Making Everything Easier!™

Teradata Special Edition

Data Discovery FOR DUMMIES®

A Wiley Brand

Learn to:

- Understand what Data Discovery is and how it helps you cope with growing data volume and complexity
- Identify platforms and architecture that make big data analytics & discovery easier

Compliments of

TERADATA®

Meta S. Brown

Dear Reader

The importance and influence of analytics continues to grow. Its breadth and depth is staggering compared to even a few short years ago. You can expect that this trend will continue, if not accelerate, in the coming years.

With so many business problems, so much data, and so many analytic options, finding an analytics solution can be overwhelming. Nevertheless, organizations must find ways to more easily, rapidly, and flexibly discover and deploy new high-value analytic processes. Luckily, combining a modern discovery platform and a data discovery methodology can make this happen.

This book provides an overview of what data discovery is and how to apply it within your organization. It also includes steps to help you succeed while avoiding common pitfalls. Every organization needs to start building a data discovery competency now — this book is a terrific starting point.

Teradata has worked for decades with the biggest and most sophisticated companies in the world. Those same companies have huge amounts of data and vast analytic requirements. What we've learned over the years, we've put back into our products and services. It is this experience that led to the creation of the Teradata Aster discovery platform and our associated analytic services.

What this book outlines isn't just what we suggest that our clients do when it comes to data discovery. It is also the approach we follow when we work with clients. We've seen the methods outlined in this book succeed across industries and around the globe. Now we're excited to be able to share what we've learned with you.

We do hope you enjoy this book and that you learn some tips that you can apply right away in your organization. We're here to help if you need it. Now go and find the next business issue that you can solve by applying the principles of data discovery!

Bill Franks
Chief Analytics Officer, Teradata
Author of *Taming The Big Data Tidal Wave* and *The Analytics Revolution*

Data Discovery For Dummies

A Wiley Brand

Teradata Special Edition

by Meta S. Brown

For Dummies®
A Wiley Brand

Data Discovery For Dummies®, Teradata Special Edition

Published by
John Wiley & Sons, Inc.
111 River St.
Hoboken, NJ 07030-5774
www.wiley.com

LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: THE PUBLISHER AND THE AUTHOR MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION WARRANTIES OF FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES OR PROMOTIONAL MATERIALS. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR EVERY SITUATION. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING LEGAL, ACCOUNTING, OR OTHER PROFESSIONAL SERVICES. IF PROFESSIONAL ASSISTANCE IS REQUIRED, THE SERVICES OF A COMPETENT PROFESSIONAL PERSON SHOULD BE SOUGHT. NEITHER THE PUBLISHER NOR THE AUTHOR SHALL BE LIABLE FOR DAMAGES ARISING HEREFROM. THE FACT THAT AN ORGANIZATION OR WEBSITE IS REFERRED TO IN THIS WORK AS A CITATION AND/OR A POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE AUTHOR OR THE PUBLISHER ENDORSES THE INFORMATION THE ORGANIZATION OR WEBSITE MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. FURTHER, READERS SHOULD BE AWARE THAT INTERNET WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ.

For general information on our other products and services, or how to create a custom *For Dummies* book for your business or organization, please contact our Business Development Department in the U.S. at 877-409-4177, contact info@dummies.biz, or visit www.wiley.com/go/custompub. For information about licensing the *For Dummies* brand for products or services, contact BrandedRights&Licenses@Wiley.com.

ISBN 978-1-118-91272-0 (pbk); ISBN 978-1-118-91303-1 (ebk)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Publisher's Acknowledgments

Some of the people who helped bring this book to market include the following:

Project Editor: Jennifer Bingham

Acquisitions Editor: Kyle Looper

Editorial Manager: Rev Mengle

Business Development Representative: Melody Layne

Custom Publishing Project Specialist: Michael Sullivan

Project Coordinator: Melissa Cossell

Table of Contents

Introduction

Computer technology and the Internet are the driving forces behind a massive and growing body of data in electronic form. Expanding data resources present new opportunities for learning and improving business processes, but also considerable challenges for data management and analysis.

Distributed computing challenges and the push for more and better analytics can make the prospect of dealing with big data seem overwhelming. You might be tempted to skip over it altogether, but that's not in your company's best interest. In the years to come, big data management and facilitating analytics will be expected core competencies for data professionals.

Data Discovery to the rescue! Built from the ground up to smooth integration of large and complex data sources, simplify data analysis, and make powerful analytics accessible, Data Discovery is your survival kit for the big data territory.

About This Book

This book gives you the scoop on Data Discovery: Get a handle on what it does and how it makes dealing with big data simpler. Here are a few things you can expect:

- Find out how a Data Discovery platform opens the door to loads of useful analytics, what they do, and how they're used (no formulas and no math included in this book!).
- Pick up pointers on fostering discussion and information sharing to help everybody in your organization get more involved with data-driven decision making.
- Be forewarned about pitfalls that could bog down your Data Discovery program — and find advice to help you sidestep problems.

This book was written with and for Teradata.

Icons Used in This Book

These icons alert you to information that is particularly important or handy.

TIP

The Tip icon denotes a handy hint. Look here for ideas for making your Data Discovery experience easier.

REMEMBER

The Remember icon denotes information that is particularly important, and especially things that are often overlooked or forgotten.

Beyond the Book

The Teradata Aster Discovery Platform gives you easy access to sophisticated big data analytics and discovery. Learn more about Teradata's Aster platform for Data Discovery at `www.teradata.com`.

And check out an animation that shows some practical applications of using Aster at `www.teradata.com/NextGenAnalytics`.

Chapter 1

Living Up to New Business Expectations through Data Discovery

In This Chapter

- Recognizing changing expectations
- Adding to your core competencies
- Making sure there's something in it for everyone

There's talk in the air, talk about analytics, data-driven decision making, and big data. Everybody stands to benefit by leveraging data resources. Yet it takes effort and know-how to make that happen. It also takes time, and everybody is busy already.

Don't worry: Something new is on the horizon. You can live up to rising expectations and get behind bigger and better initiatives while using the skills and the staff you already have. Data Discovery to the rescue! This chapter discusses how things are changing and how you can adapt.

Facilitating Big Data Analytics

The growing body of electronic data is bringing many organizations into the world of big data. *Big data,* which simply means you're working with very high volumes of data, is now a huge buzzword, as you probably know. Not everyone agrees on how much data is required for something to be considered big data, but something in the range of 100 Terabytes to 1 Petabyte is a reasonable starting point. (This is a moving target. Although this estimate makes sense for 2014, the threshold will surely rise with each passing year.) And volume isn't the only issue. Traditional, structured data is only a part of what's being collected and analyzed today. Text, audio, video, and sensor data are increasing in importance.

Add "dealing with big data" to your list of new expectations for core competencies. And understand that the data won't be kept in cold storage. It will be used routinely, analyzed, and mined for information that management can use to make the business more efficient and competitive. That means you'll not only need to see that data is retained and secured, but also that it is easily available to those who need it, and you'll have to facilitate the growing use of analytics.

Today's IT and analytics processes often involve many manual steps — far too many. And each time that a task is done manually, an opportunity for making errors is created. The labor alone is costly, yet the cost associated with making and propagating errors is far greater. When you hear that the ability to manage big data and facilitate use of analytics will be a required core competency, don't expect to address that expectation by developing lots of code and ramping up skills in new programming languages. You'll need a different approach, one that minimizes the need for specialty skills and dramatically reduces opportunities for introducing errors into business processes.

So, new expectations for core competencies will include:

- Integrating data from disparate systems
- Giving more people more access to data while maintaining proper controls
- Managing big data
- Encouraging use of new and sophisticated analytics
- Reducing need for manual coding
- Simplifying processes
- Rapidly iterating through the Data Discovery process, failing fast, and discovering more

You may think it is impossible. It isn't. But you won't be able to meet these dramatic new expectations for core competency by doing things the way you've always done them. As the business world changes, you must change too.

Addressing Challenges through Data Discovery

Data Discovery enables you to expand your core competencies and live up to demanding new expectations. It's centered in an integrated platform for data management and analysis. The platform operates rapidly even with extremely large data sets. It incorporates a full range of capabilities — from data access to analytics. Between these ends of the spectrum lie a data store and intermediate services, such as optimization and analytics execution, that are vital to the platform's function, yet largely invisible to the end user.

Think about this: Chances are that your computer has an integrated suite of office applications for everyday needs, such as word processing, scheduling, email, and basic computation (in other words, a spreadsheet) installed right now. Each application has a user interface that enables you to think about the task at hand, such as writing and sending a message or scheduling a meeting and inviting participants, and not to think about computer programming.

REMEMBER

Data Discovery does for big data analytics what the integrated office applications suite does for those common tasks. It enables users to concentrate on data analysis and its significance to the business, instead of programming and other underlying tasks.

The following sections discuss issues that hold most organizations back from using big data analytics, and explain the role of Data Discovery in addressing them.

Lack of capacity for data and computing intensive tasks

Transactional systems, such as those used for sales, customer service, or technical support, are often stretched to the limits of available capacity, even with each user accessing only one or just a few records at any given time. Even very simple analytics require many data records to operate, and some demand considerable computing resources to execute calculations. Running an ordinary report has been known to slow transactional systems to a standstill.

A key element of Data Discovery is putting resources in place — hardware and software — reserved specifically for exploratory analytics. This prevents interference with everyday operations while allowing analysis to be conducted at any time.

Limited availability of specialized programming talent

Typical big data analysis methods require a great deal of programming, and depend on special datacentric programming tools such as Pig and MapReduce. Programmers who have strong skills with such tools are few, and they're in high demand.

The structure and simplified interface of a Data Discovery platform significantly reduces the amount of programming needed to perform data analysis. What's more, it replaces arcane languages with easier or more common ones, such as SQL.

Resistance to adopting new business methods

People hesitate to change the way they do business. They may be pushed to their limits with current obligations and find it difficult to try new things, even if they believe that change could benefit them in the long run. They may be suspicious and doubt that change is in their interest. They may lack the skills needed to change. They may fear failure.

Data Discovery can't eliminate everything that makes people reluctant to change, but it can tremendously reduce the effort and risk involved.

Data Discovery enables you to expand your core competencies to facilitate big data computing, yet you and the data users throughout your organization will still be able to concentrate on business and not get caught up in computing minutia.

Use case: Reducing customer support call volume in telecommunications

Most businesses need a call center. A phone call and a conversation with a real, live person are often a customer's preferred way to make a purchase, resolve a problem, or get information.

On one hand, customer support calls present opportunities to please customers by resolving issues and building customer satisfaction and loyalty. On the other hand, providing live customer support is costly, and a call to customer support often reflects a problem that should have been prevented.

How important is it to identify and correct process flaws that leave customers in need of support? First, consider the cost of providing live customer support. Cost per minute, average call length, and first-call resolution rates vary by industry, type of call, and the individual call center. Costs-per-call range from $2 to $4 for simple inquiries to $10, $20, or more for technically complex issues. Longer calls or more calls imply higher costs, but not greater customer satisfaction.

Call center costs aren't the only issue. For each customer who calls to resolve a problem, there are others who are dissatisfied, but don't call. Dissatisfied customers may leave. They complain to others. When they don't call, their problems may worsen, and when the problems are finally addressed, the costs to fix them may be much higher.

In the telecommunications industry, competition is cutthroat. Many customers change providers frequently to get new phones or lower rates. Pricing pressure and equipment costs mean that managing support expenses is absolutely necessary to keeping a company profitable. Because individual customers often spend less than $100 per month for services, the cost of even a single support call cuts significantly into returns.

One telecommunications company used Data Discovery to bring together records from differing systems such as online payment history and customer support interactions — including unstructured data as audio recordings and text. This was all data that the company collected in the routine course of business, but it had been splintered into different forms and locations before.

Through Data Discovery, the company added new analysis methods to its mix, such as text analysis, which makes it possible to extract information from written text without requiring a human being to read it all, and path analysis, which is used to identify frequently occurring sequences of events. Path analysis is useful for troubleshooting processes because it can reveal patterns of unexpected

(continued)

(continued)

or undesired behavior, which may be traced to preventable causes such as website design issues.

One pattern that drew attention in this case was that customers frequently called customer support immediately after making an online payment. Perhaps those were cases where the customer had not been able to complete the transaction? No! The payments were completed properly. However, they were often last-minute or late payments.

Investigating the specifics of the follow-up calls revealed an issue. Customers were calling to verify that their service would not be interrupted. Because the callers were already paying online, all that was needed was a simple message at the end of the payment process to verify that service would not be interrupted.

Companies like this often have millions of customers and handle millions of support calls each year. This single discovery opened the door to millions of dollars saved for the telecommunications company each year.

