Data Warehouse Design for E-Commerce Environment

Il-Yeol Song and Kelly LeVan-Shultz
College of Information Science and Technology
Drexel University
Philadelphia, PA 19104
(Song, sg963pfa)@drexel.edu

ABSTRACT

Data warehousing and electronic-commerce are two of the most rapidly expanding fields in recent information technologies. In this paper, we discuss the design of data warehouses for e-commerce environment. We discuss requirement analysis, logical design, and physical design issues in e-commerce environments. We have collected an extensive set of interesting OLAP queries for e-commerce environments, and classified them into categories. Based on these OLAP queries, we illustrate our design with data warehouse bus architecture, dimension table structures, a base star schema, and an aggregation star schema. We finally present various physical design considerations for implementing the dimensional models. We believe that our collection of OLAP queries and dimensional models would be very useful in developing any real-world data warehouses in e-commerce environments.

1. Introduction

In this paper, we discuss the design of data warehouses for the electronic-commerce (e-commerce) environment. Data warehousing and e-commerce are two of the most rapidly expanding fields in recent information technologies. "E-commerce provides for sharing of business information, maintaining business relationships, and conducting business transactions by means of telecommunication networks [Zwas96]." Forrester Research estimates that the e-commerce business in USA could reach to $327 billion by 2002 [Forr], and International Data Corp. estimates that at more than $400 billion [IDC]. Therefore, business analysis of e-commerce will become a compelling trend for competitive advantage. A data warehouse is an integrated data repository containing historical data of a corporation for supporting decision-making processes. A data warehouse provides a basis for online analytic processing and data mining for improving business intelligence by turning data into information and knowledge. Since technologies for e-commerce are being rapidly developed and e-businesses are rapidly expanding, analyzing e-business environments using data warehousing technology could enhance significant business intelligence. A well-designed data warehouse would feed business with the right information at the right time in order to make the right decisions in e-commerce environments.

Designing a data warehouse is time-consuming. Pressures from business, both internal and external, are forcing data warehouse projects to show their usefulness to the business quickly. The data warehouse designers must be certain that the data warehouse

will contain all of the information necessary for the business executives to make informed decisions about the direction of their business. These internal pressures to prove the worth of the data warehouse have made the requirements definition phase extremely important and time-consuming. External pressures can also be seen affecting data warehouse design such as the rapidly changing technologies available and the expanding field of businesses. These requirements could be even more significant in an e-commerce environment.

The addition of e-commerce to the data warehouse brings both complexity and innovation to the project. E-commerce is already acting to unify once standalone transaction processing systems. These transaction systems such as the sales system, the marketing system, the inventory system, and the shipments system all need to be accessible to each other for the e-commerce business to function smoothly over the internet. In addition to typical business aspects such as customers, sales, shipments, and payments, e-business now needs to analyze additional factors unique to the Web environment. For example, it is known that there is significant co-relation between the web site design and sales and customer retention in e-commerce environment [LS98]. Other unique issues include capturing the navigation habits of its customers [Kimb99], customizing the Web site design or Web pages, and contrasting the e-commerce side of its business against catalog sales or actual store sales. Data warehousing could be utilized for all of these e-commerce specific issues. Another concern in designing a data warehouse in the e-commerce environment is when and how we capture the data. Many interesting pieces of data could be automatically captured during the navigation of web sites.

In this paper, we present requirement analysis, logical design, and physical design issues in building a data warehouse for e-commerce environments. We have collected an extensive set of interesting OLAP queries for e-commerce environments, and classified them into categories. Based on these OLAP queries and Kimball's methodology for designing data warehouses [KRRT98], we illustrate our design with a data warehouse bus architecture, dimension table detail diagrams, a base star schema, and an example of aggregation schema. We finally present various physical design considerations for implementing the dimensional models. We don't claim that our model could be universally used for all e-commerce businesses. However, we believe that our collection of OLAP queries and dimensional models could provide a framework for developing a real-world data warehouse in e-commerce environments.

We assume our reader is familiar with the basic terminology of dimensional modeling such as star schema, fact table, dimensions, and aggregation. For the details of the terminology, we recommend the work in [CD97, KRRT98]. We also assume our target environment is ROLAP, rather than MOLAP that uses hypercube structures [DSHB98].

The remainder of this paper is organized as follows: Section 2 discusses our data warehouse design methodology. Section 3 presents requirement analysis aspects and interesting OLAP queries in e-commerce environments. Section 4 covers logical design and development of dimensional models for e-commerce. Section 5 discusses physical

design aspects of the logical dimension models. Section 6 concludes our paper and discusses further research issues in designing data warehouses for e-commerce.

2. Our Data Warehouse Design Methodology

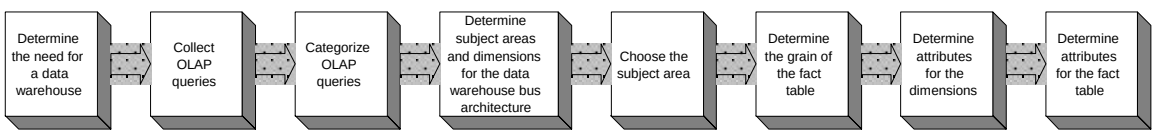
The objective of a data warehouse design is to create a schema that is optimized for decision support processing. OLTP systems are typically designed by developing entity-relationship diagrams (ERD). There are some researches that show how to represent a data warehouse schema using an ER-like model [Kimb97, GR98, SBHD98, AS97] or how to use the ER model to verify the star schema [KS97]. Ceri, Fraternali, and Paraboschi propose ten design principles for data intensive web sites [CFP99]. The data schema for a data warehouse must be simple to understand for a business analyst. The data in a data warehouse must be clean, consistent, and accurate. The data schema should also support fast query processing. The dimensional model, also known as star schema, satisfies the above requirements [Kimb96, KRRT98, AV98]. Therefore, we focus on creating a dimensional model that represents data warehousing requirements.

The data warehousing design methodologies are still evolving as data warehousing technologies are evolving and we do not have a thorough scientific analysis on what makes data warehousing projects fail and what makes them successful. According to a study by the Gartner group, the failure rate for data warehousing projects runs as high as 60%. Our extensive survey shows that one of the main means for reducing the risk is to adopt an incremental developmental methodology [MC98, AM97, KRRT98]. These methodologies let you build the data warehouse based on architecture. As we gain more experience with data warehousing projects, the design methodologies will also become mature. There are several design methodologies discussed in the literature. Meyer and Cannon [MC98] present a detailed 22 steps to develop a data warehouse from team building to the implementation phase. Anahory and Murray [AM97] also presents an architecture-based methodology for a data warehouse development. In this paper, we adopted the data warehousing design methodology suggested by Kimball and others [KRRT98]. The methodology to build a dimensional model consists of the following four steps:

1. Choose the data mart
2. Choose the grain of fact table
3. Choose the dimensions appropriate for the grain
4. Choose the facts.

Our design methodology is based on [KRRT98] and can be summarized in Figure 1. We have collected an extensive set of OLAP queries for requirement analysis. We used the OLAP queries as a basis for the design of the dimension model.

Figure 1. Dimensional Modeling Process for E-commerce Data Warehouse



3. Requirement Analysis for Data Warehouses in E-commerce

In this section, we present our approach for requirement analysis based on an extensive collection of OLAP queries.

3.1 Requirements

Requirements definition is an unquestioningly important part of designing a data warehouse especially in an electronic commerce business. The data warehouse team usually assumes that all the data they need already exists within multiple transaction systems. In most cases this assumption will be correct. In an e-commerce environment, however, this may not always be true since we also need to capture e-commerce unique data as summarized below.

