Communications of the Association for Information Systems

Volume 8 Article 1

January 2002

Recent Developments in Data Warehousing

Hugh J. Watson
University of Georgia, hwatson@terry.uga.edu

Recommended Citation
Watson, Hugh J. (2002) "Recent Developments in Data Warehousing," *Communications of the Association for Information Systems*: Vol. 8 , Article 1.
DOI: 10.17705/1CAIS.00801
Available at: https://aisel.aisnet.org/cais/vol8/iss1/1

Communications of the Association for Information Systems

RECENT DEVELOPMENTS IN DATA WAREHOUSING

Hugh J. Watson
Terry College of Business
University of Georgia
E-mail: hwatson@terry.uga.edu

ABSTRACT

Data warehousing is a strategic business and IT initiative in many organizations today. Data warehouses can be developed in two alternative ways -- the data mart and the enterprise-wide data warehouse strategies -- and each has advantages and disadvantages. To create a data warehouse, data must be extracted from source systems, transformed, and loaded to an appropriate data store. Depending on the business requirements, either relational or multidimensional database technology can be used for the data stores. To provide a multidimensional view of the data using a relational database, a star schema data model is used. Online analytical processing can be performed on both kinds of database technology. Metadata about the data in the warehouse is important for IT and end users. A variety of data access tools and applications can be used with a data warehouse – SQL queries, management reporting systems, managed query environments, DSS/EIS, enterprise intelligence portals, data mining, and customer relationship management. A data warehouse can be used to support a variety of users – executives, managers, analysts, operational personnel, customers, and suppliers. Data warehousing concepts are brought to life through a case study of Harrah's Entertainment, a firm that became a leader in the gaming industry with its CRM business strategy supported by data warehousing.

Keywords: Data warehouse, data warehousing, data mart, operational data store, data warehousing architecture, development methodology, ETL processes, metadata, relational and multidimensional databases, star schema, online analytical processing (OLAP), data access tools and applications, end users, customer relationship management (CRM), Harrah's Entertainment

I. INTRODUCTION

Data warehousing is one of the most important strategic initiatives in the information systems field [Eckerson, 1999]. These repositories of data play critical roles in understanding customer behavior in customer to business e-commerce, connecting trading partners along the supply chain, implementing customer relationship management strategies, and supporting comprehensive performance measurement systems, such as balanced scorecards. The most recent data that we found was a prediction by the Palo Alto Management Group that the data warehousing market will grow to $113.5 billion in 2002, including the sales of systems, software, hardware, services, and in-house expenditures [Eckerson, 1998].

Most fundamentally, a data warehouse is created to provide a dedicated source of data to support decision-making applications [Gray and Watson, 1998]. Rather than having data

scattered across a variety of systems, a data warehouse integrates the data into a single repository. It is for this reason that a data warehouse provides "a single version of the truth." All users and applications access the same data. Because users access better data, their ability to analyze data and make decisions improves.

This article is a tutorial on data warehousing that introduces the newest concepts in the field. First, key definitions and concepts are presented (Sec. II and III). The two strategies or approaches to building a data warehouse -- the data mart and enterprise-wide data warehouse strategies are discussed next (Sec. IV). A data warehouse requires that data be extracted from source systems, transformed, and loaded into the warehouse. These ETL processes, as they are commonly called, are described in Section V. Metadata (Sec. VI), which is data about data, is important to warehousing, because it helps IT personnel and users understand and work with the data in the warehouse. Next, the options for storing data, which include relational and multidimensional databases, are considered (Sec. VII). The value of a warehouse occurs when users and applications make use of the data in the warehouse. Warehouse users include executives, managers, analysts, operational personnel, suppliers, and customers. Applications include reporting, online analytical processing, DSS/EIS, and data mining. All of these are discussed in Section VIIII. To illustrate how organizations employ data warehousing, a case study of Harrah's Entertainment, a leader in data warehousing, is presented in Section IX. A conclusion is provided in Section X.

II. BASIC DEFINITIONS AND CONCEPTS

Most simply, a **data warehouse** is a collection of data created to support decision making. Users and applications access the warehouse for the data that they need. A warehouse provides a data infrastructure. It eliminates a reason for the failure of many decision support applications – the lack of quality data.

A data warehouse has the following four **characteristics** (Inmon, 1992):

Subject oriented. Warehouse data is organized around specific subjects, such as sales, customers, or products. This arrangement is different from transactional systems where data is organized by business process, such as order entry, inventory control, or accounts receivable.

Integrated. Data are collected from multiple systems and are integrated around subjects. For example, customer data may be extracted from internal (and external) systems and integrated around a customer identifier so that a comprehensive view of the customer is created.

Time variant. A warehouse maintains historical data (i.e., it includes time as a variable). Unlike transactional systems, where only recent data, such as for the last day, week, or month, are maintained, a warehouse may store years of data. Historical data is needed to detect deviations, trends, and long-term relationships.

Nonvolatile. A warehouse is nonvolatile – users cannot change or update the data. Non-volatilty makes sure that all users are working with the same data. The warehouse is updated, but through IT controlled load processes rather than by users.

Whereas a data warehouse is a repository of data, **data warehousing** is the entire process. As shown in Figure 1, data warehousing encompasses a broad range of activities: all the way from extracting data from source systems to the use of the data for decision-making purposes. Specifically, it includes data extraction, transformation, and loading, the access of the data by end users, and applications.

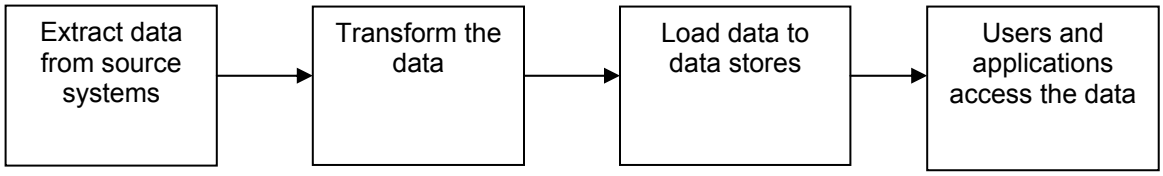

Figure 1: The Data Warehousing Process

A **data mart** is similar to a data warehouse, except a data mart stores data for a limited number of subject areas, such as marketing or sales data. Because it is smaller in scope than a data warehouse, it also tends to contain less data and support fewer applications.

Data marts can be either independent or dependent.

- An **independent data** mart is built directly from source systems. Just like a data warehouse, data are extracted, transformed, and loaded to the data mart.
- A **dependent data mart** is created with data drawn from a data warehouse. It provides a copy of data extracted from the data warehouse (although additional transformations, such as computing summaries may be performed prior to loading).

An independent data mart is often a "point solution" which, while solving an immediate problem, creates a new problem long term. For example, a department may need data for particular applications and have the resources to build a data mart on their own. While the independent data mart may prove successful, it can plant the seeds for problems later when the organization tries to create a compatible enterprise-wide data infrastructure.

A dependent data mart is created to give users a customized view of specific data that they need. For example, financial data may be placed on a dependent data mart for use by a company's financial analysts. Often data from the corporate warehouse is supplemented with data used locally only by the user group. A dependent data mart typically provides faster response times to queries than the data warehouse from where the data originally came. From an enterprise perspective, dependent data marts are much preferred over independent marts because the data comes from a source that was built from an organization-wide perspective. the "single version of the truth" is maintained. Everyone is working with the same data, even though it may be housed in different places.

An **operational data store** (ODS) consolidates data from multiple source systems and provides a near real-time, integrated view of volatile, current data. The extraction, transformation, and loading processes for an operational data store are the same as for a data warehouse. An ODS differs from a data warehouse, however, in that historical data are not maintained. Seldom does an ODS maintain data that is more than 30 or 60 days old. The purpose of an ODS is to provide integrated data for operational purposes. For example, an ODS may integrate customer data which allows customer service representatives to have a full understanding (e.g., preferences, profitability) of customers at any customer "touchpoint." Some companies develop an ODS to avoid a full-blown Enterprise Requirements Planning (ERP) implementation. One purpose of ERP is to integrate data, which can be done less expensively with an ODS.

A recent development is the emergence of **oper marts** (Imhoff, 2001). Oper marts are created when current operational data needs to be analyzed multidimensionally. The data for an oper mart comes from an ODS. Only a small subset of the data in an ODS is used – enough to support the desired analysis. The oper mart data is stored in a multidimensional manner (e.g., a star schema, which is discussed in Section VII). The data in the oper mart is updated by transactions occurring in the ODS. When the desired analysis is completed using the oper mart, the mart is dismantled.

For example, consider a situation where a hurricane is poised to strike south Florida and an insurance company wants to analyze its exposure. To perform this type of analysis, it is necessary to have up-to-date data on customers in south Florida, the policies they hold, and the coverage of the policies. This data can be created for an ad hoc analysis by extracting it from an ODS and placing it in an oper mart.

III. THE ARCHITECTURE FOR DATA WAREHOUSING

The **architecture** for data warehousing includes the component parts and the relationships among the parts. Figure 2 shows a typical, comprehensive architecture for data warehousing. Appendix A lists the vendors mentioned in Figure 2 and elsewhere in this article, provides their website URLs, and briefly describes their product offerings. Appendix B lists and explains the acronyms used in this article.

