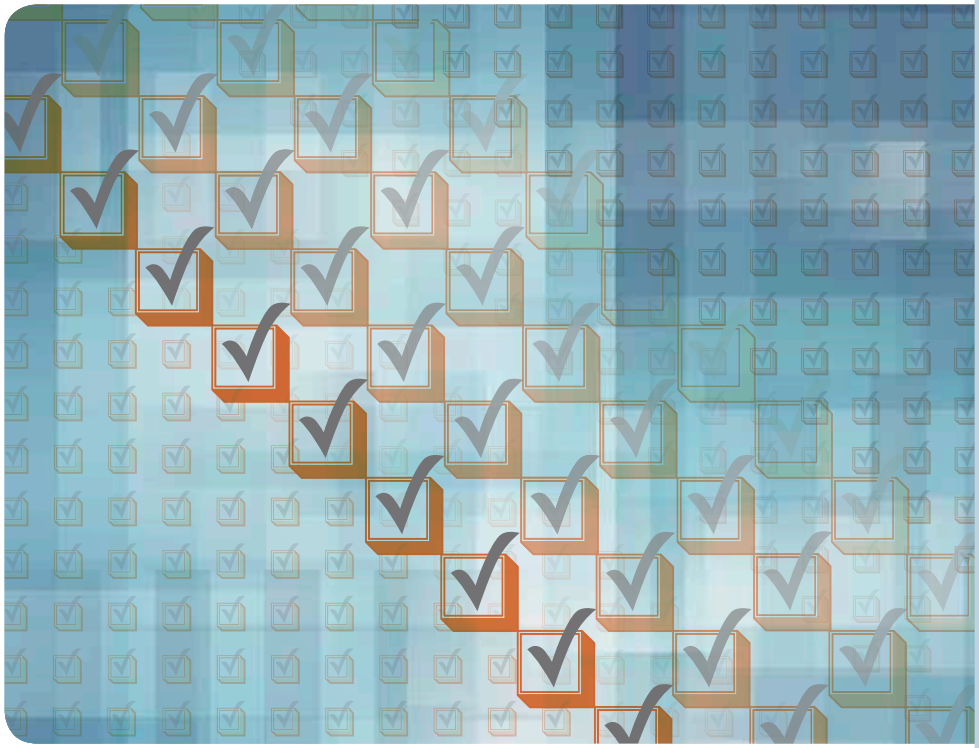
TDWI RESEARCH

TDWI CHECKLIST REPORT

Data Requirements for Advanced Analytics

By Philip Russom



www.tdwi.org

tdwi
THE DATA WAREHOUSING INSTITUTE

SEPTEMBER 2009

TDWI CHECKLIST REPORT

Data Requirements for Advanced Analytics

By Philip Russom

TABLE OF CONTENTS

1201 Monster Road SW, Suite 250
Renton, WA 98057

T 425.277.9126
F 425.687.2842
E info@tdwi.org

www.tdwi.org

FOREWORD

According to a recent survey from TDWI Research, 38% of organizations surveyed are practicing advanced analytics today, whereas 85% say they'll be practicing it within three years. Why such a dramatic upsurge in advanced analytics now? The use of advanced analytics is driven up by organizations' need to understand constantly changing business environments (as seen in the current recession and the resulting market turmoil), as well as to discover opportunities for cost reductions and new sales targets (which are key to surviving and thriving in a down economy). To meet these business goals, organizations are stepping up their use of two forms of advanced analytics: *query-based analytics* (which relies on complex SQL statements to define recent business events) and *predictive analytics* (which uses data mining and statistical methods to anticipate future events).

Organizations will face challenges as they move into advanced analytics. Many don't understand that reporting and analytics are different practices, often with different data requirements. Many have designed a data warehouse to fulfill the requirements of reporting and online analytic processing (OLAP), and they will soon need to expand the warehouse (or complement it with analytic databases) to fulfill the data requirements of advanced analytics, whether query-based or predictive. Being new to advanced analytics, they don't know what its various forms are, much less when to use which form. Most of these organizations are experienced in data integration, data quality, data modeling, and so on; yet, they don't know how to adjust these data management practices to fit the needs of advanced analytics.

This TDWI Checklist Report seeks to clear the confusion by listing and explaining data requirements that are unique to advanced analytics. The assumption is that it's hard for organizations to succeed with analytics when they haven't given it the right data in the right condition. Hence, this report focuses on the data requirements of advanced analytics so organizations may become better equipped to populate a data warehouse or analytic database with data and data models that ensure the success of advanced analytic applications.