The major problems of designing a data warehouse for e-commerce environments are:

- Handling of multimedia and semi-structured data
- Translation of paper catalog into a Web database
- Supporting user interface at the database level (e.g., navigation, store layout, hyperlinks)
- Schema evolution (e.g., merging two catalogs, category of products, sold-out products, new products)
- Data evolution (e.g., changes in specification and description, naming, prices)
- Handling meta data
- Capturing navigation data within the context [Kimb99]

In this paper, we focus on data available from transactional systems and navigation data. To our knowledge, there has been no detailed and explicit dimensional model on e-commerce environments shown in literature. Kimball discusses the idea of Clickstream data mart and a simplified star schema in [Kimb99]. But it shows only schematic structure of the star schema. Buchner and Mulvenna discuss the use of star schema for market research [BM98]. They also show only a schematic structure of star schema. We show the details of dimensional models.

3.2 OLAP Queries for E-commerce

Many business analysts are spending the majority of their time finding the data they need and formatting it for their statistical applications, and only a small fraction of the time actually analyzing the data for the business. The data warehouse needs to provide these business analysts with the useful data that they need in a usable format, therefore the requirement specifications should start with the business analysts. The best place to start is by interviewing the business analysts and determining the questions that are being asked about the business. Most business analysts can rattle off at least twenty questions that have been asked of them in the past week or month depending upon the business.

Our approach for requirement analysis for data warehousing in e-commerce environment was to go through a series of brainstorming sessions to develop the OLAP (online analytic processing) queries. We visited actual e-commerce sites to get

experience, simulated many business scenarios, and developed these OLAP queries. After we have captured the business questions and OLAP queries, we assigned them to categories. Most often these categories will become a subject area that the data mart will be designed around. For instance, consider the OLAP questions clustered underneath the Sales and Market Analysis category. As the data warehouse designers analyze these business questions they will see patterns emerging. The business wants to analyze its sales according to time, products, customers, vendors, promotions, sales channels, shipping, and locations. An astute data warehouse designer will look at this list and determine that each of those items can become a dimension table. The fact table would then consist of the important values the business wants to analyze associated with their product sales. When you combine the aforementioned dimension tables with this newly designed fact table then we now have a Sales subject area for the data warehouse.

Once the OLAP queries were collected, the designers needed some form of classification in order to categorize the queries. This classification was accomplished by consulting business experts to determine processes throughout the business. Some of the processes determined by the warehouse designers in cooperation with business experts were the following seven categories: *Sales & Market Analysis, Returns, Website design & navigation analysis, Customer service, Warehouse/Inventory, Promotions,* and *Shipping*. Another form of classification was proposed to classify the OLAP queries according to the primary dimension they would access. However, this type of classification format was considered to have too narrow a scope since many of the OLAP queries crossed the dimension boundaries. The former classification scheme based upon the business's processes translated better to the formation of data mart subject areas rather than trying to tie the OLAP queries to a single dimension..

The following seven categories and OLAP queries are provided to demonstrate only a few of the possible questions that an e-commerce business executive may be asking of his analysts. We note that these query classifications are not mutually exclusive and queries are not exhaustive. Most queries were identified by our brainstorming and simulations.

Sales & Market Analysis

- What is the purchase history/pattern for repeated users?
- What type of customer spends the most money?
- What type of payment options is most common? By size of purchase? By socio-economic level?
- What is the demand for Top 5 x's based on the time of year and location?
- List sales by product groups, ordered by IP address.
- Compared to the same month last year, what are the lowest 10% items sold?
- Of multiple product orders is there any correlation between the purchases of any products?
- Establish a profile of what products are bought by what type of clients.
- How many different vendors are typically in the customer's market basket?
- How much does a particular vendor attract one socio-economic group?

- Since our last price schedule adjustment which products have improved and which have deteriorated?
- Do repeat customers make similar product purchases (within general product category) or is there variation in the purchasing each time?
- What types of products do repeat customers most often purchase?
- For each vendor, what are the top three products offered that are most often purchased?
- What are the top 5 most profitable products by product category and demographic location?
- What are the top ten products that customers purchased in conjunction with product X?
- Which products are also purchased when one of the top 5 selling items is also purchased?
- What products have not been sold online since X days?
- In what zip codes do the highest number of sales occur?
- What day of the week do we do the most business by each product category?
- What is our average volume of business per product category per sales channel?
- What items are requested but not available and how often and why?
- What is the best sales month for each product?
- What is the average number of products per customer order purchased from the website?
- What is the average order total for customer orders purchased from the website?
- How well do new items sell in their first month?
- What season is the worst for each product category?
- What % of first-time visitors actually make a purchase?
- How many items are in the average order?
- What products attract the most return business?
- Is there a geographic correlation to with time of the year for the sales pattern of a certain product?
- Which customers who have previously ordered by phone are now using the web site(s)?
- Of the customers who access the e-commerce site(s) and don't make a purchase, how many call and order via the phone?
- Are new product offerings being introduced to established customers?
- Based on history and known product plans, what are realistic, achievable targets for each product, time period and sales channel?
- What is the sales to plan percentage variation for this year? What is the planning discrepancies?
- Have some new products failed to achieve sales goal? And should they be withdrawn from online catalog?
- Do we on target to achieve the month-end, quarter-end or year-end sales goals, by product or by region?

Returns

- How often was product X returned?

- How often did a customer request a refund and how often did they request an exchange for another product?
- What are the top 5 products which have been returned by customers after purchasing?
- Do customers who complain or return items make future purchases?
- Do certain customers repeatedly return items?

Website design & navigation analysis

- At what time of day does the peak traffic occur?
- At what time of day does the most purchase traffic occur?
- Which types of navigation patterns result in the most sales?
- How often are purchasers looking at detailed product information by vendor types?
- What are the top ten most visited pages? (Per day, weekends, months, seasons)
- How much time is spent on pages with banners and without banners?
- How does a non-purchase correlate to web site navigation?
- Which vendors have the most hits?
- How often are comparisons asked for?
- Based on website page hits during a navigation path what products are inquired about the most but seldom purchased during a visit to our website?
- Do products with pictures and extended descriptions sell better than those without pictures?
- Where are high-spending customers surfing to our website from?
- How often do customers arrive at the website from their ISP's home page?
- How often do customers arrive at the website from a site containing an ad banner?
- How often do customers make a purchase when arriving from a website containing an ad banner?
- How often do customers arrive at the website from links contained in e-mail notification?
- Do most customers use the search engine or just browse the site?
- Do items highlighted on the main page sell better?
- What % of customers who leave items in shopping basket return later and purchase them?
- How does Internet traffic bandwidth affect the number of clients?
- What is the most popular search engine through purchaser's access?
- What are the top complaints about the web site(s)?
- What groups of customers find the web site(s) hardest to use?
- Make recommendations for future purchases to the client, based on what the client purchased in the past.

Customer service

- What are the top 5 complaints about the products or services?
- Does e-mail notification of new products or price reductions to regular customers increase sales?
- How many people immediately "unsubscribe" when sent an e-mail notice?
- Did sales decrease after requiring users to register?

Warehouse/Inventory

- Which locations provide a cost-effective restock to which locations?
- What is the average back-order time, i.e. the time, when a product is out of stock, from when a customer orders the item until it is back in stock and shipped?
- Do we have adequate inventory for a particular product to meet anticipate demand?

Promotions

- After 10% discount promotion, what is the increase of sales for the products?
- To what extent did a promotion of a product effect sales of that product?
- Do sales incentives like "limited time offer" increase sales?
- Do discounts based on multiple purchases of an item increase sales?
- Do specials offered to best customers' result in increased sales?
- Is there a correlation between promotions and sales growth?
- Are some sales group achieving their monthly or quarterly targets by excessive discounting?
- What average discounts are being given for different products or channels?
- Is our advertising budget properly allocated? Do we see a rise in sales for products and in areas where we run campaigns? How much is the rise?

