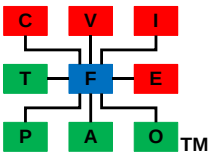


The Data Organization

1251 Yosemite Way
Hayward, CA 94545
(510) 303-8868
info@thedataorg.com

Data Mart/Data Warehouse Strategies

By Rainer Schoenrank

Data Warehouse Consultant

January 2018

Biography

Rainer Schoenrank is the senior data warehouse consultant for The Data Organization. He has degrees in physics from the University of Victoria and computer science from the University of Victoria and California State University Hayward. He has built data warehouses for clients such as Pacific Bell, Genentech, GE Leasing, SGI, PPFA, Brobeck, BofA, Clorox, Leapfrog and Intuitive Surgical. He can be reached at rschoenrank@computer.org.

1. Abstract

When embarking on a data warehousing project, there are four problems that set the project apart from other software development projects:

- Large scope – this is a project control and budget issue, not a technical problem.
- Physically large database – this is a technical implementation issue and has been solved many times, for example, Amazon.com, Google, Yahoo, etc.
- Complex database structure – this is a database design issue and the key technical problem.
- Unknown architecture – this is a strategy issue and is discussed in this paper.

"Data Mart/Data Warehouse Strategies" explores the problem of architecting the framework that encompasses the business data processing applications, the data marts and the processes that connect them in order to transfer data from the business applications to the data marts. It traces the origins of business intelligence reporting and explains why data marts are beneficial.

It enumerates the strategies to build data marts. Looking into all possible framework architectures, it shows why those architectures are considered, and reveals how the recommended architecture is determined. It explains the advantages that the recommended framework offers and identifies critical factors to successfully implement it.

The data mart/data warehouse framework determines the requirements of the extract transform and load (ETL) process required to move the business data into the data warehouse and into the data marts. These process requirements can serve as the quality assurance reference document to ensure that the ETL process implementation conforms to the architecture selected.

2. Introduction

The problem that this paper addresses is how many different ways there are to build and load data marts so that an intelligent choice can be made. The wrong choice results in data redundancy among databases, an inability to trace detailed data back to its original source data, an inability to analyze data across multiple data sources, or unnecessary data processing complexity. The wrong choices have scalability and complexity problems. As new data sources are added, errors in the data organization propagate, compounding the problems over time and resulting in unnecessary data management support and maintenance. The right choice avoids all these problems.

3. Background

In 1494, Luca Pacioli published the textbook *Summa de arithmetica, geometria, proportioni et proportionalita* (Everything about arithmetic, geometry, and proportions) that included accounting and bookkeeping practices (ten Have, 1986). It described a financial record keeping system using journals, registers and a general ledger. The daily business transactions were systematically recorded in the journals that were summarized into the monthly registers and the general ledger. The general ledger recorded the complete financial state of the business.

The textbook was the first to codify the double-entry accounting method, which enabled others to study and use it. Businesses that used it gained a competitive advantage. Also, instead of simply providing products and services to customers, they now had a way to systematically keep records of how their businesses were doing.

Today, businesses that want to know how they're doing gather data about their transactions. The transactions capture data that show how, where, when, by whom and to whom products are sold and how, where, when, by whom and from whom inventory is purchased. This data capture is used to answer a myriad of questions that are inherent in the question, "How is the business doing?" This question encompasses a myriad of detailed questions concerning sales, marketing, finance and other key performance indicators. Examples of these underlying questions are:

- Which products are the most profitable?
- Which salesperson is the most productive?
- Which customer orders the most?
- Which customer is the most profitable?

The ability to answer these kinds of questions can change the direction of a business.

The first attempt to answer these questions is typically addressed by gathering the data for each question as it arises. Each answer is provided by a unique data acquisition process that is a time-consuming one-off, resulting in few questions being asked and the answers being accepted, since they are the only answers available.

In the next attempt to answer these questions, the data gathering is typically automated within financial or customer relationship management applications. As this data becomes available, more questions are asked, and the answers are no longer one-offs, but each business application may give a different answer to the same question. Now the time-consuming process becomes reconciling the different answers. When the reconciliation process becomes too costly, the business looks for a better solution – the data mart.

4. Why Use Data Marts?

If a business had only one application to collect its data and keep its records, there would be no need to use a data mart. The business application would deliver all the reports required. But businesses today use many applications, and data marts consolidate data from many applications into a single structure. This consolidation enables the data mart users to transform the data into business intelligence and knowledge, and gives the business a way to manage and improve its operations. By examining data marts, users can determine the state of the business and detect trends to:

- Improve business processes.
- Improve sales effectiveness.
- Respond rapidly to internal changes.
- Respond rapidly to external pressures.
- Anticipate customers' needs.
- Identify hidden business opportunities.

