Enterprise Strategy Group | Getting to the bigger truth.™

WHITE PAPER

OPTIMIZING AN ENTERPRISE DATA WAREHOUSE FOR A MULTICLOUD STRATEGY

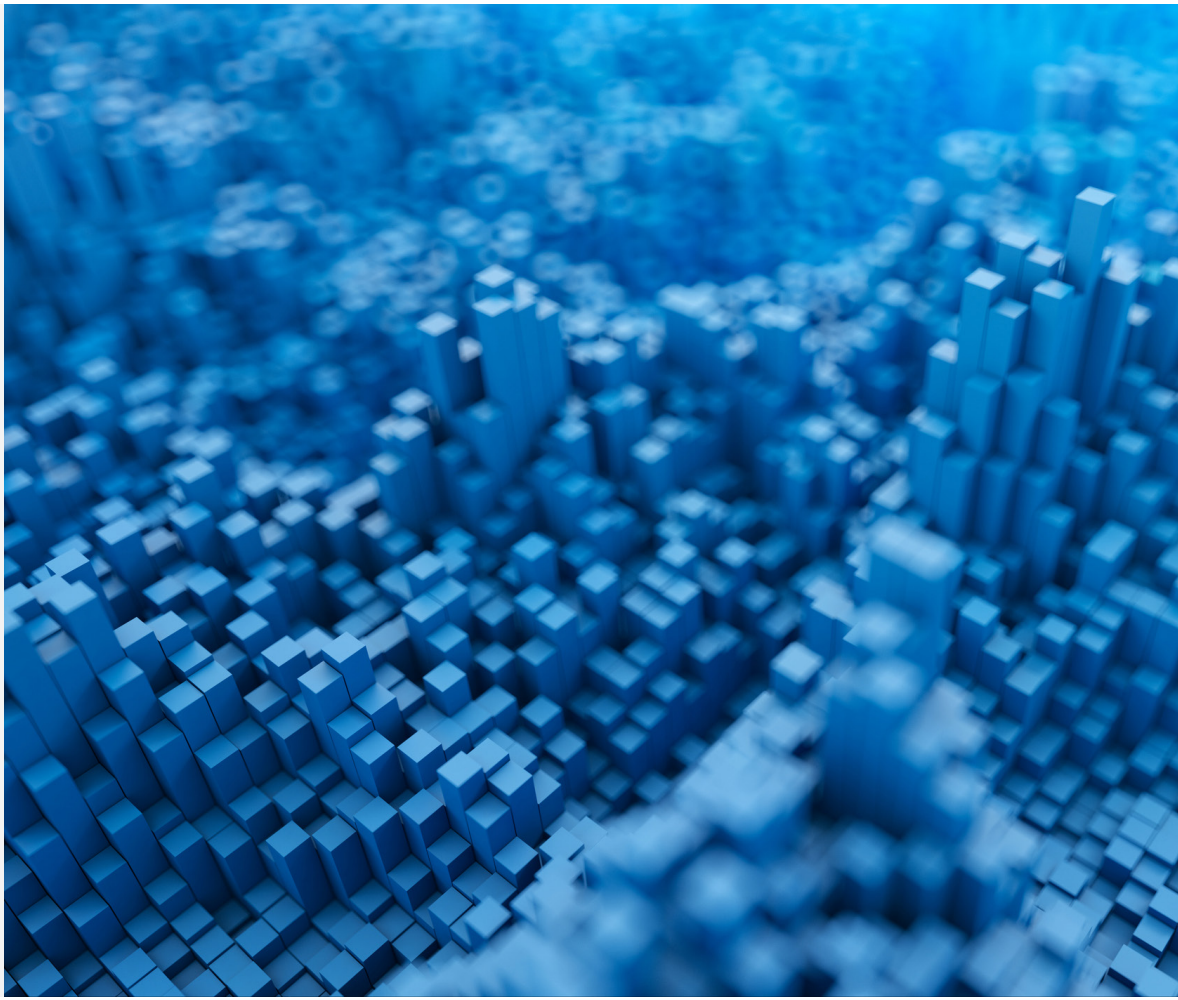
Best practices to ensure trusted, business-ready data and the benefits that come with it

By Mike Leone, ESG Senior Analyst, and Leah Matuson, Research Analyst

FEBRUARY 2019

This ESG White Paper was commissioned by IBM and is distributed under license from ESG.