Shippings

- Is there a change in delivery type at different times of the year, i.e. preceding major holidays?
- What is the average time from ordering date to shipping date? Does this vary by product?
- What types of delivery options are requested per each category and region?

4. Logical Design

In this section, we present the details of logical design of data warehouse for e-commerce environments. Our methodology has been inspired by the work of Kimball [KRRT98].

4.1 Data Warehouse Bus Architecture

From analyzing the above OLAP categories, the data warehouse team can now design an overview of the electronic commerce business, called the data warehouse bus architecture as shown in Figure 2. The data warehouse bus architecture is a matrix that shows dimensions in columns and data marts as rows. The architecture was first proposed by Kimball [Kimb98] in order to standardize the development of data marts into an enterprise-wide data warehouse rather than becoming stove-pipe data marts that were unable to be integrated into a whole. By designing the warehouse bus architecture the design team determines before building any of the dimensions which dimensions must conform across multiple subject areas of the business. This lessens the impact of changes later in the development life cycle when an enterprise-wide data warehouse is being built.

Data Warehouse Bus Architecture for E-commerce Business

	Dimensions																		
Data Mart	Date	Time	Customer	Product	Promotion	Website	Navigation	Advertising	Warehouse	Vendor	Communication	CustComplaints	Cust Ship-to	Ship-from	Ship Mode	Deal	Employees	Location	Equipment
Customer Billing	x	x	x	x	x								x		x				
Sales	x	x	x	x	x	x	x	x	x				x		x			x	
Promotions	x	x	x	x	x	x	x	x											
Advertising	x	x		x	x	x	x	x											
Customer Service	x	x	x			x	x				x	x					x		
Network Infrastructure	x					x	x										x	x	x
Inventory	x			x					x	x							x	x	
Shipping	x		x	x					x	x			x		x		x		
Receiving	x			x					x	x				x	x	x	x		
Accounting	x		x	x	x											x			
Finance	x		x	x	x			x	x	x					x	x	x	x	x
Labor & Payroll	x																x	x	
Profit	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Figure 2. Data Warehouse Bus Architecture for E-commerce

The matrix allows us to identify so-called *conformed dimensions*. The conformed dimensions are those that are used by multiple data marts. Designing a single data mart still needs to examine other data mart areas to create a conformed dimension. This will ease the integration of multiple data marts into an integrated data warehouse later.

The E-commerce business has several significant differences from regular businesses. In the case of E-commerce, you don't have a sales force to keep track of, just customer service representatives. You can't influence your sales through a sales force incentives in e-commerce therefore the business must find other ways to influence its sales. This means that e-commerce businesses must pay more attention to their promotions and advertisements to determine their effect upon the business. This means tracking coupons, letter mail lists, e-mail mailing lists, banner ads, and ads within the main website.

The e-commerce business also needs to keep track of clickstream activity, something that the average business need not worry about. Clickstream analysis can be compared to physical store analysis where the business analysts determine which items in which locations sold better as compared to other similar items in different locations, such as analyzing sales from endcaps against sales from along the paths between shelves. There is similarity between navigating a physical store and navigating a website in that the e-commerce business must maximize the layout of its website to provide friendly, easy navigation and yet guide consumers to its more profitable items. Tracking this is

more difficult than in a traditional store since there are so many more possible combinations of navigation patterns throughout a website.

Also, the very nature of E-commerce makes gathering demographic information on customers somewhat complicated. The only information customers provide is name and address, and perhaps credit card information. For statistical analysis, DW developers would like demographic information, such as gender, age, household income, etc., in order to be able to correlate it with sales information, to show which sorts of customers are buying which types of items. Such demographic information may be available elsewhere, but gathering it directly from customers may not be possible in e-commerce. The data warehouse designers must be aware of this potential complication and be able to design for it in the data warehouse.

4.2 Dimension Models

4.2.1 Grain of the Fact Table

By analyzing the OLAP queries and creating the data warehouse bus architecture, the designers now have requirements they need to start building the data warehouse. The focus of the data warehouse design process is on a subject area, which is most often determined by the business before assigning a project. With this in mind the designers return to analyzing the OLAP queries with a different purpose in mind. Before they decide about dimensions, dimension attributes, or fact attributes, they need to know the grain of the fact table. The *grain* determines the lowest, most atomic data that the data warehouse will capture. The grain is usually discussed in terms of time such as a daily grain or a monthly grain. The lower the level of granularity, the more robust the design is since a lower grain can handle unexpected queries and additional new data elements later [KRRT98]. For a sales subject area, we want to capture daily line item transactions. The designers have determined that from the OLAP queries collected during the course of the requirement definition, the business analysts want a fine level of detail to their information. For example, one OLAP query is "What type of products do repeat customers most often purchase?" In order to answer that question the data warehouse must provide access to all of the line items on all sales because that is the level that associates a specific product with a specific customer.

4.2.2 Dimension Table Detail Diagrams

After determining the fact table grain, then the designers proceed to determine the dimensional attributes. The dimensions themselves were already determined when the data warehouse bus architecture was designed. Once again, the warehouse designers return to analyzing the OLAP queries in order to determine important attributes for each of the dimensions.

The warehouse team starts documenting the attributes they find throughout the OLAP queries in dimension table detail diagrams. See the Product Dimension detail diagram in Figure 3. The diagram models individual attributes within a single dimension. The diagram also models various hierarchies among attributes and properties such as slowly changing attributes and cardinality. The cardinality is shown on the top of each box inside the parentheses.