Data marts enable users to analyze the data multi-dimensionally. Each user has a unique way of organizing and viewing the data. In the example of Figure 1, for instance, the data mart's sales transaction data is organized by the dimensions of market, product and time.

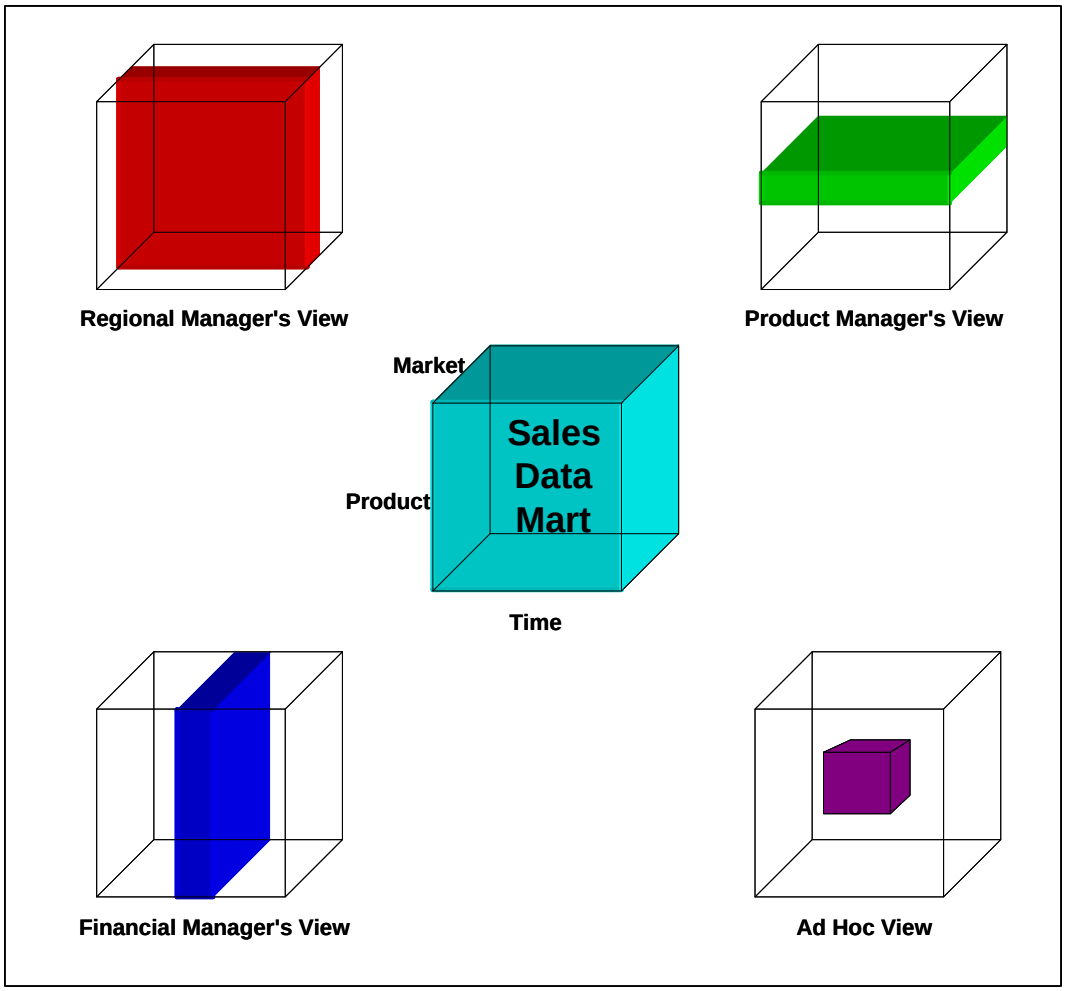


Figure 1: Different Views of a Sales Data Mart's Data

The regional manager's view of sales is what occurred in his or her region, when it occurred, and what products were involved. The product manager's view of sales is what occurred for a given product, when it occurred, and in which regions it occurred. The financial manager's view of sales is what occurred in all regions for all products in a given fiscal year. Using the same sales data mart, the regional manager, product manager and financial manager can all quickly retrieve and view the accurate information that is pertinent to their interest, and the details of the data will be identical where the dimensions match, i.e., this region, this product, this fiscal year.

With a well-designed data mart and a data analysis tool, most business users can create ad hoc views of the data, to accommodate the business' need to respond to changing demands.

