TDWI CHECKLIST REPORT

Big Data Analytics

By Wayne Eckerson

Sponsored by

aster data
more data. big insights.

tdwi
THE DATA WAREHOUSING INSTITUTE

AUGUST 2010

TDWI CHECKLIST REPORT

Big Data Analytics

By Wayne Eckerson

TABLE OF CONTENTS

1201 Monster Road SW, Suite 250
Renton, WA 98057

T 425.277.9126
F 425.687.2842
E info@tdwi.org

tdwi.org

FOREWORD
THE EVOLUTION OF BIG DATA ANALYTICS

There are two major trends causing organizations to rethink the way they approach doing analytics.

Big data. First, data volumes are exploding. More than a decade ago, I participated in the formation of the Data Warehouse Terabyte Club, which highlighted the few leading-edge organizations whose data warehouses had reached or exceeded a terabyte in size. Today, the notion of a terabyte club seems quaint, as many organizations have blasted through that threshold. In fact, it is now time to start a petabyte club, since a handful of companies, including Internet businesses, banks, and telecommunications companies, have publicly announced that their data warehouses will soon exceed a petabyte of data.

Deep analytics. Second, organizations want to perform "deep analytics" on these massive data warehouses. Deep analytics ranges from statistics—such as moving averages, correlations, and regressions—to complex functions such as graph analysis, market basket analysis, and tokenization. In addition, many organizations are embracing predictive analytics by using advanced machine learning algorithms, such as neural networks and decision trees, to anticipate behavior and events. Whereas in the past, organizations may have applied these types of analytics to a subset of data, today they want to analyze every transaction. The reason: profits.

For Internet companies, the goal is to gain insight into how people use their Web sites so they can enhance visitor experiences and provide advertisers with more granular targeted advertising. Telecommunications companies want to mine millions of call detail records to better predict customer churn and profitability. Retailers want to analyze detailed transactions to better understand customer shopping patterns, forecast demand, optimize merchandising, and increase the lift of their promotions.

In all cases, there is an opportunity to cut costs, increase revenues, and gain a competitive advantage. Few industries today are immune to the siren song of analyzing big data.

This TDWI Checklist Report is designed to provide a basic set of guidelines for implementing big data analytics. The analytical techniques and data management structures of the past no longer work in this new era of big data. This report will help you take the first steps toward achieving a lasting competitive edge with analytics.

NUMBER ONE
EDUCATE YOUR ORGANIZATION ABOUT THE DRIVERS OF BIG DATA ANALYTICS.

There are many reasons organizations are embracing big data analytics.

Data volumes. First, they are accumulating large volumes of data. According to TDWI Research, data warehouse data volumes are expanding rapidly. In 2009, 62% of organizations had less than 3 TB of data in their data warehouses. By 2012, 59% of those organizations estimate they will have more than 3 TB in their data warehouses and 34% said they would have more than 10 TB. (See Figure 1.)

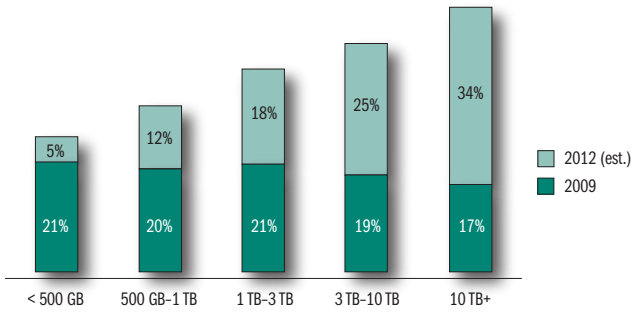

Figure 1. *Current and projected data warehousing data volumes, based on 417 respondents. Source: TDWI Research, 2009.*

Data volumes are expanding because the price-performance of data management systems is rapidly increasing. This enables organizations to collect more detailed transaction data, including Web clicks, point-of-sale records, and claims data. For example, a telecommunications company can now store call detail records for months or years instead of summarizing and archiving the records after a day or week. Analysts can use this historical detail to better understand traffic patterns and customer behavior, for instance.

