

Data Analytics— A Practical Approach

Abstract

More than ever, the world is one of data and rules—business, human and environmental. The use of daily, weekly and monthly analytical monitoring to determine patterns in business data is common to identify what we are doing well, determine how we can do it better and recognize problems before they result in material damage. Data analytics (DA) can be as simple as finding duplicate payments in accounts payable or evaluating sales patterns to determine the best location for a warehouse. DA can also be as complicated as identifying statistical outliers for potential fraudulent activity. DA can use a single spreadsheet or extract data from multiple platforms and formats. For some people, there is a joy in learning the secrets that data hold. This white paper was written to illuminate what DA can offer an enterprise.

DATA ANALYTICS—A PRACTICAL APPROACH

ISACA®

With 95,000 constituents in 160 countries, ISACA (www.isaca.org) is a leading global provider of knowledge, certifications, community, advocacy and education on information systems (IS) assurance and security, enterprise governance and management of IT, and IT-related risk and compliance. Founded in 1969, the nonprofit, independent ISACA hosts international conferences, publishes the *ISACA® Journal*, and develops international IS auditing and control standards, which help its constituents ensure trust in, and value from, information systems. It also advances and attests IT skills and knowledge through the globally respected Certified Information Systems Auditor® (CISA®), Certified Information Security Manager® (CISM®), Certified in the Governance of Enterprise IT® (CGEIT®) and Certified in Risk and Information Systems Control™ (CRISC™) designations. ISACA continually updates COBIT®, which helps IT professionals and enterprise leaders fulfill their IT governance and management responsibilities, particularly in the areas of assurance, security, risk and control, and deliver value to the business.

Disclaimer

ISACA has designed and created *Data Analytics—A Practical Approach* (the “Work”) primarily as an educational resource for security, governance and assurance professionals. ISACA makes no claim that use of any of the Work will assure a successful outcome. The Work should not be considered inclusive of all proper information, procedures and tests or exclusive of other information, procedures and tests that are reasonably directed to obtaining the same results. In determining the propriety of any specific information, procedure or test, security, governance and assurance professionals should apply their own professional judgment to the specific control circumstances presented by the particular systems or information technology environment.

Reservation of Rights

© 2011 ISACA. All rights reserved. No part of this publication may be used, copied, reproduced, modified, distributed, displayed, stored in a retrieval system or transmitted in any form by any means (electronic, mechanical, photocopying, recording or otherwise) without the prior written authorization of ISACA. Reproduction and use of all or portions of this publication are permitted solely for academic, internal and noncommercial use and for consulting/advisory engagements, and must include full attribution of the material’s source. No other right or permission is granted with respect to this work.

ISACA

3701 Algonquin Road, Suite 1010
Rolling Meadows, IL 60008 USA
Phone: +1.847.253.1545
Fax: +1.847.253.1443
E-mail: info@isaca.org
Web site: www.isaca.org

Data Analytics—A Practical Approach

CRISC is a trademark/service mark of ISACA. The mark has been applied for or registered in countries throughout the world.

Acknowledgments

ISACA wishes to recognize:

Project Development Team

Duffie Brunson, KPMG, USA
Scott Gilbert, SLG Enterprises, USA
Pam Kammermeier, CISA, Federal Deposit Insurance Corp., USA
Parm Lalli, CISA, ACDA, Sunera LLC, Canada
Rich Lanza, CFE, CPA, Cash Recovery Partners LLC, USA
Anthony P. Noble, CISA, CCP, Viacom Inc., USA
Thomas Steeves, CISA, ACDA, Canada
Stephen Valance, CISA, CPA, Movado Group Inc., USA

Expert Reviewers

Steven Lacoursiere, CISA, ACDA, USA
Phillip J. Lageschulte, CGEIT, CPA, KPMG LLP, USA
Anthony P. Noble, CISA, CCP, Viacom Inc., USA
Owen B. Rockentine, CISA, Comerica Inc., USA

ISACA Board of Directors

Emil D'Angelo, CISA, CISM, Bank of Tokyo-Mitsubishi UFJ Ltd., USA, International President
Christos K. Dimitriadis, Ph.D., CISA, CISM, INTRALOT S.A., Greece, Vice President
Ria Lucas, CISA, CGEIT, Telstra Corp. Ltd., Australia, Vice President
Hitoshi Ota, CISA, CISM, CGEIT, CIA, Mizuho Corporate Bank Ltd., Japan, Vice President
Jose Angel Pena Ibarra, CGEIT, Alintec S.A., Mexico, Vice President
Robert E. Stroud, CGEIT, CA Technologies, USA, Vice President
Kenneth L. Vander Wal, CISA, CPA, Ernst & Young LLP (retired), USA, Vice President
Rolf M. von Roessing, CISA, CISM, CGEIT, Forfa AG, Germany, Vice President
Lynn C. Lawton, CISA, FBCS CITP, FCA, FIIA, KPMG Ltd., Russian Federation, Past International President
Everett C. Johnson Jr., CPA, Deloitte & Touche LLP (retired), USA, Past International President
Gregory T. Grocholski, CISA, The Dow Chemical Co., USA, Director
Tony Hayes, CGEIT, AFCHSE, CHE, FACS, FCPA, FIIA, Queensland Government, Australia, Director
Howard Nicholson, CISA, CGEIT, CRISC, City of Salisbury, Australia, Director
Jeff Spivey, CRISC, CPP, PSP, Security Risk Management, USA, ITGI Trustee