Getting Familiar with Data Discovery Technology

An integrated Data Discovery platform (which is one of the best solutions for Data Discovery) will have three major parts: the data store, the processing framework, and analytic engines. Like the parts of an office applications suite, the parts of the Data Discovery platform are built from the ground up to work smoothly together.

Separating Data Discovery from operations

You already have a data warehouse. Why would you need yet another place to put data?

Your data warehouse was planned with multiple purposes in mind. Supporting daily operations has top priority. Maintaining information for legal and other archival purposes might be next. And of course the data is available for analysis. However, there may be some issues.

Even simple reporting can be rather data intensive. You may already divert reporting to quiet hours to avoid interference with operations. Analytics such as behavior analysis, statistical analysis, and machine learning are data intensive and also processor and memory intensive. You can't very well ask the new Chief Data Officer to put all the data scientists on the night shift, can you? So you'll need to put data in a place where it can be used for analysis without disturbing everyday business. That's where the Data Discovery platform comes in play. The Data Discovery platform provides an integrated solution to analyze structured and unstructured data using newer analytic techniques like MapReduce and graph.

But the Data Discovery platform isn't just a safe playground for the analytics team. Big data analytics don't mix well with the average relational database. The right data store will be *multitype*. That is, it will have the ability to contain all types of data, including unstructured data. The primary form for unstructured data today is text. Expect video and audio to rise in importance in the years to come.

Take a look at Figures 1-1 and 1-2 to see how diverse sources can be united with a Data Discovery platform.

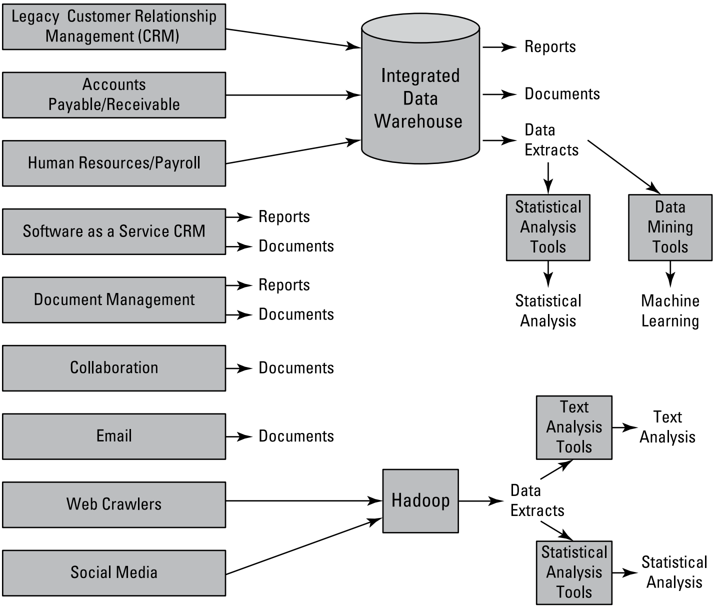


Figure 1-1: Typical enterprise data landscape.

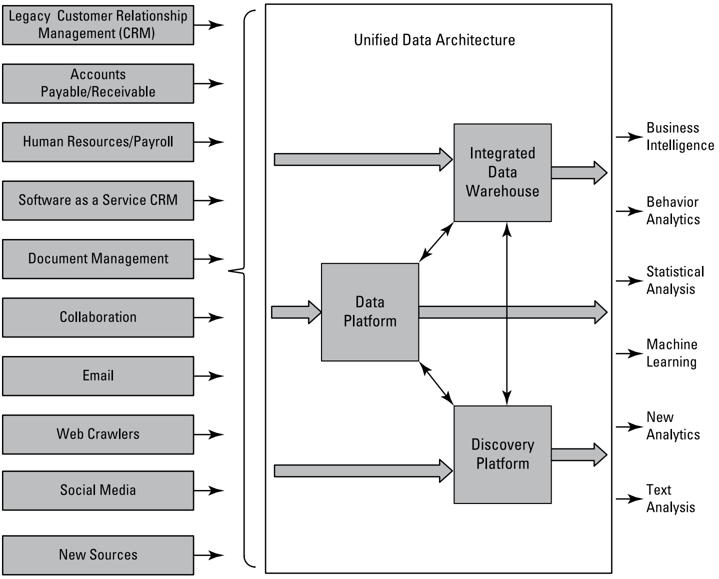


Figure 1-2: Next-generation analytic architecture.

In order to complete analysis quickly with extremely large quantities of data, the Data Discovery platform must have massively parallel processing capability. That is, it must be able to distribute work across many processors. Individual processors each take on different portions of the work and all operate at the same time — virtual teamwork. And it must have appropriate architecture — not simply to store data, but also to facilitate processing. *Shared-nothing architecture* (in which individual nodes are independent and don't share memory or disk storage) is the right choice.

Structure that makes analysis simpler and more efficient

Most conspicuous of the processing framework's elements is the *interface,* a means for users to interact with the Data Discovery platform. An effective interface must allow the user sufficient control to specify the exact range of data to be used and analysis methods to be applied without the complexity of elaborate programming.

TIP

To make data discovery pervasively available within the organization, complex coding languages such as Java, MapReduce, and Python aren't good options. Yet virtually all statisticians and data scientists have some programming skills, as do many (though certainly not all) other types of data analysts. So a simple or familiar code interface is the most-used interface for Data Discovery. For example, Teradata Aster's SNAP framework leverages widespread familiarity with SQL and enables users to invoke analytics via SQL statements.

You probably know that the way a query is written can have a dramatic effect on how long it takes to execute. So, there might be two or more ways to write a query that end up with the exact same result, but one would take much longer than the other to produce that result. The same is true of analytic formulas — there may be different versions that suggest different steps, but amount to the same thing in the end. The optimization process converts your code to the equivalent that is most efficient, while the execution process carries out those orders.

Teradata's SNAP framework features an integrated optimizer and executor for efficient query processing across multiple analytic engines and data stores.

Broad and deep analytics options

Analytics engines power specific analytic capabilities. The methods available to you vary with the platform and options that you select. At a minimum, most analytics teams expect access to a range of classical statistical analysis methods, as well as popular machine learning techniques for classification and market basket analysis. Text is a predominant form for big data, so expect text analysis to be important as well. (Read more about analytics in Chapter 3.)

For example, Teradata Aster Discovery Platform features a SQL, MapReduce, and Graph analytic engine to enable multiple analytics like path, pattern, text, statistical, and machine learning.

Confronting the Vs

Big data — there certainly is a lot of it. But quantity alone isn't proof of value. There are times when all you need is a little data to make that magic, and that's fine. Some business challenges, though, absolutely demand dealing with data on a grand scale.

When you have a big problem, ask big questions like these:

- How is this problem affecting the business — are we wasting time, materials, or cash?
- What are the losses associated with the problem?
- What information would help us to discover causes and potential solutions?
- What data is most relevant for our needs?
- What is the best way for us to get useful information from that data?
- How are you going to operationalize this new insight?
- Do we have the authority to take action to alleviate or eliminate the problem?

You must understand that there are significant challenges that are unique to dealing with data resources in the Petabyte size range. When you can solve a problem with a small amount of data, you should! But when you encounter the big problem that can only be addressed with big data, you've got to be ready.

A Data Discovery platform helps you put big data to work on your own business problems by reducing the complexity of the process. With a data discovery platform in place, you will be able to manage and analyze data with fewer steps, execute analyses more quickly, and get up to speed more easily than you could with other big data technologies. Data Discovery puts big data power in the hands of you and the staff you have today, without the need for near-mythical specialized talent.

The unique nature of big data is often characterized by a handful of words that begin with V. Most important among these are value, volume, velocity, and variety.

Value

Just because you have a lot of data, that doesn't mean it's valuable. Put a key business problem together with the right data and thoughtful analysis — that's how you get value.

Focus on the business problem's impact and your potential for solving it. Value isn't determined by the magnitude of your data, but by the magnitude of the problem you solve with it. Learn about big data not for its own sake, but for what you can do with it.

Volume

The big thing about big data is that it's big. Literally big. There's a lot of it. There's no official minimum, but a starting point somewhere in the 100 Terabyte to 1 Petabyte range is reasonable.

You will surely encounter some people describing much smaller quantities as *big data.* Don't argue with them; their words reflect the reality of their own challenges in dealing with that data, or the potential they see in it. But extreme quantity poses unique challenges.

You can't put a 100 Terabyte database on an ordinary computer, or even a single computer — it must be spread across multiple servers. Ordinary data management software isn't designed for that, so you must use specialized tools. Those tools, in turn, call for specialized skills to use them properly.

Velocity

Having big data is something like having dandelions in your garden. It seems that each time you look, you see more. Although big data doesn't reproduce, per se, much of comes from machine sensors, social media activity, and other pathways that produce large amounts of new data throughout the day, every day.

Variety

Data isn't just columns of numbers these days. The biggest component of many data sources is text, such as posts from social media, comments given with warrant claims, or notes in medical records. Images, audio, and video can all be forms of data. These so-called *unstructured* data formats can't be analyzed directly through conventional statistical analysis or machine learning methods.

More V-words

Value, volume, velocity, and variety are the words most often used to characterize big data, yet others are sometimes included in the V-word paradigm. Veracity, a reference to data quality issues such as inconsistencies, incompleteness, latency, and simple errors, is the one most often mentioned. Others — including validity, visibility, variability, visualization, volatility, and even victory — are sometimes mentioned.

Each of these terms speak to legitimate issues, and may be more serious where big data is concerned. However, it is still value, volume, velocity, and variety that best characterize the unique nature of big data and its challenges.

Rewarding All Stakeholders

Anybody can grab a bit of data, put it in a spreadsheet, and explore. When the stakes are low, it's easy to get involved. Then again, it's also easy to avoid getting involved. Somebody wants access to the data warehouse, to explore? That might not be a good enough reason to provide access or support.

REMEMBER

Big data raises the stakes for analytics. When the magnitude of the data rises to Terabytes and more, the data isn't just lying around somewhere handy. It can't be loaded into spreadsheets, let alone analyzed that way. With big data, everybody has to get involved, because teamwork is an absolute requirement.

But to get everybody involved, there has to be something in it for everybody. Each stakeholder invests in Data Discovery, and each gets something in return. When you get into Data Discovery, somebody is bound to ask, "What's in this for me?" Now you're ready to answer.

Stakeholder: Business management

Investment: Authority and budget.

Returns: Better information for the decision process. Greater accuracy, richer detail, enhanced ability to identify factors affecting business outcomes, and the connections among them.

Stakeholder: IT leadership

Investment: Manage the process and provide staff for implementation and support.

Returns: Eliminate stream of individual support requests for analytics. Free IT staff for other responsibilities. Improve internal customer satisfaction. Leverage existing investments in people and technology.

Stakeholder: IT staff

Investment: Hands-on work to implement the Data Discovery platform, integrate with data sources, and ensure that end users have the access they require.

Returns: Dispense with oodles of requests for data extracts.

Stakeholder: Distinguished analysts

These are people whose primary job responsibility is data analysis. Their job titles and areas of strength vary from place to place. Some of the most common titles for this group include statistician, data scientist, and quantitative analyst.

Investment: Get used to the new platform.

Returns: Greater productivity, fewer errors, and more time to spend on the most interesting aspects of data analysis. Better and easier data access, better integration among data sources, opportunity to probe data in greater detail. Broader range of analytic methods.

Stakeholder: Other data analysis roles

There may be many people in your organization who perform some data analysis, but don't make that the primary focus of their work. Needs, interests, and skillsets vary considerably within this group. Treat them as individuals, not a uniform block.

Investment: Depends on the individual. Some will learn new code; others may need a new application or integration with a familiar tool.

Returns: Access to greater data variety and depth of data and analysis methods.

Stakeholder: Everybody else

Investment: Yes, everybody invests something. Even those not directly involved may find another project waiting until Data Discovery gets underway — or be inconvenienced somewhere along the line.

Returns: Better information to support decision making at all levels of the organization, leading to a healthier, better-run workplace.

Chapter 2

Building Data Discovery Capabilities for Your Business

In This Chapter

- Facing up to competitive pressure
- Letting people focus on their own responsibilities
- Keeping workloads under control

Data-driven business practices affect everyone in the business. Some businesses have grown up with a data-driven culture, but that's not typical today. So, for most organizations, and most of the people within them, the transition to data-driven business calls for changing everyday practices, and for changing culture.

Every business can benefit from analytics. The challenge lies in making the transition with the resources available, and without disrupting business or causing needless friction.

Engaging the Business

Certain sectors have a long history of using analytics effectively. Direct marketers, for example, were routinely using formal and effective testing methods for advertising in the early 20th century. The methods used then are still valid today, and of course many more have been added to the mix since then.