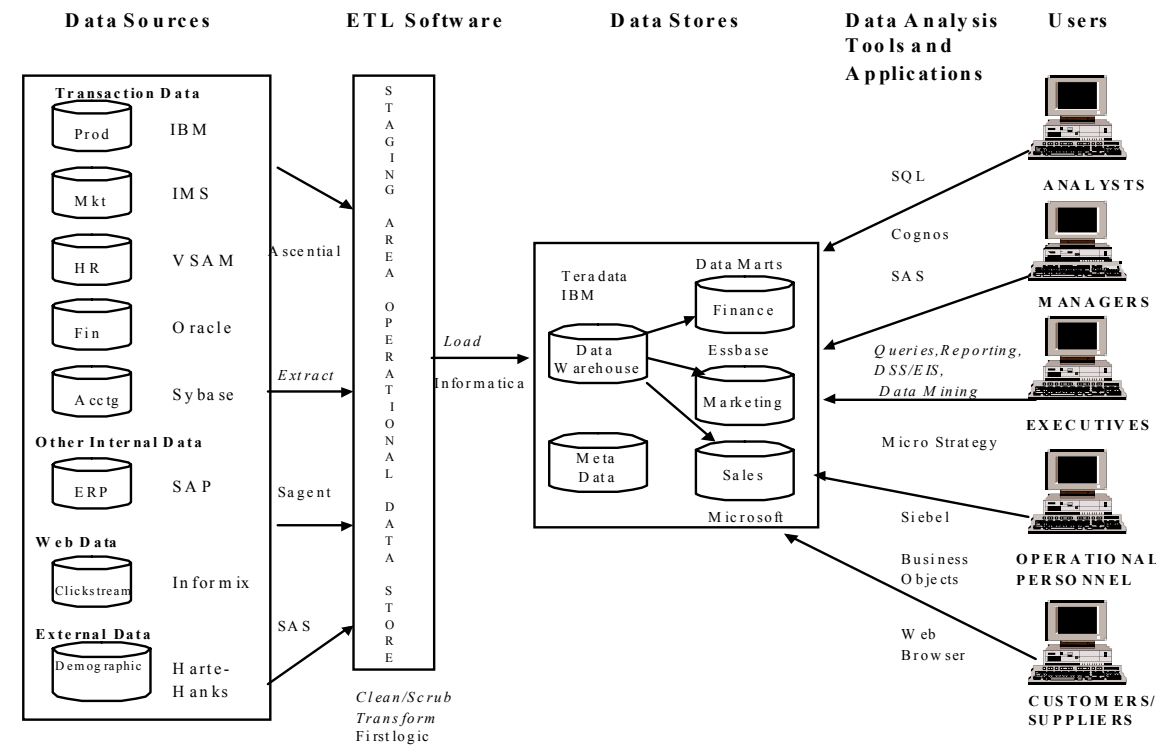

Figure 2: A Comprehensive Data Warehousing Architecture

The left-hand side of Figure 2 shows the various data sources. Much of the data comes from transactional (i.e., operational) systems such as production, accounting, and marketing. Data may also come from an ERP, such as those produced by SAP or PeopleSoft. Web data in the form of web logs may also feed into the data warehouse. And finally, external data, such as US census data may be included. These data sources often use different hardware and software, and a mixture of hierarchical, network, and relational data models for storing the data. It is not unusual for a data warehouse to draw upon over 100 source systems.

The data are extracted from the source systems using custom-written software (such as COBOL routines) or commercial extraction, transformation, and loading (ETL) software (Section V). The data is then fed into a staging area (e.g., an Oracle database) where it is transformed. Special-purpose software may be used to facilitate data cleansing processes. The processed data may then be used to support an operational data store. The data is also ready for loading into the data warehouse.

The data warehouse provides the repository of data used for decision support. Subsets of the data may be used to create dependent data marts that support specific kinds of users, such as financial analysts or quality control specialists. Typically the dependent marts provide a multidimensional view of the data, such as by customer, by product, by location, and over time. A multidimensional database, such as Essbase, that is designed to represent data

multidimensionally and provide fast response times, may be used. Metadata about the data in the warehouse (e.g., where and when the data are extracted, and the scheduled reports for users) are maintained so that it can be accessed by IT personnel and users. Metadata is discussed in Section VI.

Many people access the warehouse using data access tools and applications that are appropriate for their needs. Power users (e.g., analysts) who understand the underlying data model for the warehouse and how to write SQL queries, may write their own SQL queries. Many users (e.g., analysts, managers) employ a managed query environment such as Business Objects or Cognos to access data. These products provide a Windows-like interface for designing queries, which then automatically generate the needed SQL code. Specially trained analysts may also perform data mining using warehouse data (either within the warehouse, downloaded to a server, or on the desktop). The data warehouse may also be used to support specific DSS and EIS applications. Products like MicroStrategy Agent from MicroStrategy and Holos from Crystal Decisions are used for the latter purposes. The warehouse is also used with specific application software (e.g., Seibel for sales force automation).

A recent development is to give customers and suppliers access to warehouse data. A web browser is normally the client of choice for this purpose. In general, web browsers are used to access data warehouse data by a wide variety of users for many different purposes.

IV. ALTERNATIVE DEVELOPMENT APPROACHES

Even though data warehouses are widespread, there is no common agreement about the best development methodology to use.[1] In fact, two competing approaches are used. The first is associated with Bill Inmon, who is recognized as "the father of data warehousing," for his early consulting and writings [Inmon, 1992]. He advises companies to use a top-down, enterprise data warehouse approach. The second approach is associated with Ralph Kimball [1992], another highly respected consultant and writer on data warehousing. He recommends that companies use a bottom-up, data mart approach. Both approaches provide benefits but involve limitations. When successfully executed, both strategies result in an integrated enterprise data warehouse. In the next two subsections, these two approaches are described, compared, and contrasted.

THE DATA MART STRATEGY

The data mart strategy is a "start small, think big" approach. It typically begins with a specific business need for data, usually in the sales or marketing areas. A business unit manager (e.g., VP for Marketing) often sponsors the project and provides the necessary resources and support. The initial data mart contains data for only a single or a limited number of subject areas and draws data from a small number of source systems. Because of its limited scope, a data mart can be developed quickly, at a relatively low cost, and provide a fast return on investment.

If the data mart is successful (thus providing a "proof of concept" for data warehousing), the project team expands the data mart by adding more subject areas, users, and applications. It is at this point that great care needs to be exercised. In a worst-case scenario, separate data marts are developed and maintained by different business units with little or no attention paid to integrating the data either logically or physically.

The better case is to recognize at the outset that the initial data mart will grow and that this growth should be anticipated and planned. From the beginning, the project team should develop consistent data definitions, an integrated data model, and common dimensions and measures (i.e., "conformed" dimensions); implement a scalable architecture that accommodates growing data, users, and network traffic; and select an appropriate portfolio of end user data access tools. These are challenging tasks because they require planning and foresight; and cross-departmental participation, agreement, and governance.

The data mart approach is appealing because it provides usable data faster, at a lower cost, and with less financial risk. The difficulty, however, is in successfully growing a data mart,

[1] An earlier version of this section was originally published in Watson et al. (2001).

integrating new subject areas, data, users, and applications along the way. Some firms end up with multiple data mart "silos" that only perpetuate their data integration problems.

THE ENTERPRISE DATA WAREHOUSE APPROACH

In most organizations, the development of an enterprise data warehouse is a desired end goal; however, the data mart approach may not be able to accomplish that objective. Bill Inmon argues that it is unlikely to do so because the architectures of marts and warehouses are "genetically" different. Illustrating this point, in a personal conversation with the author, he said, "You don't plant a seed, see it grow into a tumbleweed, and then become an elm." Unless companies address up front the need to handle a large volume of data, integrate data from multiple sources, and provide for data governance, they are unlikely to develop an enterprise data warehouse successfully. The enterprise data warehouse strategy does not preclude the creation of data marts. The marts, however, are created after the warehouse is built, and they are populated with data pulled from the warehouse instead of from source systems. Users access these dependent data marts instead of the warehouse, resulting in faster system response time and a simpler and more customized data view that meets their needs.

With the enterprise data warehouse approach, the data in the warehouse is normalized. It only becomes denormalized when it is organized in a star schema (Section VII) in the dependent data marts. This approach is different from the data mart approach, where each data mart uses denormalized data in a star schema and integrates the marts through conformed dimensions.

When executed successfully, the enterprise approach results in an integrated data warehouse that contains many subject areas and supports multiple users and applications. However, as is the case with most large IT projects, there is the risk that it is never completed, or the end result fails to meet user and organizational needs.

V. EXTRACTION, TRANSFORMATION, AND LOADING PROCESSES

Data extraction, transformation, and loading (ETL) takes data from **source systems**, prepares it for decision-support purposes, and places it in the **target data base(s)**. It is the "plumbing" work of data warehousing – dirty, complex, time consuming, and expensive. More than one data warehousing project failed because appropriate ETL processes could not be put in place, either because of problems working with source systems, inadequate technology for the task at hand, inadequate data warehousing expertise on the project, and/or organizational issues.

DATA SOURCES

Over the years, organizations developed a large number of **applications** to support business processes. Many of these applications were written in COBOL and continue to be used today. Many of the applications are poorly documented and use cryptic table and attribute names. The data in these applications are often difficult to access for decision-support purposes, and even if the data can be accessed, doing so creates performance problems for the applications because accessing the data slows the processing of transactions.

Many companies replaced their legacy, operational systems with **enterprise resource planning (ERP)** systems from vendors such as SAP, PeopleSoft, Oracle, and J.D. Edwards. These ERP systems are an important source of decision support data, but the software often stores data in complex, proprietary data structures. As a result, extracting data from them is difficult. In response, ERP vendors created data mart solutions (e.g., Business Warehouse from SAP). Using the vendor's software, ETL processes are performed on the ERP data files, the data is maintained in a data mart, and predefined and *ad hoc* data analyses are performed on the mart's data. Database vendors also responded by creating data extraction routines that can be used with ERP software.

