GLOBAL EDITION

Business Intelligence, Analytics, and Data Science

A Managerial Perspective

FOURTH EDITION

Ramesh Sharda • Dursun Delen • Efraim Turban

Pearson

FOURTH EDITION
GLOBAL EDITION

BUSINESS INTELLIGENCE, ANALYTICS, AND DATA SCIENCE:

A Managerial Perspective

Ramesh Sharda
Oklahoma State University

Dursun Delen
Oklahoma State University

Efraim Turban
University of Hawaii

With contributions to previous editions by

J. E. Aronson
The University of Georgia

Ting-Peng Liang
National Sun Yat-sen University

David King
JDA Software Group, Inc.

Pearson

Harlow, England • London • New York • Boston • San Francisco • Toronto • Sydney • Dubai • Singapore • Hong Kong
Tokyo • Seoul • Taipei • New Delhi • Cape Town • Sao Paulo • Mexico City • Madrid • Amsterdam • Munich • Paris • Milan

CHAPTER

3

Descriptive Analytics II: Business Intelligence and Data Warehousing

LEARNING OBJECTIVES

- Understand the basic definitions and concepts of data warehousing
- Understand data warehousing architectures
- Describe the processes used in developing and managing data warehouses
- Explain data warehousing operations
- Explain the role of data warehouses in decision support
- Explain data integration and the extraction, transformation, and load (ETL) processes
- Understand the essence of business performance management (BPM)
- Learn balanced scorecard and Six Sigma as performance measurement systems

The concept of data warehousing has been around since the late 1980s. This chapter provides the foundation for an important type of database, called a *data warehouse*, which is primarily used for decision support and provides the informational foundation for improved analytical capabilities. We discuss data warehousing concepts and, relatedly, business performance management in the following sections.

3.1 OPENING VIGNETTE: Targeting Tax Fraud with Business Intelligence and Data Warehousing

Governments have to work hard to keep tax fraud from taking a significant bite from their revenues. In 2013, the Internal Revenue Service (IRS) successfully foiled attempts, which were based on stolen identities, to cheat the federal government out of $24.2 billion in tax refunds. However, that same year the IRS paid out $5.8 billion on claims it only later identified as fraud.

States also lose money when fraudsters use stolen Social Security numbers, W-2 forms, and other personal information to file false refund claims. This kind of crime has increased in recent years at an alarming rate. "Virtually all Americans have heard of identity theft, but very few are aware of this explosive increase in tax return fraud," says Maryland Comptroller Peter Franchot. "This is an alarming problem, affecting every state. It is, literally, systematic burglary of the taxpayer's money."

In Maryland, the people charged with rooting out false refund claims are members of the Questionable Return Detection Team (QRDT). Like their counterparts in many other states, these experts use software to identify suspicious returns. They then investigate the returns to pinpoint which ones are fraudulent.

Challenge

In the past, Maryland used metrics that examined tax returns one by one. If a return displayed specific traits—for instance, a certain ratio of wages earned to wages withheld—the software suspended that return for further investigation. Members of the QRDT then researched each suspended return—for example, by comparing its wage and withholding information with figures from a W-2 form submitted by an employer. The process was labor intensive and inefficient. Of the approximately 2.8 million tax returns Maryland received each year, the QRDT suspended about 110,000. But most of those turned out to be legitimate returns. "Only about 10% were found to be fraudulent," says Andy Schaufele, director of the Bureau of Revenue Estimates for the Maryland Comptroller.

In a typical year, that process saved Maryland from mailing out $5 million to $10 million in fraudulent refunds. Although that's a success, it's only a modest one, considering the resources tied up in the process and the inconvenience to honest taxpayers whose returns were flagged for investigation. "The thought that we were holding up 90,000 to 100,000 tax refunds was tough to stomach," Schaufele says. "We wanted to get those refunds to the taxpayers faster, since many people count on that money as part of their income."

Solution

Maryland needed a more effective process. It also needed new strategies for staying ahead of fraudsters. "All the states, as well as the IRS, were using the same metrics we were using," Schaufele says. "I don't think it was hard for criminals to figure out what our defenses were." Fortunately, Maryland had recently gained a powerful new weapon against tax fraud. In 2010, the Maryland Comptroller of the Treasury worked with Teradata

of Dayton, Ohio, to implement a data warehouse designed to support a variety of compliance initiatives.