If you and a friend each get what appears to be the same piece of advertising in the mail, compare the pieces carefully — you may notice some small differences. Perhaps the words on the envelope are a bit different, or the layout isn't quite the same. Maybe you're offered two different prices for the same item! These differences aren't mistakes; they're tests to see what works best. The best catalog marketers test every element of their material — images, copy, layout, pricing, and more. Online retailers have even more opportunity for this kind of testing and measurement.

Marketers use information from tests to decide what products to offer, what prices to charge, what copy, colors, and images to use, and so on. This is data-driven decision making. Broadly speaking, it's nothing new.

But even in sectors with a history of working with analytics, not everybody does business the same way. Some direct marketers send a lot of mail without much study of the response. Many businesses have gone bankrupt doing that, but others remain. Every insurance company uses statistical analysis to set rates, yet often those companies don't approach marketing with the same rigor. Some manufacturers have been known to set strict standards for data analysis in manufacturing quality, yet not adhere to their own standards.

Rising competition changes business

Mounting competitive pressure means that businesses that have gotten by without serious use of analytics are questioning that approach. Consider these issues confronting businesses today:

- Many brands are contracting or disappearing in the face of nimble competition.
- Even the hottest of businesses must be watchful for changes in the landscape. (Myspace, once the world's most popular social networking site, has since been overtaken by hundreds of others.)
- Brands that once seemed to be out of the running may be revived. (Abercrombie and Fitch, more than a century old and extremely popular today, was once bankrupt. Apple Inc. pioneered personal computing and is the largest corporation in the world today, yet it lost money in the 1990s.)

Changing business situations are nothing new. Mail order catalogs once struck fear in the hearts of rural shop owners, much as Internet retailers challenge bricks and mortar stores today. Shopping malls, big box retailers (such as Target and Best Buy), and warehouse clubs each drove significant changes in retailing consumer product manufacturing. Yet today's business world is arguably more international, more price-driven, and more competitive than ever before. Competitive pressure is forcing businesses to run on tight margins and squeeze more from every dollar invested.

Analytics and your competition

Aggressive and consistent use of analytics is central to the way that thriving brands run. Leading online retailers report great returns from data-driven decision making at all levels. Businesses that have grown up with an analytics culture have a competitive advantage. Others are working to catch up.

Businesses that don't succeed in moving to data-driven practices will put themselves at risk in a variety of ways:

- ✓ Information available to support decision processes will be inferior to the information available to competitors.
- ✓ Lack of adequate documentation and measurement will prevent management from learning which business practices are working and which aren't.
- ✓ Costs will be incurred for data collection and storage for daily business and legal obligations, but valuable information will remain trapped in that data, doing nothing to help reduce costs or drive revenue.

REMEMBER

Every business incurs costs for data collection and maintenance. It makes sense to use data assets to your own advantage.

Data Discovery facilitates change

The first move toward Data Discovery usually comes from the business side of the house.

Marketers are often the driving force behind a shift toward data-driven business, particularly *direct marketers* (that is, those who offer products for sale directly to the end user, rather than distribution through wholesalers or retailers). These marketers know what's tested, how, and with what result. They're actively involved with analytics and use the information directly to design advertising materials, packaging, even products.

Others may be involved in a much less active way. Next time you buy groceries, pay attention to what happens after you pay. Your cashier may hand you a coupon (or several coupons) along with your receipt. Look at those coupons, and you'll see that they weren't printed at random. The offers will be matched to your purchases, geared to influence you to buy larger quantities, or perhaps to try a competing brand. This specificity is the result of the marketing team's thoughtful analysis and partnerships that enable them to instantly create coupon offers based on the purchases you have made. The cashier doesn't have to think about it because the thinking has already been done.

Marketers want to think about what marketing options are available and how well they work. Cashiers don't; they have other things to do. So the object is to establish effective systems that let each person see and do what's relevant to her own job, and not have to think about the rest.

Data Discovery is all about letting each person think about, well, just what she's supposed to think about. A marketer needs to have access to all the data relevant to marketing. That might include a mix of conventional numeric data from a marketing campaign, text from social media, sample documents used in the campaign, and other documents used internally to document the process. The same marketer generates new information that must be available for use, such as a list of target customers, a predictive model or business rule, or a new coupon, artwork and all. The Data Discovery platform provides access to all the information resources the marketer needs, and a pathway for adding new information and making that available to others.

The unique capability of a Data Discovery platform to draw in data from all sources and make both data and analytics available to users has a natural appeal to business users. But don't neglect the last mile — business people are often somewhat familiar with analytics, but rarely programmers. Business users must supplement the Data Discovery platform with familiar tools or new applications that leverage the platform and suit their needs.

Easing the Support Burden

The business side of the house wants and needs more — and better — analytics. That means they need data from a number of disparate sources. These might include:

- A legacy customer relationship management (CRM) system on local servers
- Accounts payable and receivable
- Human resources and payroll
- A newer, Software as a Service (SaaS) CRM system
- A document management system
- A collaboration system
- Email
- Web crawlers
- Social media
- And more

Some of these sources may already be integrated into a data warehouse. Others might be little worlds apart. More could be added at any time. (Refer to Figures 1-1 and 1-2 to see how diverse sources can be united with a Data Discovery platform.)

Safety first!

TIP

Your business users may not have direct access to most of the data sources listed in the previous section, so each time a business user needs data, it has to be requested, and somebody from IT must get involved. That's time consuming for everyone, and it's not scalable. So a key element of Data Discovery is self-service. Easy access encourages business people to use data routinely, and in a timely manner.

Self-service? Sounds risky!

Don't panic. Data Discovery doesn't bypass identity access management (IAM) or other security protections in your data sources. To reach any data, the user must have proper credentials and permissions.

Although upfront effort is required from IT to get Data Discovery underway, the effort required isn't overwhelming. Even though data sources must be integrated with the platform, this shouldn't cause much shock to the system, because it can be done with the same skills and tools already used for similar tasks. Users must be given proper credentials for data access, and it will take some thought to ensure that each user can reach the appropriate data and nothing else. But once the setup is done, the routine burden of extracting data for analytics will be lifted from the shoulders of IT.

See Figure 2-1 for a high level architecture view of a Data Discovery platform.

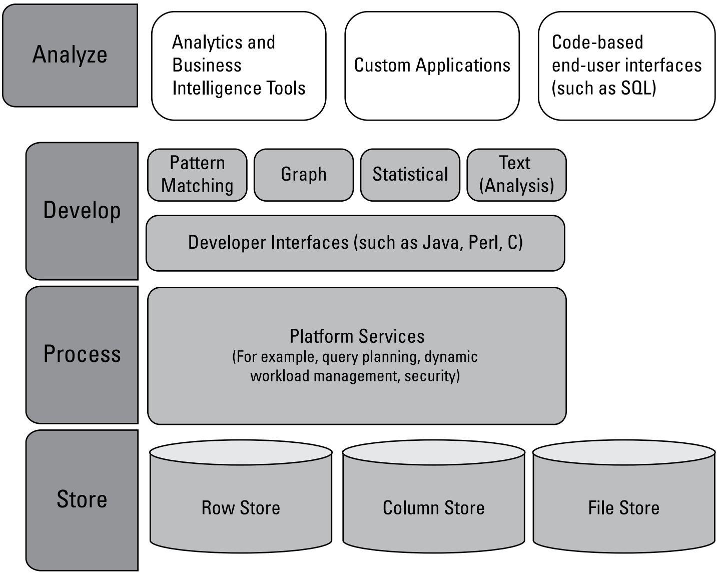

Analyze
Analytics and Business Intelligence Tools
Custom Applications
Code-based end-user interfaces (such as SQL)
Develop
Pattern Matching
Graph
Statistical
Text (Analysis)
Developer Interfaces (such as Java, Perl, C)
Process
Platform Services
(For example, query planning, dynamic workload management, security)
Store
Row Store
Column Store
File Store

Figure 2-1: A high level architecture view of a Data Discovery platform.

Leaving Old Limitations Behind

Some people don't need anyone to convince them that business should be data driven. They already focus most of their energy on data analysis. They know how to use the data sources at hand, and some that are more difficult to track down. They understand data analysis methods. They aren't dependent on spreadsheets or simplified tools with menus and icons. When the task calls for writing code, they do it. In fact, many of them do so every day.

Should these motivated and highly skilled people change the way they work? They should.

Heavy lifting for big data

Internet-based businesses with massive distributed databases gave birth to a new job title: *data scientist.* There's no official definition of this role, and some of the job descriptions in circulation call for combinations of skills that are nearly mythical. As the term has risen in popularity, many people (with differing skillsets) have come to identify themselves as data scientists. For the purpose of this discussion, a data scientist can be anyone whose work is primarily devoted to data analysis, and whose skills include writing queries or programming of some kind.

Data scientists already have methods for accessing data and analyzing it. If their methods work, you might wonder: Why mess with success? The truth, though, is that data scientists may have more to gain from adopting Data Discovery than anyone else.

It takes effort to extract data from several different sources, especially distributed sources such as Hadoop, and pull the extracts together. The process might involve two or more programming languages, such as SQL for databases (perhaps a slightly different version of SQL for one database than another) and Python or Pig for Hadoop.

The code must be written thoughtfully, because it isn't enough just to write code that looks logical. It must also be efficient. If not, a query may take much longer than necessary to run. Taking all these steps may not be easy, but the data scientist has the know-how. So far, so good.

Now the data scientist can move on to analysis! But you can't just plop big data into a spreadsheet. Big data analysis calls for yet more programming, and very detailed programming at that. It may be necessary to prepare detailed instructions for splitting a text document into individual words, or to specify just how many processors to devote to a task (in contrast, a Data Discovery platform might have similar text handling functions prebuilt, and the user would probably never need to think about processors). It takes time to get the job done.

REMEMBER

Programmers who develop commercial software can expect the help of a quality assurance team to detect problems in their code. Lucky them! Data scientists don't usually have that kind of help. So data scientists carry the responsibility of preventing errors from creeping into their processes, and for finding and correcting any errors they happen to make. That's a lot to ask. Data scientists are only human; they make mistakes.

Making the data scientist's life easier

Time is the data scientist's most precious resource. Computers get cheaper and databases get bigger with every passing moment, but there are still only 24 hours in a data scientist's day.

A Data Discovery platform enables data scientists to:

- Increase productivity
- Shift time from mundane, repetitive tasks to more interesting analysis
- Prevent errors
- Expand their portfolio of analytic methods
- Combine multiple analytic methods to deliver rich information tailored for decision support

Every data scientist aims to be productive, to do interesting (and valuable) work, and to keep that work free from errors. But a good data scientist doesn't take every claim at face value. They tend to ask a lot of questions; that's how you know you have a good data scientist on your hands!

Data scientists who work with big data have certain powers that they won't want to give up. They can access data sources without going through middlemen. They use powerful languages that give them flexibility and fine control over their work. They're free to devote some time to data exploration guided by their own curiosity, to let the data reveal its stories.

Here are examples of questions that are likely to come up, along with answers:

> **Question:** Will I still have access to all the data that I've been able to use before?
>
> **Answer:** Yes. In fact, you may gain access to additional sources if that is appropriate to your job. With a Data Discovery platform, you can access any data source for which you have appropriate permissions.
>
> **Question:** Real data scientists use [insert favorite programming language here]. Can I still use that?
>
> **Answer:** You will absolutely be able to use [insert favorite programming language here]! But that will be an option, not the only way, and usually not the best way for you to work in the future. Your value isn't in coding, but rather in solving business problems.
>
> **Question:** Will I be forced to use the simplified interface?
>
> **Answer:** We're not going to be looking over your shoulder to check. However, you should use the new interface, not because it is the only option for working with the Data Discovery platform, but because it will enable you to complete tasks in less time. It also gives you access to analytics functions provided in the Data Discovery platform. (Others on the team will be using it, getting work done faster, and getting into cool new analytics. You should, too.)
>
> **Question:** I'm really worried about this interface!
>
> **Answer:** I appreciate your concern. Let me address that. The Data Discovery SQL-based interface gives you

the ability to do what you've been doing, such as using MapReduce to work with massive data sources, in an easier way. (Although some users may not use the SQL-based interface, instead choosing a custom application or third-party tool.) Advanced analytic techniques such as MapReduce and Graph will still be working for you, but you'll be able to write your code faster and with less risk of error. You'll be able to iterate even faster (I know you won't take my word for that, but you don't have to. Try it yourself, hands-on, and you'll see.)

Question: Will I be limited to using prebuilt analytics functions?

Answer: No, but you should take advantage of them whenever possible. They were developed for your benefit by a team of behavioral, statistics, and machine learning experts and have been subjected to formal quality assurance processes to prevent errors. And you won't find them limiting. The range of analytics provided is substantial, and it includes some great stuff you don't have available today.