Another important data source is **clickstream data** gathered from customers' visits to a company's web sites. This data is important to understanding customer behavior, buying habits, preferences, and the use and effectiveness of the web site. Clickstream data includes the client IP address, hit time and date, download time, target page, user agent, query strings, server, IP address, and cookie data. If the user found the site through a search engine such as Yahoo, the referrer page and search words entered are provided. Clickstream data is collected in Internet

log files and is voluminous. Because of the volume, the records must be filtered selectively by cleaning up and removing unneeded data before the data is passed on to downstream processes (Werner and Abramson, 2001). Clickstream data is used to support operational processes (typically through an ODS) and is analyzed (in a warehouse) to find relationships, such as market segments.

Some companies include **external data** provided by third party organizations, including business partners, customers, government organizations, and organizations, such as credit bureaus, that choose to make a profit from selling their data.

DATA EXTRACTION

The two major options for extracting data from source systems are:

1. **Custom-write data extraction programs**. Because of the prevalence of COBOL legacy applications, the extraction programs are often written in COBOL with embedded SQL queries. Companies that know their source systems well, understand their complexities, have excellent in-house technical skills, and want to avoid the cost of purchasing ETL software, may want to write their own ETL software. While this option is taken by some companies, the trend is toward using commercial software, the second option.

2. **Purchase commercial ETL software**. This software is available from the major database vendors (e.g., IBM, Teradata, Oracle) and companies that specialize in ETL software, such as DataStage from Ascential Software and SAS System from the SAS Institute. The specialized ETL software allows companies to specify source systems relatively easily, indicate the tables and columns that are to be used, move the data to specified targets (e.g., databases used as a warehouse or marts), and automate the process.

ETL software is becoming more sophisticated. For example, Metagenix is a new company that addresses the difficulty of understanding the content, relationships, and data quality of legacy systems. Using AI techniques, their software:

- assesses the quality of data in columns and tables,
- determines whether the data in columns is consistent (e.g., the column is being used to store the same data rather than being used to store different things),
- identifies relationships between columns and tables,
- suggests possible primary and foreign keys for joining data, and
- generates ETL code.

The Metagenix software can be used to help a company understand its legacy systems and the problems that are likely to be encountered before making a sizable investment in data warehousing.

DATA STAGING

Once data is extracted, it is commonly placed in a **staging area** where it is transformed. The data can be imported through native interfaces, flat files, FTP sessions, or other processes. The staging area can be thought of as a “work-in-process” area. Here the data is processed prior to being loaded into the warehouse. A relational database management system is used to store staged data. Users are not given access to data in the staging area because the data is not yet ready for user consumption.

DATA TRANSFORMATION

The data from the source systems are transformed in a variety of ways, including cleansing the data, integrating the data, and other transformations. The starting point is with data cleansing.

Data Cleansing

In an ideal world, "**dirty data**" would not exist. The data in operational systems would be perfect. Unfortunately, perfection is virtually never the case. The data in these source systems is the result of poor data quality practices and little can be done about the data that is already there. While organizations should move toward improving data quality at the source system level, nearly all data warehousing initiatives must cope with dirty data, at least in the short term.
Table 1 describes some of the many reasons for dirty data.

Table 1: Sources of Dirty Data

Dummy values	Inappropriate values were entered into fields. For example, a customer service representative, in a hurry and not perceiving entering correct data as being particularly important, might enter the store's ZIP code rather than the customer's ZIP, or enters 999-99-9999 whenever a social security number is unknown. The operational system accepts the input, but it is not correct.
Absence of data	Data was not entered for certain fields. This is not always attributable to lazy data entry habits and the lack of edit checks, but to the fact that different business units may have different needs for certain data values to run their operations. For example, the department that originates mortgage loans may have a federal reporting requirement to capture the sex and ethnicity of a customer, whereas the department that originates consumer loans does not.
Multipurpose fields	A field is used for multiple purposes; consequently, it does not store the same quantity consistently. This problem can happen with packaged applications that include fields that are not required to run the application. Different departments may use the "extra" fields for their own purposes, and as a result, what is stored in the fields is not consistent.
Cryptic data	It is not clear what data is stored in a field. The documentation is poor and the attribute name provides little help in understanding the field's content. The field may be derived from other fields or the field may have been used for different purposes over the years.
Contradicting data	The data should be the same but it isn't. For example, a customer may have different addresses in different source systems.
Inappropriate use of address lines	Data was entered incorrectly into address lines. Address lines are commonly broken down into, for example, Line 1 for first, middle, and last name, Line 2 for street address, Line 3 for apartment number, and so on. Data is not always entered into the correct line, which makes it difficult to parse the data for later use.
Violations of business rules	Some of the values stored in a field are inconsistent with business reality. For example, a source system may record an adjustable rate mortgage loan where the rate on the loan is lower than the minimum interest rate.
Re-used primary keys	A primary key is not unique; it is used with multiple occurrences. This problem can occur in many ways. For example, assume that a branch bank has a unique identifier (i.e., a primary key). The branch is closed and the primary key is no longer in use. But two years later, a new branch is opened, and the old identifier is reissued. The primary key is the same for the old and the new branch.
Non-unique identifiers	An item of interest, such as a customer, is assigned multiple identifiers. For example, in the health care field, it is common for providers to assign their own identifier to patients. This practice makes it difficult to integrate patient records to provide a comprehensive understanding of a patient's health care history.

Data integration problems	The data is difficult or impossible to integrate. This problem can be due to non-unique identifiers, or the absence of an appropriate primary key. To illustrate, for decades customers were associated with their accounts through a customer name field on the account record. Integrating multiple customer accounts in this situation can be difficult. When we examine all the account records that belong to one customer, we find different spellings or abbreviations of the same customer name. Sometimes the customer is recorded under an alias or a maiden name. Occasionally two or three customers have a joint account and all of their names are squeezed into one name field or several customers have the same name.

The solutions to cleansing dirty data are:

1. Rely on the basic cleansing capabilities of ETL software.
2. Custom-write data cleansing routines.
3. Use special-purpose data cleansing software.

Special-purpose software is often used to cleanse names and addresses. Leading products and vendors include Integrity from Vality, Trillium from Harte Hanks, and i.d.Centric from Firstlogic. The data cleansing process is shown in Figure 3 (Lyon, 1998).

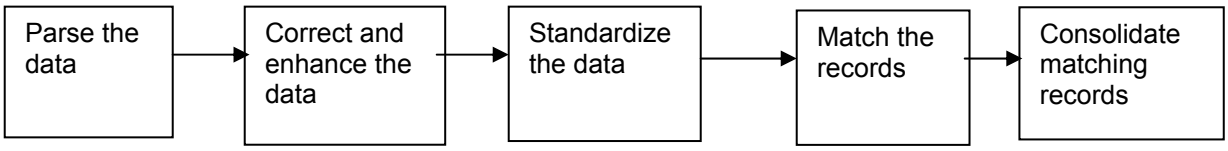

Figure 3: The Data Cleansing Process

The first step is to **parse** the individual data elements that are extracted from the source systems. For example, a customer record might be broken down into first name, middle name, last name, title, firm, street number, street, city, state, and ZIP code.

Data algorithms (possibly based on AI techniques) and secondary, external data sources (such as US Census data) are then used to **correct** and enhance the parsed data. For example, a vanity address (like Lake Calumet) is replaced with the "real" address (Chicago) and the plus four digits are added to the ZIP code.

Next, the parsed data is **standardized**. Using both standard and custom business rules, the data is transformed into its preferred and consistent format. For example, a prename may be added (e.g., Ms., Dr.), first name match standards may be identified (e.g., Beth may be Elizabeth, Bethany, or Bethel), and a standard street name may be applied (e.g., South Butler Drive may be transformed to S. Butler Dr.).

The parsed, corrected, and standardized data is then scanned to **match** records. The matching may be based on simple business rules, such as whether the name and address are the same, or AI based methods that utilize sophisticated pattern recognition techniques.

Matched records are then **consolidated**. The consolidated records integrate the data from the different sources and reflect the standards that were applied. For example, source system number one may not contain phone numbers but source system number two does. The consolidated record contains the phone number. The consolidated record also contains the applied standards, such as recording Ms. Elizabeth James as the person's name, with the appropriate pre-name applied.

Data Integration

A major purpose of a data warehouse is to integrate data from multiple systems. Consider, for example, a financial institution's customer data. Customer data is generated from a large number of touchpoints – at tellers and ATMs, through the mail, over the phone, and through

PC banking. It is also available from multiple systems – checking and savings accounts, credit cards, and mortgage loans. External demographic and financial data may be purchased. All of this data should be integrated around a common identifier, such as a SSN or an assigned ID, so that it can be used to provide an integrated understanding of the bank's customers.

Other Transformations

Other transformations include replacing values, calculating derived values, and preparing aggregates. Values that are used in operational systems may not be comprehensible to warehouse users. For example, the codes A, B, and C may be used in an operational system to designate direct, catalog, and Internet sales. Replacing them with terms that are more easily understood is a good practice.

Some values needed by users are the result of calculations performed on other values. For example, a customer's average account balance is the result of dividing the sum of a customer's daily account balances by the number of days in the accounting period. Numbers that are required by users can be computed before the warehouse is loaded.

It is also common to pre-calculate aggregates or summaries, such as total sales by all sales personnel in all regions, if they are going to be required by warehouse users. This practice can significantly improve system response time, because users do not calculate summaries "on the fly."