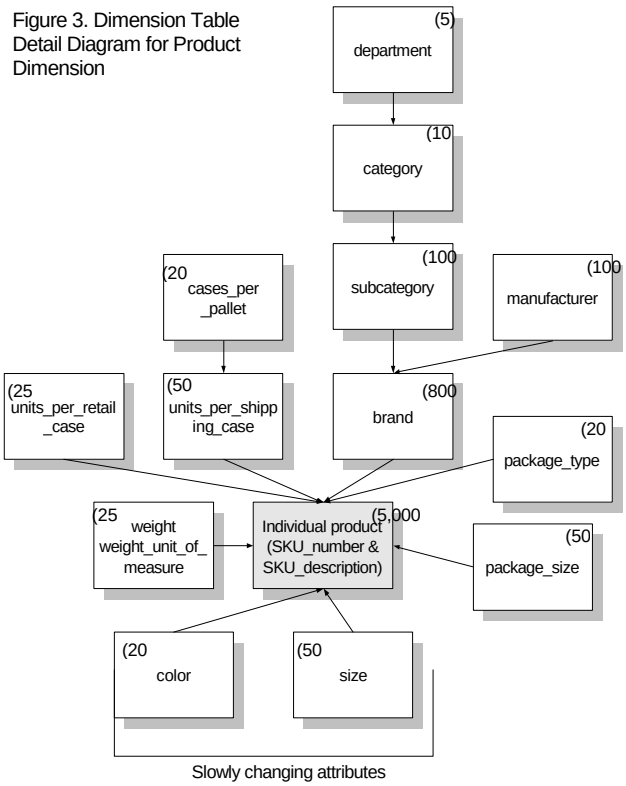


Figure 3. Dimension Table Detail Diagram for Product Dimension

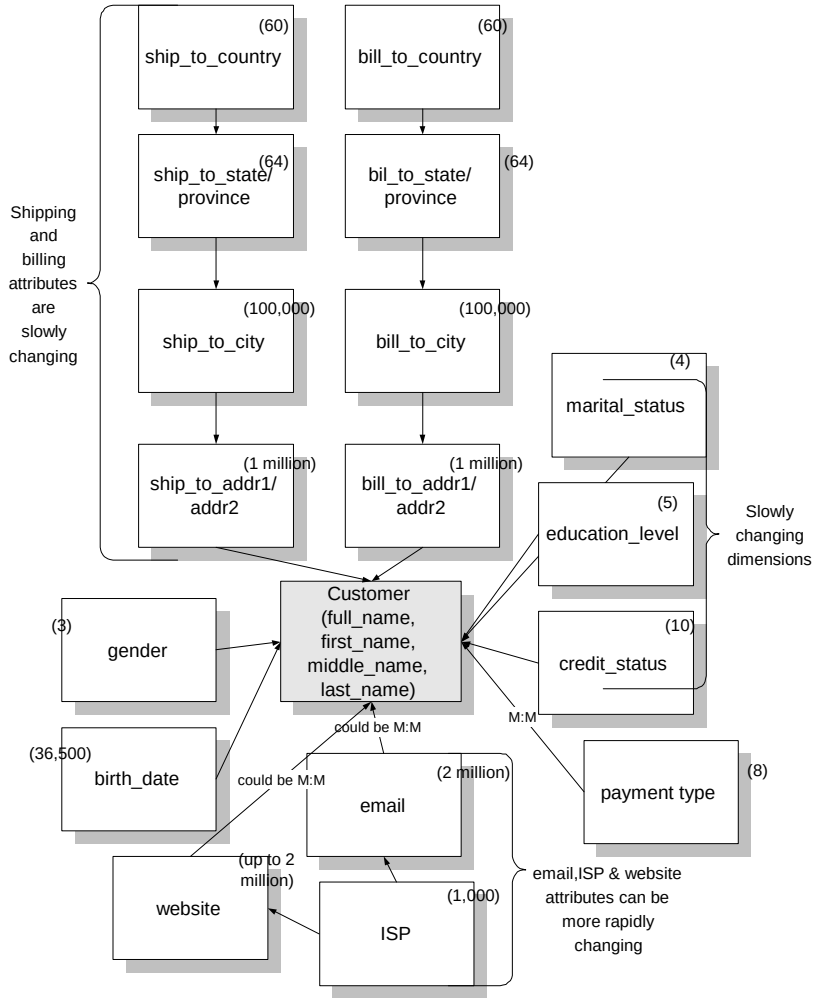
They will look closely at the nouns in each of the OLAP queries for the nouns are the clues to the dimension's attributes. For example the query "List sales by *product groups*, ordered by *IP addresses*." The designers ignore sales since that refers back to a fact to be analyzed. The designers already have product as one of their dimensions so product group would be one of the attributes. Notice that there isn't a product group attribute. One analyst uses product groups in his queries while another uses subcategories. Both of them has the same semantics. Hence, subcategory replaced the product group. One lesson to be remembered is that you don't want to finalize the design using the first set of attributes you come up with. Designing the data warehouse is an iterative process. The designers present their findings to the business with the use of dimension table detail diagrams. In most cases the diagrams will be revised in order to maintain consistency throughout the business.

Returning to our example query, the next attribute found is the IP(internet protocol) address. This is a new attribute to be captured in the data warehouse since a traditional business doesn't need to capture IP addresses. The next question is "which dimension do we put it in?"

The customer dimension detail diagram is where the ISP(Internet Service Provider) is captured. There is some argument around placing it directly within the customer dimension. In some cases, depending upon where the customer information is derived from then the designers may want to create a separate dimension that collects e-mail addresses, IP addresses, and ISPs. Design arguments such as these are where the data warehouse design team usually starts discussing possible scenarios. There is a compelling argument for creating another dimension devoted to on-line issues such as e-mail addresses, IP addresses, and ISPs. Many individuals surf to e-commerce websites to look around. Of these individuals, some will register and buy, some will register but

never buy, and some individuals will never register and never buy. The e-commerce business needs to track both individuals that become customers, those that buy, who the business will then be able to generate a more complete customer profile on. The business

Figure 4. Dimension Table Detail Diagram for Customer Dimension



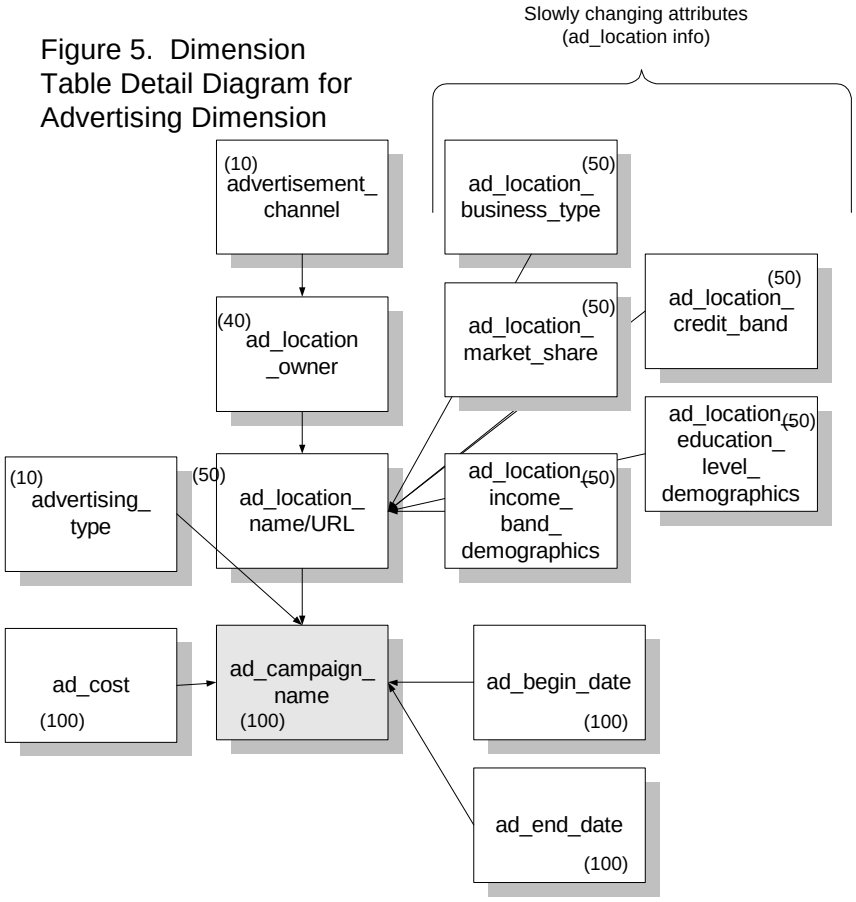
also needs to track the second set of individuals, those that don't buy, which the business will have incomplete or sometimes completely false information on.

The question that needs to be asked of the business is how they want to capture and use this information. Are they willing to tolerate incomplete and sometimes false information from potential customers in their customer dimension? If not, then a separate dimension should be created to maintain this data. Here we are only analyzing data as related to customers who have actually purchased products so the separate dimension is not depicted. The rest of the customer dimension contains the standard information captured by most businesses as it pertains to their customers.

The advertising dimension detail diagram (Figure 5) shows more of the complexities that an e-commerce data warehouse must be prepared to deal with.