These questions sound like they're grounded in worry about losing something, rather than excitement about new opportunities on the horizon, and they are. That's natural. Concern comes before enthusiasm. Most people express concern when a major change affects the way they work.

TIP

Nobody likes losing capabilities when a new system is introduced. A good onboarding process for a Data Discovery platform will include interviews with data scientists early in the game to review what each does, and how. When the new platform is introduced, first show users how to do what they've done before. Once they have reassurance that nothing has been lost, they'll be able to move forward and appreciate the benefits gained.

The thrill of discovery

The fun part, when data scientists start to feel the benefits of Data Discovery, will come quickly. It's nice when a query that used to take forever to write is finished in a fraction of the time. It's even nicer when the structure and optimization capabilities of the Data Discovery platform significantly reduce the time it takes to run a complex analytic process. And it's downright exciting when the data scientist gets all the usual stuff done early and has time to try out a cool (and informative) new visualization.

Nothing equals a hands-on experience to prove the value of Data Discovery for data scientists. But don't let the momentum stop there. Successes should be shared, and not just within the data science team.

TIP

Data Discovery exists to provide valuable information to the business. Share the good stories, not just to get the information to individual decision makers, but also to inspire everyone to try new things, to ask questions they might not have asked before, and to take better advantage of data to support their own work.

Use case: Preventing customer churn in banking

Every business needs customers. In fact, customers may be the one thing no business can do without. So it's surprising that so many businesses don't seem to care if you stay with them or not. Smart business people care a lot!

People in banking know the value of a customer, and their customers are worth plenty. The average household in the United States has over $15,000 dollars in credit card debt, and nearly $150,000 in mortgage debt, not to mention student loads, car loans, and other business that makes money for banks.

Some banks understand their customers better than others. They get involved in consumer research and use statistical analysis to predict customers who are likely to take their business to a competitor. Some of the more adventurous banks explore machine learning to widen their predictive modeling options.

As the data resources available to businesses expand, a world of new possibilities opens up. Could information available from public or commercial data vendors add predictive power? Or provide better intelligence on what action would be most effective for retaining individual customers? To find out, you have to get some data and see.

The challenge of pulling together multiple data sources prevents many banks (and other businesses as well) from exploring these possibilities. The more data you have available, the harder the challenge may seem. But think of the rewards! A single banking customer can easily pay thousands of dollars in interest and

(continued)

(continued)

fees to a bank in a single year. If your average customer is worth $1,000 a year to you, and you can retain 1,000 extra customers next year, that's a million dollars in saved revenue.

One bank, already involved in predictive modeling to prevent customer churn, has gone deeper by implementing Data Discovery. Adding new data sources and new analytics to the modeling mix through Data Discovery made big, big bucks for the bank.

How is it done?

First, data from disconnected silos is integrated into a single platform, giving analysts access to data for all channels through the Data Discovery platform. This makes it possible to follow the experience as the customer moves from ATM to teller, from email to online, and so on.

Then, expand the bag of analytics tricks. Experiment with the predictive modeling methods provided in the Data Discovery platform. Integrated data makes it possible to try these new tricks, like the data mining favorites of looking for agreement in results of two or more models and using ensemble models (two or more conventional models used in combination). And this is a great fit for path models, which allow you to discover common sequences of events.

If you knew the most common sequences of events leading up to an account closing, what could you do? How about watching more carefully after the first couple of steps? Offering a promotion after the third? Having the manager call the customer as it gets down to the wire?

You could try a variety of things, collect data, and find out what works, and what doesn't.

Banking has a great value opportunity in customer retention because banking customers are so valuable. But then, aren't your customers valuable, too? Customer retention is a big issue in many industries, especially retail, telecommunications, and healthcare. Your business could be the next to save thousands of customers — and millions of dollars in business with them.

Chapter 3

Understanding a Data Discovery Platform

In This Chapter

- Putting structure first
- Pumping up your analytics muscles
- Enjoying the fruits of your labor

If you want to make some business magic with Data Discovery, you've got to understand how it works and why you need it. This chapter discusses the working parts of your Data Discovery platform and how they work with business processes and people. It also explains the analytics toolkit and what benefits you can reap from using it.

Employing the Right Structure

Structure means so much. A sturdy house needs a solid foundation, a handsome physique depends on strong bones, and a great education begins with reading, writing, and arithmetic. Your Data Discovery platform's foundation provides structure that supports your business's goals and fits in with its processes.

Data Discovery platform structure has three major elements: the data store, processing framework, and analytic engines. Successful Data Discovery architecture places these in the proper context to serve the business with efficiency and simplicity.

Data store

Why take data you've already stored, then store an additional copy, even for a brief time? This question carries particular significance when the quantity of data involved is very large.

Adding an additional data store to a system may seem unnecessary or undesirable. Indeed, some people resist this and advocate for data virtualization. In theory, virtualization ought to minimize resource requirements. It ought to be simple.

In practice, though, devoting a data store to Data Discovery provides rewards that make it well worthwhile. An appropriate dedicated data store

- Allows Data Discovery to be conducted without taxing operational data stores and interfering with everyday business
- Ensures effective management of all types of data at scale
- Facilitates rapid execution of operations

Avoiding interference with routine business is reason enough to use a dedicated data store for Data Discovery. Another reason is that existing data storage systems weren't selected with Data Discovery in mind; after all, Data Discovery requires a multitype data store capable of massively parallel processing. Existing systems are probably not up to this task although they may have been optimized for the business processes in which they're normally used. They may now be outdated and no longer be the best-fitting options for the original purpose, let alone Data Discovery.

Processing framework

Between data storage and specific analytic techniques lie many general-purpose capabilities, most of which should go unnoticed by users most of the time. The processing framework is the workhorse that performs all those necessary capabilities. The framework carries out the operations that the user requests, it makes those operations run on the most efficient manner possible, it communicates with the dedicated data store and analytic engines, and interacts with other systems as well.

TIP

Some aspects of the framework are visible to users, most notably through user interfaces. Sophisticated analysts who have programming skills should use an interface designed with them in mind, something that allows them the breadth and depth of options that they're accustomed to, but which is easier and faster to use than languages such as Python and Pig. The right interface will provide them the same power with less fuss.

Business users and some analysts will be better off with other types of interfaces, perhaps familiar tools or applications built to fit specific needs. In some cases, these needs can be met by integrating the Data Discovery platform with another product, and the vendor may offer such integration through a partnership. When no such option exists, another type of interface is needed, one that allows software developers to develop the required applications.

Analytic engines

The Data Discovery platform should include the capability to perform a variety of analytic functions. It's desirable to have as many of the analytics that you require come from within the Data Discovery platform itself so you can take full advantage of the efficiencies provided by the framework. An added benefit is limiting the effort required for integration. The fewer parts involved, the less work in putting it all together. (Many types of analytic techniques and their uses will be explained later in this chapter.)

Architecture

The Data Discovery platform doesn't stand alone. It communicates with data sources and applications. It serves users who also must fit in with business processes and other people. Your architecture must put Data Discovery in the right context to operate efficiently and as simply as your circumstances allow.

Figure 3-1 shows an example of Data Discovery in context. It includes a variety of original data sources and data storage structures capable of supporting a range of business purposes. These, as well as the Data Discovery platform, share information with applications for marketing, business intelligence, and other purposes, which support end users in wide-ranging roles.

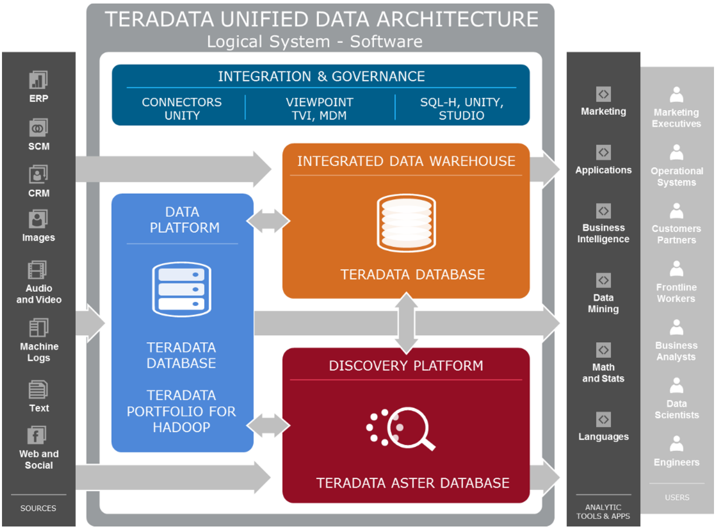


Figure 3-1: Reference Deployment Architecture: Teradata's Unified Data Architecture.

Anatomy to fit your business

The right structure to support Data Discovery for your business is one that you can adapt to fit your systems, people, and goals. You can't afford to shut down the business waiting for queries and analytics to run. You need to make the most of the staff you have now. And relevant information must be available where it's needed, when it's needed.

Put Data Discovery in the right context by examining your business goals and requirements to make sure a solution fits your business, not the other way around. Here are examples of the issues you'll want to keep in mind as you explore Data Discovery:

- **Protecting operations:** It's important to separate everyday business from analysis. The solution must shelter operations, and not allow resource requirements for analysis to interfere with other systems.
- **Service-level agreements:** Data Discovery doesn't call for the same stringent uptime requirements as data warehousing or other operational systems.

- **Demands on your staff:** In the short run, you may need to supplement your current resources with transition and training help. Primary goals are to increase data analysis productivity and depth while sparing analysts from dealing with data management complexity, and preventing IT from being overwhelmed by new and changing requirements.
- **Minimal effort:** Choose a solution that will minimize requirements for data modeling, integration, analysis, and other activities.

Process flow

The Data Discovery platform should enable an ongoing and iterative process flow, a cycle of activity that is repeated again and again, developing depth and revealing new information with each pass through the cycle. Here are some of the steps:

1. **Acquire:** Import data from operational, archival, or other sources and load it into an analytical data store.
2. **Prepare:** Before analysis, data requires examination, cleaning, and manipulation to suit the requirements of the specific business problem and analytic methods to be used.
3. **Analyze:** Perform calculations for statistical and other sophisticated data analysis techniques.
4. **Visualize:** Providing results in graphical form makes it easier to identify and understand valuable information.

Learn more about this cycle and how it fits into everyday business in Chapter 5.

Streamlining visualization

Visualization is a key element of Data Discovery, because images transcend numbers and formulas for communication with executives, clients, and others who aren't expert analysts. Data Discovery platform should support visualization options like:

- Choose a platform that offers a variety of purpose-built data-aware visualizations as part of the platform.

- Web-based application capabilities ease viewing and sharing of visualizations.
- Integration with familiar (or desired) business intelligence, data analysis, and other visualization tools to allow expansion of the user base and visualization options.

Exploiting New and Powerful Analytic Methods

Getting your data together doesn't mean much until you do something with it! Your Data Discovery platform should provide a generous range of analytic capabilities. This section describes a selection of analytic functions to give you a sense of the range and depth that are open to you through Data Discovery. It's not totally comprehensive. That would take a much longer, and perpetually updated, book! If you're not an expert, this is where you can get a handle on the range of analytics that Data Discovery can provide. And you analytics experts can skim this section for things you've never tried before — valuable goodies to add to your bag of tricks.

Basics for data preparation and transformation

Think analysts spend all their time analyzing? Think again. Most data analysts spend more time on data preparation than the fun stuff. Joining, aggregating, and reorganizing data and other simple operations will always be part of the process. Data Discovery helps make the process simpler, and reduces the number of steps involved, but the basic stuff still needs to be done.

TIP

A lot of today's data comes from web activity, and it's important to have the right tools for getting web data into the format you need.

Look for web data handling functions like these:

- **XML parsing:** Web and many other computer functions are grounded in markup languages such as HTML, the code that defines the content of web pages. XML parsing enables you to extract specific functional parts of markup, such as a name, price, or title, for further analysis.

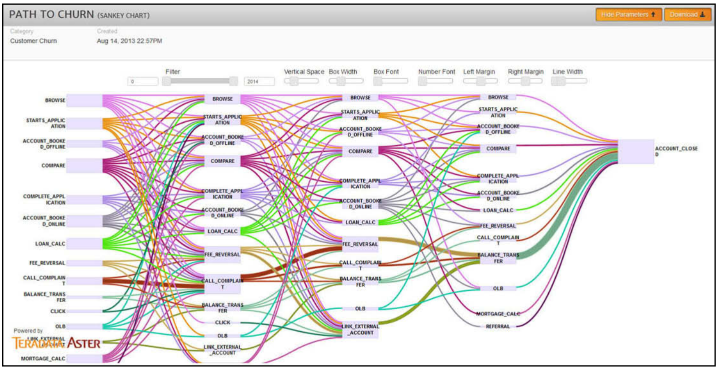
- **IP geomapping:** IP geomapping converts an Internet Protocol (IP) address, the code that identifies an individual computer on the internet, to a geographic location. The information returned (and its accuracy) varies, but it may include details such as country, region, city, latitude, longitude, ZIP code, time zone, connection speed, internet service provider (ISP), and domain name.
- **Sessionization:** Sessionization groups a series of individual clicks (which appear in the web log as file request records) into sessions using a unique session identifier (in other words, a session number).