Data Loading

Once the data is transformed, it is ready for **loading** into the warehouse. The first loading provides the initial data for the warehouse. Subsequent loadings can be done in one of two ways. One alternative is to **bulk load the warehouse** every time. With this approach, all of the data (i.e., the old and the new) is loaded each time. This approach requires simple processing logic but becomes impractical as the volume of data increases. The more common approach is to **refresh** the warehouse with only newly generated data. This process is more complicated because the changes in the source system data must be captured as the changes are made. The process is commonly called **change data capture**.

Another issue that must be addressed is how frequently to load the warehouse. Factors that affect this decision include the business need for the data and the business cycle that provides the data. For example, users of the warehouse may need daily, weekly, or monthly updates, depending on their use of the data. Most business processes have a natural business cycle that generates data that can be loaded into the warehouse at various points in the cycle. For example, a company's payroll is typically run on a weekly basis. Consequently, data from the payroll application is most likely loaded to the warehouse on a weekly basis.

The trend is for continuous updating of the data warehouse. This approach is referred to as **"trickle" loading** of the warehouse. Several factors cause this near real-time updating. As data warehouses are increasingly used to support operational processes, having current data is important. Also, when trading partners are given access to warehouse data, the expectation is that the data is always up-to-date. Finally, many firms operate on a global basis and there is not a good time of day or week to load the warehouse. Users around the world need access to the warehouse on a 24X7 basis. A long "**load window**" is not acceptable.

VI. METADATA

It is important to maintain data about the data (i.e., **metadata**) in the data warehouse. Both the IT personnel who operate and manage the data warehouse and the users who access the warehouse's data need metadata. IT personnel need such information as data sources and targets; database, table and column names; refresh schedules; and data usage measures. Users need include entity/attribute definitions; the report/query tools that are available; report distribution information; and help desk contact information.

In general, metadata has not received the attention that it deserves. The reasons for this neglect include insufficient attention from system developers, uncertainty about what metadata should be stored, and a lack of methods for sharing metadata across different vendors' products.

Two developments are improving a firm's ability to create metadata and then to exchange the metadata across different vendors' products (Soschin, 2001). The first are attempts to create standards for metadata models. These models define what types of metadata should be stored to support interoperability among system components. Texas Instruments and Microsoft have been at the forefront of these efforts and created the Open Information Model (OIM). The OIM was the first standard to be widely accepted by both vendors and companies. Eventually, the OIM standards were rolled into parallel efforts by the Object Management Group, which resulted in standards known as the Common Warehouse Model (CWM).

The second development is the use of application program interfaces (APIs). Vendors are increasingly providing comprehensive, platform-independent APIs within their products that define what metadata their products store and how it can be accessed. Extensible markup language (XML) is quickly becoming the standard for developing APIs. Systems only need to be able to interpret and process XML to access the metadata.

Vendors are concentrating their efforts on the exchange of metadata through APIs. The trend, however, is for vendors to adhere more to metadata standards, such as the CMW.

While data warehousing software (e.g., data access tools, database software) internally handles metadata, other products are specifically designed to manage metadata. They include Platinum Repository from Computer Associates, Rochade Repository by ASG, MetaStage by Ascential, and MetaCenter from Data Advantage Group.

VII. DATA STORES

The data can be loaded to a variety of **data stores** – a data mart, an operational data store, or data warehouse. As discussed in Section V, loading data to an independent data mart is not recommended. It does not lead to a "single version of the truth" or support cross-organizational analyses. The role of an operational data store is to support operational applications that need access to current data. After the data "ages," it is passed on to the data warehouse. In this way, the ODS becomes a source for the data warehouse, which is the target.

The data stores can use either relational and/or multidimensional database technology. **Relational technology** stores data in tables with attributes and rows. It is the norm for data storage. **Multidimensional database technology** stores data organized around the dimensionality of the data. The data is stored in a multidimensional cube. Taking a simple example, a product is sold at a retail outlet at a particular point in time. A likely question might be, "How many television sets were sold at Best Buys in 2001?" The three dimensions of business interest are product, retail outlet, and time. The advantage of a multidimensional database is that it provides a view of the data that is tailored to the way that users want to analyze the data. The database is also designed to optimize for response time for users' multidimensional queries (Van Dyk, forthcoming).

Many data warehouse implementations use only relational technology. The source systems are relational (mixed with flat files, VSAM, etc), as are the staging area, operational data store, data warehouse, and data marts. These implementations mix and match a variety of vendors' products – Teradata from NCR Teradata, DB2 from IBM, Oracle 8i from Oracle, and SQL Server 2000 from Microsoft.

Multidimensional databases, such as Essbase from Hyperion, are niche products. When used, they store data in a dependent data mart for direct end user access. Though there are advantages to using a multidimensional database, such as fast response times to users' multidimensional queries, there are disadvantages as well: the cost; the need to introduce, learn, and support new database technology; and the amount of data that can be stored compared to relational databases. However, this problem is fading away as the storage capacity of multidimensional databases increases.

Relational databases can provide a dimensional view of the data through the use of a star schema data model, rather than the more usual entity- relationship model. The **star schema data model** resembles a star and is made up of a fact table in the center of the star and dimension tables as the points of the star. A star schema provides a non-normalized data model. Figure 4 shows a star schema for heath care.

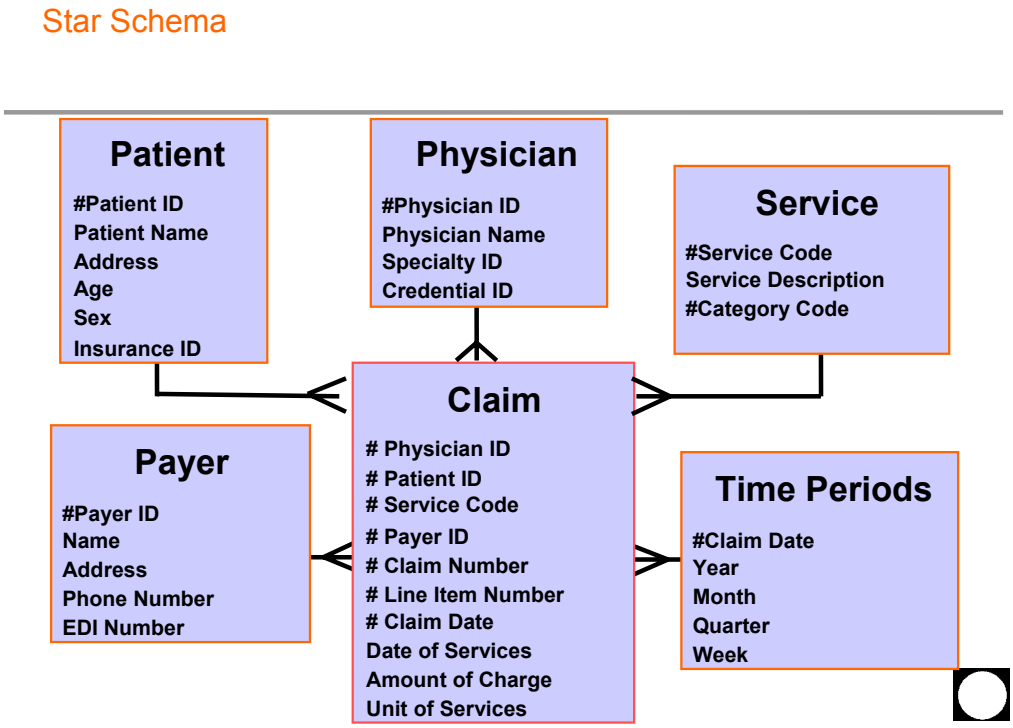

Figure 4: A Star Schema for Health Care (courtesy of Arthur Andersen)

The **fact table** stores the numerical data that were recorded for each transaction. Examples in different industries include:

- Retail -- number of units sold, sales amount
- Telecommunications -- length of call in minutes, average number of calls
- Banking -- average monthly balance
- Insurance -- claims amount

The **dimension tables** provide the dimensions through which the data in the fact table can be analyzed. The dimensions include **measures** that are further refinements of how the data can be analyzed. For example, a time dimension is likely to include daily, weekly, monthly, and annual as measures. Because a data warehouse stores historical data, time is always one of the dimensions. Examples of dimensions in different industries include:

- Retail -- store name, zip code, product name, product category, day of week
- Telecommunications -- call origin, call destination
- Banking -- customer name, account number, branch, account officer
- Insurance -- policy type, insured party

An organization will have multiple star schemas – at least one for every subject area. It is important to have many of the same dimensions and measures across the star schemas. Doing so provides **conformed dimensions** and allows meaningful queries to be made across different subject areas.

SQL queries are run against the data stored in a star schema. The primary keys in the dimension tables are also the primary keys in the fact table. In a query, the fact and dimension tables are joined through a concatenation of the keys in the fact and dimension tables. Table 2 shows an SQL query that answers the question: how many television sets were sold to Best Buys on January 15, 2001?

Table 2: An SQL Query against a Star Schema

```
SELECT CLIENT.CUSNAME, SALES.NOSOLD
FROM CLIENT, PRODUCT, TIME, SALES
WHERE CLIENT.CUSNAME=SALES.CUSNAME AND
      PRODUCT.PRODNAME=SALES.PRODNAME AND
      TIME.DATE=SALES.DATE AND CLIENT.CUSNAME="BEST  BUYS"
      AND PRODUCT.PRODNAME="TELEVISION" AND
      TIME.DATE=#01/15/2001#
```

Online Analytical Processing (OLAP) is the term usually associated with multidimensional analysis. It includes "slicing and dicing" data, drilldown, and rollup. Depending on the technology, OLAP comes in a variety of forms. For example, **ROLAP** is OLAP performed on a relational database. **MOLAP** is OLAP against a multidimensional database. **DOLAP** is desktop OLAP and involves the transfer of data cubes to the user's desktop for analysis.

