TDWI RESEARCH

TDWI CHECKLIST REPORT

THE MODERN DATA WAREHOUSE

What Enterprises Must Have Today and What They'll Need in the Future

By Philip Russom

Sponsored by

Microsoft

tdwi.org

tdwi

NOVEMBER 2013

TDWI CHECKLIST REPORT

THE MODERN DATA WAREHOUSE

What Enterprises Must Have Today and What They'll Need in the Future

By Philip Russom

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

In recent surveys by TDWI Research, roughly half of respondents report that they will replace their primary data warehouse (DW) platform and/or analytic tools within three years. Ripping out and replacing a DW or analytics platform is expensive for IT budgets and intrusive for business users. This raises the question: What circumstances would lead so many people down such a dramatic path?

It's because many organizations need a more modern DW platform to address a number of new and future business and technology requirements. Most of the new requirements relate to big data and advanced analytics, so the data warehouse of the future must support these in multiple ways. Hence, a leading goal of the modern data warehouse is to enable more and bigger data management solutions and analytic applications, which in turn helps the organization automate more business processes, operate closer to real time, and through analytics learn valuable new facts about business operations, customers, products, and so on.

For organizations that need a modern data warehouse that satisfies new and future requirements, we offer a checklist of our top six recommendations. These can guide your selection of vendor products and your solution design.

1. Deploy a modern data warehouse as a foundation for leveraging big data.
2. Make sense of multi-structured data for new and unique business insights.
3. Implement advanced forms of analytics to enable discovery analytics for big data.
4. Empower the business to operate in near real time by delivering data faster.
5. Integrate multiple platforms into a unified data warehouse architecture.
6. Demand high performance and scalability of all components of a data warehouse.

Users facing new and future requirements for big data, analytics, and real-time operation need to start planning today for the data warehouse of the future. To help them prepare, this TDWI Checklist Report drills into each of the six recommendations, listing and discussing many of the new vendor product types, functionality, and user best practices that will be common in the near future, plus the business case and technology strengths of each.

NUMBER ONE

DEPLOY A MODERN DATA WAREHOUSE AS A FOUNDATION FOR LEVERAGING BIG DATA.

The goal of managing big data is to provide a business with actionable analytic insights. In TDWI's view, the greatest business value drawn from big data comes from analyzing it. This fact is so apparent that there's even a name for it: *big data analytics*. For example, a common analytic application today is the "sessionization" of website log data, which reveals the behavior of site visitors—information that helps marketers and Web designers do their jobs better. As another example, trucks and railcars are loaded with sensors and GPS systems nowadays so logistic firms can analyze operator behavior, vehicle performance, onboard inventory, and delivery route efficiency. In these examples, collecting big data from Web applications or sensors is almost incidental. The real point is to elevate the business to the next level of corporate performance based on insights gleaned from the analysis of big data.

Leverage big data, don't just manage it. It costs money, time, bandwidth, and human resources to collect and store big data. Therefore, no one should be content to simply manage big data as a cost center that burns up valuable resources. For example, an organization of any size or sophistication will have a website that generates Web logs. Many firms have hoarded logs for years, sensing their value but unsure how to extract it. This is a great starting place for a new big data program because it identifies business value (through log analytics) based on an existing big data resource.

Big data comes in many forms, both old and new. Much of the hype around big data stresses new sources of unstructured big data, such as websites, machines, sensors, and social media. These are important because new sources tend to yield new insights or add greater detail and accuracy to older analytics. Let's not forget older forms of structured big data. For example, for decades telcos have been collecting millions of call detail records (CDRs) daily and analyzing them to detect fraud, understand customer behaviors, and plan network capacity. Likewise, thousands of organizations get most of their analytic insights from the terabytes of relational data that originated from their transactional and operational applications.

Over time, what we call *big data* today will be assimilated into the broader category of *enterprise data*. This means that enterprise data will evolve into an eclectic mix of multi-structured data types instead of its current state as mostly structured and relational data. That's a good thing for analytics, which can make more analytic correlations in a more granular fashion as the range of data types and sources broadens.

(Continues)

(Continued)

ENABLING TECHNOLOGIES

No matter where you start, the long-term trend in big data analytics is to collect and analyze all data, regardless of its size, type, model, source, or vintage. Organizations that have matured into managing and analyzing such a wide range of data typically do it with two types of data platforms—relational database management systems (RDBMSs) and Hadoop tools, especially the Hadoop Distributed File System (HDFS).