Knowledge Board

Gregory T. Grocholski, CISA, The Dow Chemical Co., USA, Chair
Michael Berardi Jr., CISA, CGEIT, Nestle USA, USA
John Ho Chi, CISA, CISM, CBCP, CFE, Ernst & Young LLP, Singapore
Jose Angel Pena Ibarra, CGEIT, Alintec S.A., Mexico
Jo Stewart-Rattray, CISA, CISM, CGEIT, CSEPS, RSM Bird Cameron, Australia
Jon Singleton, CISA, FCA, Auditor General of Manitoba (retired), Canada
Patrick Stachtchenko, CISA, CGEIT, CA, Stachtchenko & Associates SAS, France
Kenneth L. Vander Wal, CISA, CPA, Ernst & Young LLP (retired), USA

Acknowledgments (*cont.*)

Guidance and Practices Committee

Kenneth L. Vander Wal, CISA, CPA, Ernst & Young LLP (retired), USA, Chair
Kamal N. Dave, CISA, CISM, CGEIT, Hewlett-Packard, USA
Urs Fischer, CISA, CRISC, CIA, CPA (Swiss), Switzerland
Ramses Gallego, CISM, CGEIT, CISSP, Entel IT Consulting, Spain
Phillip J. Lageschulte, CGEIT, CPA, KPMG LLP, USA
Ravi Muthukrishnan, CISA, CISM, FCA, ISCA, Capco IT Service India Pvt. Ltd., India
Anthony P. Noble, CISA, CCP, Viacom Inc., USA
Salomon Rico, CISA, CISM, CGEIT, Deloitte, Mexico
Frank Van Der Zwaag, CISA, Westpac New Zealand, New Zealand

ISACA and IT Governance Institute® (ITGI®) Affiliates and Sponsors

American Institute of Certified Public Accountants
ASIS International
The Center for Internet Security
Commonwealth Association for Corporate Governance Inc.
FIDA Inform
Information Security Forum
Institute of Management Accountants Inc.
ISACA chapters
ITGI Japan
Norwich University
Solvay Brussels School of Economics and Management
Strategic Technology Management Institute (STMI) of the National University of Singapore
University of Antwerp Management School
ASI System Integration
Hewlett-Packard
IBM
SOAProjects Inc.
Symantec Corp.
TruArx Inc.

What Is Data Analytics?

Data analytics (DA) involves processes and activities designed to obtain and evaluate data to extract useful information. The results of DA may be used to identify areas of key risk, fraud, errors or misuse; improve business efficiencies; verify process effectiveness; and influence business decisions. There are many issues to consider when starting a new DA program, including maximizing the return on investment (ROI), complying with project budgets, managing false positives, and ensuring the protection and confidentiality of the source data and results.

The results of DA may be used to identify areas of key risk, fraud, errors or misuse; improve business efficiencies; verify process effectiveness; or influence business decisions.

A variety of software tools ranging greatly in cost and features are used to perform DA. Planning the approach to DA and the execution techniques are, however, usually more important success factors than the features of the tools themselves. Enterprises go through a natural progression as their DA programs mature, beginning with the purchase of tools that provide *ad hoc* analysis capabilities. Over time, *ad hoc* techniques typically evolve to more automated and repeatable processes, which then increases the effectiveness of the analysis (**figure 1**).

Ad Hoc Data Analytics

Ad hoc DA is a one-use process—a starting point that may be used to help identify patterns or potential risk areas within a business system. It is typically used for an initial investigation as a way to begin to understand the business processes while becoming familiar with the data.