Paths to discovery

If you watched people moving around a supermarket all day, every day, you'd begin to notice patterns. You might notice that some people move from the entrance to the checkout in a linear fashion, going from produce to bakery to deli, and so on, just as the store is laid out. Others might bypass perishables and head to the staples first, later returning to produce, finally stopping at the freezer case just before leaving. You might discover several common patterns, and you might even observe a connection between the shopper's preferred path and other visible things, such as whether the shopper is alone, with another adult, or with children.

Maybe you'd notice that some of the people shopping with kids are carefully bypassing the candy and cookie aisle. Those same shoppers might be willing to buy children's books, magazines, or toys, but they won't see those items if you've only displayed them next to the cookies. So you can use your observations about shopping paths to help you put those items where parents will see them.

The same types of observations can be made with data, and they apply just as well to online and multichannel behavior as to shopping in a supermarket. Frequent paths (also called *path analysis* or *sequence analysis*) do with data what you would do by observation. It identifies common behavior sequences. The visualization shown in Figure 3-2 is one approach to presenting information about such sequences.


PATH TO CHURN (SANKEY CHART)

Category
Customer Churn

Created
Aug 14, 2013 22:57PM

Filter 0 2014

Figure 3-2: Visualizing different paths leading to a single event.

Attribution is a related method that assigns value to individual steps in a path leading to a specific outcome. For example, the customer who buys a toaster probably didn't just see the sale price on the shelf and dash to checkout with the toaster in hand. That action might have been preceded by the catastrophic failure of the old toaster, reading toaster reviews online, posting toaster questions in social media, seeing your newspaper circular with the sale price, and so on. Each step had some part in the final outcome: a toaster purchase. Toaster marketers understand this, so they use attribution.

Connecting the dots with graph analysis

An old shampoo commercial helped visualize the impact of word of mouth in a distinctive way. It showed a woman washing her hair with the shampoo, smiling as she rubbed her sudsy head. She told friends, and they told friends, and the screen split again and again to show more and more images of the sudsy women's smiling faces, tiling the screen with tiny happy shampoo customers. It's a powerful image, but doesn't fully capture the complexity of real social networks.

Social networks, online and off, are created by people and the interactions between them. Individuals have connection: to friends, family, colleagues, and others. Each person has a unique mix of connections, and differing levels of influence upon each of those connections. Influence is two way and may change with the subject matter and the point in time.

There are two types of graph technologies now available: navigational and analytical. Both use the concept of connected networks of individual people or other entities, represented as elements (like points) called *nodes* or *vertices.* Relationships (connections) among the entities are represented as elements (like lines) called *edges*. The network itself, with all its entities and relationships, is a *graph*.

Navigational

Navigational graph technology, for example, graph databases or RDF/SPARQL stores, are built for data storage and management. This can be excellent for transactional queries, but not for data analysis, and can be unacceptably inefficient for that purpose. Navigational graph technology isn't suitable for Data Discovery.

Analytical

Analytical graph technology is designed and built expressly for data analysis. This is the right graph technology for Data Discovery. It can accommodate analysis involving large and complex networks. It doesn't require special formats or indexing to perform analysis. With analytical graph technology, you can analyze whole graphs at scale to get an understanding of effects such as influence and belief propagation.

Statistics plain and fancy

Maybe you took Statistics 101 at school. Whether that was just last semester or decades ago, the content was pretty much the same: descriptive statistics (things like averages and standard deviations), perhaps a few histograms, sampling concepts, basic inferential statistics (hypothesis tests) and linear regression (fitting a line to data).

In the last half century or more, the only big change in basic statistics is that nobody does the calculations by hand any more. The stuff you learned at school is still the backbone of data analysis.

And then there is the fancy stuff (translation: anything you didn't cover in Statistics 101 at school). Many useful statistical techniques that aren't brand new still may be new to you. Advanced statistical techniques that you may not have tried include methods for identifying meaningful groupings (such as customer segments or behavior patterns), modeling to predict outcomes and understand how you can influence them (also known as *predictive analytics*), simplifying problems that involve large numbers of variables, and more. With Data Discovery, they'll be easily accessible, too.

The following sections go over some of these more advanced techniques.

Clustering (K-means)

In the days before cable TV and computers, the coolest thing at grandma's house was the box of buttons. Kids would dump the buttons on the ground, look them over, touch them and sort them into little groups. Groups could be organized in many ways: by size, color, shape, or texture. Fancy buttons might get more attention than plain ones. Organizing the buttons kept kids busy all afternoon.

Marketers still play that game, but now it is called *customer segmentation,* the buttons have been replaced by consumers, and the sorting is based on demographics and behavior. The groups and the possibilities are a lot bigger, too! You could spend a lifetime on it, but you'd probably rather not. Clustering techniques, such as k-means clustering, bring the job down to size.

Time series

Some things, like shopping behavior, weather, and outbreaks of influenza, follow fairly consistent cycles in time. Time series let you get a handle on those. Common applications include sales and econometric forecasting, but time series sometimes crop up in lesser known uses, such as fingerprint analysis.

Generalized linear models (GLM)

The godmother of all statistics is a comprehensive model (or rather, a family of models) that can do all the stuff you did in Statistics 101, plus a whole lot more. This transformer can be an ordinary linear regression, or a complex function based on non-normal distributions such as Poisson or multinomial. It's used for predictive modeling applications, and is particularly popular in manufacturing and the insurance industry.

Principle components analysis

Principle components analysis (PCA) simplifies problems that involve a heck of a lot of variables. If you've used an online dating service that used an exhaustingly long questionnaire, principle components analysis was part of the process that made your matches.

Outlier identification

Another kid's pastime is a learning game where the child looks at a group of things and identifies one that doesn't fit. The TV show *Sesame Street* has a special song for the game, "One of These Things (Is Not Like the Others)." The different item might be an orange among apples or a box among balls. Outlier identification is like that, but the answers aren't as clear, the math is more complex, and instead of just three or four things, there are millions and millions.

Basket generation

A basket generator isn't a machine that makes containers out of straw. No, no, no. This kind of basket is a set of items that go together, most often products commonly purchased together. This information can be used to better organize a store, or to plan profitable promotions. For example, it might be worth your while to offer a low price on an item if you know that people who buy that item are likely to also buy one or two highly profitable items at the same time. (You may also hear of this by the names *market basket analysis* or *association rules*.)

Collaborative filter

When shopping online, adding an item to your cart triggers the retailer to offer you a few more items. Not just any items, of course, but items that the retailer has reason to believe you're likely to buy. The engine behind the offer is probably a collaborative filter, a way of identifying others whose tastes are most similar to yours, with the aim of suggesting additional items that those people bought or liked.

Classifiers

Classifiers are related to clustering methods, in that they're used to characterize groups, but in classification the groups are formed before the analysis begins. A retailer knows who bought a product and who didn't, a medical researcher knows who died and who didn't, an educator knows who was expelled and who wasn't, but they might not understand why, or what characteristics differed between the two groups.

Many classification methods exist. The most popular is the traditional statistics method: logistic regression. The popularity of logistic regression is mainly due to its age and familiarity; the results can be hard to interpret. Newer, easier-to-understand alternatives may be just as accurate, or better.

TIP

Your Data Discovery platform must include some up-to-date classifiers, especially decision trees (see Figure 3-3 for an example). Decision trees can be represented as branched diagrams (yes, that why they call them *trees*) that are easy to understand, even for people who have absolutely no interest in statistics or math (or trees). You can use them for a presentation to executives, or customers, or the general public, and everyone will understand. (Try that with logistic regression, hah!)

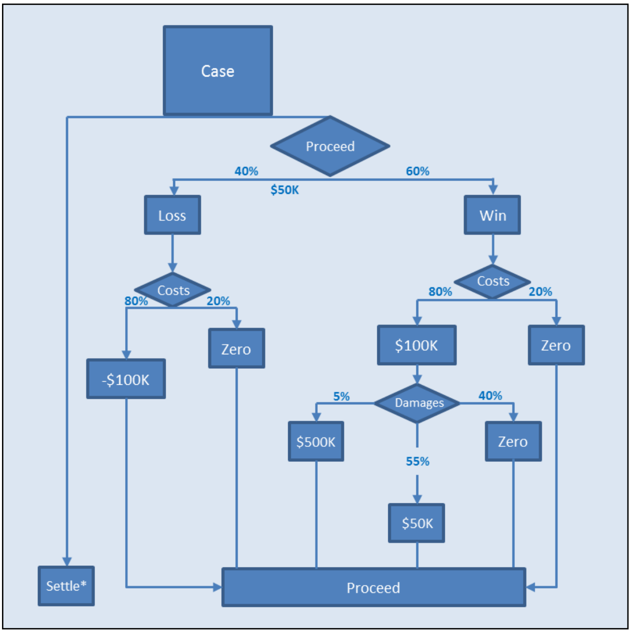


Figure 3-3: A typical decision tree.

This is an absolute must-have, so if you don’t see it, ask whether a Data Discovery platform offers it. Decision trees may be there, but called by the names of individual types such as classification and regression trees, chi-square automated interaction detection (CHAID), random forest, or naive Bayes.

Taking on text

A lot of today's data is text, and what a mixed blessing that is. Text is loaded with rich and nuanced information, but it's mighty hard to get that information out. Text analytics is still rather new, and it is far from perfect, but it is darned useful, and has already built multibillion dollar industries including search and document management.

You might make your own fortune with text analytics techniques like those discussed in the following sections.

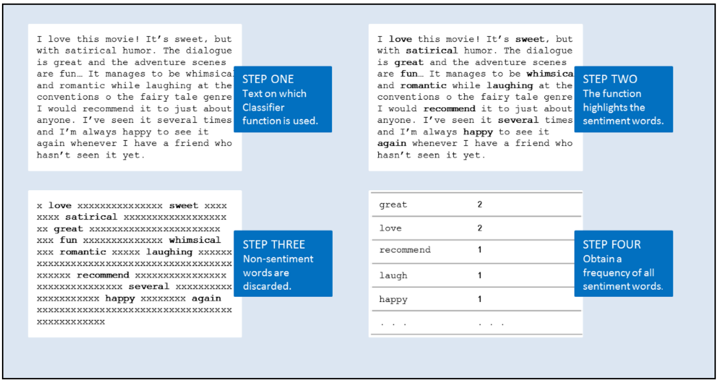
Text classification

You can't read everything, can you? If you have more text to deal with than you could ever dream of reading (and who hasn't?), you need help to organize it. Enter text classification, the way to categorize bits of text. Some of the most important ways to form categories in text are by:

- Subject matter, such as finance, sports, or politics.
- Topic, such as exchange rates, credit card debt, or reverse mortgages.
- Genre, such as the fiction styles of steampunk, cyberpunk, or elfpunk.
- Sentiment. Does the writer like your brand, or hate it?

Sentiment analysis is a hot item these days. This popular text classification application often uses only a few simple categories: positive, neutral, negative. Sometimes the categories are greater in number, and more subtle: angry, happy, sad, and so on. Figure 3-4 shows steps in an example sentiment analysis process. Because human beings often have difficulty interpreting sentiment in written text, it comes as no surprise that sentiment analysis isn't a perfect science.

Nonetheless, it can be very useful. Sentiment analysis has been used to research themes for antismoking campaigns and to measure the results. It's been used for customer retention to identify really unsatisfied customers. And it's been used for fundraising to understand why people support a cause. Real-life applications rarely require perfection; it's often invaluable just to be a little better than what you had before.



I love this movie! It’s sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun… It manages to be whimsical and romantic while laughing at the conventions o the fairy tale genre I would recommend it to just about anyone. I’ve seen it several times and I’m always happy to see it again whenever I have a friend who hasn’t seen it yet.

STEP ONE
Text on which Classifier function is used.

I **love** this movie! It’s **sweet**, but with **satirical** humor. The dialogue is **great** and the adventure scenes are **fun**… It manages to be **whimsical** and **romantic** while **laughing** at the conventions o the fairy tale genre I would **recommend** it to just about anyone. I’ve seen it **several** times and I’m always **happy** to see it **again** whenever I have a friend who hasn’t seen it yet.

