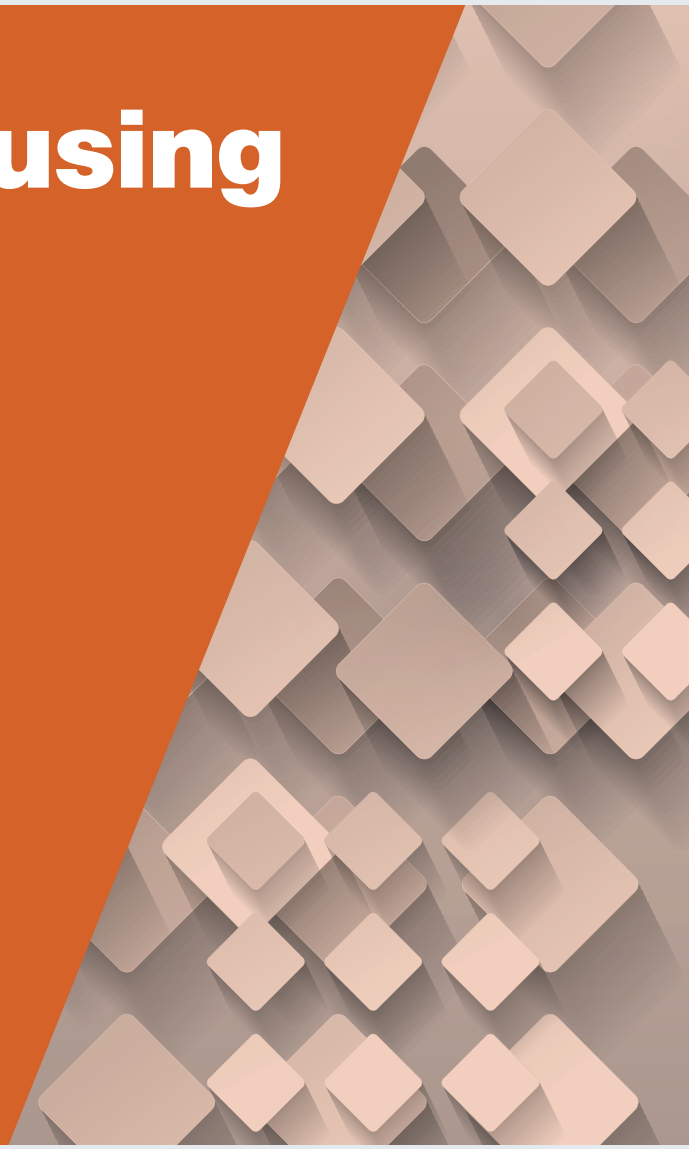
TDWI RESEARCH

TDWI CHECKLIST REPORT

Data Warehousing in the Cloud



By David Loshin

Sponsored by:

IBM

tdwi.org

tdwi

JULY 2015

TDWI CHECKLIST REPORT

Data Warehousing in the Cloud

By David Loshin

TABLE OF CONTENTS

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

An impact of the burgeoning business interest in analytics is increased interest among small and midsize businesses to adopt a business intelligence (BI), reporting, and analytics strategy. In the past, larger organizations may have been willing to invest in hardware, software, and expertise to establish an enterprise data warehouse environment. However, smaller businesses have often lacked the budget, access to skilled professionals, or the determination to design, build, and support the necessary dedicated platforms.

Expanding the pool of organizations able to embrace a data warehousing and BI capability depends on:

- Reducing the complexity of instantiating an analytics environment
- Lowering the cost barrier to entry
- Unshackling business users from the information technology (IT) department by providing rapid accessibility to data using self-service analytics tools

Fortunately, as data professionals have gained experience in implementing common design patterns and integration models, the data warehousing industry has matured, resulting in a general agreement about the architecture of an end-to-end BI and analytics framework.