NUMBER ONE

USE ADVANCED ANALYTICS TO DISCOVER RELATIONSHIPS AND ANTICIPATE THE FUTURE

There are many applications of advanced analytics, but most of them involve discovering relationships, anticipating the future, and adapting to change. Working with the right data in the right condition is key to achieving these goals.

Discover relationships. Whether advanced analytics is based on data mining, statistics, artificial intelligence, or complex queries, it can help you discover and quantify important relationships that you may have been unaware of. These relationships can reveal fraud, define customer segments, group products of affinity, and link field conditions that lead to product failures. The newly discovered relationships, in turn, help you reduce fraud and its costs, target marketing campaigns more accurately, develop effective merchandizing strategies, and improve product quality.

Anticipate the future. Predictive analytics can produce scores and statistics through which you can predict the likelihood of various outcomes of certain situations. For example, predictive models quantify a customer's proclivity to churn, thereby giving you an opportunity to retain the customer. Predictive models can assist with various types of forecasting. Likewise, predictive analytics can quantify future risk for pragmatic applications in actuarial tables or loan approvals.

Understand and adapt to change. On the one hand, advanced analytics can help you understand change in the form of rising costs or new customer behaviors. On the other hand, the discoveries made through analytics can lead to positive changes that help your business adapt to an evolving world.

These goals are worth pursuing from a business standpoint, but they require specialized analytic tools and analytic databases from a technology standpoint. This means that organizations new to advanced analytics will need to reach beyond their current reporting and data warehouse infrastructures.

NUMBER TWO

SCALE UP DATA INTEGRATION TO HANDLE LARGE ANALYTIC DATA VOLUMES

Many analytic databases regularly begin an analytic cycle with multiple terabytes. Hence, whether the data is heading into an EDW or a stand-alone analytic database, data loading must scale up to handle large data volumes that are loaded very quickly. Likewise, large data extracts from operational systems must be as non-intrusive as possible.

Sometimes advanced analytics involve a few large and sporadic data extracts and loads. In fact, a number of large table dumps and other broad extracts may occur one or more times, until the data set seems right to the business analysts. This is very different from the recurring nightly batch runs that are typical of ETL for EDWs. Even so, once analysts learn what they need from the ad hoc data sets they've collected, there may be a need to infuse some of this data into an EDW, updated with regular data integration jobs. Over time, analytics should help evolve and tune data collection strategies so that only the right data is collected.

Extract, transform, and load (ETL) continues to be the preferred method of data integration for data warehousing because of its unique ability with data transformations. Even so, with data sets for advanced analytics, it often makes sense to rearrange the acronym to ELT. This way, data is available for preliminary analysis immediately. Incremental transformations, when available, can be done as needed within the EDW or analytic database. When these databases are on an MPP platform that's purpose-built for data warehousing, the platform is better equipped from a hardware viewpoint than the average data integration server, and this makes ELT even more compelling.

NUMBER THREE

REALIZE THAT REPORTING AND ANALYTICS HAVE DIFFERENT PURPOSES AND NEEDS

Reporting and analytics are two different practices that have different goals, methods, sponsors, funding, and enabling technologies. Yet many people confuse the two, perhaps because most vendors' platforms for business intelligence (BI) include functions for various types of reporting and summarized analysis in the form of online analytic processing (OLAP).

By comparison, predictive analytics (which includes techniques for data mining and forecasting) is far more exploratory and forward-looking than reporting and OLAP. The value of predictive analytics is the discovery of unknown facts and relationships, the confirmation of known or suspected relationships, and the leverage of those relationships for better decision making. For example, predictive analytics can be used to develop new products and attract new customers, as well as to reduce cost, risk, and fraud. Predictive analytics achieves this through data-sampling best practices that ensure predictive models "generalize" to new data—future data that is yet to be collected—thereby giving the models broad usage over a long lifecycle.

Achieving these discovery-driven goals through reporting is unlikely, since most reports are based on a small amount of summarized information that's already well known and studied. Likewise, OLAP is usually implemented as a form of parameterized reporting, where the parameters represent dimensions. In such implementations, the available parameters limit the breadth of the analysis, and the analysis cannot be broadened without technical personnel developing more parameters. Table 1 summarizes the differences between reporting and analytics.

TABLE 1.