On the one hand, RDBMSs and SQL are more critical to DW and analytics than ever. On the other hand, they excel with relational data, whereas much of the valuable big data that's emerging is not relational or even structured. That's where HDFS complements an RDBMS. HDFS can serve as a massive and scalable operational data store (ODS) and/or data staging area that feeds an RDBMS. Furthermore, when combined with other Hadoop tools such as MapReduce, Hive, and HBase, HDFS can be a powerful platform for non-SQL, algorithmic analytics, plus large-scale reporting or ETL functions.

Hadoop's strengths lie in areas where most data warehouses are weak, such as unstructured data, very large data sets, non-SQL algorithmic analytics, and handling file-based data. Conversely, Hadoop's limitations are mostly met by mature functionality available today from a wide range of RDBMS types, including OLTP databases, columnar databases, DW appliances, and new cloud-based warehouses.

In that light, Hadoop and RDBMS-based data warehouses are complementary, despite a bit of overlap. More to the point, the two together empower user organizations to perform a wider range of analytics with a wider range of data types, with unprecedented scale and favorable economics. For these reasons, most modern data warehouses will be built on a foundation consisting of both relational and Hadoop technologies in the near future.

NUMBER TWO

MAKE SENSE OF MULTI-STRUCTURED DATA FOR NEW AND UNIQUE BUSINESS INSIGHTS.

In addition to very large data sets, big data can also be an eclectic mix of structured data (relational data), unstructured data (human language text), semi-structured data (RFID, XML), and streaming data (from machines, sensors, Web applications, and social media). The term *multi-structured data* refers to data sets or data environments that include a mix of these data types and structures.

Multi-structured types of big data have compelling value propositions. For example, human language text drawn from your website, call center application, and social media can be processed by tools for text mining or text analytics to create a sentiment analysis, which in turn gives sales and marketing valuable insights into what your customers think of your firm and its products. As another example, organizations with an active supply chain can analyze semi-structured data exchanged among partners (in, say, XML, JSON, RFID, or CSV formats) to understand which partners are the most profitable and which supplies are of the highest quality.

ENABLING TECHNOLOGIES

Hadoop empowers organizations to finally crack the nut of multi-structured data. For years, most BI and DW professionals have known there's business value in processing multi-structured data. Even so, few have done anything of consequence, due to a lack of credible support in data platforms, but the situation is changing.

Being a data-type-agnostic file system, HDFS manages the full range of file-based data that's structured, unstructured, semi-structured, or a mix of these. The introduction of Hadoop into the modern data warehouse provides low-cost and scalable mechanisms for capturing and managing diverse multi-structured data and analyzing it for business value.

Natural language processing (NLP) is key to getting business value from multi-structured data. NLP takes many forms, including tools or services for text mining, text analytics, sentiment analysis, fact clustering, and search.

NUMBER THREE

IMPLEMENT ADVANCED FORMS OF ANALYTICS TO ENABLE DISCOVERY ANALYTICS FOR BIG DATA.

Business managers assume that the best route to business value from big data is through advanced analytics. Online analytic processing (OLAP) continues to be the most common form of analytics today. OLAP is not going away due to its value serving a wide range of end users. The current trend is to complement OLAP with advanced forms of analytics based on technologies for data mining, statistics, natural language processing, and SQL-based analytics. These are more suited to exploration and discovery than OLAP is. Note that most DWs today are designed to provide data mostly for standard reports and OLAP, whereas the modern data warehouse will also provide more data and functionality for advanced analytics.

Most applications of advanced analytic methods enable discovery. This is especially true of data mining technologies and similar clustering or correlation algorithms. The user is usually looking for facts about the business or related entities (customers, products, locations) that were previously unknown, or they may be looking for fraud, new customer segments, hidden costs, correlations among people, affinities among products, and so on.

Big data can enable new analytic applications. For example, in recent years, a number of trucking companies and railroads have added multiple sensors to each of their fleet vehicles and train cars. The big data that streams from sensors enables companies to more efficiently manage mobile assets, deliver products to customers more predictably, identify noncompliant operations, and spot vehicles that need maintenance.

Big data can extend older analytic applications. For example, so-called 360-degree views of customers are more complete when based on both traditional enterprise data and big data. In fact, some sources of big data come from new customer touchpoints such as mobile apps and social media. Big data can also beef up the data samples parsed by analytic applications, especially those for fraud, risk, and customer segmentation.

