# *Chapter 1*

# The Big Picture: An Introduction to Data Warehousing

## Introduction

In 1977, Jimmy Carter was President of the United States, *Star Wars* hit the big screen, and Apple Computer, Inc. introduced the world to the first personal computer. Four years later, Ronald Reagan was president, Prince Charles and Lady Diana married, and IBM began selling its IBM PC. From that beginning, computing power started moving from the mainframe to the desktop. Soon thereafter, spreadsheets and word processing applications began their journey to replace clipboards and typewriters. By 1985, Mikhail Gorbachev was the leader of the Soviet Union, New Coke hit the shelves, and data was going everywhere. Even inside the mainframe computers, data was finding its own home. Supervisors, managers, and executives alike were no longer able to look at a single clipboard to find out how the business was performing. The data was hopelessly entrenched in applications and nooks and crannies that would never again see the light of day. Such was the origin of Decision Support Systems.

## Decision Support Systems

Decision Support Systems allowed managers, supervisors, and executives to once again see the clipboard with all its information. The information, which previously had been on a clipboard, had become a report, either printed on paper or displayed on a screen. One report revealed one single business area. Another report revealed a different business area. By 1992, Windows® 3.1 was in the stores and Ralph Kimball and Bill Inmon were figuring out how to gather data from two business areas and figuring out how to warehouse the data of an enterprise (Figure 1.1).

Ralph Kimball was a co-creator of the Xerox Star Workstation, the world's first commercially viable GUI application. Ralph was the founder and CEO of Red Brick Systems, the group which created an extremely fast RDBMS targeted specifically for data warehousing. When he authored *The Data Warehouse Lifecycle Toolkit*, Ralph introduced the Dimensional Data Model (discussed in Chapter 5, Database Design).

Kimball and Inmon, working separately, arrived at a common set of guidelines (or principles). These principles are:

Bill Inmon was the creator of the Corporate Information Factory and Government Information Factory. In so doing, Bill also established many of the principles of Data Warehousing.

- **Subject Orientation**: Data will be grouped by subject, rather than author, department, or physical location. So, all manufacturing data goes together, and the sales data, and the promotions data, etc., regardless of where it came from.
- **Data Integration**: Even though data comes from separate applications, departments, etc., differences should be smoothed out so they have the same look and feel.
  - Form: When two data elements (e.g., phone numbers) have different layouts (e.g., 123-123-1234 and (123) 123-1234), one layout will be superimposed on both of them.
  - Function: When two data elements identify the same thing (e.g., a hammer) with two different names (e.g., part 32G and part B49), these two names will be replaced with one name.
  - Grain: When two data elements apply different hierarchies (e.g., region and district) to the same thing, or different levels of detail (e.g., miles and feet), the two data elements will be resolved to the same level of hierarchy or detail.
- **Nonvolatility**: Unlike the data in operational applications, which is discarded once the company is finished using it, the data in a data warehouse will remain in the warehouse.