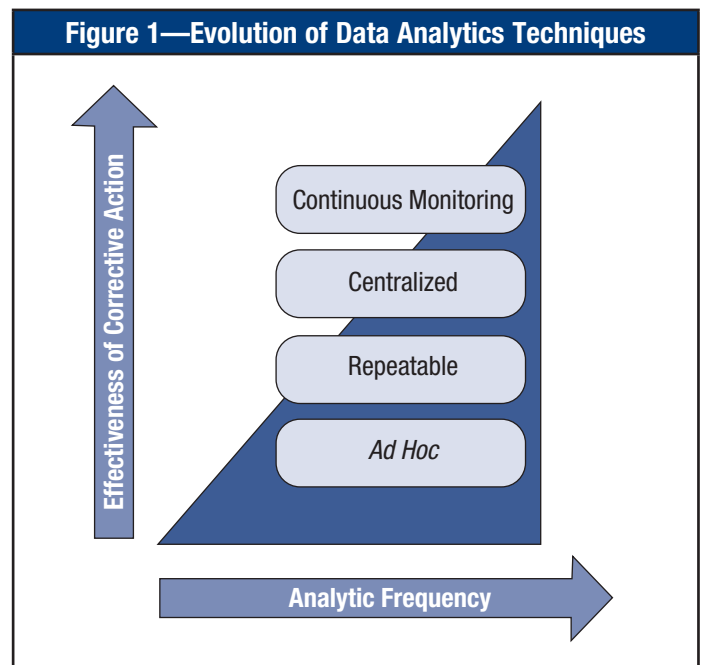
The effort to develop and document *ad hoc* DA is time well spent because that information may be used going forward to automate the process for repeatable results.

Ad hoc DA, which can be outsourced or performed in-house, is typically run to support specific projects. The data may be supplied by IT or obtained by the data analyst. Significant time may be spent acquiring, importing and verifying data, so it is useful to retain the data knowledge for later projects.

Ad hoc DA is rarely performed on production systems, but it may include production data, data warehouses and historic files.

Ad hoc DA may be difficult to repeat if the steps performed are not well documented or if there are a number of complicated steps to perform. The documentation can serve as a starting point to move up the maturity scale for all repeatable DA and continuous monitoring (CM).

Typical *ad hoc* DA may begin with exploratory data mining, benchmarking/trending or data quality testing, but it may also include specific management-directed requests from any business area or function. The results of *ad hoc* DA vary from the increased knowledge of systems and processes to system improvements and cost savings.



Repeatable Data Analytics

The use of *ad hoc* DA often requires an enterprise to rely on selected skilled individuals to perform the testing; such reliance may increase risk, due to attrition and the lack of knowledge transfer. In addition, enterprises that have seen the benefits of DA tend to begin to look for ways to improve the efficiency and increase the frequency of running analytics. For these reasons, *ad hoc* DA inevitably leads to the next maturity stage: repeatable DA.

Repeatable DA is predefined and scripted; it is designed to perform the same tests on similar data (e.g., data from a different time period) on a scheduled basis. The benefits include consistency, efficiency and more effective corrective actions. Analytics are often stored on the personal workstations of analysts or analytic “librarians.” Source data files can still be supplied by IT, but data access tools such as Open Database Connectivity (ODBC) may also be used to import data directly from production systems.

Capturing the logic within program scripts moves the necessary analytics knowledge from individuals into the DA tools themselves. These program scripts become members of logic libraries that can be run repeatedly, used for training purposes or used as the basis for new DA projects. The quality of analysis is improved and remains consistent from run to run, as the data acquisition process is partially or fully automated.

Centralized Analytics

The next step up the maturity scale from repeatable DA is a centralized approach for the development, storage and operation of repeatable DA. In this approach, a central repository is established for repeatable DA programs and standard data files; standards for DA development are documented and DA applications are set up and scheduled to run against the centralized data on a regular basis or on demand.

Data to be analyzed may either be pushed to the repository or extracted directly from different sources as needed, and the analytic results themselves are stored in the repository. This centralized approach has the following advantages:

- The process is more consistent, efficient and repeatable.
- The results are more reliable and consistent.
- The chance of multiple variations being scattered across individual machines is reduced.
- The potential negative impact on the performance of production applications is minimized.
- Data security is improved.
- Backups are more easily performed.
- The use of a centralized repository increases the overall performance of workstations.
- Access to analytics and results is available to more people, increasing productivity and improving the use of supporting reference materials, analytic sample logic and source data.

Continuous Monitoring

CM marks the highest point on the maturity scale. At this stage, analytics are fully automated and running at regularly scheduled intervals and may be embedded directly into a production system. A continuous run of analytics enables the immediate identification of potential exception transactions. Many commercially available CM packages include sophisticated web interfaces, e-mail notifications, workflows, remediation tracking, dashboards and/or heat maps.

A continuous run of analytics enables the immediate identification of potential exception transactions.

Because it is primarily management’s responsibility to assess control effectiveness, CM is often developed and owned by operations management. The benefits to the enterprise include improved efficiency, reduced errors and timely identification of problems. CM is often similar to the analytics used by auditors when testing internal controls.