Managing big data for analytics differs from managing other DW data. Analytic discoveries often depend on highly detailed source data. If the source data is merged, transformed, or standardized, the details can be lost. Furthermore, data standards and data models imposed on data can limit discovery. This is why the modern data warehouse maintains source data in its original state when the data will apply to discovery analytics. This also allows data to be repurposed at analysis time, thus enabling data analysts to go any direction a new discovery mission suggests.

ENABLING TECHNOLOGIES

RDBMS optimized for SQL-based analytics. TDWI surveys show that SQL-based analytics is the most common form of advanced analytics at the moment, being more popular than analytics based on data mining, statistics, artificial intelligence, or natural language processing (NLP). This is natural because most DW professionals know SQL well and have many BI and analytic tools that are compatible with standard SQL.

In SQL-based analytics, a data analyst, data scientist, or similar user starts with ad hoc queries and builds them up through numerous iterations until the resulting data set yields the desired epiphany. With each iteration, the SQL code gets longer and more complex. That's why SQL-based analytics is best done with tools and RDBMSs that are built and optimized for highly complex SQL.

RDBMS integrated with Hadoop. SQL-based analytics assumes relational technology, even when applied to data managed in Hadoop. This is another reason the modern data warehouse is built with both relational and Hadoop technologies. RDBMSs and relational tools provide SQL support, productive development tools, and optimization for standard SQL that's far more high performance, feature rich, and user friendly than what's available from Hadoop products such as Hive and HBase. Conversely, Hadoop provides a massive data store with support for multi-structured data types.

Hadoop for a wide range of algorithm analytics. SQL aside, a growing number of analytic tools based on mining, statistics, and NLP are available as modules that run against HDFS data, usually through layers above HDFS, such as MapReduce, Hive, and HBase. These algorithmic approaches complement SQL's set-based approach.

Self service for end users. Accessing HDFS data from Excel (if you have the proper interface) provides the high ease of use that empowers a wide range of users to do their own data exploration, discovery analytics, and data visualization with Hadoop data.

NUMBER FOUR

EMPOWER THE BUSINESS TO OPERATE IN NEAR REAL TIME BY DELIVERING DATA FASTER.

Sometimes the term *real time* literally means that data is fetched, processed, and delivered in seconds or milliseconds after the data is created or altered. However, most so-called real-time data operations take minutes or hours and are more aptly called *near real time*. That's fine, because near real time is an improvement over the usual 24-hour data refresh cycle, and few business processes can react in seconds anyway.

Business managers need near-real-time reports and analyses for time-sensitive business processes. This is already the case with established BI practices, such as operational BI and metrics-driven performance management. Both assume fresh data fetched and presented in reports and dashboards to end users in real time or near to it. This enables managers to make tactical and operational decisions based on very fresh information.

The practices of operational BI and performance management have brought operational reports from overnight refreshes to near-real-time refreshes multiple times during the business day. Managers now need analytics to likewise evolve from its current offline latency toward real-time analytics, sometimes called operational analytics. A few organizations need to capture streaming big data and analyze the stream in real time or close to it. Examples include applications for financial trading systems, business activity monitoring, utility grid monitoring, e-commerce product recommendations, and facility monitoring and surveillance.

ENABLING TECHNOLOGIES

The modern data warehouse must support a number of capabilities that fetch, process, and deliver data in real time or close to it.

SQL queries that return results in seconds. One of the reasons that refreshing management dashboards in near real time has become a technical reality is that recent versions of RDBMSs can execute report queries in sub-second time frames. For the same performance with the complex queries of SQL-based analytics, look for RDBMSs that are purpose-built for DW and analytics.

Columnar data stores. When data from each column is stored in close physical proximity, retrieving data from a single column has far less I/O overhead than with the row-oriented methods RDBMSs followed for decades. Because most queries in DW and analytics focus on columnar data, a columnar data store greatly accelerates query speed. Plus, a column store compresses data dramatically with little overhead, which is important for big data's big volumes. Hence, columnar data stores have become popular in recent years, and should be included in a modern data warehouse.

In-memory databases. One way to get high performance (in the sense of fast data access) from a database is to manage it in server memory. This accelerates processing by avoiding time-consuming I/O. However, you need an RDBMS that can manage in-memory data well by updating memory (but keeping it in sync with a persistent copy on disk) and by recognizing hot data (and giving it priority for scarce memory resources). In many applications today, metric data in support of operational BI is managed in memory so managers can refresh dashboards on demand. An emerging practice involves in-memory databases for advanced analytics, typically to speed the scoring of analytic models.

