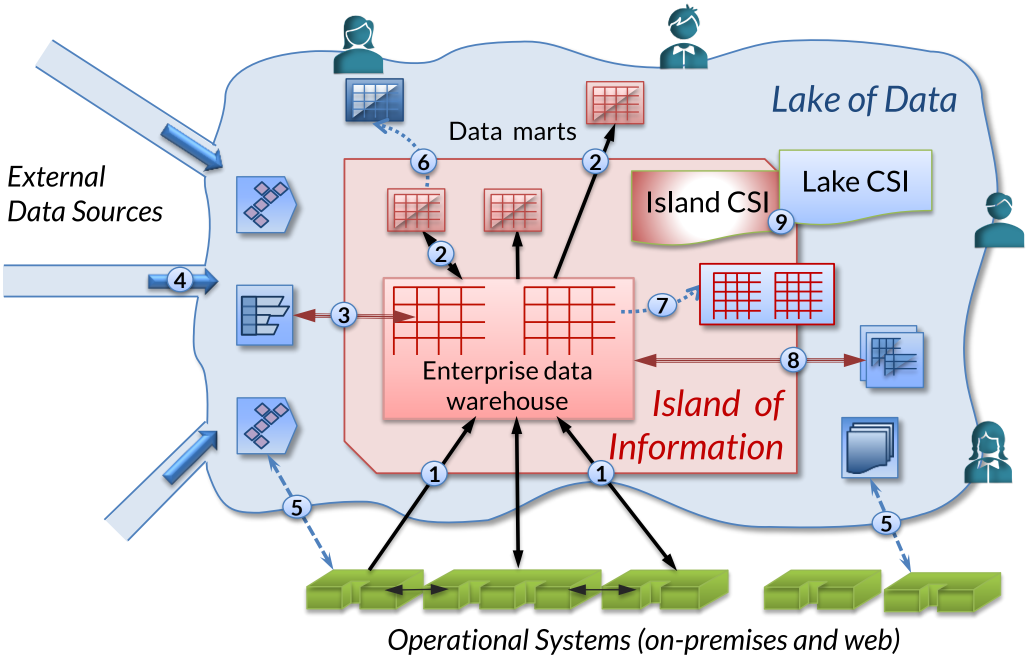

Figure 1: A warehouse on an island in a lake

dwarfed in volume by big data and often seen as somewhat old-fashioned, it remains central to running and managing a digital business.

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 should be hidden and managed by technology as discussed in the section *"Reinventing warehouse and lake for a new era"*.

These functional and illustrative concepts, with their opposing data characteristics and uses define the shores of the lake of data and of the island of information. They do *not* imply that data should never cross those boundaries. 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.

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

Further note that this architectural picture does not imply any physical placement of any component on premises, in the cloud—private or public—or any combination of these. In fact, with the increasing focus on the cloud environment combined with a long history of legacy on-premises systems in many enterprises, a common arrangement is a hybrid approach of on premises and cloud depending on the main sources of data involved.

Introducing data fabric, data lakehouse, and data mesh

Recently, several vendors and consultants have tried to address the various challenges of combined data warehouse / data lake environments in different ways. This has resulted in three architectural patterns of varying scope, approach, and practicality. We'll examine them in order of their current ease of implementation.

Data fabric

Data fabric is principally a function/product framework defined by major analyst firms, first in 2016 by Forrester[5] and, more recently, highlighted by Gartner in their 2021 Top 10 Data & Analytic Trends. Numerous vendors support the framework, offering products to implement all or parts of the function. An enhancement of the *logical data warehouse* concept of the past decade, Gartner's 2021 definition[6] of data fabric is: "a design concept that serves as an integrated layer (fabric) of data and connecting processes. A data fabric utilizes continuous analytics over existing, discoverable, and inferenced metadata assets to support the design, deployment, and utilization of integrated and reusable data across all environments, including hybrid and multi-cloud platforms."

Data fabric is, in essence, a logical data warehouse, enhanced with improved CSI and AI automation.

In the context of figure 1, data fabric proposes that the various data stores (EDW, marts, and data lake) and functions stand as shown and are all discoverable, accessible, and automatically managed in a virtualized environment. Active, real-time metadata is central to this pattern and forms the basis for AI-based automation of discovery and management of all the data in the environment. In essence, your existing systems and data stores are overlaid with an AI-enabled access and governance fabric. The data fabric thus elaborates on the function required to create a lake with a well-managed island of information at its core and vendors supporting the data fabric pattern are beginning to fill in the products and tools required to implement it.