Microsoft became a major force in the data warehousing market with SQL Server 2000 with Analysis Services. This server product combines both database and basic OLAP functionality. OLAP vendors have scrambled to avoid competing directly with Microsoft by either moving out of the OLAP tools business to packaged business intelligence applications, or by providing functionality beyond that which is provided by Microsoft (e.g., the vendors' products utilize the data cubes that can be created with Analysis Services). Currently, SQL Server 2000 with Analysis Services is primarily implemented with data marts, but as its storage capacity expands over time, it is likely to compete well in the large data warehouse market.

VIII. DATA ACCESS TOOLS AND APPLICATIONS FOR END USERS

Many tools and applications can be used to access warehouse data. While IT groups prefer to implement and support a single tool, it is seldom possible to do so. The information needs of users, their job requirements, and their willingness and ability to work with computers differ in too many ways. As a result, organizations implement multiple data access tools and applications.

When selecting a tool, it is important to recognize that a flexibility/ease of use tradeoff always exists. Applications that are easy to use (like an Executive Information System (EIS)) are relatively inflexible. Highly flexible tools (like SQL queries) are more difficult to use. It is a mistake to give a user a tool that provides more functionality than is needed because the tool may be too difficult to use. It is also a mistake to implement a tool that does not have the functionality that the user needs.

A large variety of data access tools and applications can be used with a data warehouse. Business intelligence (BI) is a term that is commonly used to describe the set of tools and applications that is used for decision support purposes. The most important data access tools and applications are listed and described in Table 3.

Table 3. Data Access Tools and Applications

SQL queries	They require the user (either an application programmer who embeds SQL in an application or an end user) to understand the warehouse's data model and know how to write SQL code. While this expectation is reasonable for IT personnel, it is not for the vast majority of end users. With the exception of only the most advanced power users, SQL queries are not an option.
Management Reporting Systems	Organizations have used management reporting systems for many years -- first paper based and now more commonly computer based. A continuing problem, however, is the inconsistencies among the reports. For example, the reports that sales and marketing use may differ because their data comes from separate source systems. A data warehouse, with its "single version of the truth" helps management reporting systems because it provides a single, consistent, integrated source of data to draw upon.
Managed query environments	Products like Business Objects, Brio, ProClarity, and Cognos are used by analysts and other organizational personnel who need to generate information as part of their job responsibilities. Users are presented with a graphical interface that allows them to specify relatively easily what needs to be done with the data, such as "slicing and dicing" the data according to specifications or identifying the top ten customers. Managed query environments can also be used to create reporting systems and simple EIS. Once an analysis is performed, it can be saved and made available as a report or an icon on a screen. Managed query environments are available with both thin (i.e., browser) and fat clients. The fat clients must be loaded on users' desktops but they provide the added functionality that some users need. In most cases, however, a web browser is the client of choice because of its lower cost, lower administration expenses, reduced training requirements, and portability. It is part of a four tier architecture, where the browser communicates with a web server, which talks to an application server, which communicates with a backend database.
DSS/EIS	Decision support systems and executive information systems (DSS/EIS) have been key decision support applications for many years. DSS provides information to support a specific decision-making task and EIS provides information targeted at the broad information needs of senior management and other organizational personnel (Watson, Houdeshel, and Rainer, 1997). Many EIS include the ability to "slice and dice" data and embed DSS applications. Depending on the application, the DSS/EIS may support a few users or many users throughout the organization. There are a variety of vendors that offer DSS/EIS software, such as MicroStrategy. As with managed query environments, DSS/EIS software provides both fat client and browser alternatives. Prior to data warehousing, some DSS/EIS failed because of an inadequate data infrastructure to support them. It is interesting to note that some of the early conceptual writings on DSS discussed the need for a specialized database to support DSS (Sprague and Watson, 1975). Today's data marts and warehouses have many similarities to what was conceptualize for DSS over 25 years ago.
Enterprise intelligence portals	At present, unstructured data, such as documents, email, video clips, and images are not included in data warehouses, even

	though their potential for supporting decision making is well recognized and the technology for doing so exists. Instead, unstructured data, if it is managed at all, is located in document or knowledge management systems. While one might speculate that data warehousing and knowledge management systems might merge, at present, this does not seem to be the trend. In addition to storing different kinds of data, they also use different technologies (e.g., Lotus Notes for knowledge management systems), are built and maintained by different organizational personnel, and involve different organizational issues. Although both structured and unstructured data needs to be delivered to users' desktops, this is not being done through the integration of data warehousing and knowledge management systems. Rather, the integration is being accomplished through enterprise intelligence portals (or enterprise information portals), which seamlessly access and display data from a variety of data sources. Portals are customizable to meet each user's information needs. Enterprise intelligence portal products and vendors include Jasmine from Computer Associates and Brio Portal from Brio.
Data mining	The "data mining" term is used both broadly and narrowly. The broad interpretation is that data mining includes almost any kind of data analysis, including, for example, OLAP. The narrow interpretation is the preferred use, and refers to an automated search of data to discover hidden relationships. Data mining is used for example, to discover market segments and the characteristics of customers who are likely to leave. Data mining involves the use of specialized algorithms (e.g., neural networks, genetic algorithms), most of which were created by academic researchers years ago but are only now finding their way into software products. Data mining software is available from the large software companies (e.g., Teradata Warehouse Data Mining from Teradata, Intelligent Miner from IBM) and from a large number of small, specialized firms (e.g., KnowledgeSeeker from Angoss). Teams of analysts who combine three important skills – knowledge of data mining algorithms and software, the ability to work with large amounts of data, and domain knowledge -- are needed to perform data mining.
Customer relationship management	One of the most popular corporate initiatives is customer relationship management (CRM) (Dyce′, 2001). Its purpose is to attract, retain, and enhance customers by maintaining relationships to increase revenues and profits. Critical to a comprehensive CRM initiative (as opposed to a point CRM solution) is moving to "a 360 degree view" of the customer, which means providing information about every interaction the company has with each of its customers. A 360 degree view requires the integration of data into a data warehouse from every "customer-facing application" or "touchpoint," such as transactions at an ATM, calls received at a call center, and responses to marketing campaigns. CRM applications are both operational and analytic. The operational applications use data, typically from an operational data store, to support operational activities. For example, a call center that responds to customer queries may provide customer service representatives with comprehensive information about customers – such as the products that they purchased, products that they might want to purchase, and the customers' value (i.e., profitability) to the firm – to help guide the representatives in their interactions with

	customers (scripts are also commonly given to the service representatives). With analytical CRM, the data is analyzed, often using data mining methods. In our example, the list of products that a caller might want to purchase would be the result of analytical CRM.
Specialized applications	Many specialized applications rely on a data warehouse -- revenue management, supply chain optimization, and connecting customers and suppliers along the supply chain.

Users of a data warehouse can be categorized and thought about in many ways. One way is to consider whether they primarily produce or consume information. Another view is by organizational position, such as whether they are a manager or an analyst. Yet another perspective is how they move through, analyze, and use warehouse data. Each view helps triangulate on the nature of warehouse users, what kinds of data access software and applications they need, how data should be organized, how they should be supported, and how value is derived from the warehouse.

Organizations contain both information producers and information consumers. **Information producers**, like analysts, create information for their own use and for use by others (such as reports for managers). **Information consumers**, such as executives, consume the information that is prepared for them (like is available in an EIS). Tools and applications that are appropriate for each type of user need to be provided. For example, an information consumer may only need web access to scheduled reports while an information producer may need the power of a fat client, managed query environment.

Organizations consist of executives, managers, analysts, and operational personnel. Organizations also interact with customers and suppliers. Each group has general characteristics (allowing for considerable variation depending on the organization and the person) in terms of information needs, job responsibilities, and willingness and ability to access and manipulate warehouse data. These differences influence the tools and applications they use and how they are supported. For example, executives need easy access to satisfy well-defined information needs. They have neither the time nor the inclination to do detailed *ad hoc* data analysis. When they do need additional information, an analyst is available to create it. Analysts, on the other hand, create their own information with relatively little assistance once trained on using the data access tool and data warehouse.

Inmon (2001) categories the data warehousing user community as tourists, explorers, and farmers. Each community has its own set of characteristics that differentiate it from the other communities. **Tourists** do not know what they want. They survey as much territory as possible, looking at an overview of many things but with little deep data analysis. The design of the database is not an issue, because they do little data analysis. They often look at a data set once, never to return. They love the Internet and its search capability. Metadata is very important to the tourist.

Explorers have an idea of what data they want and where to look for it, but are not exactly sure how they are going to meet their objective of finding a “priceless gem” in the data. Often they find nothing. Explorers do not go to as many places as tourists go because they know where the potential gems are located. The problem is the vast amount of data between the explorer and the gem. Explorers need to work with large amounts of detailed, historical data. They require metadata to get started. Explorers also have to understand the underlying data model.

Farmers know what data they want and where and how to get it. They may not look at a lot of data because they know what they are looking for. Farmers normally find “flakes of gold” rather than a gem. Because they use the same data regularly, farmers do not rely on metadata. Understanding the underlying data model is important to their work.