Contents



Back to Contents

Modern Challenges of an Enterprise Data Warehouse

Enterprise Data Warehouses (EDWs) have existed for about 20 years. They serve as the foundations of insight-driven organizations, delivering timely analysis and reporting of structured data, handling large analytic workloads, and supporting the high levels of concurrency that these organizations demand (i.e., many users simultaneously accessing the EDW). But while EDWs have been a familiar presence in many organizations, as companies look to reduce their data center footprints, increase organizational agility, and incorporate as much data as possible into their analytic workflows, the architectural rigidity, complexity, and cost of a traditional EDW are becoming increasingly apparent. Due to their long-established presence, EDWs have been relegated to catchall status, with organizations utilizing them for activities for which they weren't originally created. This unwieldly scenario is pushing the EDW to become more of a cost center than an insight enabler.

Without a separate infrastructure to perform data integration tasks, such as the processing of large volumes of data through extract, transform, and load (ETL) jobs, the data integration workload gets pushed into the EDW, consuming valuable computing resources in the platform."

The Cost of Data Integration

The EDW is one of the most expensive parts of the data center infrastructure, combining proprietary hardware and software into a structured data platform meant for analysis and reporting. The challenge is that organizations are turning to the EDW to satisfy workload requirements for which the EDW was not intended, with data integration workloads being the main culprit.

Without a separate infrastructure to perform data integration tasks, such as the processing of large volumes of data through extract, transform, and load (ETL) jobs, the data integration workload gets pushed into the EDW, consuming valuable computing resources in the platform. Not only does this steal resources from analysis and reporting workloads, but it compromises the EDW's normal workload performance and the EDW users' ability to retrieve information in a timely manner.

Unfortunately, this scenario has become common, resulting in organizations having to scale their EDWs up and out by purchasing more licenses, compute, and storage to satisfy the burden placed on them by running workloads that were never meant to be run on EDWs. It should be noted that, since the EDW is likely run on specialized hardware, the cost of that compute and storage is significantly higher than that of commodity hardware. This is one of the reasons ESG research shows that more than one-third (38%) of organizations view the complexity of data integration as a top data analytics challenge.[1]

[1] Source: ESG Research Report, *Enterprise Big Data, Business Intelligence, and Analytics Trends*, January, 2015.
All ESG research references and charts in this white paper have been taken from this report unless otherwise noted.

Back to Contents

More Data, Additional Data Sources

While organizations understand the value of incorporating as much high-quality data as possible into their BI and analytics workflows, two key challenges exist when it comes to enabling this capability in a traditional EDW.

The first challenge arises from the rigidity an EDW imposes on the type of data that can effectively be used: an EDW is unable to natively process unstructured data. Because many organizations look to incorporate data from multiple data sources, it is likely they will be incorporating semi-structured and/or unstructured data. These data types, from various sources, such as IoT devices or web data (e.g., social media, consumer sentiment, etc.), must be included to ensure the most accurate insight. When ESG research respondents were asked about the data sources they currently use or plan to use for BI and analytics purposes, their responses ran the gamut from structured to unstructured data sources. In fact, 44% cited IT system data (server, storage, network, and other logs), while 40% cited data from content management systems, and 36% said email or text documents. In addition, more than one-quarter (27%) cited web and/or clickstream data, while 26% said sensor and machine data (RFID tags, geological readers, etc.).

The second key challenge revolves around the length of time organizations can store data in an EDW. Due to capacity limitations, companies leveraging traditional data warehouses are struggling with the ability to retain data for as long as necessary. This untenable situation is forcing companies to make tradeoffs—either deal with much higher operational expenses to increase compute and storage capacity (i.e., keep data in the EDW for a longer period of time); or risk inaccurate or incomplete insights due to a limited view into historical data.

> **“Organizations are cutting corners to ensure the timely delivery of data. It’s not surprising this leads to data quality issues.”**

Back to Contents

Data Quality

Organizations must use the highest quality data for analysis to ensure their insights are accurate. When they leverage EDWs, they must employ complex data cleansing routines and hand coding to assure the data is high quality. Herein lies the challenge. Due to data having to be properly structured before being loaded into the EDW, organizations are unable to push the cleansing activity to an EDW, meaning they need another tool and likely, infrastructure, to support data cleansing. And even if the EDW infrastructure was able to handle cleansing, it would eliminate too many resources from the actual analytics and reporting jobs.