STEP TWO
The function highlights the sentiment words.

x **love** xxxxxxxxxxxxxxx **sweet** xxxx xxxx **satirical** xxxxxxxxxxxxxxxxxx xx **great** xxxxxxxxxxxxxxxxxxxxxxx xxx **fun** xxxxxxxxxxxxxx **whimsical** xxx **romantic** xxxxx **laughing** xxxxxx xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx xxxxxx **recommend** xxxxxxxxxxxxxxxx xxxxxxxxxxxxxxx **several** xxxxxxxxxx xxxxxxxxxxx **happy** xxxxxxxx **again** xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx xxxxxxxxxxxx

STEP THREE
Non-sentiment words are discarded.

great	2
love	2
recommend	1
laugh	1
happy	1
. . .	. . .

STEP FOUR
Obtain a frequency of all sentiment words.

Figure 3-4: Steps in a sentiment classification process.

Entity extraction

You may be interested in specific elements in text such as names, places, or products mentioned. You can get them! That’s what entity extraction is all about.

Tagging

Want to make it easy to find particular parts of documents later? Maybe you’d like to develop an application to help people find research papers based on the authors, the funders, the institutions involved, or all of the above, but all you have to start with is unstructured text. It would be easier if those things were marked, but you’re not going to do that by hand.

Let the Data Discovery platform do it! Tagging is an automated process for marking specific parts of text to make subsequent use easier. Tagging is usually done with some form of extensible markup language (XML). XML is familiar to most developers today, and anyone who knows a bit about web development will find XML tags easy to interpret.

Reaping Benefits

Functional architecture and sophisticated analytics aren’t ends — they’re means. Data Discovery supports your business by providing valuable information for decision support, saving time in numerous ways, and by operating in a way that integrates smoothly with existing systems and business processes.

Knowledge is power

Data Discovery's most important benefit is information. When you have information that is accurate and relevant to your business, you have a solid basis for rational decision making. If you're lucky enough to have information that your competitors don't have, you have a distinct competitive edge.

Data Discovery, together with your own unique data resources and the insight of your staff, provide that edge.

Beyond the sum of parts

A cup of flour, by itself, doesn't make a satisfactory meal. It lacks many necessary nutrients; it's bland and dry and unappealing. Yet, combine flour with the right accompaniments, using the right techniques, and you can create bread, cake, and a world of other satisfying foods. When you have few data analysis options available, you don't have a satisfying basis for decision support. Bits and pieces of data and analysis just don't satisfy. It takes a full spectrum of analysis methods to develop a complete and satisfying story.

Customer retention is a significant concern for any customer-based industry (see Chapter 2 for a case study that discusses customer retention in banking). Telecommunications providers often face crushing challenges with customer churn. People switch providers to get new equipment, lower prices, or better service. But matching these concerns to individual customers can be difficult. And knowing what to offer to an individual customer to keep them happy has been almost impossible.

At least until Data Discovery comes on the scene. Data Discovery opens the door to a wide range of analytics to form a complete picture. So, when a telecommunications company confronts customer churn, it can use data from every available channel (such as the company website and call center, bricks and mortar store transactions, and social media), and extract meaningful information from many angles to form a clear understanding of the customer's point of view. See Figure 3-5 for a visual.

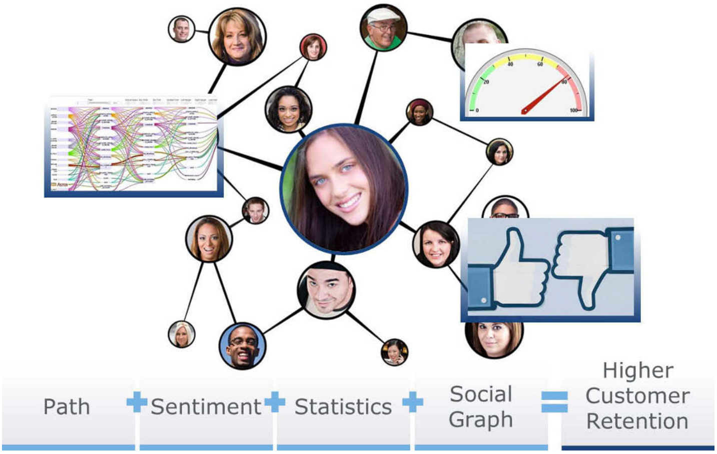


Figure 3-5: An ensemble of analytic methods provide a thorough understanding of your customer's behavior.

Just a few of the many elements that can be integrated into an ensemble of analytic methods to build customer understanding include:

- **Path analysis** to follow individuals as they move within and between channels, interacting with the telecommunications company and investigating options.
- **Text analysis** to reveal what customers are saying. Text forms a huge portion of today's big data sources, and can be found in call center records, social media, help requests, and other areas.
- **Classical statistics and data mining techniques** to build predictive models based on behavioral and other data.
- **Graph analysis** (also called *social graph analysis* or *social network analysis*) to unearth relationships among customers and identify those who are most influential.

You need speed

Time is money! You can save time through Data Discovery in many ways:

- Simple implementation
- Your staff already has the skills and tools needed to get Data Discovery off the ground

- ✓ Light maintenance
- ✓ Once implemented, Data Discovery relieves staff from constant requests for data
- ✓ Fast learning
- ✓ Users of all backgrounds can leverage what they know to become competent quickly
- ✓ Greater productivity
- ✓ The right interfaces reduce the time needed to set up analysis tasks
- ✓ Rapid execution

Play nicely with others

Wanna be heroic? Data Discovery lets people in differing jobs, with differing priorities and differing skills, make peace in the workplace.

Could you cobble together the same results with a mix and match solution? Nothing's impossible. But how long would it take to assemble? Are you certain those parts would work together? If they don't, what's your back-up plan? How hard would it be for the users to learn several new tools? And then there's the matter of optimization. Good luck perfecting that. A solid Data Discovery platform is a far more straightforward approach, and lower risk.

Chapter 4

Putting Data Discovery into Action

In This Chapter

- Talking about Data Discovery
- Finding new ways to operate
- Extending the reach of analytics

An ocean of data, the finest Data Discovery platform, and an armada of data analysts aren't worth a dime to your business unless you actually use the information they produce for decision making.

Just because you introduce a Data Discovery initiative, that doesn't mean that you can guarantee everyone in the organization will use it. This chapter goes over some ways to introduce the idea of Data Discovery and influence people to love it, including sweet talking and bribery.

Spreading the Good Word

Sometimes, companies introduce a new tool or process, expect the world of it, tell everyone it's the greatest thing since indoor plumbing, and it flops.

Perhaps you heard about the time tracking system for the sales department, the one nobody uses anymore? Nobody talked with the sales team about it, so the system wasn't set up to meet their needs. Maybe you remember a document management system that was so hard to use that people stopped sharing documents. Experiences like these make people wary of new things.

If the first thing people hear about Data Discovery is you telling them that they're going to use it, don't expect an enthusiastic response. Enthusiasm must be earned. The following sections give you some tips on how to talk to your co-workers about Data Discovery to make them more interested and invested right from the start.

Frame it right

Most people give little thought to Data Discovery; they think about their own work and problems. You can begin to interest people in Data Discovery even before you implement it or mention it by name. The key is to show interest in people and their problems, gather information about how those problems affect them, and use that information to tailor the messages you share. Developing interest early means more people will welcome Data Discovery when it comes time for the roll out.

When you talk about Data Discovery, try not to mention Data Discovery. Talk about helping the business. Address the needs and wants of the individual person whenever possible. Frame the information in terms of a business problem that concerns the listener. Remind the listener of the problem and its impact, then explain how the business problem can be addressed with the help of Data Discovery. This kind of reminder engages the listener, who will be pleased to hear you addressing his concern. And it sets the stage for introducing Data Discovery and motivating people to take advantage of the new capabilities it provides.

Ask before you tell

Before you can tell, you have to ask. You must approach each and every stakeholder in a spirit of cooperation and ask about something that's important to all of them: themselves. This is a technique frequently used in consultative sales.

Give people the chance to tell their stories and you'll learn things you need to know. An executive might reveal a need for better sales forecasts to avoid stock-outs that result in lost sales when customers don't find the items they want in stores. In an independent conversation, a marketer might explain that the sales team's forecasting process doesn't account for individual marketing promotions, leading to stock-outs on advertised items. The marketer and the executive have somewhat different concerns, but the two are clearly related, and each provides some depth in your understanding of the problem and what might address it.

Knowing who to talk to

Meet with people throughout your organization. Make sure to touch base with people from each of the major stakeholder groups. Use this sample list as a starting point and adapt it to fit your own situation:

- Business management
- Business operations
- Sales and marketing
- Customer service
- IT leadership
- Data analysts and business users
- Data scientists and statisticians
- Everybody else (people who don't use analytics directly, but whose work may be affected by the results)

TIP

Prepare a list of starter questions to begin your discussions. Aim to understand each person's everyday goals. Learn how different roles interact with one another, and how a delay or other problem in any one area affects others elsewhere in the business.

Moving forward

When the time comes to implement Data Discovery, tell people about it. Whenever you can, tell the story in terms of the issues uncovered in your meetings with stakeholders.

TIP

Analytics provides the real power in Data Discovery, but the term *analytics* can be a bit alarming and technical. Instead of discussing something that may put your listeners off, tell stories drawn from any analytics success that you company has already had, and say how they helped the business. Build on that by presenting issues uncovered in your interviews. And you can borrow stories, such as the use cases in this book, and others that you read. Don't oversell borrowed stories, though. People like to hear about things that seem close to home.

Emphasize the benefits of Data Discovery first:

- Business results and outcomes
- Better and more useful information
- Less drudgework and more time for other things
- Faster results
- Easier and more accessible analytics
- Reduced risk of error

If your stories convince people that these benefits are attainable, they will willingly listen to the details.

Nothing succeeds like success

After you launch your Data Discovery initiative, touch base with your users often. Gather success stories, even small ones, and spread the word.

REMEMBER

The most convincing case for the value of a new thing is the endorsement of a trusted person. People trust someone they know more than any stranger. The little wins of the person in the next cubicle are more enticing than the boldest news report or the fanciest marketing presentation.

Changing Habits

Your colleagues have invested years in developing the work methods that they use now. They have seen initiatives come and go. They've all had a few bad experiences. So don't expect them to drop everything they've done before and leap at the chance to change. Don't think ill of them for having doubts; it's healthy to have doubts. But people will make changes when they see good reasons to do so.

For instance, even devoted traditionalists use computers, because they can't resist the rewards: access to information, contact with like-minded people, easier ways to perform common tasks, and so on. People can and will change habits if the conditions are right. Your job is to make Data Discovery seem enticing.

Understanding habit

A habit has three elements: a cue, a routine, and a reward. If you want Data Discovery to become a routine, first investigate the cues and rewards associated with the routines your people have now.

A statistician gets a request from the marketing department. The marketers want information about the customers buying a particular product. The statistician requests an extract of relevant data from the IT department, waits for the data to arrive, then analyzes the data by writing code in a specialty programming language.

Although this is a work task, it's also a habit. First look at the elements of the statistician's current habit. The cue is an information request from the marketing department. The routine is requesting the data extract from IT and performing analysis in a particular way. The reward is a happy supervisor.

TIP

The object is to tie the cue, an information request, to new behavior: using the Data Discovery platform and process. The key is to associate an appealing reward to the desired behavior. This might sound like a psychology experiment. In fact, you may have to experiment to find the cues and rewards that work. The right mix differs from person to person. Data Discovery enables people to replace drudgework with more interesting and valuable activities, and many people will find that to be all the reward they need, but not everyone shares the same motivations.

Forming habits

What kinds of cues and rewards can form the basis of a Data Discovery habit? Cues usually take one of these forms:

- Location
- Time
- Emotional state
- Other people
- Actions that precede the routine

Emotional state might not be a viable (or desirable) choice of cue for developing a work habit (however it may very well be the cue for existing habits that you'll hope to eliminate). Time and actions can be good choices. Devoting the same time each day or week to Data Discovery makes sense, as does performing Data Discovery immediately after an appropriate action takes place, such as an update to key data.

A reward can be as natural as the enjoyment of interesting work. But people are different, and many people are more easily motivated by other things. The right reward for one person might be recognition; another might be more enthusiastic at the prospect of finishing work quickly and being able to go home on time.

TIP

Don't rule out the possibility of offering some concrete rewards, especially during the early phases as Data Discovery is introduced. It's amazing what people will do for a perk such as a crate of steaks, a gift card, or the latest electronic gadget. (Such gifts don't necessarily need to be expensive, but they should be thoughtfully matched to the tastes of the people involved. You might feel that fine Nebraska beef is the best treat ever, but if there are vegetarians on your team, gift cards might be the way to go.)

Infiltrating Everyday Business

You want Data Discovery to seep right into the fabric of the organization. You want to see healthy new work habits cropping up all around you, and everybody using data and analytics in their decision processes. Be prepared to work for it.