IX. CASE STUDY: HARRAH'S HIGH PAYOFF FROM CUSTOMER INFORMATION[2]

To illustrate the concepts in this article, a case study is presented. It tells the story of Harrah's Entertainment, which assumed a leadership role in the gaming industry through its innovative CRM strategy (Section VIII) supported by data warehousing. In 2000, Harrah's won the prestigious Leadership Award from The Data Warehousing Institute for its work. In the case, you will see the linking of business and IT strategies; the building of WINet, which provides Harrah's data warehousing infrastructure; Harrah's Total Rewards program; and how Harrah's practices operational and analytical CRM.

INTRODUCTION

Harrah's Entertainment, Inc. (or simply Harrah's) assumed a leadership role in the gaming industry through a business strategy that focuses on knowing their customers well, giving them great service, and rewarding their loyalty so that they seek out a Harrah's casino whenever and wherever they play. The execution of this strategy involved creative marketing, innovative uses of information technology, and operational excellence. These component parts first came together in 1997 and resulted in many benefits, including:

- A doubling in the response rate of offers to customers;
- Consistent guest rewards and recognition across properties;
- A brand identity for Harrah's casinos;
- An increase in customer retention worth millions of dollars;
- A 72 percent increase in the number of customers who play at more than one Harrah's property, increasing profitability by more than $50 million; and
- A 62 percent internal rate of return on the information technology investments.

In the following sections, Harrah's business strategy is described, focusing on the branding of the Harrah's name and customer relationship management. To execute their business strategy, substantial investments in information technology were required to integrate data from a variety of sources for use in Harrah's patron database (an operational data store) and the marketing workbench (a data warehouse). This infrastructure supports operations, offers, Total Rewards (a customer loyalty program), and analytical applications. The case study illustrates many of the concepts introduced in the tutorial.

COMPANY BACKGROUND

In October 1937, Bill Harrah opened a bingo parlor in Reno, Nevada. He focused on customer comfort, running fair games, and ensuring that customers had a good time. In 1946, Harrah purchased The Mint Club, which took him from the bingo parlor business to full-scale casinos. After renovating the club, it was reopened as Harrah's Club and began the Harrah's style of casino entertainment. Harrah's was the "friendly casino," where employees knew the customers' names. In 1955, Harrah opened another renovated casino, this time on the south shores of Lake Tahoe. The gaming clubs at Harrah's Reno and Lake Tahoe were prosperous throughout the 1960s and 70s as Harrah continued to expand and improve these properties. By 1971, Harrah recognized that the practice of going to local bankers or competing gamblers to borrow money for supporting growth was limiting. He took his company public and became the first purely gaming company to be listed on the New York Stock Exchange.

Bill Harrah's vision for growth was continued by Philip Satre who was named president in 1984 and led Harrah's entry into the Atlantic City market. In 1993, legislation was passed that allowed gambling on Indian reservations and riverboats. Seizing the opportunity, Harrah's quickly expanded into these new markets, through the building of new properties and the acquisition of Showboat casinos, the Rio All-Suite Casino, and Players International. Entering the new millennium, Harrah's had 25 casinos, making it one of the world's largest gaming companies.

[2] This section was written by the author and Linda Volonino of Canisius College and is partially based on a case study that was originally published in Eckerson and Watson (2000). It is reprinted by permission.

Harrah's sites are in every major U.S. market where gambling is allowed. These casinos and supporting hotels employ over 40,000 people, serve over 25 million customers, provide 13,200 hotel rooms, 41,648 slot machines, 1,065 table games, and over 1.4 million square feet of gaming space (www.Harrahs.com, 2000).

HARRAH'S BUSINESS STRATEGY

The decision to expand into additional gaming markets was a critical part of Harrah's business strategy. The growth of these markets was considered to be inevitable and helpful to Harrah's and the industry. As management thought about how it could create the greatest value for its shareholders, it was decided that a brand approach should be taken. With this approach, the various casinos would operate in an integrated manner rather than as separate properties. This idea was a radical paradigm shift in the gaming industry where casino managers historically ran their properties as independent fiefdoms and marketing was done on a property by property basis. The new approach created commonalties in the gambling experience for customers across the various casinos. Advertising and offers would promote the Harrah's brand. Recognition and reward programs for customers who cross-played at more than one of Harrah's properties would be instituted. Harrah's mission was to build lasting relationships with its customers.

Also motivating the strategy were the experiences of some of the new Las Vegas hotels and casinos (e.g., the Bellagio and Paris) that invested vast sums of money in lavish hotels, shopping malls, and attractions such as massive dancing water shows and a replica of the Eiffel Tower. While these malls and attractions are highly popular, their costs cut investment returns in half. Harrah's wanted to take a different, more cost-effective route that not only attracted customers, but also maintained and enhanced customer relationships.

Critical to their strategy was the need to understand and manage relationships with their customers. Harrah's believed that strong customer service relationships build on a foundation of customer knowledge. To build this foundation, Harrah's had to learn about their customers' behaviors and preferences. They had to understand where their customers gambled, how often they gambled, what games they played, how much they gambled, and what offers would entice them to visit a Harrah's casino. Armed with this information, Harrah's could identify specific target customer segments better, respond to customers' preferences, and maximize profitability across the various casinos.

A key addition to the Harrah's management team was Gary Loveman who was named Chief Operations Officer (COO). This former Harvard professor had the understanding and skills needed to analyze customer behavior and preference data and to put programs in place to capitalize on this knowledge. He helped make Harrah's customer relationship management (CRM) strategy a reality.

To generate the necessary data, Harrah's made a substantial investment in information technology. It captured data from customer touchpoints, integrated it around the customer, and stored it for later analysis. To understand customers' preferences, Harrah's mined the data, ran experiments using different marketing interventions (i.e., special offerings), and learned what best met customers' needs at the various casinos. From these requirements, Harrah's Winners Information Network (WINet) emerged.

WINET: CREATING A SINGLE CUSTOMER VIEW

In 1994, Harrah's began work on WINet under the leadership of John Boushy who at the time served as Harrah's CIO and Director of Strategic Marketing. The purpose of WINet was to collect customer data from various source systems, integrate the data around the customer, identify market segments and customer profiles, create appealing offers for customers to visit Harrah's casinos, and make the data available for operational and analytical purposes. The repository for this data is a patron database (PDB) that serves as an operational data store. It provides a cross-property view of Harrah's customers. In 1997, Total Gold, a patented customer loyalty program was put in place, through which customers could earn points for their gambling activities (e.g., playing slot machines) and redeem their points for free retail products, rooms,

food, and cash. The marketing workbench (MWB) was also implemented to serve as a data warehouse for analytical applications.

The development of WINet was not without problems. For example, some complicated queries on MWB, originally an Informix database, took so long to run that they never finished within the computing window that was available. NCR Teradata, which was providing bench marking services for Harrah's, offered to run the queries on their Teradata database software and hardware. The performance improvement was so great that Teradata was brought in to redesign the system on NCR Teradata software and an NCR WorldMark 4700 UNIX System.

At the same time that Harrah's was considering moving to NCR Teradata, a decision was made to review the data access tools that the marketing department used. The outcome was a switch to Cognos Impromtu and SAS. Marketing analysts at the corporate and individual property levels use Impromtu to run predefined reports and queries and to execute ad hoc queries. Analysts use SAS for market segmentation analysis and customer profiling.

Figure 5 shows the architecture for WINet. The component parts of WINet are described in the following sections.

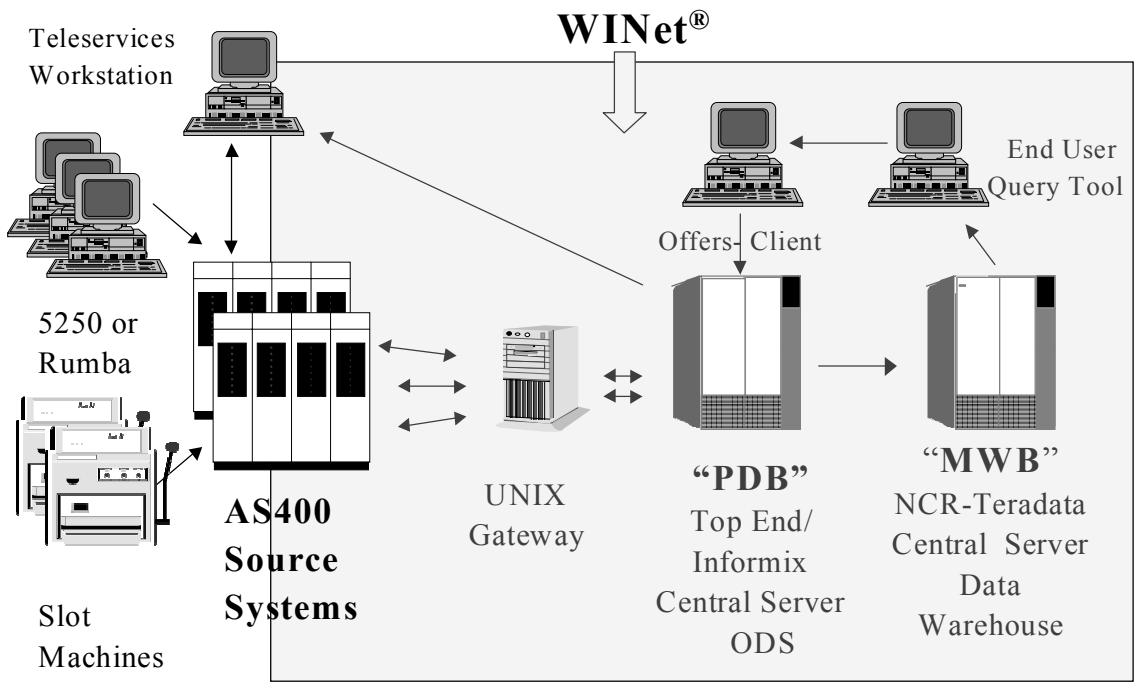

Figure 5: WINet Architecture

Data and Source Systems

Data is captured and collected from a variety of source systems. The hotel system records the details of a customer's stay, demographic data (e.g., home address), and preference data (e.g., smoking or non-smoking room). Data recorded from tournaments and special events (e.g., wine tasting weekends, slot machine tournaments) are included. Most players obtain a loyalty card (e.g., Total Gold) which they use to obtain points that can be redeemed for rewards (e.g., free meals, tickets to shows). In the case of slot machine play, the customer inserts the loyalty card into the machine and every play is recorded. With table games (e.g., blackjack), the player gives the card to the dealer and the pit boss enters into a PC the game played and the minimum, average, and maximum amount bet over a period of time (e.g., typically every two hours).