Data lakehouse

Data lakehouse is a technology-driven architectural pattern introduced[7] by software developer Databricks in January 2020. It proposes to unify data warehouse and lake on a common technology platform, based on modern open-source / cloud-based technology. Its proponents argue that this should be the sole future platform for a variety of reasons, including lower technology costs and the assertion that a large percentage of data in the warehouse comes from the lake anyway.

The lakehouse therefore proposes a common object store for both warehouse and lake data, with ACID[§] transaction support for ongoing, real-time data loading, as well as schema—such as star or snowflake—enforcement and evolution. The technology can thus support both (so-called) structured and unstructured data use cases, and diverse workloads, including data science, BI, and analytics.

> Data lakehouse is a technology-driven consolidation of data warehouse into a data lake.

In essence, data lakehouse is a direct, centralized, and unified implementation of the logical island-in-a-lake architecture, although developed completely independently from it via technology-driven thinking. Its provision of both warehouse and lake function on a single platform offers synergies in implementation and data management. However, this same consolidation presents the risk of sub-optimized support of the differing business needs and a lack of comprehensive infrastructure components in some implementations. The approach will be of particular interest to those with strong, well-managed data lakes and a comprehensive data platform. It is particularly interesting for businesses where much of their operational data comes from the web.

Data mesh

Data mesh was unleashed on an unsuspecting world in May 2019 by Thoughtworks' consultant, Zhamak Dehghani[8], and has since become one of the most hyped topics in data management in recent years. In complete contrast to the lakehouse, the avowed aim of this approach is to remove centralized bottlenecks to the delivery of data for BI and analytics. Two key bottlenecks are identified for elimination. First is the centralized architecture/technology of data warehouse, data lake, and the population infrastructure—ETL pipelines. Second are centralized organizational units, such as ETL development, data modeling, and data governance.

> Data mesh tries to remove bottlenecks to analytics delivery by decentralizing both technological and organizational approaches with data as a product.

In their place, data mesh offers[9] a "domain-driven analytical data architecture where data is treated as a product and owned by teams that most intimately know and consume the data, applying the principles of modern software engineering and the learnings from building robust, internet-scale solutions…" Two key principles of the approach are domain-driven design, where data decomposition is aligned to business/organizational domains, and data as a product, where data, its maintenance code, and metadata/CSI are bundled into a single "quantum" that is owned and maintained entirely within a data do-

[§] ACID = Atomicity, Consistency, Isolation, and Durability. ACID transactions ensure the highest data reliability and integrity, avoiding inconsistent data because of incomplete operations.

main by a cross-functional, business and IT team. The approach borrows heavily from microservices thinking and applies it to data, as well as heavily emphasizing distributed data governance as vital to a modern digital business.

Of the three patterns, data mesh maps most awkwardly to the picture shown in figure 1. There is no concept of an information island and even the data lake is seen as more virtual than physical. Data mesh represents a classical "paradigm shift" in data management for BI and analytics away from physical centralization of data for reconciliation and management and, in its purest form, is both product agnostic and poorly supported by current products. It is also evolving in different directions: in some descriptions, for example, data warehouses are accepted as valid data products. Nevertheless, data mesh is founded on worthwhile principles and much of its thinking, especially data governance aspects, is applicable in more centralized implementations.

Reinventing warehouse and lake for a new era

With an integrated picture of the data warehouse and lake, as well as operational systems and emerging architectural patterns, we are ready to consider how to build a better and more unified solution to old data warehouse and data lake problems and take advantage of the opportunities presented by new and evolving platforms.

I hear you say, this image of a warehouse on an island in a lake is simple and 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 warehouses and lakes and opportunities to address the needs of modern digital business:

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 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.

Reinventing preparation and enrichment of data across all feeds into and within the lake of data is a key benefit of a conjoined warehouse and lake.

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 via the data lake:* 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.

4. *External data ingestion:* populating the data lake has evolved dramatically since the early days when simply loading large files sufficed. Today, incremental loading, updating existing files, and streaming are needed. The management of these processes is complex and error prone, and increasingly requires significant IT involvement.

5. *Transfers between data lake and operational systems:* predictive and prescriptive models developed in the lake by data scientists must be operationalized in the runtime systems of the business and further learnings there shared with the model development environment. Such model (and data) management is a key and complex aspect of analytics.