Actions speak loudly

Lead by example. Dive into Data Discovery yourself, as a hands-on user and to inform your own decisions. Your actions will send the loudest and clearest possible message about your own commitment to Data Discovery.

Your involvement will be much more than an exercise in public relations. By using Data Discovery yourself, you'll deepen your own understanding of it. You'll share the user experience, discover what's easy and what's challenging for you, and better empathize with the experience of others with differing skills and experience. You'll learn about analytic methods you've never encountered before and why they mean so much to business. And you'll confront the issues surrounding data-driven decision making head on.

REMEMBER

Your business will probably not yield to the siren call of Data Discovery overnight. You will have to reinforce your message of commitment to it every day over time to drive change. You committed to expand your core competencies through Data Discovery, and that means you must be actively involved for the long term.

Smoothing the way

Even people who are interested in Data Discovery and motivated to use it may not leap in right away. They may encounter difficulty using the Data Discovery platform hands-on. Don't let enthusiasm and interest fade away! Offer help.

REMEMBER

An SQL interface looks like a breeze to a data scientist who's accustomed to programming in complex languages such as Python or Pig, but it's a mystery to lots of other people. So although the primary users of the Data Discovery platform will work through an SQL programming interface, it's also valuable to have options for integrating Data Discovery with third-party or custom applications that are accessible to others.

Cultivating awareness

Imagine how difficult it might be to hang a picture on your wall if you'd never seen or heard of a hammer or nail. Think of how much time you might spend in failed attempts, and how long it might be before you succeeded. You might never succeed. And you might hurt yourself or damage tools that were never meant for the task.

Another thing: You can't fix a problem if you don't know it exists.

TIP

Your people can't solve problems with Data Discovery unless they're aware of problems facing the business and the range of capabilities available through Data Discovery. To build that awareness, you've got to get people talking and thinking.

Facilitating communication

Data Discovery is designed to bring analytical thinking into the workplace, to encourage the use of data in decision making, and to enable the involvement of people with differing skills and roles throughout the organization. It's powerful stuff, tailored to the complex, high-volume data challenges of today. It's the latest and the greatest, but it wasn't the first.

Data-driven decision making has been evolving for thousands of years. Thousands! The ancient Chinese used statistical groupings and economic forecasting more than 2,000 years ago. Egyptians kept detailed accounting records thousands of years before that. And that's good! It means you can borrow valuable ideas and practices from many people and places.

The field of manufacturing quality assurance has enjoyed many successes in involving diverse communities in data analysis. This is one of many great areas to look for ideas that you can adapt to your own workplace. You can draw inspiration from examples like these:

- Posting metrics and simple analyses (such as charts) in conspicuous places to build awareness and encourage discussion.
- Empower nonmanagement staff to make some decisions. Some automotive manufacturers empowered skilled laborers to shut down a production line if they felt quality issues necessitated it.
- Establishing certification systems based on educational attainment and experience, and providing recognition and professional opportunities for certified professionals.

You're not paid to be original, but to get things done, so collect ideas wherever you can find them! Look to professional societies, vendors, books, and other businesses for practices that might fit your situation. Even an old concept may be one that you haven't yet put into action yourself. Here are a few that are new, yet you may not have tried them for Data Discovery yet:

- Encourage the formation of user groups to develop Data Discovery skills.

- Introduce birds-of-a-feather gatherings, where groups of people with a shared business application, professional discipline, or other specific interests meet to network and share information.
- Sponsor contests that add an element of fun to the learning process. The object is to encourage people to try things and share experiences, so keep the spirit light. You might even offer little rewards for the best stories about data flops. Sharing these stories helps everyone accept that you can't win them all — there can be valuable lessons learned in everything you try.

Use case: Increasing customer lifetime value with product recommendations in retail

When you eat out, your server will always offer something extra. Would you like fries with that? You can upgrade that drink to the larger size for just 25 cents! Have a look at our dessert tray! These little extras make a lot of money for the restaurant, and it's not just restaurants that do this.

Retailers of all types look for add-on sales opportunities to build revenue and profits. When the retailer sells online, and offers thousands of products, add-on selling can be even more lucrative, but also far more complicated.

Penelope puts something in her cart. It's an electric pressure cooker. Ahana puts something in her cart. It's the latest fantasy novel from a bestselling author. Kyle puts something in his cart. It's a CD reissue of an obscure Beethoven recording. You can't just offer each these people a side of fries and a bucket of soda.

How would it be done face to face? The sales person would consider factors like these:

- What item has this person selected?
- Who is this person — an adult or a child, industrious or not, price conscious or extravagant?
- What do I have to offer — items similar to the selection, or popular with people like this?

Experienced salespeople often get very good at matching the offer to the person and building sales. How can this be replicated, or better yet, improved, online? One approach centers on the shopping cart. What's in it? What other things are

(continued)

(continued)

frequently purchased in combination with the items you've selected?

Ahana has put a popular fantasy novel in her cart. The web retailer's e-commerce system may use navigation aids to lead Ahana to other books by the same author. Enabling this kind of navigation may be just a simple matter of tagging listings with information such as an author name, brand, or product category (books, music, housewares, and so on).

Another tactic compares items in the cart to other carts. Many people who buy the new novel are also buying earlier works by the same author, or recent releases in the same genre, some are buying licensed products tied to the book's release, and others are buying items that aren't obviously related. Anything that is frequently purchased in combination with what's in Ahana's cart is a good choice for an add-on offer.

Perhaps your nephew in Cincinnati has a birthday coming up. You haven't seen him in a year. He's turning 10. He likes gaming and fighter planes, but he doesn't like to read.

You want him to read. And you want him to like your present. You might take the case to a children's librarian. The librarian would draw on years of experience with boys of the same age and similar tastes. She would suggest some titles with the right appeal (maybe *Horrible Histories* or an insider's guide to his favorite game).

Recommendations like that can also be simulated online. The principle is the same. Take the information you have about the person (not the cart) and match it to other people with similar histories, based on past purchases and product reviews, then suggest items that similar people bought and liked.

The cart-based approach is basket analysis. The process for matching with similar people is usually collaborative filtering. Both are widely used and good matches for Data Discovery.

Chapter 5

Avoiding Pitfalls

In This Chapter

- Thinking realistically
- Getting information into circulation
- Building credibility with easy wins

Make your life easier by thinking ahead about potential Data Discovery problems and taking preventive action early in the process. Prevention is much easier than remediation. This chapter gives you some tips on nipping problems in the bud.

Resisting Wishful Thinking

Big data success stories can be inspiring. And there are plenty of them! Just remember that stories have the icky bits edited out, and real life doesn't. Businesses that have great success with analytics may have stumbled many times on the way. And not every business that makes an investment in analytics is successful. What goes wrong? In most cases, the problem is wishful thinking.

Sounding like a plan

You want to retire one day, so you set goals, prepare, and save for that. You don't simply wish for a nice retirement, you plan and act accordingly.

TIP

You can maximize your chance of success with analytics by planning and then acting on your plan. This may not seem like front page news, yet it's the main difference between analytics success and failure. Don't just wish for success, make a plan and follow it.

Spreading a little sunshine

Data Discovery won't fly if the users won't use the platform. Hoping won't make it so. But there's an old saying, *you'll catch more flies with honey than vinegar*. People are much more likely to do what you want if you make it pleasant for them.

Make the interface fit the user

A person who has never written a line of code isn't going to be very motivated about Data Discovery if the only user interface you provide is SQL. But the same person might be delighted to use the platform through a menu-driven front end.

Sweeten the deal with automation

A power spreadsheet user feels comfortable with those tools, but spreadsheets have high error rates that create unnecessary business risks. Ease the transition to alternative tools by providing training and possibly consulting support to replace the spreadsheets with less risky alternatives. Make the proposition more attractive by lightening the workload; replace spreadsheets with automated processes when possible.

Don't ask the impossible

Transitions, even those that make life easier in the end, can seem frightening. Your people have devoted years of effort into learning the development processes that they use today, and they may have had some previous experiences with system changes that weren't successful, so you must respect their doubts. Also understand that even motivated and educated people need support to help them adopt new work methods. Help them through the transaction by

- Identifying opportunities for small wins to build confidence
- Providing training
- Allotting adequate time for learning and exploration with the Data Discovery platform

Data Discovery is an ongoing cycle

Data Discovery is an ongoing process. It's not a straight line that begins with a problem and ends with a permanent solution, but rather, a cycle. An issue may be examined, analyzed, and addressed again and again, with gradual improvement on each pass.

One of the important risks in Data Discovery is the possibility of becoming so involved in one part of the process, such as analysis or data gathering, that other necessary steps, such as measurement and evaluation, are neglected. Data Discovery isn't the individual steps — it's the ongoing cycle of activity, from identifying a problem to taking corrective action to assessment and reevaluation, as shown in Figure 5-1.

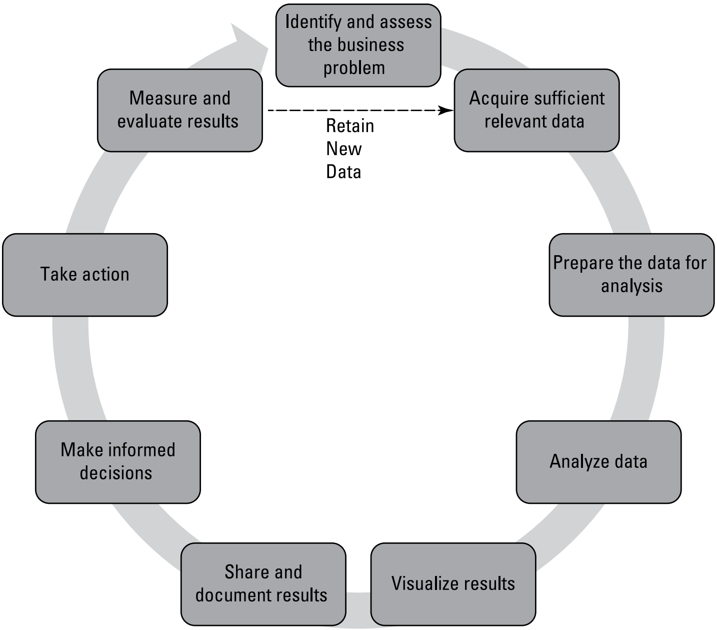

Identify and assess the business problem
Acquire sufficient relevant data
Prepare the data for analysis
Analyze data
Visualize results
Share and document results
Make informed decisions
Take action
Measure and evaluate results
Retain New Data

Figure 5-1: The Data Discovery process.

Identify and assess the business problem

Begin with a single, well-defined problem. Detailed information about choosing appropriate starter problems is given later in this chapter.

TIP

Review the problem and surrounding issues with input from all stakeholders, if possible. How is the problem affecting the business? How much are the potential savings? Have possible solutions been proposed? How could you evaluate those alternatives quantitatively? Involving stakeholders early in the process builds comfort and commitment.

Gather sufficient relevant data

Often, the data you need to investigate a problem is already on hand. Even if it has been collected through disconnected activities and stored in disparate systems, the Data Discovery platform lets you bring it together.

Even a massive data store doesn't necessarily contain everything you need. If gaps exist, consider your options.

TIP

You might be able to scale back your investigation and work with what you have, or obtain what you need by conducting an experiment or survey. Data vendors offer a remarkable range of information, so you may be able to buy the data, especially if what you need is demographic or consumer-behavior data.

Prepare the data for analysis

Most data analysts spend most of their time preparing data for analysis. Data preparation doesn't disappear in Data Discovery, but it can be much easier than with traditional methods.

Data Discovery can remedy data preparation problems by making the process simpler and reducing steps. Users may even be able to combine data preparation steps with analysis in the same query, and leave the details to the utilities within the Data Discovery framework.

Explore and analyze data

This is where the fun comes in. Analysts of every stripe, whether they call themselves data scientists, data analysts, or any of a hundred or more other titles, find the data analysis phase a lot more enjoyable than data preparation.

Simplified data access and preparation leaves more time for analysis. And with more time, analysts are able to work more carefully and in greater depth. Specifics on a wide variety of analytics available for Data Discovery and some popular uses for them appear in Chapter 3.

Share and document results

Valuable discoveries must be shared, particularly with those who are in a position to put the information to work. This is a place where the process often breaks down.

Ensure that a pathway exists for sharing the results of data analysis with the appropriate people. Creating these pathways should be a part of the implementation process for Data Discovery.

When communication breaks down, the problem may not be lack of effort or lack of access to decision makers. Many data analysts, indeed many people, just don't have great communication skills or the right style for the audience.

Data scientists, in particular, are often asked to be nearly superhuman in their range of skills. Some people describe them as part programmer, part statistician, part storyteller, and more. But real human beings are rarely gifted in every way. So consider taking advantage of some of the time and money that you save with Data Discovery, and investing it in communication training for data analysts.

