research

CHECKLIST REPORT 2019

The Modern Data Warehouse and Analytics Stack

Six keys for success

By Fern Halper, Ph.D.

Sponsored by:

amazon web services

MATILLION

+ableau

tdwi | TRANSFORMING DATA WITH INTELLIGENCE™

OCTOBER 2019

TDWI CHECKLIST REPORT

The Modern Data Warehouse and Analytics Stack

Six Keys for Success

By Fern Halper, Ph.D.

TABLE OF CONTENTS

2 FOREWORD

3 NUMBER ONE
Consider a cloud data warehouse

5 NUMBER TWO
Make use of diverse data types

6 NUMBER THREE
Utilize modern push-down ELT processing where possible

7 NUMBER FOUR
Integrate tooling for data discovery and analytics

8 NUMBER FIVE
Govern the data

9 NUMBER SIX
Look for an integrated stack

10 FINAL THOUGHTS

11 ABOUT OUR SPONSORS

12 ABOUT THE AUTHOR

12 ABOUT TDWI CHECKLIST REPORTS

12 ABOUT TDWI RESEARCH

tdwi

Transforming Data With Intelligence™

555 S. Renton Village Place, Ste. 700
Renton, WA 98057-3295

T 425.277.9126
F 425.687.2842
E info@tdwi.org

tdwi.org

FOREWORD

In today's competitive environment, it is more important than ever for organizations to be able to adapt to changing conditions. Organizations are looking to become more flexible and agile in their approach to business. This is driving how they approach data and analytics, as well. Data for analytics comes from an increasing number of sources, both internal and external to an organization.

TDWI research indicates that companies are typically capturing terabytes of data from dozens of data sources, with the primary business goal being to analyze the data, often in a self-service manner, to drive insight and action. Although some organizations are satisfied with the traditional warehouse that deals primarily with structured data and reporting, the vast majority of organizations we survey recognize the importance of modernizing their warehouse and analytics environment so they can scale to meet dynamic data and analytics needs.

Modernizing the warehouse comes in many forms. Some organizations augment their existing data warehouse with newer platforms (optimized mostly for analytics) in a tightly integrated, multiplatform architecture. Others re-platform their environment by migrating the warehouse data (wholly or in part) to new data platforms, such as those based on NoSQL Hadoop systems, columnar data stores, or cloud-native or new nonrelational data warehouses.[1]

We see a growing trend towards selecting data warehouse platforms that are purpose-built and optimized for cloud-native operation. Moving entirely to the cloud is typically driven by a business need to add features, scale, or performance that would be difficult to achieve in the application's existing environment.

Closely coupled with the need to modernize the data warehouse is the need to efficiently collect diverse data from multiple sources, transform this data, and analyze it. Organizations want flexibility and scalability for their analytics efforts. There is increasing pressure on BI and analytics teams for more visibility and uses for that data, often in a self-service manner.

In fact, TDWI research indicates that self-service is a top priority for a majority of organizations. That means that new data warehouse platforms must often be coupled with other tooling to form a modern analytics stack (e.g., a data collection and transformation technology).

The pillars of this architecture include the modern data warehouse, newer data integration tooling, and self-service analytics, with the data warehouse as a centralized repository to serve high-performance analytics needs at scale. This TDWI Checklist Report examines best practices for the modern data warehouse and the modern analytics stack.

[1] For a more detailed discussion of these and other modernization strategies, see the *TDWI Best Practices Report: Data Warehouse Modernization*, online at tdwi.org/bpreports.

1 CONSIDER A CLOUD DATA WAREHOUSE

Today, many organizations are moving their data to the cloud to take advantage of its flexibility and ability to deal with data at scale. TDWI research sees growing interest in cloud data warehouses for data management to support analytics. For example, in a recent TDWI survey, about 30 percent of respondents were already using a cloud data warehouse. That number is set to more than double in the next few years if users stick to their plans.

The movement to the cloud for data warehousing can take different forms. Some organizations augment their existing on-premises data warehouse with a cloud data warehouse. They might use their on-premises data warehouse for supporting traditional reporting and dashboards. The cloud data warehouse is then used for certain kinds of analytics that require a scalable and flexible environment. In other cases, organizations reimagine and redesign their entire platform in the cloud and move all of their data to the cloud data warehouse or even multiple cloud data warehouses.