Embedded CM modules are preventive solutions in which preconfigured rules are established to identify violations of the segregation of duties (SoD) or monitor high-risk transactions in real time. They can also act as advanced management tools that allow test runs of the workflow. The process involves three steps:

1. Identify transactions that match predefined criteria.
2. Copy the transactions to a data file.
3. Alert business unit managers, auditors or other stakeholders of transactions matching the criteria.

This approach is often achieved using Structured Query Language (SQL) commands against the production database and stored database procedures, including triggers to monitor accounting transactions. When the predefined criterion is met, an alert is triggered, and the transaction is held until it is approved by the person who was alerted. Embedded CM modules are particularly suitable for stable environments where there is a need for CM. They are particularly useful for monitoring and reporting on irregularities in the internal control environment and can be used to identify potentially fraudulent transactions.

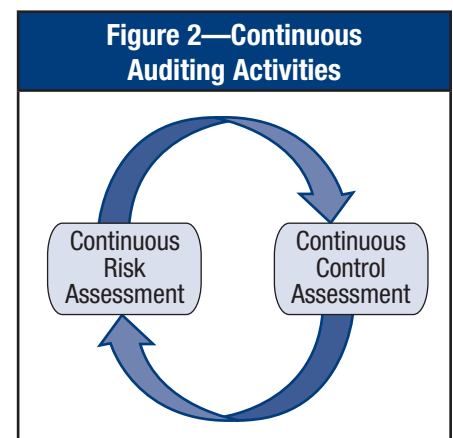
Independent CM modules are detective solutions. Preconfigured rules are set up to run against data extracted after a transaction has been processed. The rules consist of data extraction from the production system to another server/computer at regular intervals and the performance of predefined data tests. This solution allows for the use of multiple rules to reduce the likelihood of false positives (where something is flagged for review, but turns out to be a valid transaction) and results can be sent to business unit managers, auditors or other stakeholders after each run.

This approach can help reduce the ongoing cost of regulatory compliance by developing modules to meet specific regulations. This can help avoid the costs associated with spending time to manually select a sample and then verifying that the required controls are working effectively. The advantage of independent CM is that data can be brought into the process from multiple production systems. However, the tests are run after a transaction has been processed. Independent CM solutions are typically used in environments where there is a need for a solution outside the existing financial or other system.

Security incident event monitoring (SIEM) is a form of CM that focuses on monitoring security incident event logging, network activity, firewall events and system access. Examples of the type of analytics that could be performed in support of SIEM are identifying application setup and master data changes to monitor or track changes made to access rights.

Continuous auditing¹ (CA) is the process of performing audit-related activities in a continuous manner (**figure 2**). These activities can include continuous risk and control assessments in which automated analytics are an integral part of the solution. A direct relationship exists between the level of CM performed and the effort required for CA. If CM is effectively used by management, the effort required by auditors for control testing is reduced.

A direct relationship exists between the level of CM performed and the effort required for CA.



Benefits and Limitations of Different Data Analytics Approaches

As previously mentioned, leveraging DA within an enterprise can provide many significant benefits, including greater insight, improved assurance over the quality of data, and the potential for increased staff productivity. When properly executed, DA is more of a journey than a single project. Most enterprises that set out on the DA journey jump in too

¹ Coderre, Dave; *GTAG Continuous Auditing: Implications for Assurance, Monitoring and Assessment*, The Institute of Internal Auditors, USA, November 2005

far and too fast, resulting in failure, which makes it difficult to recalibrate and rejoin the path. The key to mastering the challenges and capturing the benefits is to enter into DA with a well-thought-out plan.

Like all business plans, the DA plan should focus on an end result. DA should start small and build on successes. Success seems to grow quickly for enterprises that start with the following steps:

- Perform *ad hoc* analyses in support of targeted risk areas.
- Leverage DA within many projects for greater insight.
- Move to repeatable analyses performed periodically on high-risk activities.
- Define the measures of success along the way.

EXAMPLE

A regional commercial bank was regrouping after the recent financial meltdown. As part of its overall business plan, the bank focused on understanding its credit exposure in commercial real estate—by geographic area and industry.

In partnership with the IT department, the business unit identified the correct data tables from the source data file, pulled the data into a spreadsheet via basic SQL queries and started analysis.

The pilot was successful and the plan was expanded to leverage two reporting tools owned by the bank: one for drill-down reporting and the other for statistical analysis. Training for business unit staff was the only out-of-pocket expense.

With the tools available and training provided, the business unit was able to gradually expand the volume of data and increase the sophistication of its analyses. This allowed the business unit to shift from periodic analyses to regular, monthly (continuous) analyses. The keys in this example are the gradual nature of the plan, the use of existing investments, and the working relationship between IT and the business unit.

This example seems to flow effortlessly from a basic idea to an effective tool supporting the business. In truth, there were many issues and lessons along the way that had to be addressed. The following sections describe a few of the lessons that tend to be common to many enterprises moving in this direction.