In addition, there are new sources of data that organizations would like to bring into the analytical orbit, including social media data (e.g., blogs, tweets, online discussions), sensor data (e.g., RFID chips), GPS data, various devices or appliances (e.g., SmartMeters) that call home, and traditional unstructured data, such as audio, images, video, and text (e.g., e-mail, Web sites, documents).

Business value. Second, organizations see value in analyzing this detailed information, which encourages them to collect more data. The more data an organization collects, the more patterns and insights that it can mine for.

(Continued, next page)

(Continued)

For example, AT&T Mobility now calculates the profitability of its 80 million customers nightly, enabling the company to spot critical changes in customer behavior quickly and launch highly targeted marketing campaigns in response. "If you lost 1% of your customers yesterday, wouldn't you like to know today who they are and whether there might be a common theme behind the churn? We can now pinpoint those subscribers and work proactively to win them back," says a director of financial analysis at the company. "Ten years ago, we made assumptions on samples of data and based decisions on gut feel or someone's ability to argue an opinion. Now there is more precision."

Sustainable advantage. Finally, companies see big data analytics as one of the last frontiers of achieving competitive advantage. Analytics offers a sustainable advantage because it harnesses information and intelligence, things that are unique to each organization and cannot be commoditized. The popularity of the book *Competing on Analytics* by Tom Davenport and Jeanne Harris, which is targeted to business executives, has encouraged organizations to explore how to leverage analytics for competitive advantage.

NUMBER TWO

DETERMINE THE TYPE OF ANALYTICS YOU NEED.

There is not a lot of agreement about the definition of analytics. That's because any common term gets hijacked by vendors, consultants, and experts to further their own interests. Part of the confusion stems from the fact that there are two types of analytics.

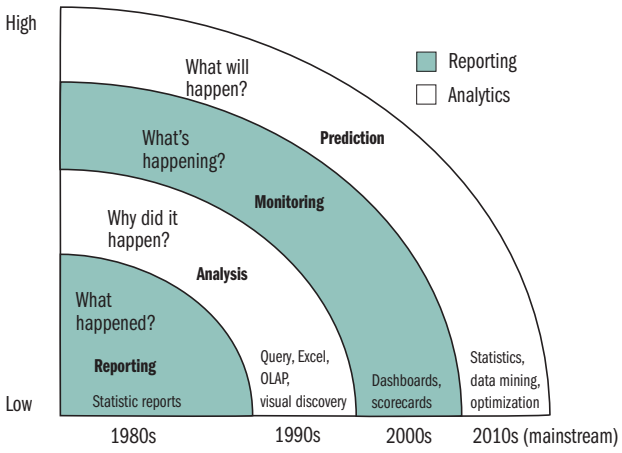

Figure 2. *Historically, there have been two waves of reporting, each followed by a wave of analytics. Reports are typically a jumping-off point for analytics.*

Exploration and analysis. One type of analytics is exploration and analysis. This approach involves navigating historical data in a top-down, deductive manner. To start, an analyst needs to have an idea of what is causing a high-level trend or alert. In other words, the analyst must start with a hypothesis and deduce the cause by exploring the data with ad hoc query tools, OLAP tools, Excel, or SQL. Here, the burden is on the business analyst to sort through large volumes of data and find the needle in the haystack. This type of analytics has been around for a long time and constitutes the bulk of activity done by business analysts.

Prediction and optimization. Another type of analytics is prediction and optimization. Although the algorithms used to power these types of analyses have existed for decades, they have been implemented only by a small number of commercial organizations. Business users model historical data in a bottom-up, inductive manner. They apply data mining tools to create statistical models to identify patterns and trends in the data that can be used to predict the future and optimize business processes. Here, the process is inductive. Rather than starting with a hypothesis, you let the tools discover the trends, patterns, and outliers on your behalf. (However, in reality, it takes some knowledge of the business process and data to apply these tools with reliable accuracy.)