There are numerous benefits for utilizing a cloud data warehouse. The cloud provides storage and compute resources suitable for big data scale and demanding analytics workloads, with minimal setup and maintenance, at relatively low cost.

Many of the newer cloud data warehouses are architected to take advantage of the compelling properties of the cloud:

- **ON-DEMAND SCALABILITY.** TDWI research indicates that elasticity is the top benefit users see from the cloud. Organizations like the fact that the cloud is elastic and can automatically allocate resources as analytics workloads ramp up and can reallocate them as processing subsides. Cloud elasticity helps a data warehouse cope with the highly unpredictable ad hoc query workloads that result from modern self-service practices. Users also like a pay-as-you-go model.
- **SEPARATES COMPUTE FROM STORAGE.** Cloud data warehouses separate computing resources from storage so that organizations can scale each independently. For instance, if an organization has a batch workload running during the weekend that needs extra computing capacity, it can provision extra computing resources for the duration of the workload and de-provision the compute after the successful execution of the job. This saves on computing costs. Meanwhile, the rest of the organization's data lives in a storage environment that is specifically built for large data sets.
- **HIGH PERFORMANCE AT SCALE.** Cloud data warehouses are architected for strong performance. For example, cloud data warehouses often make use of newer technologies such as columnar databases. Columnar databases can provide fast query performance because of the way data is stored (in columns rather than rows, which can improve I/O for analytics query processing). Additionally, cloud data warehouses are often optimized to support queries against large amounts of data via massively parallel processing (MPP) architectures to parallelize and distribute SQL operations to take advantage of all available resources.

1 CONSIDER A CLOUD DATA WAREHOUSE CONTINUED

Some cloud data warehouses even use advanced technologies such as machine learning to predict incoming query run times and assign them to the best queue for the fastest processing.

- **SECURITY.** Although many organizations worry about the security of their data in the cloud, providers have made it their business to provide tight security. This means that security measures such as access and encryption of data at rest and in motion are supported natively. For instance, some cloud providers enable organizations to set up a virtual private cloud to control network access and support end-to-end encryption.
- **INTEGRATED WITH OTHER CLOUD SERVICES.** Cloud data warehouse providers often offer tight integration with other cloud services such as data loading and transformation as well as analytics and machine learning services.
- **AUTOMATED PROVISIONING AND MANAGEMENT.** Modern data warehouses automate many administrative tasks such as provisioning, backups, and replication. This way teams can be more productive with their data, instead of spending time on repetitive tasks that add no value.
- **PAY-AS-YOU-GO MODELS.** Many cloud users like the ability to pay only for what they use. This, along with on-demand scalability and separation of computing resources and storage, allows for more cost-effective architectures.

2 MAKE USE OF DIVERSE DATA TYPES

Modernizing the data warehouse offers a number of benefits, but TDWI research indicates that a top driver is support for analytics in general, including self-service visualization and exploration. Traditionally, organizations stored only structured data, such as billing information, in their data warehouses for use in reports and dashboards. However, modern cloud data warehouses support a wide range of data types and analytics, and connecting disparate data to create a rich data set is key to better analytics. Therefore, the modern analytics stack should support a wide range of data types including semistructured and unstructured data types. These include:

- **DISPARATE TRADITIONAL DATA SOURCES.** In TDWI research, traditional structured data is often cited as a top data source for analytics. Structured data comes from multiple locations, including legacy mainframe systems, multiple relational databases, and files that are on premises. It also includes newer platforms such as a data lake that might house traditional as well as newer types of data sources.
- **NEW DATA SOURCES.** Although traditional structured data is still the primary source for data analysis, other data sources such as semistructured (JSON, XML) and unstructured (text, multimedia) data are becoming mainstream. Geospatial data is also popular, especially for self-service analytics. Sources such as machine-generated data (e.g., from sensors or IoT devices) are also gaining steam. Examples of these new data sources might include sentiments from call center notes, medical images, or streaming temperature sensor data.
- **CLOUD-GENERATED DATA.** A considerable amount of data is now generated in the cloud, which includes social media data, IoT data, and data from other cloud applications, to name just a few. Traditional external data sources such as demographic data that enrich other data sources are also often cloud based. Newer cloud data sources, such as weather data, home sales prices, and financial data (among others) are also becoming available.