Make informed decisions

Business culture romanticizes the risk-taking business leader who makes decisions from the gut. At least, that type of leader gets plenty of attention in the media, provided that the outcome is good. Leaders can and do sometimes build iconic images for themselves by spinning good stories that portray them as insightful risk takers whose intuition is superior to anything data could produce.

Managers who see themselves in this way may be reluctant to make decisions based on data. That won't change overnight. Like a bird's nest, change is built from bits and pieces. Sometimes, a particularly compelling analysis can be persuasive. The analytics-driven success of a colleague or competitor can make a difference, as can a bad experience with a decision made without (or in defiance of) analytics.

Take action

Putting a decision into action actually starts long before the decision is made, even before data is analyzed. The most important prerequisite to turning a decision into reality is choosing the right problem. Imagine possible solutions early and find out whether you can take the actions that might be proposed.

There may be legal issues to consider, and matters of safety and privacy. An action that solves one problem might create another. The manager who has the power to take a particular action may not be willing to do so, and may have good reasons for that. Give thought to all such issues from the beginning, so that you can put your Data Discovery effort into the problems that have the best potential to be solved.

Do a thorough analysis and make a convincing presentation of the case for any action you propose. Presentations should be backed up with written reports and other documentation.

Measure and evaluate results

You made a decision and took action, what were the results?

Informal evaluation of results isn't enough. Define the metrics you mean to influence, then measure and record the results. Retain the information you gather. Be willing to admit that some things don't work as you thought they would and move on.

TIP

When your actions produce great results, as they often will, you will have excellent evidence not only of success, but also of where to put the credit.

Retain the new data that you create

Make and keep detailed notes at every stage of the Data Discovery process. Make sure measurements and other new data are retained and made accessible. This information may be the basis of your future Data Discovery projects, or it may fulfill legal or business process requirements. It helps everyone explain and understand how and why particular things were done. Time invested in documentation saves a lot more time later!

The new information you create through Data Discovery may even become your next product: a data product.

Biting Off What You Can Chew

When you learn new things in your personal life, you take on modest projects, not colossal ones, at first. The first meal you cook is tomato soup from a can, not bouillabaisse. You build a lamp, not a house. You sew an apron, not a wedding gown.

TIP

If analytics is new to you and others in your company, don't go after big and complex projects first, even if they promise big returns. Start with the projects you find easiest and give yourself the best opportunity you can to enjoy a little success.

No matter what your level of experience with analytics, your best starting point in Data Discovery will be with projects just a step more complex than what you've been able to do before.

Set attainable goals (and then monitor them!)

Look for starter projects that:

- Are only slightly more complex than those you've been able to handle without Data Discovery.
- Involve data sources that have not been brought together before.
- Offer an obvious, measurable reward if solved. Cost reduction is usually the most appealing reward, so give top priority to those projects.

Refer to Chapter 1 for a use case that discusses a telecommunications company who used Data Discovery to investigate one specific customer behavior that wasn't well understood: customers calling customer support immediately after making a payment online. The behavior was already known to be a common pattern, but without the Data Discovery platform to bring together disparate data sources, the company had never been able to understand the cause.

This is a model example of a good starter Data Discovery project. It was a narrowly defined problem, yet one that the company had been unable to solve without Data Discovery. The solution depended on bringing together certain data sources for the first time, hence showing off the unique capabilities provided by the Data Discovery platform. And because the telecommunications company knew the cost of a call to customer service and the number of calls at issue, it was a simple matter to estimate the cost savings potential of a solution.

Walk before you run

Why limit yourself to small projects when bigger problems are waiting to be solved? Well, if you want to eat an elephant, you must do it one bite at a time.

When you face a problem that seems big and complex, you may find it very difficult to tackle as a whole. You could put in a lot of time and effort, yet still be overwhelmed. And you might not have any clearly measurable value to show for your time and investment.

TIP

Try to think of a big problem as the result of many smaller problems acting together. Dissect the big problem into little ones, and take those on one at a time, starting with the one you find easiest to solve. Or split the work across teams, matching the problems to the team with the skills most suitable for finding a solution.

This approach yields quick wins (successes) from the start. You'll feel good, you'll have clear results to show for your trouble, and you'll build credibility.

Credibility is your path to trust. If you've demonstrated success in Data Discovery repeatedly, you'll be free to take on more difficult and riskier projects. If sometimes you fail, you will still be trusted because you have succeeded so many times before.

Chapter 6

Ten (Okay, 11) Teradata Resources for More Information

In This Chapter

- Hearing from people who use Data Discovery
- Peeking under the hood
- Getting the scoop from a bestselling author and expert

You're salivating at the prospect of easier, faster, and deeper analytics, aren't you? You want to know more! You want to hear from people who've been there, scrutinize the details and get a better understanding of how Data Discovery could work for you. This chapter directs you to some sources where you can find just that.

Virtual Event: Data Discovery in Action

To see Data Discovery in action, check out Teradata's virtual event. You will find information about what Data Discovery does and how it can help your business. You'll also see some real-world use cases. Check it out at `www.teradata.com/discovery`.

User Story: Verizon Wireless

America's largest wireless carrier explains how a college intern uncovered valuable new customer intelligence in just a few weeks. Hear about the unexpected customer behavior that couldn't be explained until bringing it together through Teradata's Unified Data Architecture and Data Discovery Platform that enabled Verizon to actually hear what was on the customers' mind. Find the story at www.teradata.com/verizon.

User Story: Cardinal Health

A Fortune 500 healthcare services company optimizes the supply chain that serves over 60,000 locations. Hear how some users cut the time needed for working with raw data by 50 percent, leaving them time to move past reporting to do more research. Find the story at www.teradata.com/cardinal.

User Story: McCain Foods

The world's biggest maker of frozen French fries motivates everyone to get involved by giving you a view of performance data. (Next on the agenda: addressing food shortages worldwide!) Go to www.teradata.com/mccain.

Use Case: Grow Loyalty of Influential Customers

Discovering which of your customers are most influential is the start of building their loyalty. Find out more at www.teradata.com/influential.

Whitepaper: The Rise of Data Discovery

Thomas H. Davenport, President's Distinguished Professor in Information Technology and Management at Babson College is the author of the upcoming book *Big Data at Work* (Harvard Business Review Press, 2014) and many other books on IT and business, including the bestselling *Competing on Analytics: The New Science of Winning* (Harvard Business Review Press, 2007).

In a whitepaper, Davenport summarizes learning from his interviews with early adopters of Data Discovery. He outlines motivations to engage in Data Discovery and typical applications, and explains the attributes of an effective platform and process, as well as barriers that must be overcome. Find out more at `www.teradata.com/tomdavenport`.

Whitepaper: Extending Analytics to Nonrelational Data

Traditional analytics methods suit data that is structured in traditional ways. Learn how nonrelational data is different, why traditional analysis methods don't suit it, and how to approach analysis of nonrelational data. Find out more at `www.teradata.com/extendinganalytics`.

Whitepaper: Teradata Aster Discovery Portfolio

Find out about the full range of analytics available in the Teradata Aster Discovery Portfolio. Top techniques are described in depth, complete with code examples. Go to `www.teradata.com/portfolio`.

Whitepaper: Data-Driven Marketing

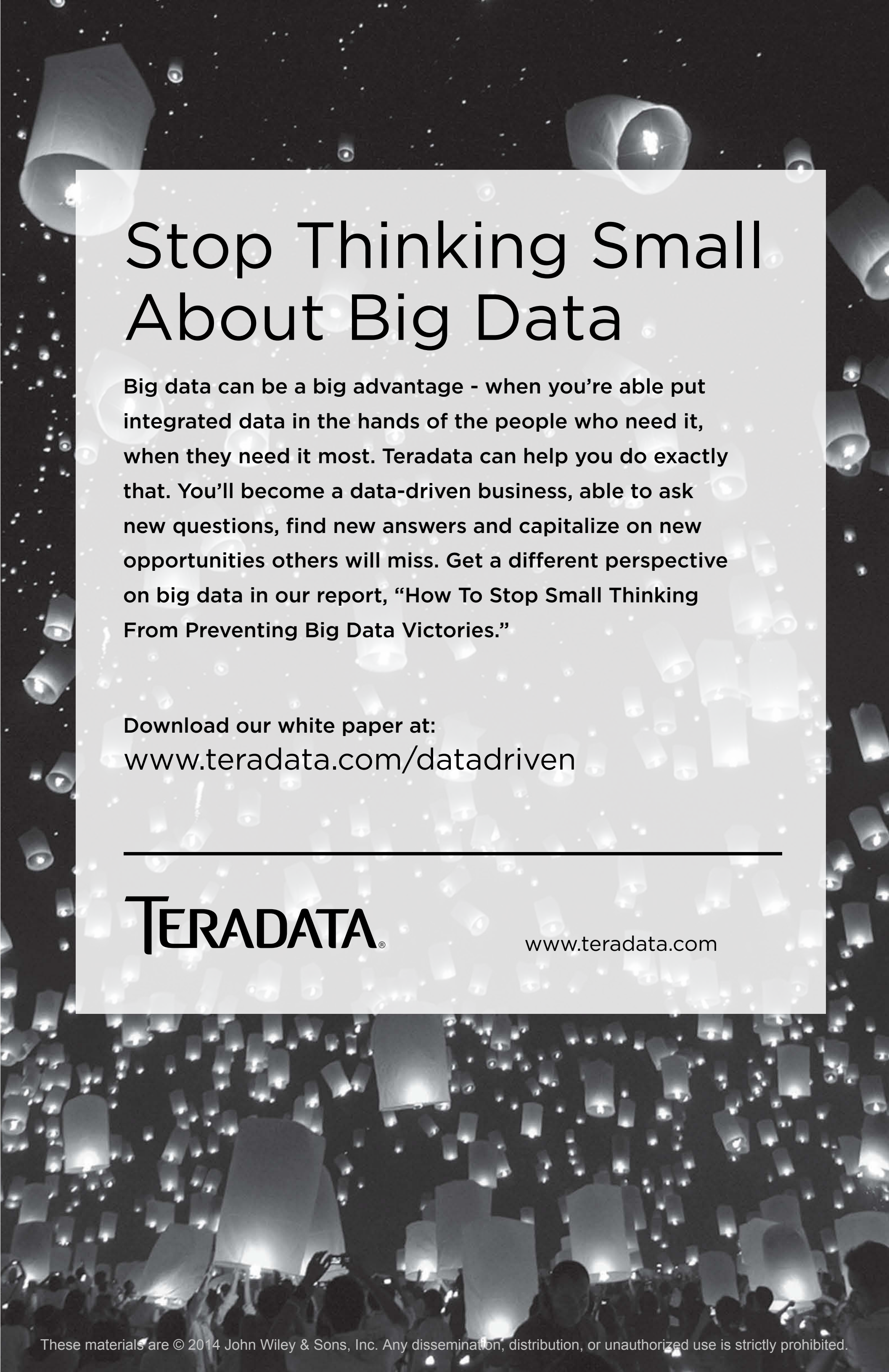
Learn why marketers can get better information about customer behavior now than ever, and the benefits of a data-driven marketing approach. Go to `www.teradata.com/marketing`.

Product Information: Teradata Aster Discovery Platform

Go to the starting point for information on the Teradata Aster Discovery Platform. Find it at `www.teradata.com/DiscoveryPlatform`.

Talk to Teradata

Not finding just the right information from these sources? Time to get on the horn and ask for what you need! Contact Teradata at 1.888.278.3732 or `www.teradata.com/contact-us`.



Learn how Data Discovery gives the power of new insights to more people, with less fuss!

Overwhelmed by the data tsunami and rapidly growing demands for big data analytics and discovery? Data Discovery to the rescue! Much more than another data analysis tool, Data Discovery lets you get that data under control and puts it to work to drive everyday business decisions.

- ***Expanding core competencies** — address new expectations surrounding big data and a changing business climate*
- ***Building Data Discovery capabilities** — empowering people throughout your organization with the right structure and preparation*
- ***Putting Data Discovery into action** — infiltrating routine business with the resources needed for data-driven decision making*
- ***Avoiding pitfalls** — steering clear of wishful thinking and other common problems*

Meta S. Brown is a consultant, speaker, and writer who helps businesses make better decisions with the help of data and the right analysis methods. She is the author of *Data Mining For Dummies*.

Open the book and find:

- **Simplifying big data analytics and discovery so you can get more value from your data**
- **Integrating Data Discovery into everyday business for maximum impact**
- **Demystifying behavior analysis and machine learning techniques and putting them to work**
- **Getting familiar with Data Discovery platforms and architecture**

Go to Dummies.com®
for videos, step-by-step examples, how-to articles, or to shop!

FOR DUMMIES®
A Wiley Brand

ISBN: 978-1-118-91272-0
Not for resale