Coupling this prototypical architecture with increasingly sophisticated software tools addresses this challenge. It simplifies launching a dedicated system for data analysis and reduces the need for specialized skills. The convergence of the standardized architecture for data warehousing and BI, advances in data integration and data discovery technologies, and cost models of utility computing via the cloud enables vendors to develop cloud-based data warehousing systems.

This paradigm allows vendors to instantiate their tool suite as a managed platform in a way that presents a special opportunity for a broad community of businesses to fulfill the objectives of their analytics strategies. Relying on a service provider to instantiate the platform simplifies the business's progression toward the advantages of analytics. Using a cloud-based data warehouse eliminates the need for capital hardware investments and IT support staff to configure and manage the platform. This frees the data scientists, analysts, and architects to focus on the analytical models that will improve business performance.

Our objective in this Checklist Report is to share best practices that will position the reader to take advantage of a cloud-based data warehousing solution.

NUMBER ONE

FIT YOUR DATA WAREHOUSE PLATFORM TO THE ANALYTICS PURPOSE

The concept of a data warehouse may conjure different things to different information consumers. In practice, data warehousing increasingly covers a broad spectrum of capabilities. These capabilities range from periodically published predefined reports reviewed by operational managers to ad hoc queries and interactive drill-through analyses performed by business analysts seeking to address particular business challenges, to complex predictive analytics models in the realm of the statisticians and data scientists. In reality, these different types of applications are bound to have different functional and system requirements for data accessibility, computation, and complex algorithms.

An effective strategy allows the data warehouse practitioners to determine both the resource requirements and the required utilities for quickly addressing each customer's needs. This suggests reconsidering the perception of the enterprise data warehouse as a monolithic system designed for continuously supporting a mixed workload and instead looking at fitting the data warehouse platform to specific business purposes. For example, routine, periodic, and canned reports may be suited to an established dimensional star-schema data mart while predictive modeling applications might be better suited to a specialty high-performance analytics appliance that can handle massive data volumes and parallel computation.

Cloud-based data warehousing adds to this evolving hybridization of the reporting and analysis ecosystem. A cloud-based approach provides nimbleness in deployment. When a partner or service provider manages a cloud data warehouse and helps in the instantiation of the schema and data loading, it frees the stakeholders to concentrate on the analysis and the results instead of building the system.

The projects targeted to a cloud deployment often have tightly-scoped requirements. Examples include short-lived projects, seasonal analyses, analyses that are relevant for a short time, limited self-service reporting systems, and even prototyping new reports and analyses. For such projects, using a cloud-based data warehouse provides specific value because there is no need to design, develop, and deploy the platform and data management framework. Aside from reducing the start-up costs, this accelerates analysis and reduces or even eliminates ongoing maintenance costs.

NUMBER TWO

USE A COST MODEL TO DETERMINE WHEN CLOUD BI MAKES SENSE

The cost of managing a data warehouse is one of the barriers that must be overcome before the benefits of a reporting and analytics environment can be achieved. However, most data practitioners are largely unaware of the many variables that contribute to the total cost of a data warehouse's operations over the system's lifetime, including:

- **Acquisition costs** associated with evaluating and buying hardware, storage, software, and network connectivity
- **Deployment costs** such as project planning, oversight and management, system design, development, configuration, testing, and implementation
- **Data development and management costs**, including data extraction, design and development of data integration applications, and design and implementation of data warehouse schemas
- **Business opportunity costs** incurred when the business is impacted by delays in getting the system running
- **Operations and maintenance costs** covering power, cooling, space, and telecommunications
- **Recurring costs** such as software license maintenance, system upgrades, and coverage for data archiving, data backup, recovery, and disaster planning

Different organizations may have variable tolerance for different cost categories. More established businesses may be willing to make a capital investment in infrastructure knowing that the benefits outweigh the start-up costs. Small or new businesses might not have sufficient capital reserved to pay the recurring costs over an extended timeframe yet may desire a short time to value.