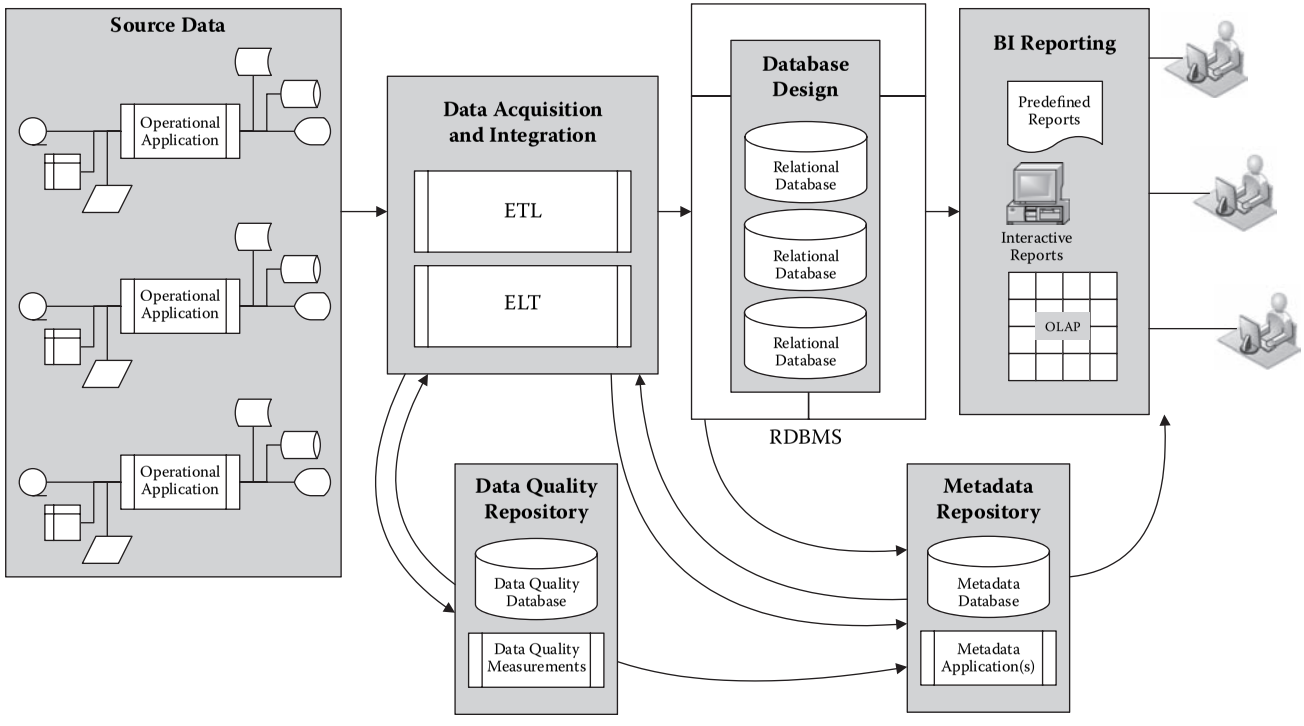


**Figure 1.1 The big picture.**

- **Time Variant**: All data has a context at a moment in time. A data warehouse will keep that context. So, all data from 1995 will retain its context within 1995.
- **One Version of the Truth**: The proliferation of data in the 1980s and 1990s yielded many copies of the same data. Only the one, true gold, standard copy of each data element would be included in a data warehouse.
- **Long-Term Investment**: A data warehouse should be flexible enough to absorb changes in the company and the world, and scalable enough to grow with the company. By doing so, a data warehouse can add value to the company for a long time.

## Dimensional and Third Normal Form Data Models

Kimball and Inmon arrived at the same set of principles, yet each used completely different designs. Kimball created the Dimensional Data Model (Figure 1.2).

Also known as the Star Schema (because it resembles a star), a Dimensional Data Model has a distinct shape. In the middle is a Fact table (a Fact is an event, transaction, or something that happens at a single moment in time). Surrounding the Fact table are Dimension tables. Each Dimension table holds all the permutations of a single hierarchy of the company (e.g., geography: city, county, state, region, district, etc.; or time: second, minute, hour, day, fiscal week, payroll week, fiscal quarter, etc.).

Bill Inmon preferred the Third Normal Form Data Model (Figure 1.3). Rather than capture hierarchies and relationships in Dimension tables, the Third Normal Form allowed the data to have the same flexibility as the company.

Within the data warehousing community, a debate emerged. Which was better, the Dimensional Data Model or the Third Normal Form data model? By the twenty-first century, the answer was clear — both. Both designs had their strengths and their weaknesses. Rather than apply a "one size fits all" mindset, data warehouse designers learned to apply the strengths and avoid the weaknesses of both in each situation.

## Storing the Data

While the debate between Dimensional Data Model and Third Normal Form Data Model was still going on, the data warehousing community was also deciding how to physically store the data. Three methods were found: a central Enterprise Data Warehouse (EDW), several distributed Data Marts, and an Operational Data Store (ODS).

The central EDW held all the data from all the business subjects in one database (Figure 1.4). The Data Mart held one subject area only (Figure 1.5). If another

subject area were needed, then that would be another Data Mart. The best method of feeding data to a Data Mart was to integrate that data into an Enterprise Data Warehouse first and then send the data on to its Data Mart.

The Operational Data Store (Figure 1.6) lives on the other side of the EDW and retrieves operational data from the business, integrates the data, and stores the data in its own database. Unlike the EDW, the data in an ODS is volatile. Volatile means the ODS only stores the value for a data element (e.g., balance on hand) that is true at real-time (e.g., balance on hand as of right now). This is different from the non-volatile data in an EDW (e.g., balance on hand for every day for the past two years). When an ODS is present, the EDW can gather its data from the ODS rather than from the business. There's no need to ask the business the same question twice.