	Reporting and OLAP	Advanced Analytics, Both Query-Based and Predictive
Business Method	Performance management for business entities, relative to a business plan.	Develop new products, customers, etc. Reduce cost, risk, fraud.
Information Purpose	Update known facts. Quantify past performance.	Infer unknown facts and relationships. Quantify future probabilities.
Output	Historical standard reports, dashboards, metrics, KPIs, cubes for OLAP, etc.	Predictive models, scores, forecasts. Results of complex queries. Insights.
Queries	Known, simple queries that are easily optimized.	Queries that become very complex as they evolve via iteration.
Volume per Query	Small (usually less than a gigabyte).	Large (possibly terabytes).

NUMBER FOUR

DISTINGUISH BETWEEN DATA WAREHOUSES, DATA MARTS, AND ANALYTIC DATABASES

TDWI defines an enterprise data warehouse (EDW) as an all-encompassing data warehouse platform that manages data about many business departments and their functions (unlike single-subject data marts), supports multiple data processing workloads (not just reporting and OLAP), and manages data in multiple data structures and models (which includes low-level, detailed data). With this definition in mind, the EDW is an optimal environment that every organization should aspire to, because it can support on a single platform just about any business intelligence (BI) and advanced analytics practice that an organization may choose to adopt. And the EDW consolidates all BI data and advanced analytics data into a single source that ensures the "single version of the truth" that most organizations demand from BI and data warehousing.

Organizations that have deployed a true EDW can depend on it as an able platform for advanced analytics. After all, an EDW can handle both query-intense and predictive-scoring workloads, plus it can manage the low-level, detailed data that advanced analytics often requires. These workloads and data have special needs, so they are usually enabled by their own special tables or databases within the EDW, thereby creating an analytic database (or analytic sandbox) managed by the EDW and resident on its platform.

However, not all organizations have an EDW, as defined here. Some data warehouses focus on managing decision data about multiple business departments and supporting workloads for reporting and OLAP. They don't support analytic workloads or massive volumes of detailed data. The missing support may be because of the design that users created for the warehouse, or it may be caused by deficiencies in the vendor-built database management system or hardware architecture that the organization selected for the warehouse's platform. Either way, organizations with a warehouse focused on reporting and OLAP will need to extend or complement it with a separate analytic database to support an analytic workload and appropriate data—if they are to provide the right data in the right condition that advanced analytics requires.

Table 2 summarizes the differences between EDWs, data marts, and analytic databases.

TABLE 2

	Enterprise Data Warehouse	Data Mart	Analytic Database
Business Method	Single version of the truth for enterprise performance.	Single subject area(s) for application-specific purposes.	Test bed for exploring change and opportunity.
Optimization	Multiple update speeds, high performance, workload management, in-database analytics.	Regularly updated data for reporting, performance management, and OLAP.	Unpredictable data sets about changing markets, costs, customers, risks, etc.
Data Attributes	High standards for production data, plus inclusion of experimental data.	Carefully transformed, cleansed, modeled, and audited.	Less cleansed and modeled. Often just raw source data.
Data Models	3NF data model to model the enterprise with views for application flexibility.	Relational models for reporting. Multi-dimensional models for OLAP.	3NF of source data. Models demanded by analytic tools. Predictive models and scores.
Data Lifecycle	Permanent history with transient, elastic logical marts.	Permanent history of enterprise performance.	Data tends to be transient, as analytic needs change.
Data Acquisition	Well-governed process with the flexibility for self-provisioning elastic logical marts.	Slow process due to data transformation, cleansing, modeling, audit trail, etc.	Load data fast with little prep and start analysis immediately, regardless of state of data.

NUMBER FIVE

DESIGN A DATA WAREHOUSE ARCHITECTURE THAT ACCOMMODATES ANALYTICS

One of the most critical design and architecture decisions adopters of advanced analytics must make is whether to store analytic data in a data warehouse or in a stand-alone analytic database.

Analytics processed within the EDW. This is what most users tell TDWI they would like to do. The catch is that many of the analytic tools based on data mining technologies require users to dump analytic data into flat files with a specific record structure or a single denormalized table, because that's the data structure the tool is optimized for. In recent years, data mining and predictive analytic tools have gotten better at processing data while it's stored in a DBMS. This is usually called "in-database analytics." (Of course, SQL-based analytics demand that data be managed by a SQL-compliant DBMS.) The trend toward in-database analytics will undoubtedly continue, because most users would rather manage data with a data warehouse or similar database and leave the data in place when analyzing it.

In-database predictive analytics requires that the data warehouse platform support the scoring of predictive models deployed to it. This is typically enabled by user-defined functions (UDFs) and SQL that can run predictive analytics programs natively in the database management system.