As officials discussed which initiatives to launch, one idea rose to the top. "We determined that we should prioritize our efforts to go after refund fraud," says Sharonne Bonardi, Maryland's deputy comptroller. So the state started working with Teradata and with ASR Analytics of Potomac, Maryland, to develop a better process for isolating fraudulent tax returns (Temple-West, 2013).

"The first step was to analyze our data and learn what we knew about fraud," Schaufele says. Among other discoveries, the analysis showed that when multiple returns were suspended—even for completely different reasons—they often had traits in common. The state built a database of traits that characterize fraudulent returns and traits that characterize honest ones. "We worked with ASR to put that information together and develop linear regressions," Schaufele says. "Instead of looking at one-off metrics, we began to bring many of those metrics together." The result was a far more nuanced portrait of the typical fraudulent return.

Instead of flagging returns one by one, the new system identifies groups of returns that look suspicious for similar reasons. That strategy speeds up investigations. The analytics system also assigns a score to each return, based on how likely it is to be fraudulent. It then produces a prioritized list to direct the QRDT's workflow. "We're first working on the returns that are more likely not to be fraudulent, so we can get them out of the queue," Schaufele says. The more suspicious-looking returns go back for further review.

Results

"With these analytics models, we're able to reduce false positives, so that we don't overburden the taxpayers who have accurately reported their information to the state," Bonardi says. Once investigators remove their returns from the queue, those taxpayers can get their refunds.

Thanks to the new technology, QRDT expects to suspend only 40,000 to 50,000 tax returns, compared with 110,000 in past years. "Of those we've worked so far, we're getting an accuracy rate of about 65%," says Schaufele. That's a big improvement over the historical 10% success rate. "Once the returns are identified which may be fraudulent, the team of expert examiners can then carefully review them, one at a time, to eliminate returns that are found to be legitimate," Maryland Comptroller Franchot says. "The entire operation is getting better and stronger all the time."

As of late March, advanced analytics had helped the QRDT recover approximately $10 million in the current filing season. Schaufele says, "Under the old system, that number would have been about $3 million at this point." Not only does the new technology help the QRDT work faster and more efficiently, but it also helps the team handle a heavier and more complex workload. As tax criminals have ramped up their efforts, the QRDT has had to deploy new strategies against them. For example, in 2015 the team received some 10,000 notifications from taxpayers whose identifications had been stolen. "So we have a new workflow: We look up their Social Security numbers and try to find any incidences of fraud that might have been perpetrated with them," says Schaufele. "That's a new level of effort that this group is now completing without additional resources."

To stay ahead of more sophisticated tax schemes, investigators now not only examine current W-2 forms, but also compare them with the same taxpayers' forms from prior years, looking for inconsistencies. "The investigations are becoming more complex and taking longer," Schaufele says. "If we hadn't winnowed down the universe for review, we would have had some real problems pursuing them."

QUESTIONS FOR THE OPENING VIGNETTE

1. Why is it important for IRS and for U.S. state governments to use data warehousing and business intelligence (BI) tools in managing state revenues?
2. What were the challenges the state of Maryland was facing with regard to tax fraud?
3. What was the solution they adopted? Do you agree with their approach? Why?
4. What were the results that they obtained? Did the investment in BI and data warehousing pay off?
5. What other problems and challenges do you think federal and state governments are having that can benefit from BI and data warehousing?

What We Can Learn from This Vignette

The opening vignette illustrates the value of BI, decision support systems, and data warehousing in management of government revenues. With their data warehouse implementation, the State of Maryland was able to leverage its data assets to make more accurate and timely decisions on identifying fraudulent tax returns. Consolidating and processing a wide variety of data sources within a unified data warehouse enabled Maryland to automate the identification of tax fraud signals/rules/traits from historic facts as opposed to merely relying on traditional ways where they have been implementing intuition-based filtering rules. By using data warehousing and BI, Maryland managed to significantly reduce the false positive rate (and by doing so ease the pain on the part of taxpayers) and improved the prediction accuracy rate from 10% to 65% (more than a sixfold improvement in accurate identification of fraudulent tax returns). The key lesson here is that a properly designed and implemented data warehouse combined with BI tools and techniques can and will result in significant improvement (both on accuracy and on timeliness) resulting in benefits (both financial and nonfinancial) for any organization, including state governments like Maryland.