NUMBER THREE

ARCHITECT FOR BIG DATA ANALYTICS.

A key question facing organizations that want to compete on analytics is how to architect for big data analytics.

Bottlenecks. Today, most companies create data warehouses to store and process data for reporting and analytics. Unfortunately, most data warehouses are already tapped out: they have reached maximum storage capacity without an expensive upgrade and can't support complex ad hoc queries without wreaking havoc on performance. In addition, the underlying data warehousing platform (database, server, storage, and network) isn't scalable enough to support new sources of data (internal or external) and maintain adequate query performance.

Meanwhile, analysts typically circumvent the data warehouse when performing their analysis. They use Excel, SAS, or ad hoc query tools to pull data directly from internal and external systems (providing they have access) and load it into a desktop spreadsheet or database, or perhaps a client/server analytical workbench such as SAS. Because of the network bottleneck and limited processing power of desktop or client/server systems, analysts typically analyze only a subset of the data. This approach also forces them to spend precious time cleaning, integrating, and preparing data for analysis—tasks that DW professionals are paid to perform.

Big data analytics architecture. To avoid these limitations, companies need to create a scalable architecture that supports big data analytics from the outset and utilizes existing skills and infrastructure where possible. To do this, many companies are implementing new, specialized analytical platforms designed to accelerate query performance when running complex functions against large volumes of data. Compared to traditional query processing systems, they are easier to install and manage, offering a better total cost of ownership and sometimes a cost as little as $10,000 per terabyte.

These systems come in a variety of flavors and sizes. There are data warehousing appliances, which are purpose-built, hardware-software solutions; massively parallel processing (MPP) databases running on commodity servers; columnar databases; and distributed file systems running MapReduce and other non-SQL types of data processing languages. Sometimes companies employ multiple types to address processing requirements. For instance, comScore, an online market research firm, uses Hadoop to acquire and transform Web log files and Aster Data's *n*Cluster database for analysis.

Today, companies typically implement these types of analytical platforms to store and process data that is not currently stored in the data warehouse (e.g., sensor data, Web logs, sentiment data) or offload analytical processing from overloaded data warehouses. In some cases, companies are using these platforms as enterprise data warehouses running multiple workloads, although this is the exception today. These platforms enable users to run queries against all the data they want without having to download a subset of data across the network to their desktops or a local environment. They also dramatically improve query performance and free companies from having to extract, prepare, and manage this data.

Where the analytic platform offloads queries from the data warehouse, it houses a subset of the data in the data warehouse and functions as a dependent data mart or analytics server. Here, the data warehouse team will push or replicate the subset of data to the analytic platform and keep the two systems in sync. However, it's likely that the DW and analytic platform manage two different data sets and don't require synchronization. The analytic platform will have its own data-loading and preparation processes that are tailored to the data set it manages. Although this requires the company to manage two analytical data platforms and processes, it improves overall efficiency by applying the optimal platform for the workload.

NUMBER FOUR

EMPLOY IN-DATABASE ANALYTICS.

As mentioned earlier, business analysts traditionally download data to desktops or local servers to explore, model, calculate, and score data using specialized analytical software, such as from SAS or SPSS. But moving data from source systems to a local environment and back again chews up a lot of networking resources and takes considerable time. As a result, analysts typically download a subset of data or create a sample from the larger data file. Although sampling is often a valid option, most modelers would prefer to work with all the data at a detailed level to optimize model accuracy.