Analytic sandboxes. Data warehouses that support in-database analytics should also support analytic sandboxes. Using workload management features, an IT administrator sets up an analytic database somewhere within the EDW environment and sources it with data requested by a user. The analytic user can then work within the sandbox without unpredictable performance hits on the EDW.

Analytic databases outside the EDW. This is a well-established practice, and it takes many forms. At one extreme, rogue data marts and spreadsheets proliferate outside the EDW until IT and DW teams are forced to rein them in through time-consuming data mart consolidation projects. At the other extreme, a new best practice is to selectively isolate disruptive analytic workloads on data marts and other analytic databases outside the EDW.

NUMBER SIX

PREPARE DATA TO MEET THE NEEDS OF THE ANALYTIC METHOD YOU'VE CHOSEN

Organizations choose different analytic methods to get the answers they need, including OLAP, query-based analytics, and predictive analytics. Each has its own general requirements for data preparation.

Online analytic processing (OLAP). OLAP's purpose is to quickly answer multi-dimensional queries of summarized data. For example, consider the three dimensions of the following question: What are (1) the sales figures for (2) the western sales region in (3) the fourth quarter? Most approaches to OLAP enable multi-dimensional queries by caching data in cubes. This provides good query performance, but limits queries to the data, dimensions, and simple summary statistics of one or more cubes. With the possible exception of relational OLAP, you can discover only what's already in a cube. Depending on the tool you're using, expanding the cube may require lengthy intervention by technical personnel.

Query-based analytics. Many organizations are choosing query-based analytic methods that depend heavily on structured query language (SQL). That's because they know and trust SQL, plus they can leverage the SQL-based tools and skills they already have. With this method, users typically gather large volumes (often multiple terabytes) of raw operational data and load it into a data warehouse or analytic database. The point is to gather data and start analyzing it as soon as possible in reaction to a sudden change in the business environment. Urgency doesn't allow time and resources for much (if any) data transformation and modeling. So, users make do with operational schema and non-cleansed data by expressing transformations and multidimensional models through complex SQL statements. This method works well when the SQL processing is executed on an MPP database platform that's built for complex queries.

Predictive analytics. The data mining and statistical algorithms of a predictive analytic tool typically demand a very specific data structure, typically denormalized. Some tools have multiple algorithms, each with a unique data requirement. Most algorithms are optimized to run fast and accurately with a flat record structure, so data flattening may be required. Note that each record may have hundreds of fields, and there may be millions of records. Many algorithms prefer range fields, so certain data values may need to be transformed into ranges in a process called binning. Furthermore, some algorithms can operate only on a flat file (as opposed to in-database processing), so generating a very large flat file is core to data preparation for many algorithms.

NUMBER SEVEN

PRESERVE ANALYTIC DATA'S RICH DETAILS, BECAUSE THEY ENABLE DISCOVERY

Analytic discovery depends on data nuggets. The kind of discovery-oriented analytic applications discussed here (both query-based and predictive) depend on large amounts of source data drawn directly from operational and transactional applications (though sometimes augmented with data from a data warehouse or other source). But it's not just the size of the data sets that matters. Even more important are the details within raw source data, because much of the clustering and relationship definitions produced by advanced analytics are based on those details. Hence, analytic discovery depends on data nuggets. Likewise, statistical accuracy depends on data's details being present. Strip these out during the data preparation stages, and analytic discovery is hamstrung from the beginning.

Analytic data can also be unstructured. Textual information can be a rich source of data for analytics. The catch is that text isn't necessarily useful in its original state. There is usually a need for text mining or text analytics tools and techniques that can discover useful facts in the text and convert these to structured data (typically a table row per discovered fact). Hence, text documents, Web pages, and text fields in databases can be useful to analytics, if converted accordingly. Note that the output of text mining or text analytics tools commonly feed into predictive analytics, thereby providing rich details that enhance the accuracy of predictive models.

Data from an EDW can be analytic, too, of course. Although advanced analytics tend to need new data sets, they can also tap the content of a data warehouse. This way, the data from the warehouse provides a historic context for newly discovered facts, plus additional dimensions and other useful details. And the insights of analytics should be incorporated into the historic record of the data warehouse. For example, an analytic study of customer behavior can reveal previously unknown customer attributes that should become metrics recorded in the data warehouse.