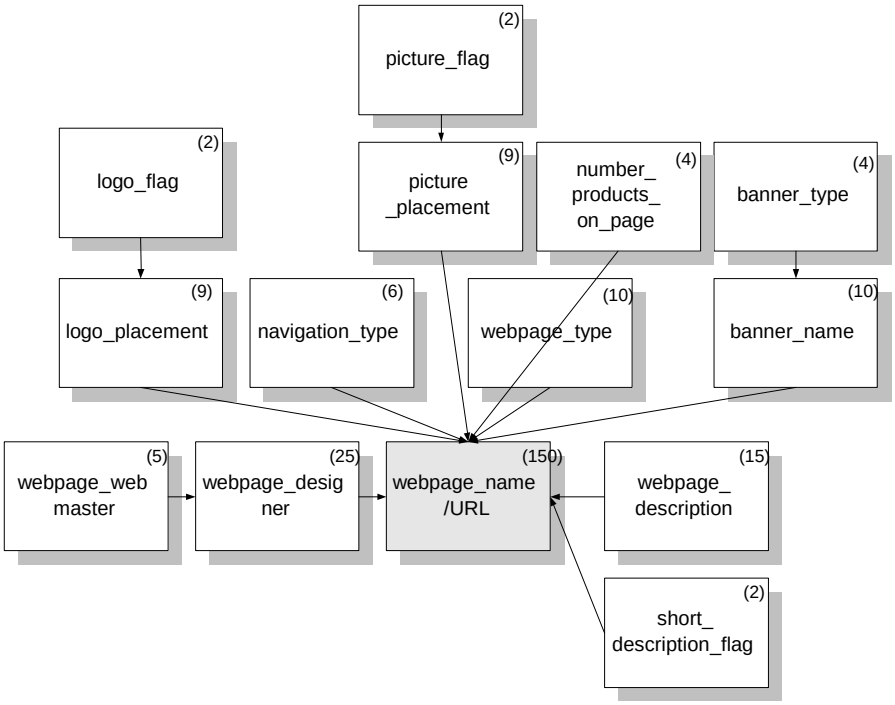
Traditional businesses that have crossed into e-commerce need to track many different forms of advertising. Specific to e-commerce, however, these businesses now must track ad banners on websites as well as the traditional means of advertising. The impact of ad banners on sales is unproven. Although businesses are spending advertising money on ad banners, many are experiencing difficulty in determining their return on investment. One of their most critical questions is whether ad banners impact their sales. By modifying the advertising dimension to accommodate banner ads, the data warehouse will provide a significant benefit to the business if it provides the means to answering that question.

Figure 5. Dimension Table Detail Diagram for Advertising Dimension



Lastly, one of the hardest decisions was where to capture website design and navigation. Does it belong in a subject area of its own or is it connected to the sales subject area? Kimball has presented a star schema for a clickstream data mart [Kimb99]. While this star schema will capture individual clicks (following a link) throughout the website, it does not directly reflect which of those clicks result in a sale. The decision to include a website dimension within the sales star schema must be made in conjunction with business experts who will ultimately be called upon to analyze the information. Many of the queries collected from brainstorming sessions centered around website navigation as well as design elements of the actual webpages. For this reason we created two dimensions, a website dimension and a navigation dimension. The website dimension (see website dimension detail diagram) provides the design information about

Figure 6. Dimension Detail Diagram for Website Dimension



the webpage the customer is purchasing a product from. Having information about webpage designs that result in successful sales is another tangible benefit of including a website dimension in the data warehouse. Much research has been invested in determining how best to sell products through catalogs and that information may not translate the same way to selling products in an interactive medium such as the internet. Businesses need more than access to data about which webpages are selling products. They need to know the specifics about that page such as how many products appear on that page and whether there are just pictures on the webpage or just a text description, or both. Much of the interest has focused on navigation patterns and website interface design, sometimes to the exclusion of the actual design of the product pages. The business needs access to both kinds of information, which the following design provides. Our model for the Website dimension is shown in Figure 6.

These are just a few examples of the dimension detail diagrams that were devised for the e-commerce star schema. After several iterations of determining attributes, assigning them to dimensions, and then gaining approval from the business experts, then the design of the dimensions stabilizes.

4.2.3 Fact Table Detail Diagram

Next, the fact table's attributes must be determined. All the attributes of a fact table is captured in fact table detail diagram as in Figure 7. The diagram shows a complete list of all the facts including derived facts. Often the facts will be directly determined from the transaction record. In this case, the record of a sale is the order and it doesn't matter whether it's a slip of paper or an electronic record. The basic facts that

we'll capture are the line item product price, the line item quantity, the line item discount amount. Some facts can be derived once the basic facts are known such as the line item tax, the line item shipping, the line item product total, and the line item total amount. Other non-additive facts may be usefully stored in the fact table such as average line item price and average line item discount, however these derived facts may be more useful calculated at an aggregate level such as rolled up to the order level instead of the order line item level. Once again the fact attributes will need to be approved by the business experts before proceeding.

Figure 7. Fact Table Detail Diagram for E-commerce Sales (Grain: Line Item)

Date_key
Time_key
Customer_key
Product_key
Promotion_key
Website_key
Navigation_key
Advertising_key
Warehouse_key
Customer_Shipto_key
Ship_mode_key
Location_key
Order_number
Order_line_number
line_item_product_price
line_item_quantity
line_item_discount_amount
line_item_product_total*
line_item_tax*
line_item_shipping*
line_item_total_amount*
average_line_item_price*&
average_line_item_discount*&
Legend: * for derived facts and & for non-additive facts