Develop a cost model to balance how key expenses impact time to value. Use the cost model to determine when cloud-based data warehousing makes the most sense. In some cases, the cost of acquiring and managing the system may be spread across several projects by leveraging the platform for other enterprise tasks. Alternatively, the agility of a managed system may pay off—if you can drive additional revenue six months earlier by using a cloud-based system, the increased revenue may more than offset the capital investment of a system acquisition.

NUMBER THREE

SHORTEN TIME TO VALUE BY SIMPLIFYING DEPLOYMENT

Cloud-based data warehousing and BI promises simplified deployment. First, much of the infrastructure work is already done—the service provider will choose the hardware platform and database management system. The selection process and the platform's management are essentially transparent to the customer.

Second, the customer will benefit from the service provider's experience with assembling the tools to support the complete process, including data ingestion, profiling, transformation, loading, reporting, and querying. Leveraging the provider's data integration, data delivery, and presentation experience can simplify development. Third, a number of cloud-based data warehousing vendors are adding value by incorporating more complex capabilities such as data discovery and visualization tools as well as integration with predictive and prescriptive modeling tools (such as those provided within the R modeling language).

Offloading underlying infrastructure-engineering tasks lets the customer focus on the qualitative parts of data analysis. By standardizing rapid-deployment processes, customers can attend to the information requirements. This approach should encompass at least these tasks:

- **Business objectives**: Articulating the objectives of organizing data for reporting and analysis and presenting those data sets to a specific community of users
- **Data requirements assessment**: Determining the data sets needed to populate the data warehouse
- **Information modeling:** Considerations for organizing the data to be represented within the data warehouse
- **Data integration:** Developing and implementing the processes for moving the required data to the cloud platform
- **Rule-based transformations:** Standardized transformations that can be parameterized within a data preparation tool
- **Business-driven analysis:** Determining the types of analyses to be performed and establishing the analytics capabilities to deliver the expected results of the project

Fortunately, the service provider's execution teams can support the mechanical aspects of many of these tasks, such as data modeling, data integration, and configuring the rule-based transformation engine. Consequently, embracing a standardized process for deploying cloud-based BI/analytics projects will increase agility and accessibility to actionable knowledge.

NUMBER FOUR

LOOK FOR CLOUD-BASED SYSTEMS WITH INTEGRATED ANALYTICS

Because the methods for business intelligence, decision-support, and decision analytics have matured over recent years, the sophistication of the business data consumers has grown along with that of the technology. Although some cloud-based data warehouse providers focus on straightforward reporting and dimensional analysis, others are rapidly integrating predictive and prescriptive analytics capabilities, including the following:

- **Clustering**, in which algorithms attempt to group entities (such as customers) based on their characteristics and behaviors
- **Segmentation**, an approach for differentiating entities (such as vendors) based on previously-created clustering models
- **Classification**, which employs iterative algorithms to assign an individual into a predefined class, such as "best customers," "good customers," "medium customers," and "undesired customers"
- **Decision trees,** which expose the most relevant criteria used for classification or for making an optimal selection when presented with choices
- **Association analysis,** which iteratively reviews relationships among events within a data set to reveal correlations that represent potential business opportunities

In the past, many of these capabilities required a segregated platform for advanced analytics calculations, but they are increasingly supported by architectural innovations such as:

- Data warehouse appliances, which are specialty platforms designed to support a mixed workload consisting of both traditional querying and reporting as well as more advanced analytics.
- In-database analytics, where the database management system vendor has engineered data mining algorithms to be integrated within a more traditional SQL-style interface, enabling the analyses to be embedded within more familiar query formats.
- In-memory computing, in which the database vendor has optimized its storage organization and its models of data access to store the most frequently used data, if not all of the data, in-memory instead of disk. This significantly speeds up both traditional and more advanced analyses.