Data Quality

Data quality is a significant issue. The pursuit of high-quality data can quickly become overwhelming, expensive and time-consuming. Because there is no such thing as perfect data, the concept of “fit for use” is important. Obtaining “fit for use” data may involve balancing data in production reports to ensure completeness; performing validation checks to identify duplicates, outliers or orphan records; and certifying the data as being “fit for use.” Shifting from the quest for “perfect” data to “fit for use” data adds a dose of practicality and reduces related time, cost and effort. It is important to note that what constitutes “good enough” data is always closely related to the specific use or purpose of the data.

Shifting from the quest for “perfect” data to “fit for use” data adds a dose of practicality and reduces related time, cost and effort.

Data Volume

It is said that in one of the hundreds of thousands of source data repositories of the largest retailer in the United States (US) is a table with nine billion rows in it—and the table is updated nightly. It holds every transaction from every store for every day. It is undoubtedly a treasure trove of information, albeit absolutely overwhelming. It is a prime example

of how the extensive scope of data available for analysis can lead to “shiny object syndrome:” the desire to capture and analyze data from all over.

Such exceedingly large data volumes cannot be analyzed with ordinary assessment methods, such as sampling or simple spreadsheets. Commercial DA tools can manage unlimited volumes of data and analyze entire data populations (with 100 percent testing).

EXAMPLE

A large credit card company faced an issue regarding the authorization of gasoline sales at convenience stores. The authorization rate charged at the gasoline pump was higher than the rate for merchandise sold inside the store. Consequently, some merchants reset their terminals to make gasoline sales look like merchandise sold in the store.

This practice represented a significant financial loss over the millions of transactions occurring every hour across the US. By developing a program that scanned every transaction for authorization and analyzing only the gasoline sales (enabled by embedded information), the data volume—while huge—became manageable.

Note: The program also identified transactions with incorrect rates, reset the rates, approved the corrected transactions and sent a message to the fraud department. The merchant was then informed so that appropriate action could be taken to correct the rates.

Project Budgets

Every project manager has had to come to grips with documenting a project plan and allocating hours (budget) to its completion. When data analysis is added to the planning stream, the risk of over- or underestimating the time required can be significant. In most instances, the major drain on time is finding, accessing and understanding the data—separate exercises that can take as much as 70 percent of the total project hours if problems occur at any phase. The next drain is performing multiple iterations of analyses and reports (e.g., to alter the appearance of results, such as charts).

Building on the earlier recommendation to develop a well-thought-out plan, a plan that includes the appropriate analyses to confirm or deny the hypothesis or accurately portray the events in question will pay big dividends in budgeting time and costs. Another huge time saver can be working with IT to understand and identify the data required before the first data request is filed.

Another huge time saver can be working with IT to understand and identify the data required before the first data request is filed.

User Proficiency and Knowledge Transfer

Whether part of a process improvement project or a CA/CM implementation, DA requires specialized skill sets among team members. Over the years, there has been an increase in computer, analytic and data skills in both college and advanced-degree graduates. In parallel, the fast-paced evolution of software makes more and more complex tasks just a click away. The combination of more technically skilled employees and solid (but basic) software tools makes it easier to launch effective and successful DA programs.

Root Cause Analysis

Root cause analysis is an important component of understanding risk and DA test results. The results of root cause analysis can be derived from multiple iterations of the analysis. The key risk is false positives, which may result in wrong overall conclusions. To mitigate against this and develop strong capabilities in root cause analysis, an in-depth understanding of the business—not just the data—is required. Collaboration between the business and IT, as has been

suggested in this paper, strengthens the results from root cause analysis. Specific root cause analysis techniques are discussed in ISACA's *Monitoring Internal Control Systems and IT*. They include the “five whys” and fishbone chart approaches.²

Strategies for Maximizing Return on Investment

For decades, surveys have consistently reported that most business software projects fail to meet business objectives. They take too long, cost too much and focus on the wrong problems. In addition to the old adage of “those who fail to plan, plan to fail,” businesses also tend to have expensive “shelfware” that no one uses. Before investing more in DA, there are five rules to follow:

1. **Do simple process development first, using existing software.** This step is the one most routinely skipped and leads to failed projects. Do not buy a DA software package and expect to do a first-rate analysis with it straight out of the box. Develop some clear scenarios for waste or fraud at the enterprise and then work with existing reporting tools (or even spreadsheets) to determine how to find duplicates, flag unusual items, etc. Work to understand these problems prior to selecting and investing in additional software.
2. **Automate data extraction and validation.** In many projects, up to 70 percent of the time allotted to the entire project can be spent getting the data into a usable format and understanding how the data relate to the business process. Often, the scope of the project is determined not by the questions that need to be asked, but by the data that have been successfully imported into the reporting tool. There is no shortcut here. Involve the various stakeholders, especially IT, early on in the process to make sure that the results sought will matter to senior management. Map out the data needed and plan on eventually using automated data extraction routines wherever possible. Investment in the right software here will save many hours. Once findings are available, report the successes to the stakeholders frequently and confirm the mandate to do more.
3. **Reduce false positives.** No matter how sophisticated the individual tests may be for identifying duplicates, rounded numbers or items entered on a weekend, the tests can often produce hundreds or thousands of “red flag” items, which can often include false positives. Yet, most projects limit sampling to a small number of items, such as from 30 to 50 items. With those odds, the project may turn up no more than one actual overpayment or error. The return on investment (ROI) from such activity is bound to be low. Such a low ROI can be improved by:
 - Combining multiple red flags into an overall score and pursuing only the highest-scoring transactions or vendors
 - Focusing on vendors with especially high ratios of red flags to dollars or red flags to transactions
4. **Prioritize by likelihood of recovery.** Apart from meeting compliance requirements, enterprises should reconsider the focus of their DA efforts. The priority should not necessarily be the largest expense area, the one least frequently reviewed, or the one suspected of having the most errors or fraud associated with it. Instead, it should be the one that will quickly yield a positive return. The problem area may be travel spending, freight, overstated revenues or inconsistencies in the general ledger. Unless the enterprise is already doing regular accounts payable recovery reviews, there is typically money that can be saved if data analysis is applied. This money can be used to fund future software developments. The goal is to get an easy win early on to create a positive attitude from the start and help make funds available for more costly and risky software development later.
5. **Refine and document the testing process over several cycles.** Once the pilot DA project has produced a successful outcome, repeat it at an appropriate interval. If the first DA project involved a specialist, then it should be repeated internally using the same agreed-on procedures before making large investments in software.

² ISACA, *Monitoring Internal Control Systems and IT: A Primer for Business Executives, Managers and Auditors on How to Embrace and Advance Best Practices*, USA, 2010, www.isaca.org/bookstore

Risk, Security and Privacy Concerns

The major concern when working with production data is privacy and/or confidentiality. A number of privacy-related laws—different in each geographic location—must be considered when performing data analysis. The responsibility for applying appropriate data privacy and security techniques falls on the individual mining the data and performing the analysis, so it is important to determine and document how data privacy and security issues will be addressed prior to performing any analysis. The following points should be considered:

- What is the goal of the data analysis project?
- How will the data be used? Anyone who will access and possibly hold data during the analysis must also adhere to appropriate data governance standards.
- Who will be able to access, review and analyze the data? It is important to be aware of the roles and responsibilities of users and user groups within an enterprise who may have access to personal information.
- How will the data be secured to prevent unauthorized access?
- How will the data be updated?

Prior to performing data analysis, the data gathering steps can reveal information that could also present privacy and security issues. For example, when working with healthcare, banking or human resources data, the following techniques should be followed:

- Always protect the enterprise by requiring signed nondisclosure or confidentiality agreements to help provide some protection against data exposure. Data analysis often calls for files to be passed to third-party sources, where it may be difficult to monitor data security.
- Start the DA program with something less sensitive—perhaps general ledger or accounts payable data. If mistakes are made, it is possible to learn from these mistakes before branching into more sensitive areas, such as payroll.
- Use data encryption techniques whenever possible because this will help protect the data. It should be noted, however, that applying these techniques will slow down the analysis itself.
- Keep sensitive files with very restricted access on a secured drive.
- Sanitize sensitive data so that personal information is not easily recognized: mask sensitive fields such as the tax identification number (TIN), the social security number (SSN) or the pay rate. Or use unique hash values to protect sensitive data from disclosure.
- Avoid sending Microsoft Excel files as attachments to e-mails; instead, consider using a network to share or hold the files—which can be accessed via a link within a notification e-mail message.
- If files are sent by e-mail, the e-mail message should be deleted from the sender's sent folder and from the recipient's in-box after the files have been downloaded and saved.

Prior to performing data analysis, the data gathering steps can reveal information that could also present privacy and security issues.

Assurance Considerations

In the past, assurance departments were limited to demonstrating benefits to their enterprises by verifying financial/operational goals or evaluating compliance with control guidance/regulations. By using DA, assurance departments may establish their value by providing assessments of potential cost savings and operational improvements identified through assurance activities.

By using DA, assurance departments may establish their value by providing assessments of potential cost savings and operational improvements identified through assurance activities.

With available data from production systems, DA enables analysis of significant processes from a cross-departmental perspective and individual task steps from a more detailed viewpoint. By using the same risk-based approach used to determine the audit schedule, potential target areas for DA may be identified in the following areas:

- Determine the operational effectiveness of the current control environment.
- Determine the effectiveness of antifraud procedures and controls.
- Identify business process errors.
- Identify business process improvements and inefficiencies in the control environment.
- Identify exceptions or unusual business rules.
- Identify fraud.
- Identify areas where poor data quality exists.