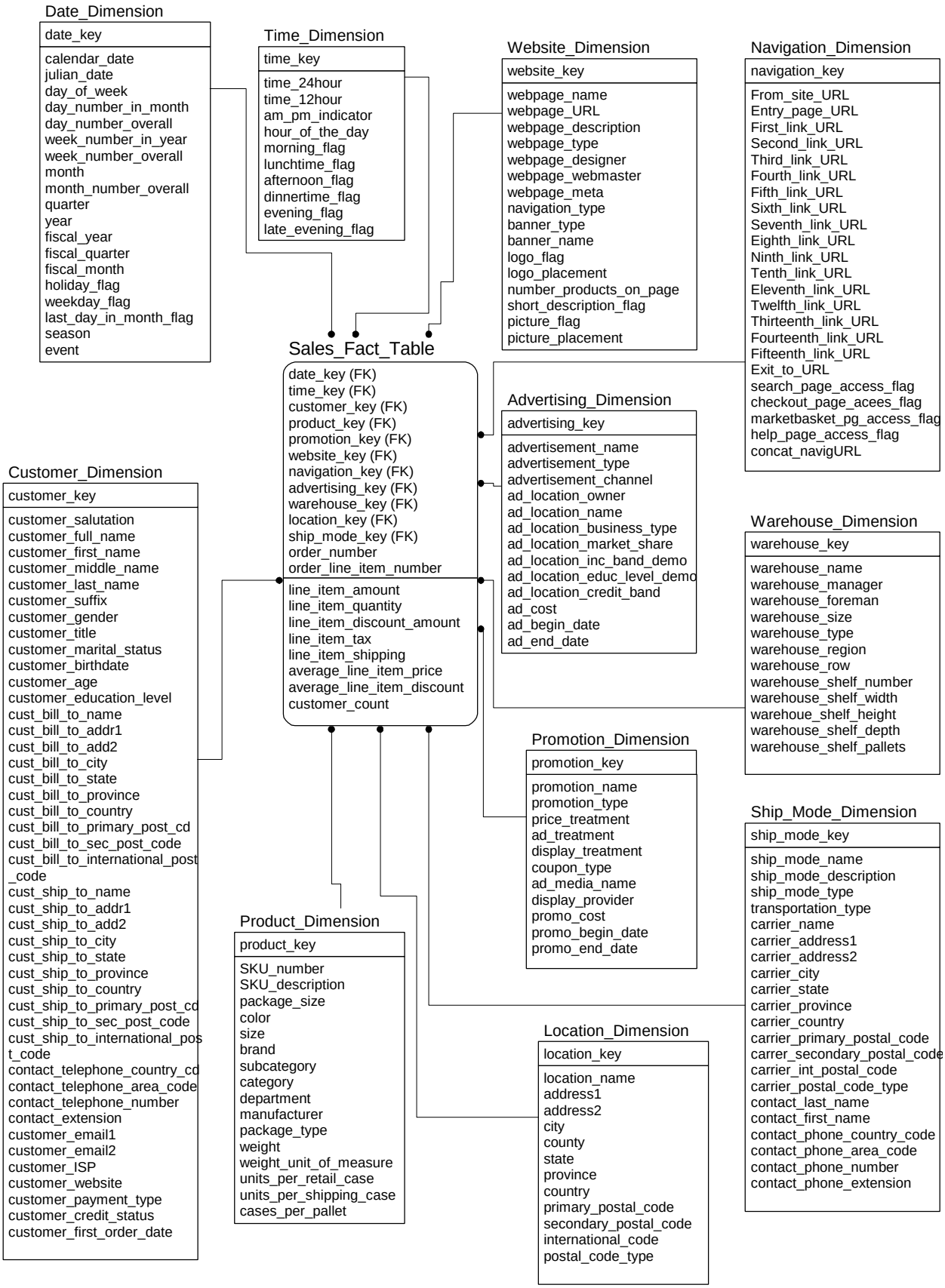


Figure 8 Base Star Schema for E-commerce Sales

4.2.4 A Complete Star Schema for E-Commerce

Finally, all the pieces have been revealed. Through analyzing the OLAP queries we started with, we have determined the necessary dimensions, the grain of the fact table at the daily line item transaction level, the attributes that belong in each dimension, and the measurable facts that the business wants to analyze. These pieces are now joined into the logical dimensional diagram for the sales star schema for e-commerce in Figure 8. As you can see from the star schema, the warehouse designers have provided the business users with a wide variety of views into their e-commerce sales. Also, by having already collected sample OLAP queries, the data warehouse can be tested against realistic scenarios. The designers pick a query out of the list and test to see if the information necessary to answer the question is available in the data warehouse. For example, how much of our total sales occur when a customer is making a purchase after arriving from a website containing an ad banner? In order to satisfy this query information needs to be drawn from the navigation and the advertising dimensions, and possibly from the customer dimension as well if the analyst wants further groupings for the information. The data warehouse design team will conduct many scenario-based walkthroughs of the data warehouse in conjunction with the business experts. Once everyone is satisfied that the design of the data warehouse will satisfy the needs of the business users then the next step is to complete the physical design.

5. Physical Design and Aggregation

In this section, we present the consideration of storage and indexing for physical design. Finally, we discuss the aggregation of the fact table and show an example aggregation schema.

5.1 Storage and Indexing

During the physical design, the warehouse designers make the decisions about how the warehouse will be implemented into the database. Many of these decisions will be based upon which commercial database they have chosen to use for the data warehouse. Regardless of the choices of database, some issues in the data warehouses' physical design are universal. In this paper, we assume we implement our design in Oracle8. We do not discuss other physical considerations such as partitioning for the lack of space.

Tablespaces

One of the first issues is where each of the tables should be created in the database. Placement of the tables should take advantage of parallel processing technology and multi-threading. Fact tables should be created in their own tablespace. If possible it should be in a tablespace with its own dedicated processor as well. The fact table will be the largest table by far in the database and will need dedicated resources since it will also be the table with the heaviest usage. Depending upon the sizes of the dimension tables they may all be lumped all together into another tablespace. If there are

one or two very large dimension tables then they may be placed into their own tablespace with the rest of the dimension tables assigned to another tablespace.

Not only are the tables themselves assigned to tablespaces, but the indexes need to be assigned as well. Because the function of a data warehouse is to provide fast and easy access to the data, the fact and dimension tables in the warehouse are heavily indexed. Again, to take advantage of parallel processing, the indexes for the dimension tables should be in a separate tablespace from the actual dimension tables themselves. The fact indexes should also be in a separate tablespace from the actual fact table. There are arguments for containing all indexes, both on the fact and dimension tables, in the same tablespace. In practice, it seems best to keep them segregated so that access to indexes on dimension tables can occur at the same time as access to indexes on the fact table do.

Indexing

Indexing schemes alone can be a hotly disputed topic on the data warehouse team. Indexing generally falls within the role of the database administrator (DBA), but the design team is also necessary at this point to guide the DBA's decision. Two main indexing techniques used for data warehousing environments are bitmap indexes [BI98] and join indexes [OG95, Vald87]. A *bitmap index* is a B tree in which each leaf node is associated with a N-bit string for N rows, where each bitmap stream is created for each value of the index. For example, a bitmap index for a gender attribute for 1 million customers will create two bitmap streams where the size of each is 1 million bits. These bitmap indexes are usually created for low cardinality attributes and perform fast AND, OR, and NOT operations. A *join index* is an index that is created based on joins between two tables. A join index can also be created among more than two tables. In that case, the join index is called a *multi-table join index*.

Each dimension table attribute that is mentioned in a query should be indexed. What kind of index used will be determined by the data values available to that field. Bitmapped indexes will be used for low cardinality attributes. The rule of thumb is that if the potential values for the attribute are less than 1% of the total records in the table then a bitmapped index should be used [CAAT98]. Otherwise, if the potential data values are greater than 1% of the total records then a B tree index can be used. To demonstrate, we'll use the product dimension as an example. At the time the product dimension detail diagram was created, the warehouse design team also documented the potential cardinality of each major attribute. From this documentation the team determines that product department, category, subcategory, and brand would be low cardinality attributes, and thus will be bitmap-indexed. The product SKU description, however, would have a B tree index rather than a bitmapped index.

The fact table provides another challenge in indexing for the data warehouse team and their DBA support. In a traditional database, the fact table would have a composite primary key index and foreign key indexes for each of the foreign keys that are contained in the table. However, the fact table has joins to every other dimension tables in the star schema. As such those indexes aren't sufficient for the accessibility necessary for a data warehouse. In order to provide the best access possible, each of the foreign keys should

have an additional individual index, either bitmapped or B tree. The type of index will again depend upon the cardinality of the attribute, as low cardinality attributes will have bitmapped indexes and high cardinality attributes will have B tree indexes. In this case, the product_key, the customer_key, and the navigation_key will all have B tree indexes while the date_key, promotion_key, website_key, advertising_key, warehouse_key, ship_mode_key, and location_key could all have bitmapped indexes.

If a join index is available in a database system, we could create join indexes between the primary key of a dimension table and the foreign key of a fact table. For example, we could define a join index between customer dimension and sales fact table using the customer_key of customer dimension and sales fact tables.

5.2 Aggregation and Materialized Views

Aggregation pre-computes a summary data from the base table. Each aggregation scheme is stored as a separate table. This subject on the aggregation is perhaps the single area that has the largest technology gap in data warehousing between research community and commercial systems. Research literature is abundant with many papers on the materialized views (MVs). The three main issues of MVs are selecting an optimal set of MVs, and maintaining those MVs automatically and incrementally, and optimizing queries using those MVs. While research on MVs focused on automating those three issues, commercial practice has been to manually identifying aggregations and maintaining them using meta data in a batch mode. Commercial systems only recently began to implement rudimentary techniques of materialized views. Microsoft OLAP Server creates aggregates based on storage and estimation of performance boost. Oracle's new version will support materialized join view, materialized aggregate view, and materialized subquery view [BDDF98].

Aggregation is probably the single most powerful feature for improving performance in data warehouses. The existence of aggregates can speed querying time by a factor of 100 or even 1000 [KRRT98]. Because of this dramatic effect on performance, building aggregates should be considered as a part of performance-tuning the warehouse. Therefore, the existing aggregate schemes should be reevaluated periodically as the business requirement changes. We note that the benefits of aggregation come with the overhead of additional storage and maintenance overheads.