Many companies are now rethinking traditional approaches to performing analytics. Instead of downloading data to local desktops or servers, they are running complex analytics in the database management system itself. This so-called "in-database analytics" minimizes or eliminates data movement, improves query performance, and optimizes model accuracy by enabling analytics to run against all data at a detailed level instead of against samples or summaries. This is particularly useful in the "explore" phase, when business analysts investigate data sets and prepare them for analytical processing, because now they can address all the data instead of a subset and leverage the processing power of a data center database to execute the transformations. It's also extremely beneficial in the "score" stage, when the model or function is applied to all incoming records. With in-database analytics, scoring can execute automatically as new records enter the database rather than in a clumsy two-step process that involves exporting new records to another server and importing and inserting the scores into the appropriate records.

To support in-database analytics, ensure that your database platform makes it easy to embed analytical functions and can run them in an MPP environment with minimal or no rework. Many analytic platform vendors are making a concerted effort to make it easy for administrators to insert custom functions into the database, and many now bundle a variety of predefined functions with their products. In addition, some analytical software vendors, such as SAS Institute, have begun to translate their core analytical functions into SQL extensions or user-defined functions that run natively in various relational databases and analytical platforms.

It's important to recognize that there are different approaches to supporting in-database analytics and not all are created equal. For instance, you should ensure that the approach you choose offers an integrated development environment for development and testing, supports multiple programming languages, provides an extensible API, and integrates with database facilities such as fault tolerance and workload management. Also, understand where the logic runs—inside the database or in a separate execution container—because the architecture of the analytic plug-ins has a major impact on the stability and reliability of the database management system.

Traditionally, IT professionals write stored procedures (SPs) or user-defined functions (UDFs) that run within the database engine to apply complex functions to data. Although database vendors offer varying levels of support for UDFs and SPs, they are often single-threaded functions that don't support MPP workloads and are difficult to write and maintain. If a memory leak or other error occurs, it can bring down a database or corrupt data records. Due to the delicacy of writing functions directly against a database, many UDFs are written by experienced database specialists using database-specific languages.

Fortunately, there are new techniques that promise to make it possible for business analysts, rather than IT professionals, to custom-code database functions that run in a parallel environment. For example, MapReduce was pioneered by Google to run custom analytics against its massive distributed computing network so that it can better understand Web site activity and user behavior. Aster Data, which is sponsoring this report, offers a version of MapReduce that runs in its *n*Cluster analytic database, supports multiple languages, and can be invoked via SQL. This makes it easier for developers to create custom functions that automatically run in a parallel environment without additional programming. And because MapReduce runs in a separate execution container, it offers better fault tolerance than previous methods of inserting custom functions into a database.

NUMBER FIVE

DON'T LIMIT ANALYTICS TO SQL.

Many analytic computations are recursive in nature, which requires multiple passes through the database. Such computations are difficult to write in SQL and expensive to run in a database management system. Thus, today most analysts first run SQL queries to create a data set, which they download to another platform, and then run a procedural program written in Java, C, or some other language against the data set. Next, they often load the results of their analysis back into the original database.

This two-step process is time-consuming, expensive, and frustrating. One architect says, "I do not like having to switch back and forth between Java and SQL. It is especially frustrating when business logic is spread across the languages. If I can have one language that does it all, it's a big win."

Fortunately, techniques like MapReduce make it possible for business analysts, rather than IT professionals, to custom-code database functions that run in a parallel environment. Aster Data has integrated MapReduce into its SQL-MapReduce framework, which lets analysts write reusable functions in almost any language they desire—Python, Java, C, C++, Perl—and invoke them with simple SQL calls. And out of the box, Aster Data bundles a number of predefined MapReduce functions in its *n*Cluster database management system, including market basket analysis, time-series analysis, sessionization, and various statistical functions. It also offers an API for customers and partners to create their own MapReduce functions and embed them in the database.

Clearly, as analytical tasks increase in complexity, developers will need to apply the appropriate tool for each task. No longer will SQL be the only hammer in a developer's arsenal. With embedded functions, new analytical databases will accelerate the development and deployment of complex analytics against big data.