Sources: Teradata case study. (2016). Targeting tax fraud with advanced analytics. http://assets.teradata.com/resourceCenter/downloads/CaseStudies/EB7183_GT16_CASE_STUDY_Teradata_V.PDF (accessed June 2016); Temple-West, P. (2013, November 7). Tax refund ID theft is growing "epidemic": U.S. IRS watchdog. Reuters. http://www.reuters.com/article/us-usa-tax-refund-idUSBRE9A61HB20131107 (accessed July 2016).

3.2 Business Intelligence and Data Warehousing

Business intelligence (BI), as a term to describe evidence/fact-based managerial decision making, has been around for more than 20 years. With the emergence of business analytics as a new buzzword to describe pretty much the same managerial phenomenon, the popularity of BI as a term has gone down. As opposed to being an all-encompassing term, nowadays BI is used to describe the early stages of business analytics (i.e., descriptive analytics).

Figure 3.1 (a simplified version of which was shown and described in Chapter 1 to describe business analytics taxonomy) illustrates the relationship between BI and business analytics from a conceptual perspective. As shown therein, BI is the descriptive analytics portion of the business analytics continuum, the maturity of which leads to advanced analytics—a combination of predictive and prescriptive analytics.

Descriptive analytics (i.e., BI) is the entry level in the business analytics taxonomy. It is often called business reporting because of the fact that most of the analytics activities at this level deal with creating reports to summarize business activities to answer questions such as "What happened?" and "What is happening?" The spectrum of these reports

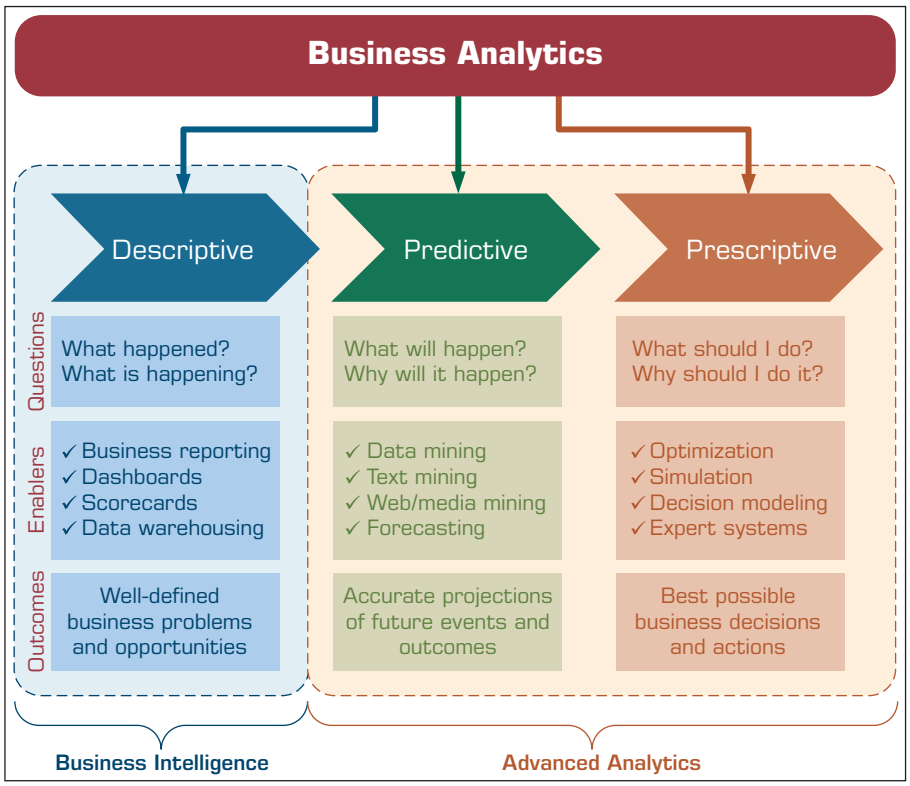

FIGURE 3.1 Relationship between Business Analytics and BI, and BI and Data Warehousing.

includes static snapshots of business transactions delivered to knowledge workers (i.e., decision makers) on a fixed schedule (e.g., daily, weekly, quarterly); ad hoc reporting where the decision maker is given the capability of creating his or her own specific report (using an intuitive drag-and-drop graphical user interface) to address a specific or unique decision situation; and dynamic views of key business performance indicators (often captured and presented within a business performance management system) delivered to managers and executives in an easily digestible form (e.g., dashboard-looking graphical interfaces) on a continuous manner.