NUMBER EIGHT

IMPROVE DATA AFTER WORKING WITH IT, NOT BEFORE

The perils of improving analytic data. As analysts understand large data sets, they may develop some data integration or data quality routines to improve the data. Likewise, they may model some data structures that assist with their analyses or constitute tables and cubes that they'll retain after their analytic work with the current data set is done. Note that improvements to the data may occur only after business analysts have worked with the analytic data set. These tasks, as already pointed out, are risky if done too early, for fear of losing the data details that discovery-oriented analytics depends on.

Data quality for analytic databases. Advanced analytics have the potential to identify useful information from data that could be perceived as having poor quality. Data anomalies, missing data, non-standard values, and so on that would be inappropriate for reporting from a standard data warehouse may hold useful information that is accessible only through the use of advanced analytics. Therefore, advanced analytics should be used to assess outliers prior to the application of standard data quality routines. Likewise, filling in missing field values (through data enhancement) may mask useful information about the data that could be determined by advanced analytics. For example, name-and-address standardization can mask fraud rings. For these reasons, some anomalies should not be fixed; instead, the anomaly should be documented in metadata, so anyone using the data set knows of its presence and how to compensate for it.

Data modeling for analytic databases. One of the conundrums of analytics is that remodeling data can speed up queries and enable multidimensional views; yet, modeling can also lose data details and limit the scope of queries. Therefore, data sets intended for analytic discovery should be treated only to data modeling that is truly required and value-adding, as seen in the flattening and binning mentioned earlier for predictive analytics. Even so, after business analysts and other power users have worked with a data set for awhile, there comes a time when remodeling is appropriate to preparing data for use in the EDW, reports, the scoring of predictive models, and other post-analytic applications.

NUMBER NINE

APPLY THE PRODUCTS OF ADVANCED ANALYTICS TO BI AND DW ACTIVITIES

At some point, the early discovery phases of advanced analytics (which this Checklist Report focuses on) often lead to later phases where the analytics becomes part of daily business intelligence (BI) activities. For instance, a business analyst may mine a data set in an ad hoc manner to understand a new customer behavior, then develop predictive models that are scored on a recurring basis to anticipate the new behavior so it can be acted on appropriately. Similarly, data sets and query results developed via iterative ad hoc queries may become "institutionalized" in the EDW, so that they can be used by reports, metrics, dashboards, and OLAP queries. Advanced analytics has its own unique set of data requirements, necessitating different handling from standard EDW data.

Ultimately, the information and insights inferred from the application of advanced analytics should be applied to standard EDW data to enhance BI activities. Standard reports and ad hoc queries are necessary to run an organization, but not sufficient to propel the organization into excellence. Advanced analytics enables continuous learning and improvement as it makes the most of organizations' data assets.

ABOUT OUR SPONSORS

Netezza appliances have revolutionized and simplified analytics for companies struggling to find the processing speed and power to analyze and understand their growing data. Today, with hundreds of customers including 24/7 Real Media, Guitar Center, Nationwide, Neiman Marcus, Ryder System, Inc., Virgin Media, Yum Brands and others, Netezza (NYSE: NZ) is a proven solution to the rising costs and complexity of data warehousing and analytics. Our global partners, including Accenture, Business Objects, Cognos, IBM, Informatica, MicroStrategy, SAS and others, and an expansive list of system integrator, reseller and developer partners worldwide, means our customers can rest assured that Netezza will live comfortably within any existing infrastructure they have in place.

SAS is the leader in business analytics software and services, and the largest independent vendor in the business intelligence market. Through innovative solutions delivered within an integrated framework, SAS offers unparalleled data integration and data quality, is the market leader in predictive analytics and provides sophisticated reporting. SAS helps customers at more than 45,000 sites improve performance and deliver value by making better decisions faster. Since 1976 SAS has been giving customers around the world The Power to Know®.

Teradata Corporation is the acknowledged global leader in data warehouse innovation and analytical solution development. Every day we raise our customers' intelligence to higher levels, making them more focused and competitive by gathering enterprise information and extracting actionable insight. Teradata elevates enterprise intelligence by giving every decision maker the insight required for smarter, faster decisions. We add value and reveal opportunity across more dimensions than any competing solution. In every industry and geography, our technologies and expertise make the difference. Simply put, Teradata solutions make companies smarter and give them the competitive advantage to win.

ABOUT THE AUTHOR

Philip Russom is the senior manager of TDWI Research at The Data Warehousing Institute (TDWI), where he oversees many of TDWI's research-oriented publications, services, and events. He's been an industry analyst at Forrester Research, Giga Information Group, and Hurwitz Group, where he researched, wrote, spoke, and consulted about BI issues. Before that, Russom worked in technical and marketing positions for various database vendors.

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.

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.