Quantitatively measuring process improvement through data analysis allows an assurance department to provide the enterprise with an internal consulting service. Documenting a summary of value-add contributions throughout the year and comparing it over a period of years may be helpful in reporting to the audit committee on an annual basis.

Strategy and Governance

When developing DA projects, it is important to measure the cost-benefit ratio, especially when starting a project without an existing query library. Long-term benefits must be evaluated slightly differently because the long-term benefits related to DA are based on each individual project's cost-benefit analysis, combined with continuous improvement benefits from repeating the same project or the ease in performing new and unrelated DA projects.

EXAMPLE

Multiple prior attempts to review an enterprise's compliance with investment policies utilized limited data extraction and provided vague results, including poor analysis of the findings. This was resolved by going back to the business owners, clearly defining the scope and objectives for the next investment compliance review, as well as documenting the data requirements. Results from this refined analysis were immediately improved. By building on the initial success as well as documenting the source code, within 18 months the analysis increased from sampling three days per month to 100 percent compliance testing for each period under review.

Each DA project requires specific objectives to determine project success and, therefore, project benefits. Success can be based on financial or nonfinancial criteria. Examples might include the identification of wasted business costs, lack of adherence to policy, total work-hours saved, fees/fines reduced or not incurred, and more accurate reporting of transaction rework not performed. All of these benefits can potentially be achieved via clearly defined objectives for DA projects.

EXAMPLE

A fraud examination may measure success in the form of both monetary and nonmonetary benefits:

- **Monetary**—Recovery of the fraudulently obtained assets
- **Nonmonetary**—Successful conviction of the perpetrator of the fraud

The cost involved to achieve success is generally quantified in monetary terms and includes project hours, software purchased and consultation fees.

Maintenance

DA programs often start out with a single DA project that focuses on a specific area of need with the purpose of clarifying reporting. Once this initial project is successful and management buy-in has been achieved, the scope is expanded to performing more extensive DA in the initial focus area or to applying DA in other areas. Enabling the transition from a one-by-one DA approach to the implementation of a DA program requires a strong DA methodology.

This methodology should leverage existing project and program management methodologies and should address specific DA needs, such as the access to data as well as documentation standards. Documentation standards will define coding standards, administration information and notes built into the code, the program-specific objectives, key contact personnel, areas of anomalies and any information that would be useful for anyone taking over the project.

The time to develop and update the documentation should be built into the project. Most key stakeholders are interested in only the final report or the action items generated based on the deliverables of the project, so these should be documented carefully.

Well-written documentation depends on having staff with the skills and training to develop the documentation. It is important to develop such expertise within the department or group; it facilitates passing on that knowledge as people move to other tasks, and ensures that the ability is not concentrated in only one person. It is also useful to ensure that information on the point of contact for obtaining data is known to a wider group so that it is not lost if one person leaves.

Conclusion

Few would argue that an enterprise's data are among its most valuable assets. Yet, without a way to obtain, cleanse, organize and evaluate the data, the enterprise is left with a vast, chaotic pool of ones and zeroes. DA coaxes order from the chaos. It helps explain patterns, which in turn help the enterprise identify what it is doing well, determine how to do it better and recognize problems before they spiral out of control. DA can be relatively simple, but it can also be extraordinarily complex. Its results can be used to identify areas of key risk, fraud, errors or misuse; improve business efficiencies; verify process effectiveness; and even influence business decisions.

When used effectively, however, DA can play an integral role in helping an enterprise unlock the treasures hidden in its massive stores of data.

As with any process, maximum benefits cannot be achieved for the enterprise if: DA is not aligned to the business, risk is not managed, or the process is not effectively planned, designed, implemented, tested and governed. When used effectively, however, DA can play an integral role in helping an enterprise unlock the treasures hidden in its massive stores of data.

Additional Resources and Feedback

Visit www.isaca.org/data-analytics for additional resources and use the feedback function to provide your comments and suggestions on this document. Your feedback is a very important element in the development of ISACA guidance for its constituents and is greatly appreciated.

Appendix A. Emerging Methodologies Within Data Analytics

Once *ad hoc* DA methods have been employed and the user understands the basic business rules within the data, more sophisticated statistical techniques can be used to uncover more complex business rules and identify, for further review, transactions that do not follow these rules.

One development methodology in common use is the Cross-Industry Standard Process for Data Mining (CRISP-DM) reference model. Data must be appropriately prepared for analysis and constitute a pool that is large enough to contain patterns, but not so large that analysis bogs down in the volume. For each technique, various parameters can be altered to adjust the number of transactions flagged and determine the number of business rules generated. It can take several iterations of adjusting these parameters before useful data can be extracted from the analysis.