  Because it makes sense to analyze data where it lives—often referred to as "data gravity"—this may explain increasing acceptance of cloud BI and analytics for big data. Moving large volumes of data can be resource intensive. If the data is generated in the cloud, it makes sense to analyze it there, especially if the platform is architected for the cloud.

One approach to make use of this data for analytics is to extract it from these multiple sources and then load it into a data store for analysis. The modern data warehouse needs to support diverse data requirements including data integration from traditional and new sources. It also must support data collection and transformation tools that provide capabilities to ingest, transform, and consume these multiple data types (see Number Three). Often, organizations will look to a centralized repository, such as a cloud data warehouse, to collect and store this data.

3 UTILIZE MODERN PUSH-DOWN ELT PROCESSING WHERE POSSIBLE

Before data is ingested into the traditional data warehouse, it is typically aggregated, cleansed, and documented to support reporting as well as compliance and audit as part of the extract, transform, and load (ETL) process. Traditional ETL requires dedicated infrastructure where data is staged and transformed. Maintaining this dedicated space can be time-consuming and costly. Additionally, some legacy tools complete only a piece of the process, such as extract and load. Finally, it can be hard to scale ETL as data volumes increase.

The trend in many modern environments is to move from an ETL approach to ELT. In ELT, data is extracted from the source system and then loaded into the target system. Typically, the tools used for this have numerous connectors that can extract data from multiple source systems, both on premises and in the cloud. Then, transformation occurs in the target system.

There are a number of benefits to this approach, especially in a data warehouse architected for the cloud because ELT in the cloud can use the processing power of the cloud platform itself. This is sometimes referred to as modern data loading, and it can involve tools purposely built for cloud data warehouses. These modern data transformation tools take advantage of the cloud to perform data transformations at scale, quickly and at low cost. These tools are optimized and designed to maximize the strengths of modern cloud data warehouses.

The idea is to use the processing engine of the platform to perform the transformations and other computationally expensive processes. The scale-out nature of the cloud makes ELT in the cloud data warehouse more performant. The platform can perform transformations quickly and scalably as more data is added from disparate sources.

Some modern data loading tools may provide a UI that makes it easy to select and connect to external data sources (both on premises and in the cloud) and build a workflow to load the data into the cloud data warehouse. The transformations, such as aggregations, replications, and transposing rows and columns, can also be built out using the data-loading tool. That can be done using either SQL or a UI that makes it easy to build SQL queries. The workflows built using these tools can be saved and reused for new data. These reusable transformations are also important for consistency. Some of these tools also follow the pay-as-you-go cloud model.

Loading entire data sets from source systems will enable data democratization because no data sets will be discarded and data will be made available to the entire user base based on the use case. Look for tools that tightly integrate with the cloud data warehouse.

4 INTEGRATE TOOLING FOR DATA DISCOVERY AND ANALYTICS

Traditional reporting and dashboards are still a popular way for organizations to gain insight and information. However, the movement in analytics is towards self-service discovery and more advanced analytics, such as predictive analytics and machine learning. As previously mentioned, in a recent TDWI survey, self-service analytics was cited as the top priority for organizations. Machine learning was close behind.

Self-service analytics enables nontechnical users to be productive with data because the tools are easier to use (thanks to GUIs), do not require coding, and do not require IT developers to set up all data access, queries, visualizations, and preparation routines. For example, users may want to visualize and explore data as it becomes available in real time. They may want to analyze vast amounts of customer data in a self-service manner to uncover insights.

Business users often want to be in control of their specific projects and needs and to leverage IT for best practices. "One-size-fits-all" enterprise BI cannot address unique requirements; self-service can. Many self-service tools can enable immersive, "speed-of-thought" experiences with data and analytics. Some vendors have even gone a step further. For instance:

- **AUGMENTED INTELLIGENCE.** Vendors are putting significant effort into software that can assist or augment human intelligence. The idea behind augmented intelligence isn't to replace humans but to help them with tasks. Solutions that bill themselves as augmented intelligence include automation, but typically the system is not completely automated. The example mentioned in Number One (machine learning infused into the cloud data warehouse to help with optimization) is one illustration of this. In analytics, some vendors have invested in augmented intelligence to help business users find insights faster. For instance, some tools will automatically find insights and provide them to the end user. Some vendors provide automated data preparation functionality to help in data cleansing. The idea behind augmented intelligence is to make the self-service experience even easier.
- **ADVANCED ANALYTICS.** Additionally, companies are looking to become more sophisticated with analytics, often using predictive analytics or machine learning (PAML) algorithms to analyze data for use cases such as anomaly detection, predicting churn, or better understanding customers. Open source tools such as R or Python are top choices for this kind of model building. Modern self-service platforms often provide external integration with these tool sets. Some even include advanced functionality such as segmentation, time series analysis, or what-if analysis.