Two important concerns in determining aggregation schemes are common business requests and statistical distribution of data. We prefer aggregation schemas that answer most common business requests and that reduce the number of rows to be processed.

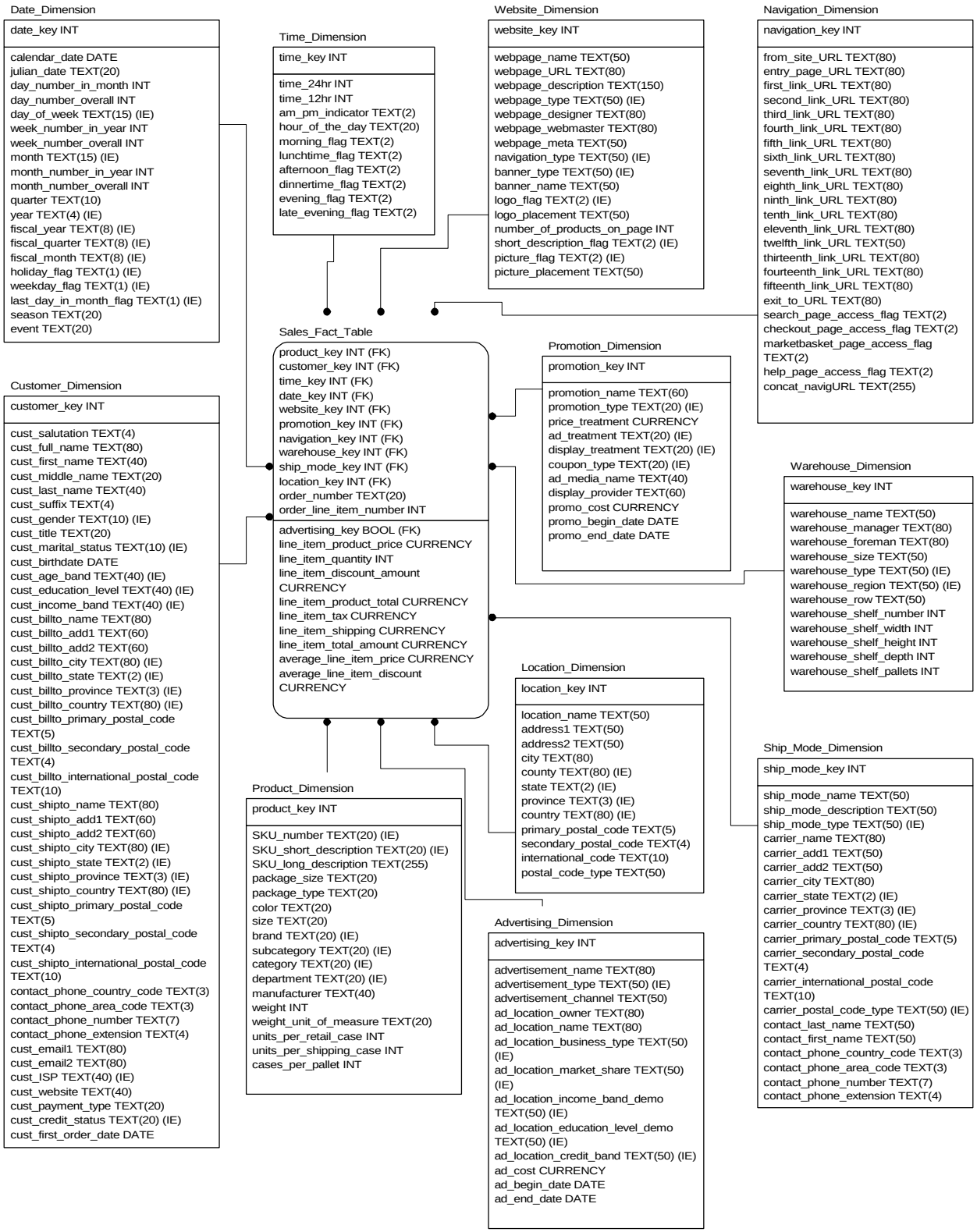


Figure 9. Physical Star Schema for E-Commerce Sales

Deciding which dimensions and attributes are candidates for aggregation returns the warehouse designers back to their collection of OLAP queries and their prioritization. With the use of the OLAP queries, the designers can determine the most common business requests. In many cases the business will have some preset reporting needs such as time periods, major products, or specific customer demographics that they routinely report. Those areas that will be frequently accessed or routinely reported become candidates for aggregation.

The second determination is made through the examination of the statistical distribution of the data. For that information, the warehouse team returns to the dimension detail diagrams that show the potential cardinality of each major attribute. This is where the potential benefits of the aggregation are weighed against the impact of the additional storage and processing overhead. The object of this exercise is to evaluate each of the dimensions that could be included in the query and the hierarchies (or groupings) within the dimensions, and list the probable cardinalities for each level in the hierarchy. Next, to assess the potential record count, multiply the highest cardinality attributes in each of the dimensions to be queried. Then determine the sparsity of the data. If the query involves the product and time, then determine whether the business sells every individual product every single day (no sparsity). If not, what percentages of products are sold every day (% sparsity). If there is some sparsity in the data then multiply the preceding record count by the percentage sparsity. Then, in order to evaluate the benefit of the aggregation, choose one hierarchy level in each of the dimensions associated with the query. Multiply the cardinalities of those higher level attributes and divide into the row count arrived at multiplying the highest cardinality attributes.

The example aggregation schema, illustrated below, is shown in Figure 10. Each aggregation schema is connected with the base star schema. The dimensions attached in the base schema is called the *base dimensions*. In the aggregation schema, the dimensions that contain a subset of the original base dimension are called *shrunken dimensions* [KRRT98].

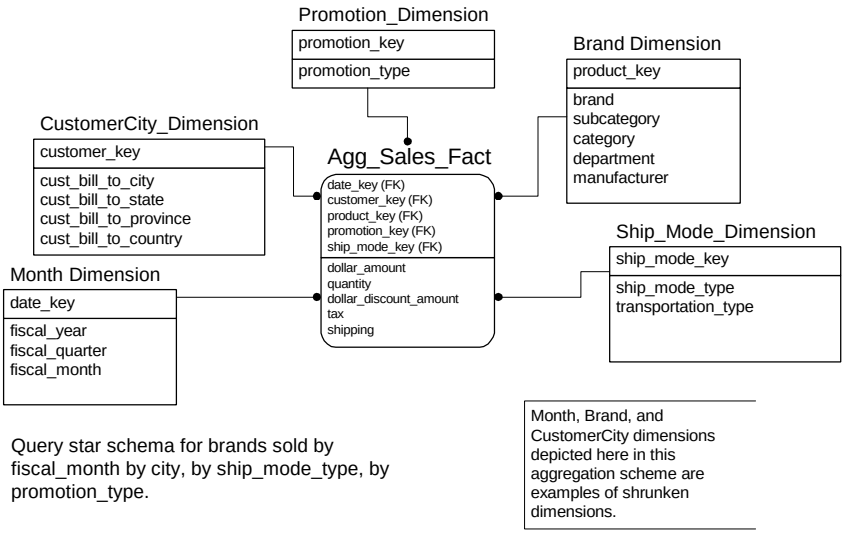


Figure 10. Aggregate Sales Star Schema

We illustrate below the use of statistical data distribution in dimensions to compute the savings by processing the number of rows. Suppose our common business requests require to process many OLAP queries *by brands sold, by fiscal month, by city, by ship mode type and by promotion type*.