Look for a provider whose cloud-based service enables the preprocessing and loading of data into a data warehouse *and* provides broad analytic functionality as part of the environment. In addition, the service provider's offerings should adapt innovative designs to meet the particular needs of its customers.

NUMBER FIVE

ENSURE THE CLOUD PLATFORM MEETS CONSISTENT PERFORMANCE REQUIREMENTS

One of the risks of any hosted application is the provider's reliance on deploying the application using virtualized environments. This may lower the customer's overall cost of operations. However, the application may be redeployed at any time on different underlying hardware and may potentially be co-located with other applications whose execution may affect your application's performance.

In most organizations, the inability to deliver reports and analyses quickly to all data consumers will impact adoption and, consequently, success. Remember that you may not get consistent performance with a virtualized environment. If your organization requires predictable performance, clearly specify the performance criteria and levels of acceptability and share these objectives with the managed service provider candidates. Evaluate the provider's methods for ensuring or improving performance. Ask such questions as the following:

- Does the cloud-based data warehouse vendor provide performance benchmarks that accurately reflect how your application will run?
- Does the vendor provide an option for deploying your project on a "bare metal" cloud platform instead of a virtualized platform?
- Can the platform be configured using architectural enhancements to speed query execution and presentation of results such as using columnar data alignment or in-memory processing?

Work with the vendor to ensure that your performance requirements are met. In addition, make sure protocols for reporting and addressing performance deficiencies are well defined.

NUMBER SIX
PROACTIVELY MANAGE DATA CONNECTIVITY AND INTEGRABILITY

If you are considering cloud-based BI and analytics, acknowledge the need to easily move the data for reporting and analysis in the cloud environment. Although the data required to populate small data marts may not seem so imposing, remain aware that the expectations and related costs of data connectivity and integration go beyond data movement and loading. The requirements necessarily incorporate the complexity of understanding, preparing, and integrating a variety of data source types. Those types can include flat file data, data in relational database management systems accessed using SQL, data managed in newer NoSQL environments, geospatial data, and HDFS files on Hadoop, among others.

Develop a plan to proactively manage data connectivity and integration. Your plan is incomplete without considering the following:

- **Network connectivity** between each of the data sources and the cloud-based data warehouse. This implies links from your organization's environment as well as accessibility to sources managed by other SaaS and cloud-based systems.
- Alternate means of **data movement** in case data warehouse volumes exceed the capacity of standard network connections, which may require faster connections with greater bandwidth.
- **Data profiling and analysis** to assess potential anomalies and to discover embedded structure, metadata, and business rules that relate to the subsequent data transformations.
- Business rules for **data standardization and transformation** as part of ongoing data preparation.
- Employing **replication and change data capture** to reduce the overhead associated with refreshing the entire data warehouse.
- **Data compression** as an alternative for reducing the time it takes to move data from any source to the cloud-based warehouse.

User demands for growing data volumes from more varied sources require more complex integration. Look for cloud-based data warehouse vendors providing tools and especially for services supporting data profiling and discovery, compression, transmission, data preparation, and efficient data loading.

NUMBER SEVEN
SATISFY SECURITY AND DATA PROTECTION REQUIREMENTS

Another perceived risk of using a hosted or cloud-based data warehouse is the potential for violation of the organization's data security policies or regulatory directives. Conventional thinking may lead to the belief that there is uncertainty in guaranteeing access security and data protection for two reasons. First, in some cases, a multi-tenant architecture allows multiple customer applications to run within the same environment, triggering fear of data leakage across the application boundary. Second, storage on virtual platforms may be distributed across multiple physical machines, which can create a fear about the ability to scrape "remnant" data should the application be migrated.