In terms of the modern analytics stack, it is important to look for tight integration with the cloud data warehouse to make data access faster and more efficient. In some cases, processing can be done in the cloud data warehouse and then cached in the analytics tool.

5 GOVERN THE DATA

Data governance generally refers to a set of processes, roles, standards, and measures that ensure important data assets are formally and consistently managed throughout the enterprise so that the people using the data can trust it. Governance rules and policies set out how an organization protects and manages its data. These rules and processes include understanding regulatory issues, dealing with data quality, maintaining standards, ensuring accountability, and keeping data secure and compliant in what is rapidly becoming a larger data ecosystem.

Because many users are most comfortable accessing and querying data in the data warehouse, most enterprises prioritize making it a trusted, curated environment, often fairly well governed. However, governance is often cited as a top challenge for organizations, especially as they move to newer platforms and environments. Traditional data governance will need to expand to address the complexities of the new cloud data warehouse and analytics stack. These include:

- **VISIBILITY.** A cloud data warehouse must enable visibility across the data life cycle. That includes the ability to monitor data-related activity in the cloud. Some cloud data warehouses provide audit logging to track information such as authentication attempts, disconnections, and queries run. This information can be used for audit and security purposes.
- **METADATA.** If the cloud data warehouse is the centralized repository for data for analysis, it will be important to include metadata about this data. This includes where the data was created, who owns it, what it is about, what it is used for, how it is organized, and where it is located (important for GDPR). Metadata helps provide data consistency and build trust. Some providers offer data catalog capabilities or partner with others to provide this capability. These catalogs help users quickly discover and understand their data in the cloud data warehouse.
- **DATA LINEAGE.** Audit logging can provide some information about what happens in the cloud data warehouse in terms of security and compliance. However, in modern environments, where transformations occur in the cloud, it is also important to track exactly what is happening to the data. Data lineage describes where data originated and how it has been changed and transformed. Modern data loading tools should enable visibility into everything that happened to the data in the new environment. This includes what happened as the data moved from source to target system and how it has changed in the target system (e.g., the cloud data warehouse).
- **COMPLIANCE REGULATIONS.** Organizations will need to ensure that their cloud providers meet the provisions of regulations such as GDPR and HIPAA.

Of course, governance involves people as well as technology. In fact, tasking data responsibilities to the right people is a core principle of data governance. Organizations will need to put processes in place to help them govern their data. TDWI also recommends that someone in the organization should own the overall data governance effort.

6 LOOK FOR AN INTEGRATED STACK

As organizations begin to build an integrated modern analytics stack that includes a cloud data warehouse, data integration, and an analytics platform, there are a number of issues to consider.

- **FUNCTIONALITY.** A good stack will include the features that your organization needs to meet its own requirements; therefore, it will be important to understand the features that are important to your own organization.

  For instance, how many data sources does your organization need to integrate? Can the data transformation tools handle the specific data sources your organization will need? What kinds of analytics are important to your organization? Will it include more advanced analytics such as machine learning? Do the platforms and services in your stack meet your company's security and governance requirements? Some required features may also be a function of the skills in your organization. For instance, your organization might be interested in tools with augmented intelligence if many business users are looking to derive insights.

- **COST CONSIDERATIONS:** Different providers have different price structures. This may include pricing per amount of time used, per query, or per cluster. Your organization will need to understand how it will use the stack to determine which cost and pricing structure best suits its needs.

- **PARTNERSHIPS.** Vendors form partnerships in order to make sure that their products are tightly integrated. Partnerships also help users by adding more capabilities. As you look for your stack, it can be helpful if the needed vendors have partnerships. Make sure these are technical partnerships rather than sales and marketing partnerships. For example, an ELT vendor may work with certain cloud data warehouse vendors at a technical level for push-down. Look for how the products connect. Is there a reference architecture for multi-tool solutions? The key to a good analytics stack is making sure that the products/services work together and are integrated.