Cloudera offerings:
Cloudera Data Platform (CDP)
CDP Data Engineering
Cloudera DataFlow
Cloudera Streaming Analytics with SQL Stream Builder
Cloudera Machine Learning
Cloudera Data Visualization

Reinventing information/data preparation: for all these processes, often called "data pipelines," reducing their technical and organizational complexity has become a major focus for lake and warehouse developers.

Technically, data lake tools, whether on-premises or cloud, offer the opportunity to create a more integrated preparation environment for both lake and warehouse. This may involve migrating from traditional ETL tools or simple scripts to (for example) Spark-based procedures or in-database ELT preparation. Such considerations form a key part of data lakehouse thinking. Extensive CSI support, both in generation and automated use, is a key aspect of this effort, as evidenced by data fabric.

Combined lake and warehouse systems raise significant organizational challenges. Warehouse and mart population has always been an IT responsibility, while the increasing complexity of lake feeds is pushing lake population beyond the capabilities of data scientists to IT. Proponents of data mesh see this centralization as a key bottleneck to flexibility and innovation in analytics environments, proposing adoption of domain-centric, data-as-a-product governance to mitigate this problem.

While differences in approach remain (incremental loads still predominate in data warehouses, for example), information/data preparation in the lake—in all five cases mentioned above—is becoming increasingly attractive and powerful to reduce the cost and impact of ETL/ELT performed in the data warehouse environment.

Furthermore, the business need is increasingly for an integrated real-time data flow from the edge to analytics and BI that includes the ability to analyze data at any point in its lifecycle, even while it is in motion in the pipeline, without any dependence on IT.

6. *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 and including machine learning, so moving some data marts onto a non-relational platform may be required.

7. *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. In particular, the scaling and elasticity of modern cloud-native relational databases, combined with their evolving functionality may become mandatory as data volumes continue to grow.

There is no suggestion here of moving the EDW to a non-relational platform. Rather the consideration is to move to an open source/cloud RDBMS. The EDW remains logically part of the island of information, but using relational technology built upon typical data lake foundations such as Spark, and with the physical storage layer underpinning the database potentially changing to use object stores, which are gaining traction here because they can also support less structured lake data formats.

Modern analytics may benefit from migration of data marts and the EDW to open-source/cloud environments.

8. *Archival:* the traditional approach to archival of cold (seldom used) data from the warehouse is to magnetic tape storage. While offering by far the lowest storage cost, tape systems often require manual IT intervention or 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.

Open source / cloud offers new opportunities to improve archival and retrieval of cold data.

Cloudera offerings:
Cloudera Data Platform
CDP Data Engineering

Warehouse and archive migration to a new platform: In the case of data mart upgrade, with many advances in analytics occurring first or only in the open-source environment, migrating some traditional data marts here provides worthwhile opportunities to build new business solutions. For technology-driven migration of both EDW and marts to the open source 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.

There is a significant attraction in having both lake and warehouse data residing on a common object store base because it allows a reduction in the number of copies of the same (shared) data that must be stored and managed, driving reduced costs in both technology and operations.

With both lake and island built on commodity hardware, a new and attractive warehouse archival environment is possible. Although clearly more expensive than tape, the added cost may be offset by the ease and speed of retrieval of archived data directly by unaided businesspeople. Using the same language (SQL) as online use, they perceive archived data as equally available (perhaps with a longer access time) as online data, enabling improved use of historical trending data.

9. *Information/data context and governance:* with information/data stored in multiple disparate locations, on-premises and in multiple clouds, partially overlapping in content and frequently duplicated, businesspeople encounter serious problems in knowing what data may be available, if it is relevant and trustworthy, and if it is the "right" data for their specific needs. The context required to use the information/data is missing or incomplete; sometimes, it is blatantly incorrect. Furthermore, as volumes—particularly from the Internet—have grown, and changes in structure and content occur ever more rapidly, manual data governance and context management have become increasingly difficult.

Context-setting information—enhanced, active metadata—is key to increasing the value of data/information for the business.

As a consequence, the real value of information/data is actually diminishing even as storage volumes and management costs grow. Businesspeople become increasingly frustrated; although promised more and better information, they often cannot even satisfy their basic needs, never mind their desires to become "data driven". IT falls further into disfavor and the business seeks more *ad hoc* and do-it-yourself solutions, adding further to the governance chaos.