In general, organizations are cutting corners to ensure the timely delivery of data. It's not surprising this leads to data quality issues that have massive ramifications—analytics and reporting jobs are unable to complete in a timely manner (or at all) or a query will yield an inaccurate insight. These insights are meant to help shape future decisions, which could impact the entire business. When employees throughout the company learn that the process has delivered bad insights, or that queries were unable to complete for whatever reason, they can lose confidence in the EDW, or the bad data might spread throughout the organization—defeating the purpose of the EDW investment in the first place.

Without ownership awareness, there is a lack of accountability."

Data Governance

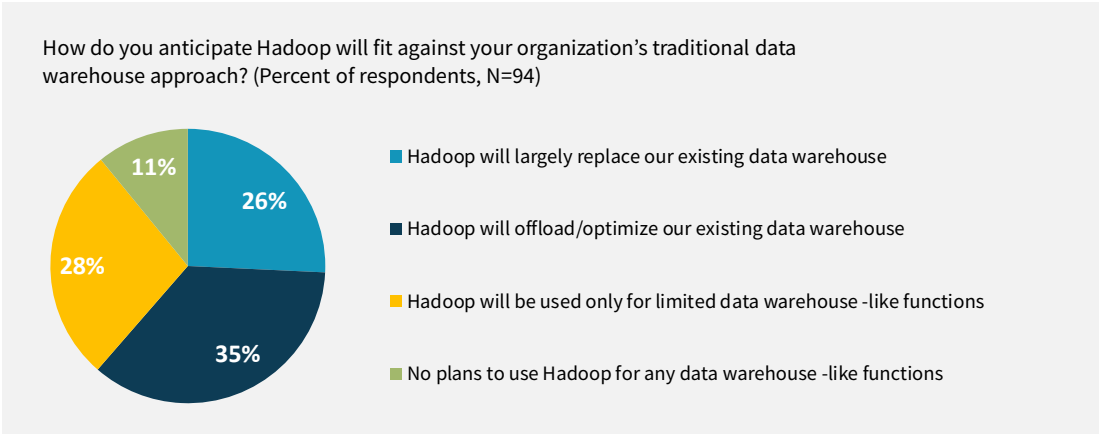
By default, EDWs lack the ability to provide data lineage, a requirement for almost all compliance regulations and a means to ensure that data hasn't been tampered with. Granted, organizations can attach tools to a data pipeline or workflow, but that just adds more complexity to the system. This is an ongoing challenge for organizations—to understand who owns the data, how it applies to the business, and who should have access. This situation leaves many companies defaulting to the concept of dumping into the EDW with a goal of tending to the governance of the data "later." As is generally the case, "later" never comes.

Without ownership awareness, there is a lack of accountability. This means that all data placed in the EDW (which is consuming costly storage capacity) will never be used—end-users will not understand how it applies to the business, or how it may be applied to solve their specific business problems.

Optimizing an Enterprise Data Warehouse

As organizations look to address the challenges encountered with their traditional EDW platform, they're exploring cost-effective solutions such as Hadoop as a means to optimize and offload those integration workloads that were never meant to run on EDW. In fact, ESG research shows that 26% of organizations surveyed will largely replace their EDWs with Hadoop, while another 36% will use Hadoop as a means of optimizing or offloading their existing EDW (see Figure 1).

Figure 1. Hadoop's Impact on Enterprise Data Warehouses



Organizations need to look at Hadoop and other platforms as more than cost-effective platforms where they can offload their data integration workloads and unused data from their expensive, scalability- and performance-challenged EDW. The Hadoops of the world are meant to be a cost-effective means to integrate, process, and analyze all data. The key word there is "data." When organizations simply move data from an EDW to a data lake—to handle more data, unstructured data, or to achieve a specific analytics goal without taking the appropriate steps and adhering to best practices—the data lake can quickly turn into a data swamp.

Optimizing your EDW is part of properly governing your data lake. The data lake serves as the trusted foundation of an organization's data-driven and, more specifically, insight-driven goals. Optimization of the EDW not only ensures security and reduces risk, but also promotes access and fosters enterprise-wide collaboration. Additionally, optimization should be performed at scale to meet the dynamic needs of the business. Read ESG's report, *Intelligent Data Governance for a Business-ready Data Lake*, for more information.