FINAL THOUGHTS

As organizations strive to become more competitive, they realize the necessity of deriving and acting on insights from more diverse data from a wider variety of sources, both new and old. This often means rethinking the company's data and analytics environment to support these new requirements. One popular option to consider is a modern analytics stack that includes a cloud data warehouse, with data loading that utilizes the power of the cloud's processing engine for transformations, and a modern analytics platform that supports self-service analytics.

ABOUT OUR SPONSORS

For over 12 years, Amazon Web Services has been the world's most comprehensive and broadly adopted cloud platform. AWS offers over 125 fully featured services for compute, storage, databases, networking, analytics, machine learning, artificial intelligence (AI), Internet of Things (IoT), mobile, security, hybrid, virtual and augmented reality (VR and AR), media and application development, deployment, and management from 55 Availability Zones (AZs) within 18 geographic regions and one Local Region around the world spanning the U.S., Australia, Brazil, Canada, China, France, Germany, India, Ireland, Japan, Korea, Singapore, and the UK. AWS services are trusted by millions of active customers around the world—including the fastest growing start-ups, largest enterprises, and leading government agencies—to power their infrastructure and make them more agile and lower costs. To learn more about AWS, visit https://aws.amazon.com.

Matillion is data transformation for cloud data warehouses. Only Matillion is purpose-built for Amazon Redshift, enabling businesses to achieve new levels of simplicity, speed, scale, and savings. Matillion products are available on the AWS Marketplace. The company has the highest-rated ELT product on AWS Marketplace, with 90 percent of customers saying they would recommend Matillion. Winner of The TrustRadius Top Rated Award in Data Integration based on unbiased feedback from customers, Matillion software is used by more than 650 customers across 40 countries. Dual-headquartered in Manchester, U.K., and Denver, Colorado, Matillion also has offices in New York City and Seattle. Learn more about how you can unlock the potential of your data with Matillion's cloud-native approach to data transformation. Visit us at www.matillion.com.

Tableau helps people and organizations become more data-driven as a trusted leader in analytics. The Tableau platform provides the breadth and depth of capabilities to serve the needs of even the largest global enterprises in a seamless, integrated experience. Tableau is designed to fit, not dictate, your data strategy, and adapts to your environment with unmatched flexibility and choice while meeting the toughest governance and security requirements. People love using Tableau because it is both powerful and intuitive—and offers a fundamentally different user experience by empowering people of all skill levels to explore and analyze data using visuals and natural language. Tableau has become the standard language of analytics for modern business users and continues to lead the industry with the most passionate and engaged user community in analytics, a customer base with millions of users at more than 86,000 organizations, and a deep commitment to customer-focused innovation. Tableau.com

ABOUT THE AUTHOR

Fern Halper, Ph.D., is vice president and senior director of TDWI Research for advanced analytics. She is well known in the analytics community, having been published hundreds of times on data mining and information technology over the past 20 years. Halper is also coauthor of several Dummies books on cloud computing and big data. She focuses on advanced analytics, including predictive analytics, text and social media analysis, machine learning, AI, cognitive computing, and big data analytics approaches. She has been a partner at industry analyst firm Hurwitz & Associates and a lead data analyst for Bell Labs. Her Ph.D. is from Texas A&M University. You can reach her by email (fhalper@tdwi.org), on Twitter (twitter.com/fhalper), and on LinkedIn (linkedin.com/in/fbhalper).

ABOUT TDWI CHECKLIST REPORTS

TDWI Checklist Reports provide an overview of success factors for a specific project in business intelligence, data warehousing, analytics, or a related data management discipline. Companies may use this overview to get organized before beginning a project or to identify goals and areas of improvement for current projects.

ABOUT TDWI RESEARCH

TDWI Research provides research and advice for BI professionals worldwide. TDWI Research focuses exclusively on analytics and data management issues and teams up with industry practitioners to deliver both broad and deep understanding of the business and technical issues surrounding the deployment of business intelligence and data management solutions. TDWI Research offers reports, commentary, and inquiry services via a worldwide membership program and provides custom research, benchmarking, and strategic planning services to user and vendor organizations.