The flexible data organization of a data mart enables users to analyze, over time, any combination of a customer, sale, employee, division, and/or product line. The data mart makes it easier to create any automated reports to managers, investors, federal, state and local governments, and makes it easier to continuously analyze the details to the question, "How is the business doing?"

The data mart requires the data collection from the business applications, such as sales, payments, and expenses applications, to be more complete and more systematic than the data collection used to answer each question as it arises. It ties data about sales, receipts, and operating expenses together with data about customers, employees, company organization and products. It also generates more questions and produces consistent answers faster than any other method.

The benefits of using data marts are that they give users analytic power and the capability to explore the business model that data marts encapsulate.

5. The Data Mart Process Problem

Once a business decides to use data marts, it must decide how to build and load them. But how many different ways are there to build and load data marts? Without knowing the possibilities and the potential pitfalls of some of these ways, a business cannot make an informed decision.

How the ETL processes transfer data from the OLTP applications (business applications) to the data marts is shown in Figure 2. Business applications on the left send their data to data marts on the right through a nebulous cloud of technology at the center.

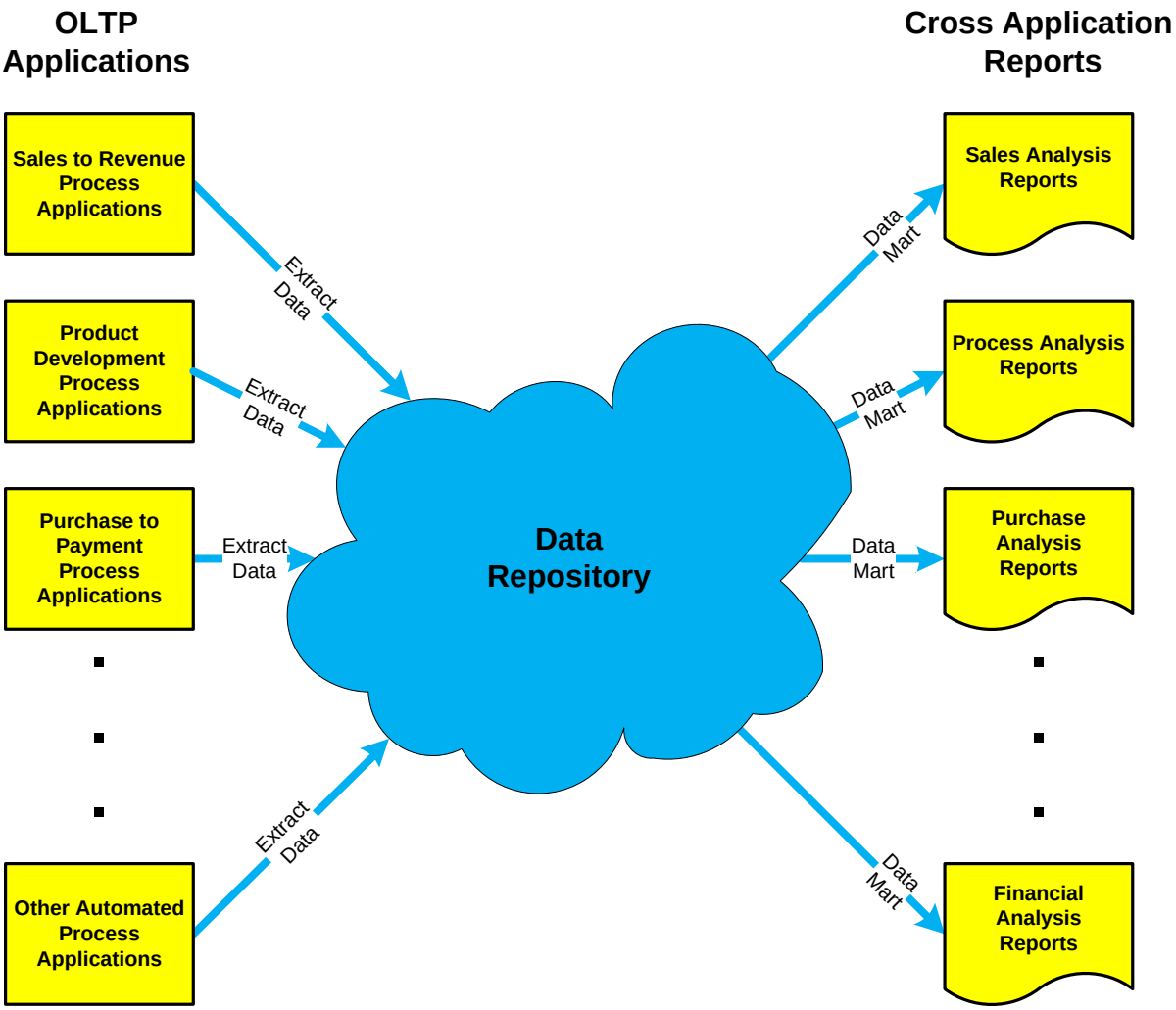


Figure 2: The Data Mart Process

For a successful data mart implementation, the cloud must be replaced by a strategy that defines a systematic process for creating data marts and loading them with data. Business users use the data mart's data to target sales, analyze costs, simplify operations, determine prices, etc.