Cloudera offerings:
CDP Shared Data Experience (SDX)

Automated management/governance based on active metadata: a recent focus on providing metadata tools for data lakes has evolved into a widespread recognition that metadata is vital for good governance and value delivery. These tools focus on collecting metadata from both the computing environment and its users, and on providing easy access to business glossaries and metadata catalogs, linked directly to BI and analytics tools.

Now, the value of pervasive and continuously refreshed (active) metadata is recognized, echoing the concept of context-setting information (CSI) introduced earlier. CSI extends far beyond the traditional boundaries of technology-centric metadata to encompass process, business, and social metadata to offer the business role- and usage-specific context for the use of information. In addition, such CSI or active metadata can readily become the basis for automation of data management and information access using machine learning and AI techniques. Data fabric, in particular, emphasizes these automation aspects.

10. *Pervasive storage, management, and access*[**]*:* as technology has advanced, cloud implementation of data lakes and warehouses make increasing technical and economic sense, even while significant levels of data and existing processing remain on premises for many traditional businesses. A comprehensive and integrated data storage and management environment spanning a pervasive on-premises and multi-cloud information environment is now needed to eliminate prior data swamps and disconnected data warehouses.

    This environment will contain all the information/data required (or desired) by the business, irrespective of structure, type, timeliness, quality constraints, and so on. Businesspeople face challenges in accessing such data. Familiar SQL and reporting tools alone will not suffice. Tools demanding programmatic approaches are more suited to IT developers and data scientists. Furthermore, the increased business need to combine externally sourced data 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.

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

Cloudera offerings:
Cloudera Private Cloud
Cloudera Data Platform

Hybrid multi-cloud computing: Hybrid multi-cloud and on-premises implementation of a combined lake and warehouse infrastructure is becoming more prevalent and vendor tools are evolving to support multiple approaches. These range from a highly consolidated lakehouse pattern to a more distributed, logical data fabric concept. Simplification of the environment is key, whether by migration to a single technological platform or via automation of management and governance using AI.

Cloudera offerings:
Cloudera Data Visualization

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 for all classes of information and data. And with data spread

[**] Point 10 spans the entire figure and no tag is shown there.

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.

Data fabric and lakehouse address this requirement in distinct ways. Data fabric offers automated, virtualized data access across multiple disparate environments. Data lakehouse, on the other hand, tackles this need—in part—through the physical integration of warehouse and lake environments and their access through a single set of tools.

Cloudera Data Warehouse

The Cloudera Data Warehouse provides an integrated, comprehensive solution to modern reporting and analytic needs by optimizing and extending existing data warehouse and data lake implementations, based on Cloudera's Data Hub, Data Flow, and other products.

The principle behind figure 1 and, indeed, the entire section "A warehouse on an island in a lake," is that the data warehouse, lake, and operational systems cannot—and must not—be considered as independent and separate entities today. 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. This direction also underpins to varying degrees all three new fabric, mesh, and lakehouse patterns. Cloudera has espoused this principle since its earliest days and pursues it through its Enterprise Data Hub product on premises and through Cloudera Data Platform (CDP) for cloud deployment.

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 five areas of focus:

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

1. *Cloud Data Reports and Dashboards:* at its most basic level, a data warehouse or mart must offer simple, direct access to data for queries, repots, and dashboards. For CDP the obvious focus is on data resident on several common cloud platforms. On-premises information is also supported using the Burst to Cloud feature that move this information (data and context) to cloud object stores for immediate query.

2. *Instant Access to Data:* with its scalable, elastic base, CDW offers excellent price-performance for timely access to data on public and private clouds for data discovery, querying and visualization independent of central IT as seen in a 2021 GigaOm report[10], showing 20% to 550% advantage over comparable offerings.

3. *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, taking advantage of open-source components such as Apache Impala and Hive, as well as Hue, underpinned by intelligent workload management with Cloudera Workload XM.

4. *Operations & Events Analytics:* growing volumes of events streaming from the IoT, clickstreams, and other sources require in-flow processing and time-series analytics. CDW, combined with Cloudera DataFlow, incorporates Apache Kudu and Druid to manage and analyze such fast-moving data at speed and with integrity.