After a customer visits a casino and is in Harrah's system, he or she is a candidate for special offers (e.g., $50 in free chips if the customer returns to a Harrah's casino within the next two weeks). Data on the offers made and redeemed are recorded for each customer.

A variety of source systems are involved. Some of them are very specific to the gaming industry, such as the slot data system, which captures data automatically from slot machine play. Others such as the hotel reservation system are more generic and involve human data entry. The systems that require human input use IBM 5250s or Rumba terminal emulation for data access or entry. All of the transactional systems run on IBM AS400s. Harrah's has no mainframe.

Patron Database

At the end of the day for each source system (the definition of "end of day" varies with the system), relevant data is extracted for loading into the PDB. First, however, validity and "saneness" checks are performed. Checking for a valid address is an example of a validity check. A saneness test checks whether the data is reasonable, such as the "drop" from a 25 cent slot machine (e.g., a $1000 drop in a hour is not reasonable). Data that fail a test are placed in a suspended file and reviewed manually. At 7:00 a.m., the data is loaded into PDB from the casino, hotel, and event management systems. The load is completed and available for use by noon. In terms of source systems, no matter which casino a customer goes to, the details of every visit are captured and ultimately find their way into PDB. The data is available by customer, casino, hotel, event, gaming product, and tracked play. Every customer is assigned an identification number, and the data about the customer are joined using the ID as the primary key. Unless needed (e.g., such as with a promotional offer), customer names and address are not used with Harrah's applications.

MARKETING WORKBENCH

Marketing Workbench (MWB) was created to serve as Harrah's data warehouse. It is sourced from the patron database. MWB stores daily detail data for 90 days, monthly information for 24 months, and yearly information back to 1994. Whereas PDB supports on-line lookup of customers, MWB is where analytics are performed. Marketing analysts can analyze hundreds of customer attributes to determine each customer's preferences and predict what future services and rewards they will want. For example, Harrah's might award hotel vouchers to out-of-state guests, while free show tickets would be more appropriate for customers who make day trips to a casino. A major use of MWB is to generate the lists (i.e., "list pulls" in Harrah's terminology) of customers to send offers to. These lists are the result of market segmentation analysis and customer scoring using MWB.

OPERATIONAL APPLICATIONS

The Patron Database supports a variety of operational applications. For example, a valued customer may be a first time visitor to a particular Harrah's property. When the customer checks in to the hotel, the service representative can look up their profile and make decisions about how to treat the customer, such as offering free event tickets or meals. Another example is a pit boss who notes that a valued customer has been gambling heavily for a long period of time relative to the customer's profile and gives the customer a coupon for a free show.

WINet Offers

WINet Offers is Harrah's in-house developed application for generating offers to Harrah's customers. To create an offer, a marketing analyst works with customer segments and profile data in MWB to create a list of IDs of customers who are in the targeted segment and fit the desired profile. These IDs are then fed into PDB, and then a program generates a customized mailing and offer for the customers. PDB also records whether the offers are accepted or not. The offers are also connected to hotel systems so that rooms can be reserved for customers who accept offers. Some campaigns are run on a scheduled basis while others are *ad hoc*. The offers can be generated at the corporate level to support the Harrah's brand or be created by an individual property (e.g., to support a mid week slot machine tournament). Harrah's makes more than 20 million customer offers annually, and tracks each offer to determine when and how offers are redeemed and how marketing activities influence customer behavior at a detailed segment level.

Total Rewards

Total Rewards is Harrah's customer loyalty program. It tracks, retains, and rewards Harrah's 25 million customers regardless of which casinos they visit over time. Total Rewards was originally introduced as Total Gold in 1997, but it was renamed in July 1999 when a three-tiered card program – Total Gold, Total Platinum, and Total Diamond – was introduced to give more recognition to Harrah's most active and profitable customers. Customers accumulate Reward Credits (points) based on their gaming and other activities at any of Harrah's properties. These Reward Credits can be redeemed for comps on hotel accommodations, meals, and shows and cash can be redeemed at any property. At specified Reward Credit thresholds, customers move to the next card level (e.g., from Gold to Platinum) and qualify for the privileges associated with that level (e.g., preferred restaurant reservations and seating, priority check-in at hotels). Customers can check their Reward Credits at any time by entering their card into a slot machine or kiosk or by logging in to www.harrahs.com. Total Rewards members are also sent offers of cash and comps for use at local Harrah's casinos and destination resorts such as Las Vegas and Lake Tahoe. Figure 6 shows a customer's view of the Total Rewards program.

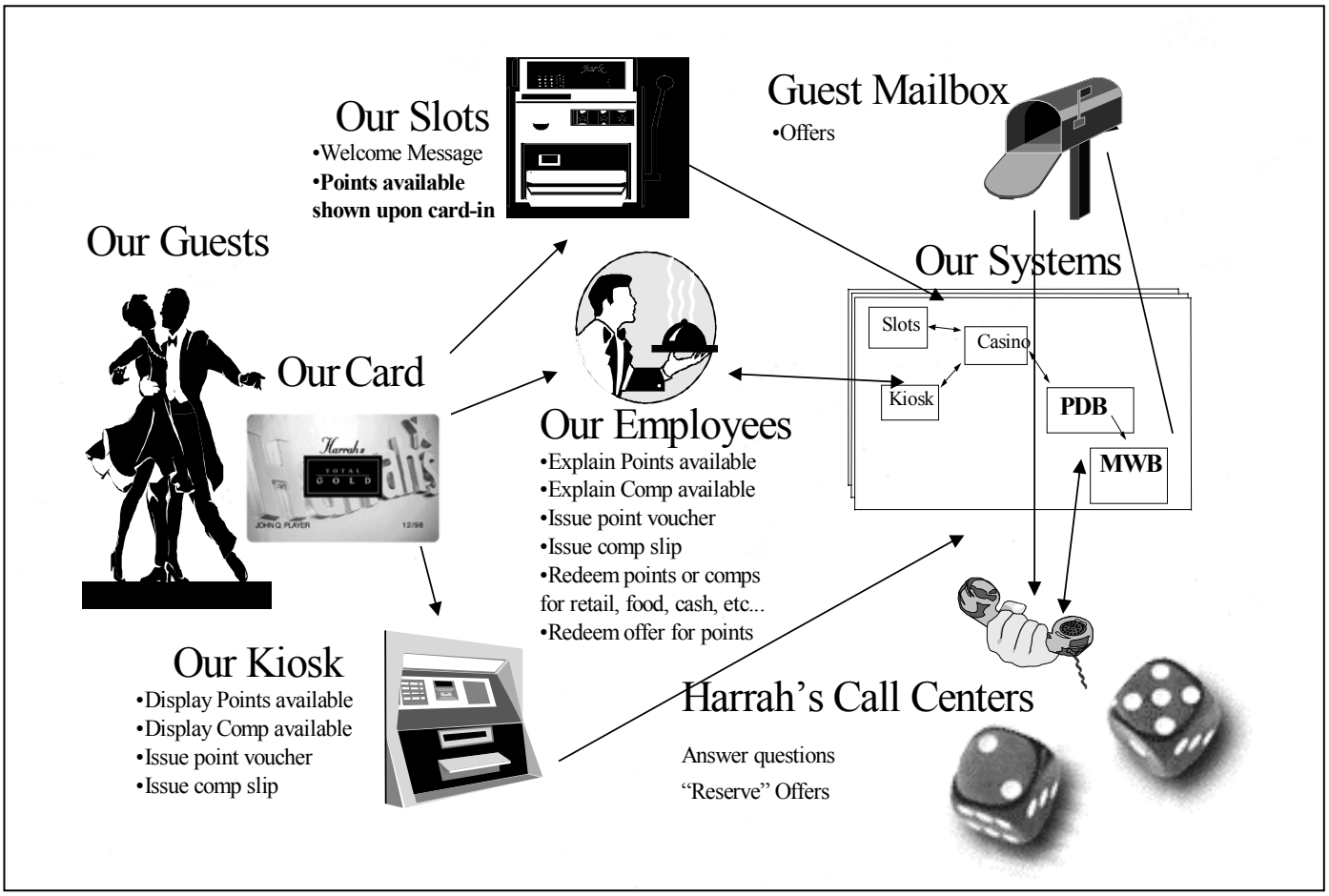

Figure 6: Customer View of the Total Gold™ Program

CLOSED-LOOP MARKETING

Like other casinos, Harrah's previously extended offers to customers based primarily on observed gaming worth. Over the years, intuition-based beliefs—called *Harrahisms*—developed for what did and did not work with their marketing campaigns. *Harrahisms* were never tested. With WINet, the foundation was put in place for a new, more scientific approach. Campaigns are designed, tested, and the results retained for future use. This data-driven testing and learning

approach is called “closed loop marketing.” Its goal is to learn how to influence positive changes in customer behavior. Harrah’s can learn what types of campaigns or treatments provide the highest net value.