Back to Contents

Using Best Practices to Ensure Success

As previously noted, while organizations can cost-effectively offload their EDW to a data lake and Hadoop, it's essential to follow best practices for optimum results, transforming the data lake into a usable asset. By prioritizing vendors and technologies that enable the key capabilities shown below, organizations will put themselves on the path to success.

Apply Data Quality and Governance

For organizations to attain the most value—from both a time-to-value, and a time-to-insight standpoint—they must make investments to ensure the data they are offloading to a more cost-effective, scalable platform is of high quality and appropriately governed. With the ultimate goal of infusing artificial intelligence (AI) and machine learning (ML) into their organization, and the understanding that data must be trusted for AI/ML projects to succeed, organizations must be on the right path to ensure success. With high-quality and governed data, organizations improve the accuracy of their insights, respond more easily to compliance audits, reduce risk, and increase their ability to more effectively protect their data. Additionally, providing trusted data on a self-service basis enables business analysts to do more of the analytics work themselves, allowing them to achieve greater results (more easily accessing relevant data based on their roles and goals) without the need for IT intervention. This can lead anywhere from improved operational efficiency and swifter insights to increased business agility and an overall improvement of businesses' performance in their respective markets.

> **With high-quality and governed data, organizations improve the accuracy of their insights, respond more easily to compliance audits, reduce risk, and increase their ability to more effectively protect their data."**

APPLY QUALITY AND GOVERNANCE

Invest in a platform that can provide data quality and data governance to ensure trust throughout the organization.

BUILD ONCE, RUN ANYWHERE

Prioritize vendors that can enable integration workloads to be run across the entire stack, regardless of technologies, with minimal to no changes required.

ENABLE SELF-SERVICE

Leverage a platform powered by trusted data with tools that increase productivity, foster collaboration, efficiently gain actionable insights, and ensure data quality remains high.

PRIORITIZE FUTUREPROOF

Ensure the platform aligns to current needs and offers flexibility to embrace future requirements, like support for open-source software projects or multicloud environments with connectors and APIs.

Back to Contents

Build Once, Run Anywhere

It's common for organizations to use manual, custom coding and scripting in an EDW environment to assist with data integration. When IT organizations migrate data integration workloads from EDW to Hadoop, their first instinct might be to continue to hand-code integration jobs. However, some solutions allow for a "build once, run anywhere" scenario, where integration jobs can be run in the EDW, in the ETL grid, in Hadoop, and in the cloud with the same code. This scenario offers faster time to value and a significant reduction in development and maintenance costs. When organizations begin evaluating approaches to optimizing their EDWs, they should prioritize vendors that are able to deliver compatibility across the entire stack: developer interfaces, metadata repositories, and processing engines for Hadoop and non-Hadoop platforms.

> **A solution that incorporates data integration, data quality, and data governance within a single platform can yield improved operational efficiency and productivity."**

Enable Self-service

A solution that incorporates data integration, data quality, and data governance within a single platform can yield improved operational efficiency and productivity. With an all-encompassing platform, organizations start on a path to better and more efficiently enable accessibility; ensure data quality remains high; shape and transform data based on user needs; and eventually empower self-service analytics through trusted data.

By supporting a broad range of technical and non-technical end-users within the same platform, virtually the entire organization can see productivity increase. For IT and other technical personnel, a single platform can help consolidate management silos, minimize maintenance windows, and reduce support time. These much-needed productivity gains enable them to focus on what matters, such as improving data pipeline efficiency, incorporating the latest and greatest technology, and providing the non-technical end-users with high quality, trusted data.

By selecting a platform with the ability to provide non-technical end-users with the right level of automation and accessibility to data, self-service is achieved. This means self-sufficient business analysts can more efficiently access and transform business-ready data for faster insight into their particular business initiatives, while data-centric roles can more effectively build analytical models to satisfy data science and AI initiatives.

Back to Contents

Futureproof the Platform

Organizations should prioritize a flexible platform that enables future growth, and doesn't impede it. Therefore, IT should understand and plan for how the selected platform can meet their current requirements, as well as align to their future business initiatives, such as leveraging hybrid or multicloud environments.