5. *Research & Discovery Analytics:* brings together all forms of predictive and prescriptive analytics from basic exploratory full-text search with Solr, to relational query via Hue to model management with CDP Machine Learning, delivering deeper integration between traditional data warehouse and data lake applications. In particular, CDP patterns automate use cases like exploratory analytics and BI at scale and real-time analytics, providing the benefits of self-service to data practitioners.

For both types of analytics, CDW integrates the Apache Iceberg open table format—based in turn on open file formats, such as Parquet and ORC—which enables seamless integration between different streaming and processing engines while maintaining data integrity between them. This allows new use cases, such as change data capture (CDC) and reproducibility of machine learning (ML) Ops.

The overall structure of the Cloudera Data Warehouse is shown in figure 2.

Underpinning this stack is 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 a hybrid, multi-cloud environment. Druid, a time series database for real-time analytics at scale and Kudu, a relational database designed for fast analytics complete the storage options.

The Shared Data Experience (SDX), provides common, 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 on-premises and cloud implementations. Shared controls provide the systems management function underpinning SDX. Workload Experience Manager

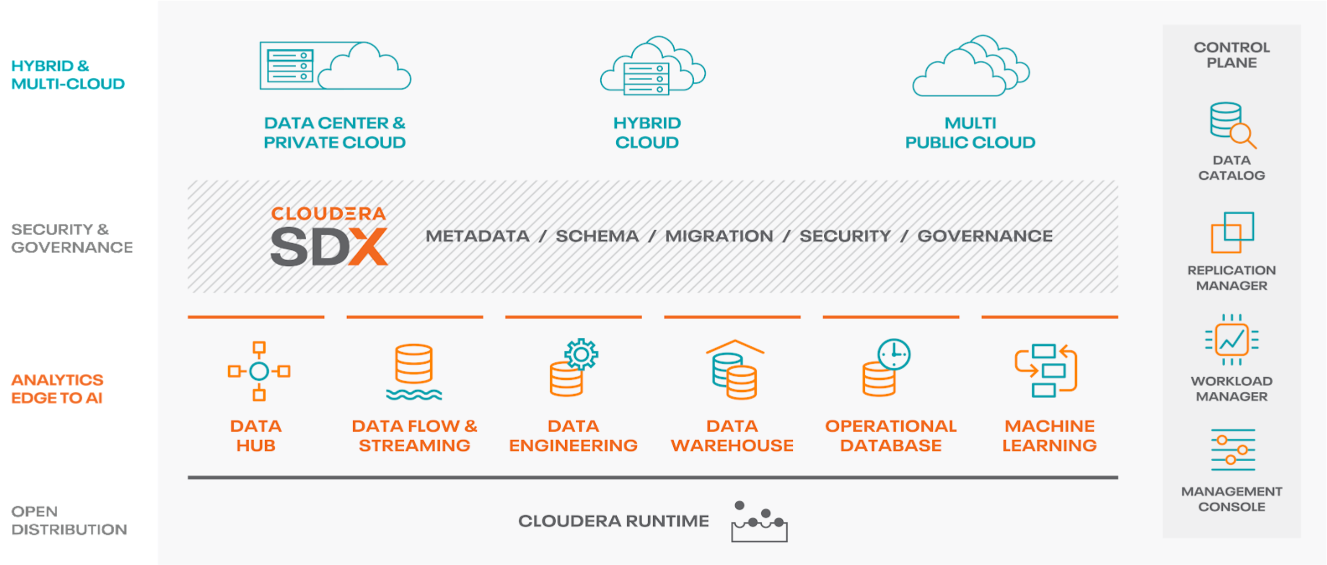

Figure 2: Cloudera Data Warehouse

(Workload XM), provides enhanced visibility and actionability to reduce migration risk, speed troubleshooting and improve uptime and resource utilization.

Separate from and independent of the storage layer, 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, Impala, and Kudu 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 here 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.

For reporting and analytics, 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.

Supporting this architectural view, Cloudera Data Patterns offer end-to-end product integrations that provide validated, reusable solution patterns to enable faster response to typical business use cases. As of this writing, two patterns—Scalable Self-Service Analytics and Business Intelligence at Scale (in Technical Preview)—are available. Two further patterns have been announced—Real Time Operational Insights and Machine Learning Exploration and Discovery.