Obviously your enterprise must perform due diligence to assess the security and data privacy protection needs and ensure that the vendor can satisfy those needs. The cloud-based DW vendor might provide the following methods:

- Authentication of user identity and authorization of the user to prevent unauthorized data access
- Finely-grained data access controls to prevent exposure of protected data attributes
- Data masking to prevent presentation of protected data attributes
- Data encryption, which might be applied to the data "at rest," or where it is stored, as well as "in motion" as the data is accessed and delivered to the user's portal
- Data wiping, which is used to completely overwrite hard drives to prevent malicious recovery

As vendors have become more proactive in identifying and addressing existing and potential security vulnerabilities, the perception of cloud-based systems as data protection risks is starting to recede. Nonetheless, one alternative to allay fears of data exposure echoes our earlier suggestions about ensuring predictable performance: seek out vendors that provide an option for deploying your project on a "bare metal" cloud platform instead of a virtualized platform. Isolation of your application will avoid the perceived risks of virtualization and multi-tenancy.

AFTERWORD

CHOOSING AND ESTABLISHING A GOOD RELATIONSHIP WITH YOUR VENDOR

The suggestions in this checklist provide context for determining if cloud-based data warehousing is right for your organization. Once you have decided to transition data warehousing and BI applications to a cloud provider, make sure to identify the right service provider. In summary, some criteria we have noted for assessing a cloud-based data warehousing service focus on the ways that their products and services supplement your reporting and analytics program, including the following:

- Reducing overall cost of development and operations
- Shortening time to value
- Reducing your reliance on internal IT resources
- Simplifying data ingestion, integration, and loading
- Empowering your data consumer community through improved ease of use
- Supporting your needs for elasticity and scalability
- Enabling business continuity through fault tolerance and managed failover
- Establishing trust in the system security and protection of private information

Once you have identified a vendor, we recommend you establish a good working relationship with a trusted cloud data warehousing provider, which is important for three key reasons:

- **Sustainability of the environment:** A trusted partner will ensure that the environment can address all your business analytics needs for all the stages of the data warehouse life cycle and the incremental needs for elasticity and scalability, security, and overall performance over the lifetime of the projects.
- **Responsiveness:** A worthy service provider can demonstrate that you can trust them to address any issues that emerge in a timely and reliable manner.
- **Engagement:** Look for providers that will help you rapidly deploy your system *and* work with you and your data consumers to continue maturing your reporting, BI, and analytics program.

Cloud data warehousing vendors can leverage their implementation experience with their customers and can marshal them along to align with the customers' short-, medium-, and long-term strategies.

ABOUT OUR SPONSOR

ibm.com

IBM Cloud Data Services provides developers with a comprehensive set of rich, integrated data services covering content, data, and analytics. Cloud Data Service offerings speed up time to market, improve uptime, and deliver higher value to developers of Web and mobile applications. For information about how IBM Cloud Data Services is changing the way services are created for and delivered to developers, follow us on Twitter at @getdashDB and @cloudant, and visit www.dashdb.com and www.cloudant.com.

ABOUT THE AUTHOR

David Loshin, president of Knowledge Integrity, Inc, (www.knowledge-integrity.com), is a recognized thought leader, TDWI instructor, and expert consultant in the areas of data management and business intelligence. David is a prolific author regarding business intelligence best practices, with numerous books and papers on data management, including *The Practitioner's Guide to Data Quality Improvement*, with additional content provided at www.dataqualitybook.com. David is frequently invited to speak at conferences, Web seminars, and sponsored websites and channels. David can be reached at loshin@knowledge-integrity.com.

ABOUT TDWI RESEARCH

TDWI Research provides research and advice for data professionals worldwide. TDWI Research focuses exclusively on business intelligence, data warehousing, and analytics issues and teams up with industry thought leaders and practitioners to deliver both broad and deep understanding of the business and technical challenges surrounding the deployment and use of business intelligence, data warehousing, and analytics solutions. TDWI Research offers in-depth research reports, commentary, inquiry services, and topical conferences as well as strategic planning services to user and vendor organizations.

ABOUT TDWI CHECKLIST REPORTS

TDWI Checklist Reports provide an overview of success factors for a specific project in business intelligence, data warehousing, 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.