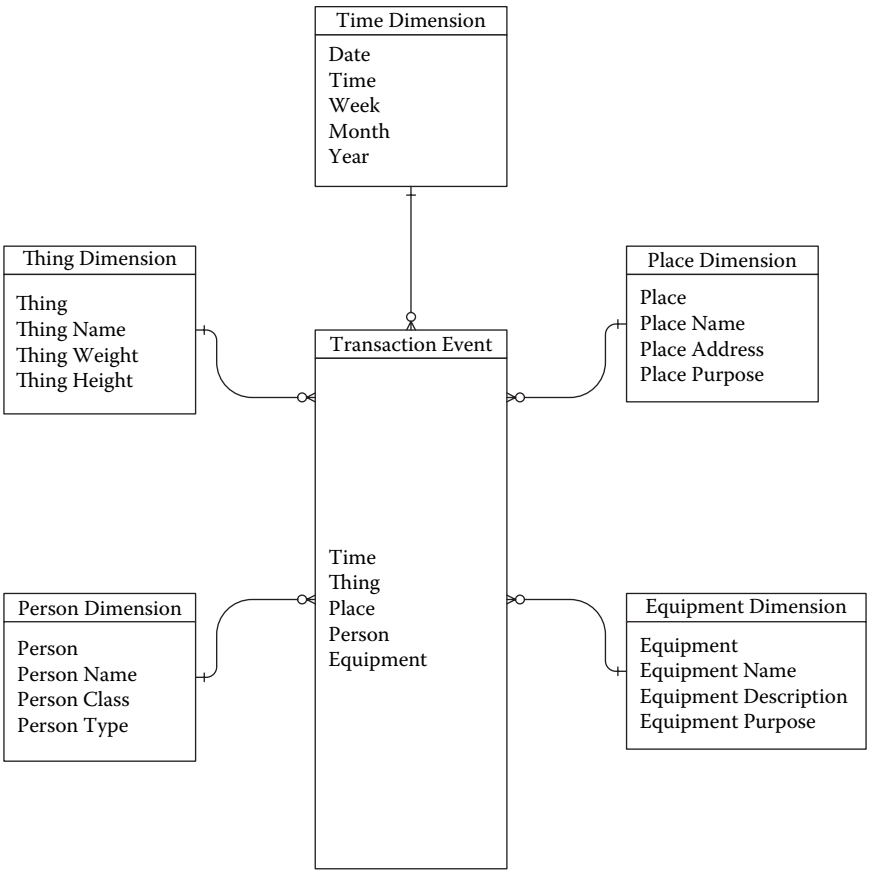


**Figure 1.2 Dimensional Data Model.**

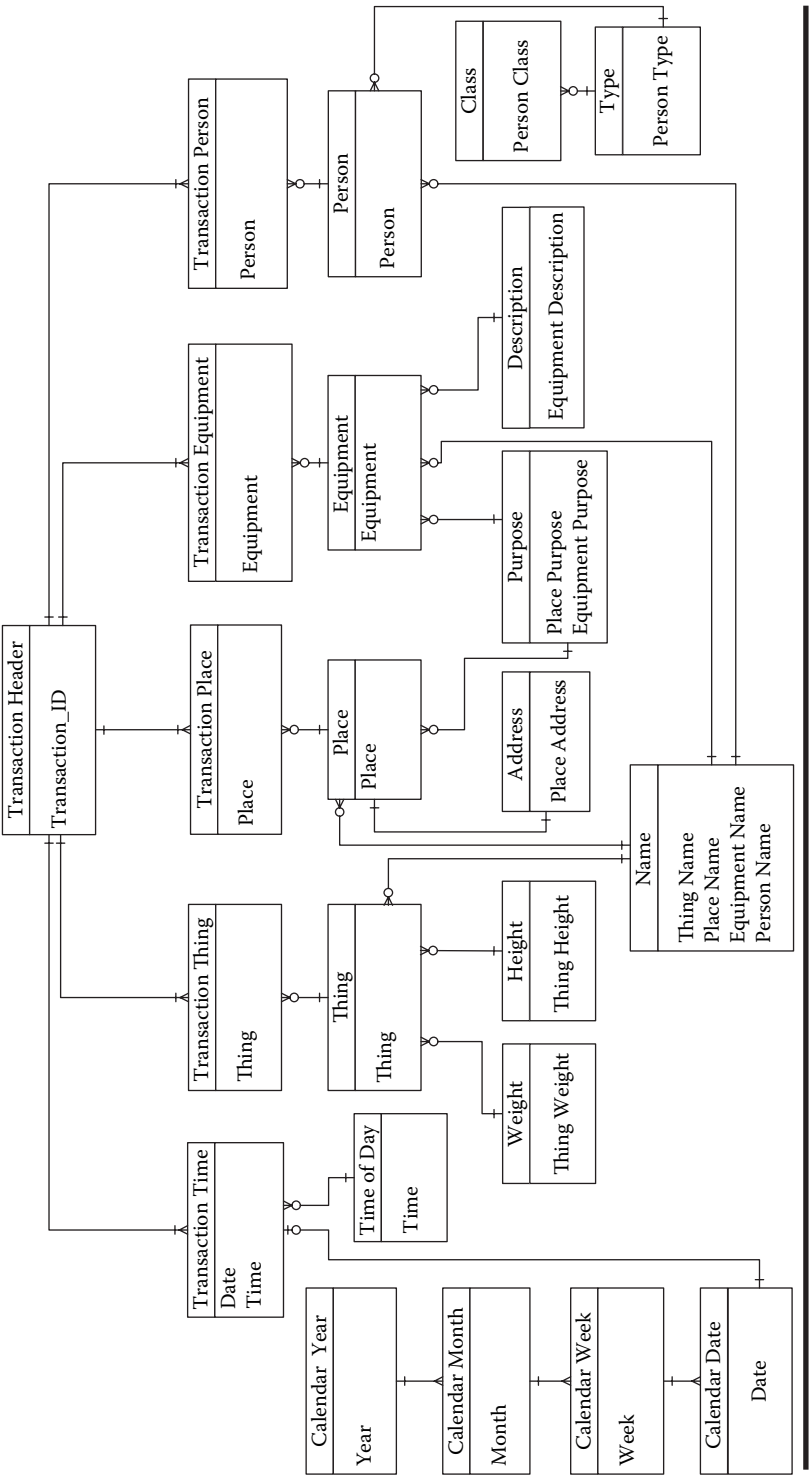


**Figure 1.3 Third Normal Form Data Model.**

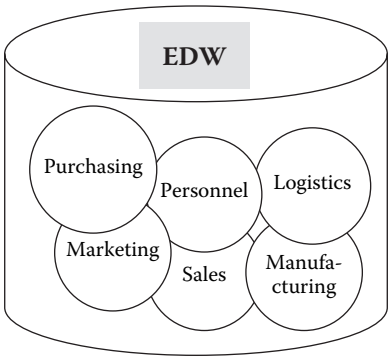


**Figure 1.4 Enterprise Data Warehouse (EDW).**

The set of applications that gather data from the business and bring that data into the data warehouse are called extract, transform, and load (ETL) applications. The ETL analyst is responsible for making the data warehouse philosophy happen.

- **Data Integration**: ETL applications integrate the data from the business, regardless of its origin, form, function, or grain.
- **Nonvolatility**: ETL applications introduce new data without destroying old data.
- **Time Variant**: ETL applications store the data with a key structure that points to a timeframe.
- **One Version of the Truth**: ETL applications reference only the one gold standard for every data element.
- **Long-Term Investment**: Populate data into a data warehouse, realizing the long-term flexibility of the data warehouse design.

## Data Availability

The data inside a data warehouse is of no use to a business without a way to use that data. Business Intelligence Reporting, also known as BI Reporting, is a set of applications by which a business can harness data and information in a data warehouse. Data is individual bits of facts and figures. By itself, data tells the business very little. Information is the compilation of individual bits of data into an observation or conclusion, which adds value to the business.

BI Reporting includes various methods by which data and information can be available to the business. Predefined reports, a staple of all information systems, can disseminate answers to the same questions (e.g., who, how many, where) on a daily basis. Interactive reports allow the business to ask a new question, or revise an existing question, and then receive the answer. OLAP (online analytical processing)