NUMBER FIVE

INTEGRATE MULTIPLE PLATFORMS INTO A UNIFIED DATA WAREHOUSE ARCHITECTURE.

Diverse big data is subject to diverse processing, which may require multiple platforms. To keep things simple, users should manage big data on as few data platform types as possible to minimize data movement as well as to avoid data synchronization and silo problems that work against the "single version of the truth." Yet there are ample exceptions to this rule, such as the RDBMS/HDFS foundation for DWs discussed in this report. As you expand into multiple types of analytics with multiple big data structures, you will inevitably spawn many types of data workloads. Because no single platform runs all workloads equally well, most DW and analytic systems are trending toward a multi-platform environment.

From the EDW to the multi-platform DWE. A consequence of the workload-centric approach is a trend away from the single-platform monolith of the enterprise data warehouse (EDW) toward a physically distributed data warehouse environment (DWE). A modern DWE consists of multiple data platform types, ranging from the traditional relational and multidimensional warehouse (and its satellite systems for data marts and ODSs) to new platforms such as DW appliances, columnar RDBMSs, NoSQL databases, MapReduce tools, and HDFS. In other words, users' portfolios of tools for BI/DW and related disciplines are diversifying aggressively. The downside is that the multi-platform approach adds more complexity to the DW environment. The upside is that BI/DW professionals have always managed complex technology stacks successfully, and users love the high performance and solid information outcomes they get from workload-tuned platforms.

ENABLING TECHNOLOGIES

A unified data warehouse architecture helps data professionals cope with the growing complexity of their multi-platform environments. Some organizations are simplifying the data warehouse environment by acquiring vendor-built data platforms that have an inherent unifying architecture and/or are based on easily deployed and extended appliance configurations.

An integrated RDBMS/HDFS combo is an emerging architecture for the modern DW. The trick is integrating HDFS and an RDBMS so they work together optimally. For example, an emerging best practice among DW professionals with Hadoop experience is to manage diverse big data in HDFS but process it and move the results (via queries or ETL) to RDBMSs (elsewhere in the DW architecture) that are more conducive to SQL-based analytics. HDFS serves as a massive data staging area or ODS for the DW architecture, whereas an RDBMS and tools for reports, queries, and analytics serve as front ends for HDFS data.

This requires new interfaces and interoperability between HDFS and RDBMSs, and it requires integration at the semantic layer, in which all data—even multi-structured, file-based data in Hadoop—looks relational. This is the secret sauce that unifies the RDBMS/HDFS architecture. It enables distributed queries based on standard SQL that simultaneously access data in the warehouse, HDFS, and elsewhere without preprocessing data to remodel or relocate it.

Pre-integration and optimization for the components of the multi-platform DW. To get this, users should consider DW appliances. An appliance includes hardware, software, storage, and networking components, pre-integrated and optimized for warehousing and analytics. Appliances have always been designed and optimized for complex queries against very large data sets; now they must also be optimized for the access and query of diverse types of big data.

Clouds are emerging as platforms and architectural components for modern DWs. Another successful scheme for simplifying the modern data warehouse environment is to outsource all or some of it, typically to a cloud-based DBMS, DW, or analytics platform. User organizations are adopting a mix of cloud types (both public and private) and freely mixing them with traditional on-premises platforms.

For many, a cloud is a solid data management strategy due to its fluid allocation and reapportionment of virtualized system resources, which can enhance the performance and scalability of a data warehouse. However, a cloud can also be an enhancement strategy that uses a hybrid architecture to future-proof data warehouse capabilities. To pursue this strategy, look for cloud-ready, on-premises DW platforms that can integrate with cloud-based data and analytic functionality to extend DW capabilities incrementally over time.

NUMBER SIX

DEMAND HIGH PERFORMANCE AND SCALABILITY OF ALL COMPONENTS OF A MODERN DATA WAREHOUSE.

Data warehouse professionals need high speed and scale from every piece of the DW environment. With both traditional enterprise data and new big data exploding exponentially, scalability for data volumes is the top priority for many teams. Other instances of scalability are pressing, too, such as scaling up to thousands of report consumers and regularly refreshing the tens of thousands of reports they depend on.

There's a need for speed. Speed and scale are related because speed for each atomic operation is required for scaling up the overall system. Speed for query performance and refreshing reports is key to meeting managers' expectations for dashboards, performance management, on-demand reports, data exploration, data visualization, and a growing range of analytics. Plus, real-time and near-real-time data operations require speed, as discussed earlier in this report.