Number of rows in the base fact table:
Assume 1% of products are sold each day by 0.01 % of customers with 10% of ship modes and 10% of promotions

Product	5000 *0.01 = 50
Ship mode	40 * 0.1 = 4
Promotion	200 * 0.1 = 20
Day	365
Customer	2000000 * 0.0001 = 200

Number of rows in base fact table for one year = $50*4*20*365*200 = 2.92*10^8$ rows

Query Types to optimize:
Analyze sales by brands sold, by fiscal month, by city, by ship mode type and by promotion type
Here only three dimensions are impacted and become shrunken dimensions.

Number of rows in the aggregation table:
The number of rows are reduced by Product/Brand, Day/Month, and Customer/Customer_city ratios.

Product brand	500	Impact	1/10
Month	12	Impact	1/30
Customer City	100000	Impact	1/20

Number of rows in aggregation table for one year = $50*4*20*365*200/(10*30*20) = 4.87*10^4$ rows

The example reduced the number of rows to be processed by a factor of 6000.
This aggregation still provides a lot of useful data in that it can be used for the analysis of any higher-level attributes in the hierarchy of Brand dimension or Month dimension. For example, the aggregation can be used for the analysis along the Brand dimension (Subcategory by Month, Category by Month, and Department by Month) and along Month dimension (Brand by Quarter and Brand by Year). This example shows the inherent power in using aggregates.

6. Summary and Discussion

In this paper, we have presented the design of data warehouses for e-commerce environment. We discussed requirement analysis, logical design, physical design issues, and aggregation in e-commerce environments. We have presented an extensive set of interesting OLAP queries, data warehouse bus architecture, dimension table structures, a

base star schema, physical star schema, and an aggregation star schema for e-commerce environments. To our knowledge, our presentation is the first detailed dimensional model specifically targeted for e-commerce environments. We don't claim that our model could be universally used for all e-commerce businesses. Our dimension model can be refined to focus on each specific subject area. Since our requirement has been focused on OLAP queries and not every OLAP queries are identified from the beginning, our dimensional model has to be modified and refined. However, we believe that our collection of OLAP queries and dimensional models would be very useful in developing any real-world data warehouses in e-commerce environments.

By building a data warehouse specifically for an e-commerce business, the company can leverage an increasingly important competitive edge in knowledge management that decision support systems provide. As has been proven true in countless data warehouse projects, the e-commerce business must provide a specific focus for the data warehouse to be built on. Some unique issues include capturing the navigation habits of its customers, customizing the Web site design or Web pages, and contrasting the e-commerce side of its business against catalog sales or actual store sales. Data warehousing could be utilized for all of these e-commerce specific issues.

Some difficulties in designing a data warehouse in the e-commerce environment are when and how we capture the data. We've shown some examples throughout Section 4 where these issues are discussed such as where to capture e-mail addresses, IP addresses, and ISPs. More research will need to be devoted to the best means of capturing those data elements. Some additional questions may need to be answered as well, such as where should the clicks (hyperlink selections) be tracked? Should that be a separate star schema as Kimball has proposed, or should it be incorporated within the most powerful and useful of star schemas, the sales star schema? Real-world data warehouses for an e-commerce should clarify these issues.

Acknowledgements

The authors thank the many students at Drexel University who participated in the discussion on the development of data warehousing for e-commerce environments. We especially thank Joe Poulin, Rich Kaznicki, Ninglan Tang, Yeon-Jin Lee, Joel A. Segal and Weining Xu for their insightful comments and their identification of OLAP queries in e-commerce.

References

[AM97] Anahory, S. and Murray, D., *Data Warehousing in the Real World*, Addison Wesley, 1997.

[AV98] Adamson, C. and Venerable, M., *Data Warehouse Design Solutions*, John Wiley, 1998.

[AS97] Axel, M. and Song, I.-Y., "Data Warehouse Design for Pharmaceutical Drug Discovery Research," *Proc. of 8th International Conference and Workshop on Database and Expert Systems Applications (DEXA97),* September 1-5, 1997, Toulouse, France, pp. 644-650.

[BD98] Buchner,A. and Mulvenna, M.. (1998). Discovering Internet Market Intelligence Through Online Analytical Web Usage Mining. *SIGMOD Record,* Vol. 27, No.4, December 1998, pp. 54-61.

[BDDF98] Bello, R.G., Dias, K., Downing,A., Feenan, J., and others (1998). Materialized Views in Oracle, *Proc. of the 24th VLDB Conf.,,* New York, 1998, pp. 659-664.

[CAAT98] Corey, M., Abbey, M., Abramson, I., & Taub, B. (1998). *Oracle8 Data Warehousing*. New York: Osborne/McGraw Hill Companies.

[CD97] Chaudhuri, S. and Dayal, U., An Overview of Data Warehousing and OLAP Technology, *SIGMOD Record*, Vol. 26, No. 1, March 1997, pp. 65-74.

[CI98] Chan, C.-Y. and Ioannidis, Y. , Bitmap Index Design and Evaluation, *Proc. of 1998 SIGMOD Conference*, pp. 355-366.

[CFP99] Ceri, S., Fraternalim P., and Paraboschi, S. (1999). Design Principles for Data-Intensive Web Sites. *SIGMOD Record,* Vol. 28, No.1, March 1999, pp. 84-89.

[DSHB98] Dinter, B., Sapia, C., Hofling, G., and Blaschka, M., The OLAP Market: State of the Art and Research Issues, *Proc. of Int'l Workshop on Data Warehousing and OLAP*, Washington, D.C., 1998, pp. 22-27.

[Forr] Forrester Research Inc. Home page at http://www.forrester.com/

[GR98] Golfarelli, M. and Rizzi, S., A Methodological Framework for Data Warehouse Design, *Proc. of Int'l Workshop on Data Warehousing and OLAP*, Washington, D.C., 1998, pp. 3-9.

[IDC] International Data Corp. Home page at http://www.idc.com/

[Kimb96] Kimball, R. (1996). *The Data Warehouse Toolkit*. New York: John Wiley & Sons, Inc.

[Kimb97] Kimball, R., A Dimensional Manifesto, *DBMS*, August 1997, pp. 58-70.

[Kimb99] Kimball, R. (1999). "Clicking with Your Customer," *Intelligent Enterprise,* Vol.2, No.1, pp. 70-74.

[KRRT98] Kimball, R., Reeves, L., Ross, M., & Thornthwaite, W. (1998). *The Data Warehouse Lifecycle Toolkit.* New York: John Wiley & Sons, Inc.

[KS97] Krippendorf, M. and Song, I.-Y, "Translation of Star Schema into Entity-Relationship Diagrams," *Proc. of 8th International Conference and Workshop on Database and Expert Systems Applications (DEXA97),* September 1-5, 1997, Toulouse, France, pp. 390-395.

[LS98] Lohse, G.L. and Spiller, P. (1998). "Electronic Shopping," Communications of the *ACM*, Vol.41, No.7, pp. 81-88.

[Maie98] Maier, D. and Cannon,C., *Building a Better Data Warehouse*, Prentice Hall, 1998.

[OG95] O'Neil, P. and Graefe, G., Multi-table Joins Through Bitmapped Join Indices, *SIGMOD Record*, Vol. 24, No.3, Sept. 1995, pp. 8-11.

[SBHD98] Sapia, C., Blaschka, M., Hofling, G., and Dinter, B., "Extending the E/R Model for the Multidimensional Paradigm," *Advances in Database Technologies (ER '98 Workshop Proceedings)*, Springer-Verlag, pp. 105-116.

[Vald87] Valduriez, P., Join Indices, *ACM Tran. on Database Systems*, Vol. 12, No.2, June 1987, pp. 218-246.

[Zwas96] Zwass, V. Electronic Commerce: Structures and Issues, *Int'l Journal of Electronic Commerce*, Vol. 1, No.1, Fall 1996, pp. 3-23.