Two examples illustrate the use and value of closed-loop marketing. Two similar groups of frequent slot machine players from Jackson, Mississippi were identified for an experiment. Members of the first group were offered a marketing package of a free room, two steak dinners, and $30 in free chips at the Tunica casino. Members of the second group were offered $60 in chips. The second, more modest offer generated far more gambling, suggesting that Harrah’s was wasting money offering Jackson customers free rooms and meals. Subsequent offers in this market focused on free chips, and profits nearly doubled to $60 per person per trip.

Another test focused on a group of monthly players who Harrah’s thought could be induced to play more frequently because they lived nearby and displayed traits such as hitting slot machine buttons quickly (i.e., “high velocity” players). To entice them to return, Harrah’s sent them free cash and food offers that expired in two weeks. The group’s number of visits per month rose from 1.1 to 1.4.

FINAL REMARKS

Harrah’s business strategy and the use of information technology are unique in the gaming industry and are more like the approaches taken in retail and financial services. Their innovative idea was to grow by getting more business from Harrah's existing customer base. This approach was in contrast to the prevalent strategy of building ever more elaborate and splashy new casinos. Gary W. Loveman refers to their success as "the triumph of software over hardware in gaming." The results are impressive and other casinos are copying some of Harrah’s more discernable methods.

X. CONCLUSION

In a few short years, data warehousing evolved from an innovation in a few leading-edge firms to a necessity in many companies. As one looks forward, the future of data warehousing is bright, for a variety of reasons. As organizations become more information intensive, the need for reliable information for decision support and other purposes continues to grow. Also supporting the growth are better understandings of how to implement data warehousing successfully, an expanding pool of people with data warehousing experience, and the availability of better data warehousing hardware and software products.

Data warehousing continues to morph in terms of what it is and how it is applied. Initially, a data warehouse was considered as only a repository of data used to support analytical applications. Today, through the use of operational data stores, it is used with operational applications. The inclusion of clickstream data to support the dialog with users and to better understand market segments and buying behavior are other important changes. So too is the practice of giving customers and suppliers access to warehouse data in order to integrate the supply chain. Comprehensive CRM operates off a 360 degree view of the customer that is provided by integrating all customer-related data in a data warehouse. Developments in business intelligence such as event, time, and algorithm-based alerts (Watson, Houdeshel, and Rainer, 1997) depend on data warehousing.

It is also interesting to note the number of universities that include considerable coverage of data warehousing in their DSS courses or offer courses in data warehousing. Data warehousing research is on the rise and there are data warehousing tracks at conferences such as the Hawaii International Conference on System Sciences (HICSS) and the America’s Conference on Information Systems (AMCIS). The *Journal of Data Warehousing* and *DM Review* are devoted to data warehousing topics.

Finally, industry is starting to provide resources to help in teaching about data warehousing. For example, several manufacturers provide data warehousing software at little or no cost to universities for teaching purposes. They include IBM (DB2), NCR (Teradata), and Oracle.

Editor's Note: This case study was received on November 20, 2001. It was with the author approximately five weeks for one revision. It was published on January 28, 2002.

REFERENCES

EDITOR'S NOTE: The text and the following reference list contains the address of World Wide Web pages. Readers who have the ability to access the Web directly from their word processor or are reading the paper on the Web, can gain direct access to these references. Readers are warned, however, that
1. these links existed as of the date of publication but are not guaranteed to be working thereafter.
2. the contents of Web pages may change over time. Where version information is provided in the References, different versions may not contain the information or the conclusions referenced.
3. the authors of the Web pages, not CAIS, are responsible for the accuracy of their content.
4. the author of this article, not CAIS, is responsible for the accuracy of the URL and version information.

Dyche′, J. (2001) *The CRM Handbook: A Business Guide to Customer Relationship Management*, Boston, MA: Addison-Wesley.

Eckerson, W.W. (1988) "Post-Chasm Warehousing," *Journal of Data Warehousing*, (3)3, pp. 38-45.

Eckerson, W.W. (April 28, 1999) *Evolution of Data Warehousing: The Trend toward Analytical Applications*, Boston, MA: The Patricia Seybold Group (April 28) pp. 1-8.

Eckerson, W. and H.J. Watson (2000), *Harnessing Customer Information for Strategic Advantage: Technical Challenges and Business Solutions* Seattle: The Data Warehousing Institute.

Gray, P. and H.J. Watson (1998) *Decision Support in the Data Warehouse*, Upper Saddle River, New Jersey, Prentice-Hall.

Imhoff, C. (2001) "Oper Marts – An Evolution in the Operational Data Store," *DM Review*, (11) 9, pp. 16, 43.

Inmon, W.H. (1992) *Building the Data Warehouse*, New York: Wiley.

Inmon, W.H. (2001) "Knowing Your DSS End User: Tourists, Explorers, Farmers," (www.billinmon.com/library/articles).

Kimball, R. (1992) *The Data Warehouse Toolkit*, New York: Wiley.

Lyon, J. (1998) "Customer Data Quality: Building the Foundation for a One-to-One Customer Relationship," *Journal of Data Warehousing*, (3) 2, pp. 38-47.

Soschin, D. (2001) "Meta Data As an IT Platform: The Strategy of Meta Data in Your Organization," *Journal of Data Warehousing*, (6) 4, pp. 30-40.

Sprague, R.H. and H.J. Watson (1975) "MIS Concepts, Part II," *Journal of Systems Management*, (26) 2, pp. 35-40.

Van Dyk (forthcoming 2002) *Journal of Data Warehousing.*

Watson, H.J., G. Houdeshel, and R.K. Rainer (1997), *Building Executive Information Systems and Other Decision Support Applications*, New York: Wiley.

Watson, H.J., B.H. Wixom, J.D., Buonamici, J.D., and J.R. Revak (2001), "Sherwin-Williams' Data Mart Strategy: Creating Intelligence Across the Supply Chain," *Communications of AIS*, (5)9.

Werner, V. and C. Abramson, (2001) "Managing Clickstream Data," *Journal of Data Warehousing*, (6) 3, pp. 11-15.

APPENDIX A. VENDORS OF DATA WAREHOUSING PRODUCTS

Ascential Software (www.ascentialsoftware.com)	A leading vendor of ETL software.
ASG (www.asg.com)	A leading vendor of metadata software.
Brio Technology (www.brio.com)	A leading vendor of data access tools (e.g., OLAP, managed query environment).
Business Objects (www.businessobjects.com)	A leading vendor of data access tools (e.g., OLAP, managed query environment).
Crystal Decisions (www.crystaldecisions.com)	A leading vendor of data access tools (e.g., reporting and OLAP)and applications (e.g., balanced scorecard).
Cognos (www.cognos.com)	Provides data warehousing data access tools (e.g., OLAP, managed query environment) and applications.
Computer Associates (www.cai.com)	Provides a comprehensive set of data warehousing tools and products.
Data Advantage Group (www.dataadvantagegroup.com)	A leading vendor of metadata software.
Firstlogic (www.firstlogic.com)	A leading vendor of data cleansing software.
Harte-Hanks (www.harte-hanks.com)	A leading vendor of CRM products and services.
Hyperion (www.hyperion.com)	Provides a comprehensive set of data warehousing tools, products, and applications.
IBM (www.ibm.com)	Provides a comprehensive set of data warehousing tools, products, and applications.
Informix (www.informix.com)	Provides a comprehensive set of data warehousing tools and products.
J. D. Edwards (www.jdedwards.com)	Provides ERP and CRM products and applications.
Metagenix (www.metagenix.com)	A new vendor of ETL products.
Microsoft (www.microsoft.com)	A leading vendor of data warehousing tools and products.
Oracle (www.oracle.com)	A leading vendor of data warehousing tools, products, and applications.
PeopleSoft (www.peoplesoft.com)	Provides ERP, CRM, and data warehousing products and applications.
SAP (www.sap.com)	Provides ERP, CRM, and data warehousing products and applications.
SAS Institute (www.sas.com)	A leading vendor of data warehousing tools, products, and applications.
Seibel (www.seibel.com)	A leading vendor of CRM applications.
Sybase (www.sybase.com)	Provides a comprehensive set of data warehousing tools, products, and applications.
Teradata (www.teradata.com)	A leading vendor of data warehousing tools, products, and applications.
Vality (www.vality.com)	A leading vendor of data cleansing software.

APPENDIX B. DATA WAREHOUSING ACRONYMS

AI	Artificial intelligence
API	Application program interface
BI	Business intelligence
CWM	Common warehouse model
CRM	Customer relationship management
DOLAP	Desktop online analytical processing
DSS	Decision support system
EIS	Executive information system
ERP	Enterprise resource planning
ETL	Extraction, transformation, and loading
FTP	File transfer protocol
MOLAP	Multidimensional online analytical processing
ODS	Operational data store
OLAP	Online analytical processing
OIM	Open Information Model
SQL	Structured query language
XML	Extensible markup language

ABOUT THE AUTHOR

Hugh J. Watson is Professor of MIS and holder of a C. Herman and Mary Virginia Terry Chair of Business Administration at the University of Georgia. He is the author of 22 books and over 100 journal articles. Over his career he has focused on the use of information technology to support decision making. Most recently, he has been studying data warehousing. Hugh is the Senior Editor of the *Journal of Data Warehousing* and a Fellow of The Data Warehousing Institute. He serves on the Editorial Board of the *Communications of AIS*.

ISSN: 1529-3181