Big data just keeps getting bigger. In fact, data volumes in the 10–99 terabyte range got the most responses in a recent TDWI survey, indicating that it's the norm for today's big data volumes. However, the survey also revealed that 100 or more terabytes will become the norm within three years, and a quarter of users surveyed anticipate breaking the 1 petabyte barrier within three years.

The DW of the future is highly available. For example, in a global firm the DW is accessed 24/7, and real-time data isn't real time if the DW is down. Look for DW RDBMSs that have dedicated functions for high availability, ranging from data replication to hardware redundancy (as in an MPP RDBMS, an HDFS cluster, or a DW appliance).

High-speed bulk data load is more important than ever given the size of big data. It's a critical success feature for "load and go" analytic methods that accumulate very large data sets within a few hours.

ENABLING TECHNOLOGIES

Depend on MPP, not SMP, RDBMSs. RDBMSs as the foundation of a modern DW are trending toward those based on a massively parallel processing (MPP) computing architecture instead of symmetrical multiprocessing (SMP). The 2013 TDWI big data management survey showed (for the first time in a survey) that more users manage big data with an MPP RDBMS than with an SMP one. SMP is still the preferred architecture for operational and transactional applications, but MPP has speed and scalability advantages for the large data set operations typical of big data analytics and data warehousing in general.

Hadoop has a strong track record for scaling to massive data volumes. HDFS clusters are known to scale out to hundreds of nodes that scale up to handle hundreds of terabytes of file-based data, all at a price that's a fraction of doing the same with an RDBMS.

Look for platforms that need little or no query optimization or data remodeling. Given the high performance of modern DWs and NoSQL platforms, there is little need now to remodel data or optimize queries for the sake of performance. Therefore, data can be left in its original schema, with all the rich details of source data. Performance is so good that many of the transformations previously done a priori in ETL are now done ad hoc in SQL-based queries and equivalent runtime processing (such as Java code running in MapReduce atop HDFS).

For scalability on demand, consider an elastic cloud. Many DW platforms, HDFS clusters, and analytic tools are now available on public clouds in a software-as-a-service licensing model. Start-up costs are low and time to use is short with a cloud-based DW. Plus, cloud-based DWs tend to be highly available. The leading benefit is that a cloud automatically provides and recovers server resources for optimal speed and scale as demanding data processing and analytic workloads ramp up and subside. Note that you needn't replace an on-premises DW wholesale with a cloud-based one to get performance benefits from a cloud. Instead, you can offload select DW and analytics workloads from on-premises to a cloud in a hybrid architectural model.

ABOUT OUR SPONSORS

www.microsoft.com

Microsoft is one of the leading providers in delivering the modern data warehouse of the future, and Microsoft has been recognized by third-party analysts like Gartner and TDWI as a leader in the data warehousing space. Beyond the ubiquitous SQL Server software technologies that most organizations leverage, Microsoft has added many enabling technologies to modernize the traditional, relational data warehouse to become the modern data warehouse. This includes technologies that realize big data and multi-structured data by offering Microsoft's Hadoop distribution, which brings the simplicity of Windows and the cloud to big data. It also includes technologies that enable in-memory performance and high scalability to handle petabytes of relational data and provide an integrated technology to query both Hadoop data and relational data in one SQL query. Finally, Microsoft delivers this experience via a traditional build-your-own deployment option, as a turnkey appliance, or as a cloud service. To find out more, go to www.microsoft.com/datawarehousing.

ABOUT THE AUTHOR

Philip Russom is director of TDWI Research for data management and oversees many of TDWI's research-oriented publications, services, and events. He is a well-known figure in data warehousing and business intelligence, having published over 500 research reports, magazine articles, opinion columns, speeches, Webinars, and more. Before joining TDWI in 2005, Russom was an industry analyst covering BI at Forrester Research and Giga Information Group. He also ran his own business as an independent industry analyst and BI consultant and was a contributing editor with leading IT magazines. Before that, Russom worked in technical and marketing positions for various database vendors. You can reach him at prussom@tdwi.org, @prussom on Twitter, and on LinkedIn at linkedin.com/in/philiprussom.

ABOUT TDWI RESEARCH

TDWI Research provides research and advice for business intelligence and data warehousing professionals worldwide. TDWI Research focuses exclusively on BI/DW 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 and data warehousing 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 THE TDWI CHECKLIST REPORT SERIES

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.