NUMBER SIX

DEFINE YOUR REQUIREMENTS BEFORE SELECTING PRODUCTS.

Know your users. First, understand who's performing the analysis, the type, and how much data they require. If a power user wants to explore departmental data, then all they might need is an ad hoc query or OLAP tool and a data mart. If it's an IT person creating a complex standard report with sophisticated metrics and functions, then it's likely they can use a scalable BI tool running against an enterprise data warehouse. If a business analyst wants to run ad hoc queries or apply complex analytical functions against large volumes of detailed data without DBA assistance, then you probably need a specialized analytical database that supports analytical functions.

Performance and scalability. Second, understand your performance and scalability requirements. What query response times are required to make various types of analyses worth doing? If you have to wait days for a result set, then you either need to upgrade your existing data warehousing environment, offload these queries to a specialized analytical platform, or reduce the amount of data by aggregating data or reducing the time span of the analysis. Also, how many concurrent users do you need to support? Many BI tools and databases experience bottlenecks with memory or threading as the number of concurrent users increases. Adding another database platform adds complexity to an existing architecture, but it may be worth doing.

In-database analytics. Third, evaluate your need for in-database analytics. If complex models or analytics drive a critical portion of your business, then it's likely you can benefit from creating and scoring these models in the DW rather than a secondary system. If so, you'll need to evaluate the combination of your analytical software and DW database. If you are using an analytical package such as SAS, SPSS, or R, then understand which functions can run inside a database and how. Do you have to rewrite functions in SQL or C, or can they be converted automatically to a plug-in function in your database of choice? If you are hand-coding functions, then evaluate whether the database supports the programming languages your analysts prefer. Finally, make sure the database can run the analytical functions in parallel without extra programming.

Other. Finally, investigate whether the analytic database integrates with existing tools in your environment, such as ETL, scheduling, and BI tools. If you plan to use it as an enterprise data warehouse replacement, find out how well it supports mixed workloads, including tactical queries, strategic queries, and inserts, updates, and deletes. Also, find out whether the system meets your data center standards for encryption, security, monitoring, backup/restore, and disaster recovery. Most important, you want to know whether or to what degree you will need to rewrite any existing applications to run on the new system.

ABOUT OUR SPONSOR

aster data
more data. big insights.

Aster Data is a proven leader in big data management and big data analysis for data-driven applications. Aster Data's *n*Cluster is the first MPP data warehouse architecture that allows applications to be fully embedded within the database engine to enable ultra-fast, deep analysis of massive data sets.

Aster Data's unique "applications-within™" approach allows application logic to exist and execute with the data itself. Termed a "data-analytics server," Aster Data's solution effectively utilizes Aster's patent-pending SQL-MapReduce together with parallelized data processing and applications to address the big data challenge.

Companies using Aster Data include Coremetrics, MySpace, comScore, Akamai, Full Tilt Poker, and ShareThis. Aster Data is headquartered in San Carlos, CA, and is backed by Sequoia Capital, JAFCO Ventures, IVP, and Cambrian Ventures, as well as industry visionaries such as David Cheriton, Ron Conway, and Rajeev Motwani.

www.asterdata.com

ABOUT THE TDWI CHECKLIST REPORT SERIES

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

ABOUT THE AUTHOR

Wayne Eckerson is the director of TDWI Research. Eckerson is an industry analyst and educator who has covered DW and BI since 1995. Eckerson is the author of many in-depth, groundbreaking reports, a columnist for several business technology magazines, and a noted speaker and consultant. He is the author of *Performance Dashboards: Measuring, Monitoring, and Managing Your Business* (John Wiley & Sons, 2005) and the creator of TDWI's BI Maturity Model and Benchmarking Assessment service. He can be reached at weckerson@tdwi.org.

ABOUT TDWI RESEARCH

TDWI Research provides research and advice for BI professionals worldwide. TDWI Research focuses exclusively on BI/DW 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 warehousing 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.