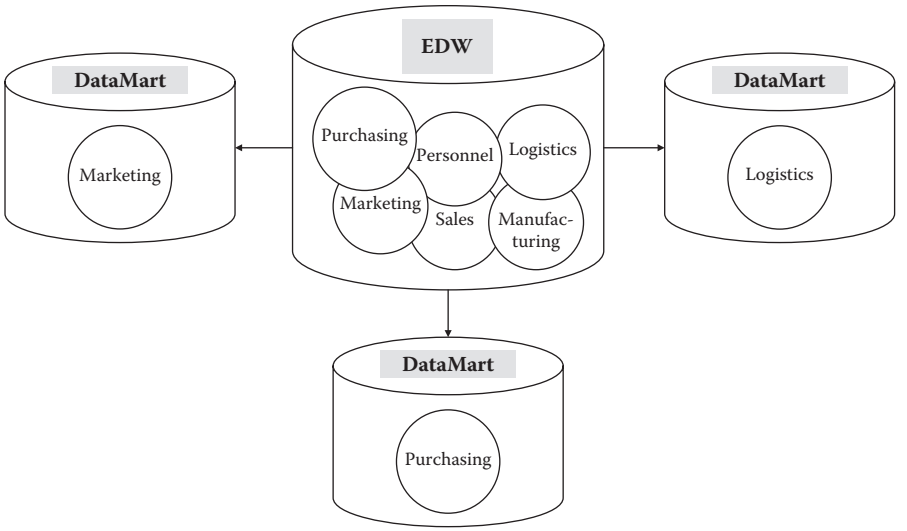


**Figure 1.5 EDW and Data Marts.**

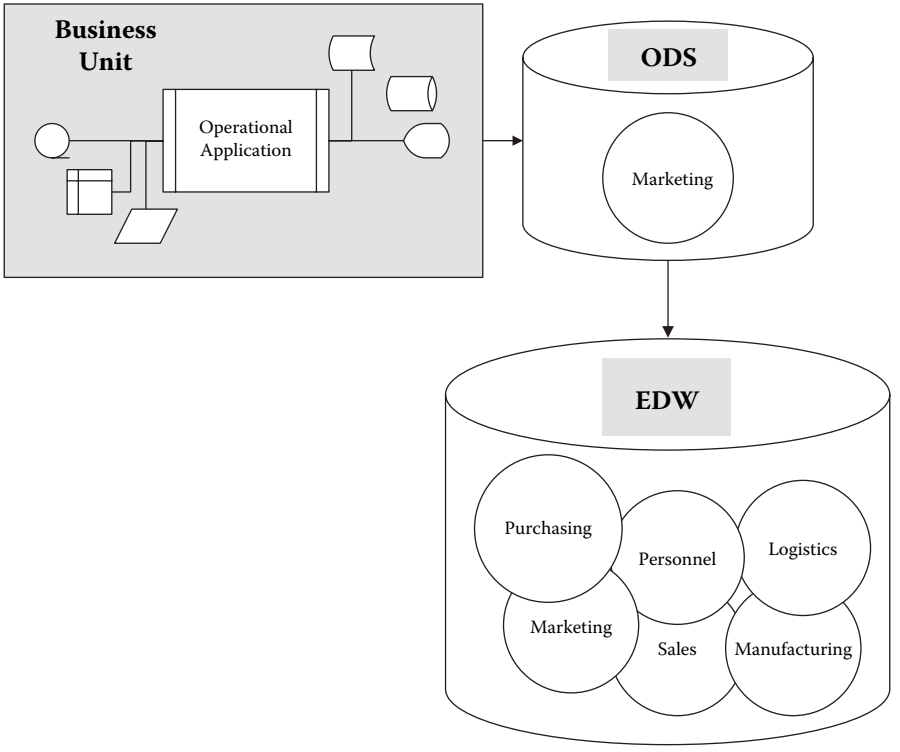


**Figure 1.6 Operational Data Store.**

reporting allows business analysts to drill up and down, left and right in stream of consciousness analysis. Using the Internet, all of these reporting options are available online. The next frontier of BI Reporting is data mining, the search for correlations in the business, which cannot be seen.

Metadata provides the background and context, which gives concrete meaning to the facts and figures in a data warehouse. Every four years, on the first Tuesday of November, all Americans use the same metadata — the number of voting precincts reporting. With 50 percent of the precincts reporting, no one believes the result. With 75 percent of the precincts reporting, we begin to take the result seriously. When 90 percent of the precincts have reported their numbers, we turn the TV off and go to bed; the election is over. Metadata in a data warehouse works the same way.

- How complete is the data?
- What is the formula for that number?
- When did the new numbers come in?

Like the number of precincts reporting, metadata in a data warehouse gives meaning and context to data.

## Monitoring Data Quality

Finally, data quality is the continuous effort to monitor the accuracy, completeness, and confidence of the data in a data warehouse. The world is full of surprises, and some of them affect the data in a data warehouse. Only the naïve assume the business and its data warehouse live in a perfect world where nothing goes wrong. Diligently monitoring data before it enters the data warehouse, the goal is to deliver data and information from which a business can derive its strategic and tactical decisions with confidence.

The explosion of data and information truly was an explosion. The facts and figures of business found their own homes in accounting systems, inventory databases, and a myriad of home-grown applications, all of which help run the business. Data warehousing gathers and integrates that disparate data so the business, through its data, can be seen in one place.