In summary, CDW offers a full function cloud-based data warehouse supporting a wide range of traditional BI and analytic function—in essence the island of information in figure 1. Combined with the additional function included in the Cloudera Data Platform, CDW consolidates access to the complete lake of data, incorporating the modern architectural patterns of data fabric and especially data lakehouse. The breadth and depth of function available in CDP actually exceeds in some ways both fabric and lakehouse thinking, offering a more comprehensive implementation of the island-in-a-lake paradigm than either or both together.

Conclusions

Data warehouse and lake have, in effect, merged in the past few years. The outcome can be positive when architecturally well-understood and implemented. The Cloudera Data Warehouse succeeds in this by combining the governance of a warehouse with the analytics of a lake in a hybrid on-premises, multi-cloud solution based on open-source software.

For thirty years, the data warehouse has remained a central component in decision-making support. A decade ago, the data lake was introduced. First seen as competitive concepts, they have evolved into equal partners. But their motivations differ. A warehouse, with its functional focus, provides the reconciled, legally founded data needed to run and manage the business. A lake offers a place to store raw data and analyze it in innovative and ever-changing ways in the illustrative paradigm. What we call the components matters less than recognizing the different but complementary roles played. However, definitional differences and implementation challenges have more recently led to three new architectural patterns—data fabric, data mesh, and data lakehouse—which, in essence, try to merge these roles using different organizational approaches and technologies.

Drawing a conceptual picture that overlays a data warehouse as an island of information in a data lake and positions traditional operational systems, shows that data and function can be positioned and moved within this joint environment to address new business needs and mitigate old data warehouse problems. It also allows us to understand these new architectural patterns. What emerges is that Cloudera Data Warehouse and Data Platform together address both data fabric and lakehouse thinking and offer the opportunity to implement them in a comprehensive way.

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 technology ecosystem to take advantage of new advances there. Some function—such as data preparation and archiving—can be 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 enables IT to rapidly and simply deliver cloud-native, self-service, analytics to BI analysts via a single, optimized SQL interface, combining traditional data warehouse and lake concepts with up to real-time data use. It addresses five distinct but interrelated approaches to modernize and extend the traditional data warehouse. These are: (1) Cloud Data Reports & Dashboards, (2) Instant Access to Data, (3) Data Warehouse Optimization, (4) Operations & Events Analytics, and (5) Research & Discovery Analytics.

CDW enables IT to rapidly deliver cloud-native, self-service, analytics to BI analysts via SQL, that combines traditional warehouse and lake concepts with up to real-time data use.

This evolution in architecture from warehouse vs. lake to warehouse ***and*** 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.

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 2013 book, ***"Business unIntelligence—Insight and Innovation Beyond Analytics and Big Data"*** *(bit.ly/BunI-TP2) is available in both hardcopy and e-book formats.*

As founder and principal of 9sight Consulting (www.9sight.com), Barry provides strategic consulting and thought leadership to buyers and vendors of BI solutions. He is continuously developing new architectural models for all aspects of decision-making and action-taking support. Now based in Cornwall, UK after a decade in South Africa, Barry's knowledge and expertise are in demand both locally and internationally.

[1] Dixon, James, "James Dixon's Blog: Pentaho, Hadoop, and Data Lakes", (October 2010), bit.ly/2BHwplU

[2] Devlin, Barry, "The Data Warehouse Lives On", (July 2019), 9sight and Cloudera, bit.ly/2tP8nr5

[3] Yuhanna, Noel, et al, *"The Forrester Wave™: Cloud Data Warehouse, Q1 2021"*, (March 2021), bit.ly/3D3FQ0d

[4] Devlin, Barry, *"Business unIntelligence"*, (2013), Technics Publications LLC, bit.ly/BunI_Book

[5] Yuhanna, Noel,, *"The Forrester Wave™: Big Data Fabric, Q4 2016"*, (November 2016), bit.ly/37qZlUB

[6] Gartner, *"Data Fabric Architecture is Key to Modernizing Data Management and Integration"*, (May 2021), gtnr.it/3gOo7zk

[7] Lorica, Ben, et al, *"What Is a Lakehouse?"*, (January 2020), bit.ly/35vEXxi

[8] Dehghani, Zhamak, *"How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh"*, (May 2019), bit.ly/3dtLrI3

[9] Thoughtworks, "Data Mesh", (2020), thght.works/34R1Oac

[10] McKnight, William and Dolezal, Jake, GigaOm, *"Cloud Data Warehouse Performance Testing"*, (Feb 2021), bit.ly/3va8M34