Data mining is the analysis of large data sets by a computer program to identify patterns (business rules) that exist within the data. This information is then used to flag records that have an unlikely probability of matching those rules. Data mining can be used by the business for root cause analysis and to identify exceptions in existing data for correction purposes. Organizations can use the analysis to validate business rules, examine data quality and identify outlying transactions for follow up. The value of using scatter-graphs and other kinds of visual analysis to present data mining results cannot be overstated. Because the relationships under study are generally abstract and involve multiple dimensions, they are difficult to visualize in two dimensions. Numerous tools are available today that allow users to present the data along multiple axes and then drill down into outliers to determine if they deserve further review.

Predictive DA involves the analysis of large data sets for the purpose of predicting future activity patterns based on past transactions. Usually this analysis is used to predict the value of a field (such as an amount) for a transaction, which is then compared to the actual value entered to ensure that it is within acceptable limits based on the other fields entered. This analysis can be performed by management in real time to flag transactions for additional review or by assurance staff to validate the predictability of recent transactions based on the rules discovered by the tool in past transactions.

Special attention should be given to transactional scoring as a way of increasing the predictive power of the tests. Many business rules may exist for identifying fields in transactions that are higher in risk. Transaction scoring can be used to reduce a sample for testing by assigning a value to each field to generate an overall score for risk based on the business rules.

Fuzzy logic matching of data helps to identify potentially fraudulent transactions or duplicate records for correctional purposes. It employs fuzzy logic to compare data values selected and flags similar records found, such as “Lime Street” and “Lime St.” Tools can compare the records within one data set to identify potential duplicate records or compare selected fields across two data sets to identify whether the same data are contained in both data sets. The tool parameters can be adjusted to avoid identifying too many false duplicates. Fuzzy matches are the ones most often missed by the existing control setup within production systems, and they tend to have a higher probability of being legitimate exceptions.

Appendix B. Top 10 Types of Analytic Logic

Duplicate transactions	<ul style="list-style-type: none"> • Exact duplicates—All fields are identical within a date range. • Fuzzy duplicates—Some fields are identical, with at least one or more fields that are similar or different.
Data quality	<ul style="list-style-type: none"> • Fields where key data elements are missing or invalid are identified. • Date ranges fall outside of normal values. • There are sequence gaps in key fields, such as the check or payment number.
Transaction limits	<ul style="list-style-type: none"> • Single and multiple accumulated values exceed limits. • Transaction amounts exceed, or are just below, the authorization limit.
File matching	<ul style="list-style-type: none"> • There is a two- or three-way match between related transactions. • Unmatched—Orphaned records occur between related files.
Character pattern matching	<ul style="list-style-type: none"> • Prohibited key words • Prohibited vendors/employees—Percent of names matched against a list of restricted names: <ul style="list-style-type: none"> – Matched to the US Office of Foreign Assets Control’s Specially Designated Nationals (OFAC SDN) list to identify terrorists: www.treasury.gov/resource-center/sanctions/SDN-List/Pages/default.aspx – Matched to the General Services Administration Excluded Parties List System (GSA EPLS) to identify parties that are excluded from receiving federal contracts: https://www.epls.gov/epls/search.do – Matched to the Office of Inspector General List of Excluded Individuals/Entities (OIG LEIE) to identify individuals and organizations blocked for federally funded healthcare providers: www.oig.hhs.gov/fraud/exclusions/exclusions_list.asp • Phonetic string match—The phonetic name is matched against the list of restricted names: SOUNDSLIKE algorithm of the New York State Identification and Intelligence System (NYSIIS) Code: www.dropby.com/NYSIIS.html • Fuzzy address match—A portion of address values are matched against the list of restricted addresses.
Segregation of duties (SoD)	<ul style="list-style-type: none"> • Performed at the security table level to identify potential conflicts • Performed at the transaction level to identify violations that occurred
Aging	<ul style="list-style-type: none"> • Single record age (number of days between Create Date and Approval Date) • Multiple files aging (Invoice Create Date prior to PO Create Date)
Numeric pattern matching	<ul style="list-style-type: none"> • Benford analysis—Transaction amounts fail to follow expected digital frequencies. • Numeric sequence or gaps—Sequences of check numbers • Frequent transactions have even dollar amounts.
Date/time matching	<ul style="list-style-type: none"> • Transaction dates occur on a weekend or holiday. • Transactions occur at odd hours.
Variance tests	<ul style="list-style-type: none"> • Comparison of the number of and amount of variances to a yearly average: <ul style="list-style-type: none"> – Is there a product price variance spike? – Is there an excessive spike in vendor invoice counts?