Generally speaking, and as depicted in Figure 3.1, BI systems rely on a data warehouse as the information source for creating insight and supporting managerial decisions. A multitude of organizational and external data is captured, transformed, and stored in a data warehouse to support timely and accurate decisions through enriched business insight. This chapter aims to cover the concepts, methods, and tools related to data warehousing and business performance management.

What Is a Data Warehouse?

In simple terms, a **data warehouse** (**DW**) is a pool of data produced to support decision making; it is also a repository of current and historical data of potential interest to managers throughout the organization. Data are usually structured to be available in a form ready for analytical processing activities (i.e., online analytical processing [OLAP], data mining, querying, reporting, and other decision support applications). A data warehouse is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management's decision-making process.

A Historical Perspective to Data Warehousing

Even though *data warehousing* is a relatively new term in information technology (IT), its roots can be traced back in time, even before computers were widely used. In the early 1900s, people were using data (though mostly via manual methods) to formulate trends to help business users make informed decisions, which is the most prevailing purpose of data warehousing.

The motivations that led to the development of data warehousing technologies go back to the 1970s, when the computing world was dominated by mainframes. Real business data-processing applications, the ones run on the corporate mainframes, had complicated file structures using early-generation databases (not the table-oriented relational databases most applications use today) in which they stored data. Although these applications did a decent job of performing routine transactional data-processing functions, the data created as a result of these functions (such as information about customers, the products they ordered, and how much money they spent) was locked away in the depths of the files and databases. When aggregated information such as sales trends by region and by product type was needed, one had to formally request it from the data-processing department, where it was put on a waiting list with a couple of hundred other report requests (Hammergren & Simon, 2009). Even though the need for information and the data used to generate it existed, the database technology was not there to satisfy it. Figure 3.2 shows a timeline where some of the significant events that led to the development of data warehousing are shown.

Later in the last century, commercial hardware and software companies began to emerge with solutions to this problem. Between 1976 and 1979, the concept for a new company, Teradata, grew out of research at the California Institute of Technology (Caltech), driven from discussions with Citibank's advanced technology group. Founders worked to design a database management system for parallel processing with multiple microprocessors, targeted specifically for decision support. Teradata was incorporated on July 13, 1979, and started in a garage in Brentwood, California. The name *Teradata* was chosen to symbolize the ability to manage terabytes (trillions of bytes) of data.

The 1980s were the decade of personal computers and minicomputers. Before anyone knew it, real computer applications were no longer only on mainframes; they were all over the place—everywhere you looked in an organization. That led to a portentous problem called *islands of data*. The solution to this problem led to a new type of software, called a *distributed database management system*, which would magically pull the

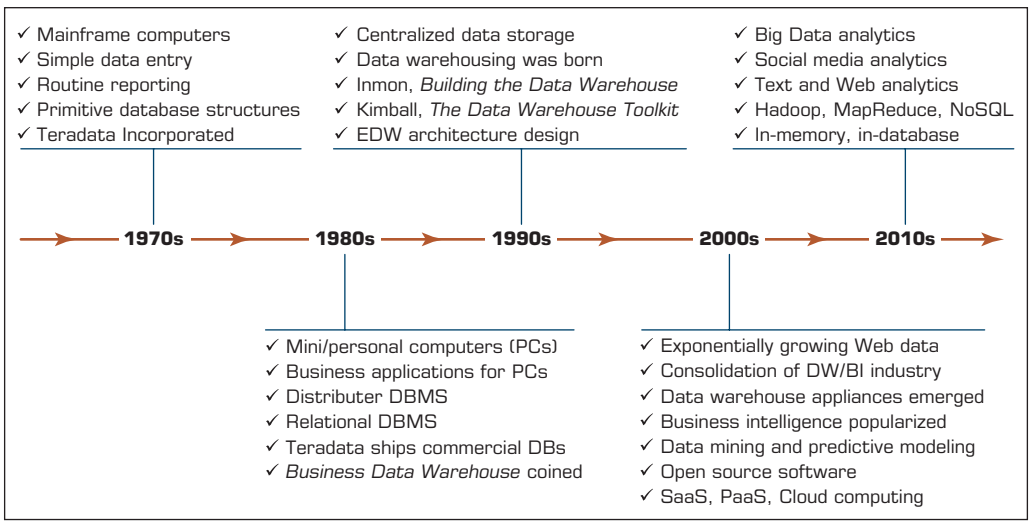

FIGURE 3.2 A List of Events That Led to Data Warehousing Development.