When organizations are on the path to optimizing an EDW, it is essential they are aware of the IT architecture to which they plan to offload, including any limitations it may impose on their current, or more importantly, future workloads. IT must fully recognize how their data integration workloads operate, as well as how the destination architecture can enable better performance, scalability, reliability, and security. Additionally, with organizations looking to infuse open source technology into their workflows, it's vital for IT to understand which software projects, other software platforms, and processing engines are supported as part of their chosen platform.

This includes knowledge regarding:

- Availability of APIs to other open source projects.
- Connections to parallel databases, conventional ETL servers, ETL grids, and Hadoop clusters.
- Integration with public or private clouds using Spark.

With cloud adoption continuing to rise, cloud-first strategies are becoming more prevalent than ever before. This is especially true for digital-native organizations that have been operating for ten years or less. Of the 58% of organizations that utilize public cloud infrastructure for infrastructure-as-a-service (IaaS), 76% use two or more public cloud service providers.[2] This speaks volumes to the need for a data platform to not only put organizations on a path to consume cloud services but to also offer them freedom of choice when integrating their public clouds of choice.

"Of the 58% of organizations that utilize public cloud infrastructure for infrastructure-as-a-service (IaaS), 76% use two or more public cloud service providers. This speaks volumes to the need for a data platform to not only put organizations on a path to consume cloud services but to also offer them freedom of choice when integrating their public clouds of choice."

[2] Source: ESG Research, *2019 Technology Spending Intentions Survey.*

Back to Contents

Benefits When Optimization Is Done Right

A successful EDW optimization can produce several benefits. Organizations that have followed best practices in optimizing and offloading their ETL workloads have experienced impressive benefits. These companies have been able to offload their ETL workloads to a more cost-effective platform, freeing up expensive EDW resources while improving warehouse performance for both the normal workload, and the analytics and reporting for which the warehouse was designed.

A successful EDW offload allows organizations to:

Leverage Hadoop as a means of storing more data for longer periods of time. This is important because organizations no longer need to pay to scale out the proprietary EDW platform storage. Instead, they can use their existing storage knowing they are offloading the cold or rarely used data to Hadoop. They can then leverage Hadoop to run analysis, analytics, and integration tasks on a growing data set of historic and new data.

Add new data regardless of type or structure. Organizations can incorporate more data and additional data sources (combining structured and non-structured data) by offloading their EDW to a platform like Hadoop.

Ensure the trustworthiness of data by directly incorporating data quality controls and correct levels of data governance. By using the proper tools and software, organizations that successfully offload their ETL workloads to Hadoop are better able to ensure the trustworthiness of data. In addition, incorporating correct levels of data governance helps ensure that the data has business meaning; transparency in lineage and ownership; the appropriate access controls; and the correct level of data stewardship. Essentially, data governance helps to ensure the data pipeline is working efficiently, and, if a problem arises, can help IT to know where it originated.

Align to future business goals, whether integrating a new technology or tool, accelerating hybrid or multicloud adoption, or infusing artificial intelligence throughout an organization to maintain a competitive edge.

Back to Contents

The Bigger Truth

Organizations are wasting no time prioritizing data-centric business strategies, enabling them to gain actionable insights that help them stay relevant or, better yet, disrupt the competition. While EDWs help deliver insight based on the analysis and reporting of structured data, organizations are being forced to improperly utilize EDW resources to satisfy workloads that were never meant to be run on it—such as data integration and ETL jobs. Additionally, modern organizations recognize the value of incorporating more data for longer periods of time on a cost-effective platform that ensures high data quality and governance. Simply put, traditional EDWs are incapable of achieving these goals in a reasonable way, whether due to architectural limitations or cost.

By leveraging a modular platform that incorporates comprehensive data integration, quality, and governance, anchored by a massively parallel processing architecture, organizations gain confidence in reliably satisfying the modern needs of a business. And by ensuring the proper levels of integration are available to other technologies, open source projects, and complementary tools, the concept of coding once, running anywhere, and deploying anywhere based on preference or mandate will enable organizations to be more agile and operationally efficient in their pursuit of growth, competitiveness, and real-time responsiveness.

LEARN MORE

For more resources to help you with EDW optimization or to talk to an expert, visit ibm.com/offloading-edw

Back to Contents

Enterprise Strategy Group is an IT analyst, research, validation, and strategy firm that provides actionable insight and intelligence to the global IT community.