6. Goals of a Good Strategy

Each solution for transforming the cloud into a data mart/data warehouse strategy should meet certain goals. The data mart/data warehouse strategy in the information technology infrastructure must provide a structure and framework for the design and implementation of the data marts.

Suppose that because of a business merger a new source of invoice data must be added on the right side of Figure 1. A comprehensive data mart/data warehouse strategy will have anticipated that adding the new data source (and adding a sales business application pictured) should not affect the product sales data mart's existing data. The new source's data can simply be added to the product sales data mart for more complete analytic capability.

Additionally, there are core requirements that a successful strategy must satisfy. These requirements are:

- Disallow any wholesale redundancy or duplication of a database's data.
- Enable tracing the detailed data back to the source application data as required by Sarbanes-Oxley.
- Enable analysis across multiple sources of data.
- Reduce data processing complexity.

Also, the data mart strategy must accommodate management's strategic goals, the recommended management goals are that:

- The data mart is the only repository used for reporting. Users should have no need to build their own reporting and data integration tools on their desktops to integrate data.
- The data presentation requires no programming intervention.
- The data mart development process is repeatable.

The first of these management goals is met by management discouraging users from building their own reporting and integration tools. The second goal is met by choosing the best OLAP tool for users. The data mart/data warehouse strategy gives you the process to build data marts, which can now be repeated, satisfying goal three.

7. All Possible Data Mart Strategies

For the problem shown in Figure 2, three questions must be answered before choosing a strategy and building a data mart.

- How many data sources will be considered?
- How do we move data from our business application(s) to our data mart(s)?
- How many data marts do we need?

Each of these questions has two possible answers. The possible answers are:

- The data sources can be:
    1. A single data source. The data mart(s) will enhance the reporting capabilities of one business application.
    2. Many data sources. The data mart(s) create an enterprise-wide view of the business with data from all business applications.
- The operational data can be moved to the data marts using:
    1. Only processing to load the data marts in a single step (as proposed by R. Kimball, 1998).
    2. A database to consolidate and organize the data before loading the data marts (a data warehouse as proposed by W. Inmon, 1989).
- For each of these choices, a business can build:
    1. One single data mart for the entire business.
    2. Many data marts, each one containing the business model for a single business subject area.

Answering these questions systematically yields eight possible strategies. Each answer corresponds to a different data mart/data warehouse strategy, as shown in Tables 1 and 2.

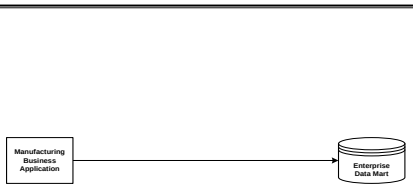
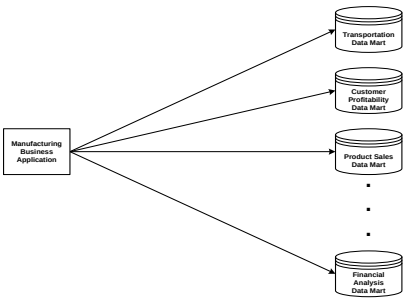
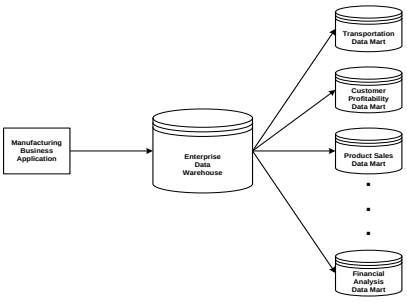
Case	Data Source	Data Warehouse	Data Mart	Architecture Diagram	Strategy Comments
1	single	no	one		This strategy is the one used in academic examples and in pilot projects to test the acceptability of data mart reporting.
2	single	no	many		The data marts replace the business application reports.
3	single	yes	one		The data warehouse database duplicates the data mart database. This data redundancy conflicts with the strategic goals.
4	single	yes	many		The data warehouse database duplicates the application database. This data redundancy conflicts with the strategic goals.

Table 1: Data Mart/Data Warehouse Strategies for a Single Data Source

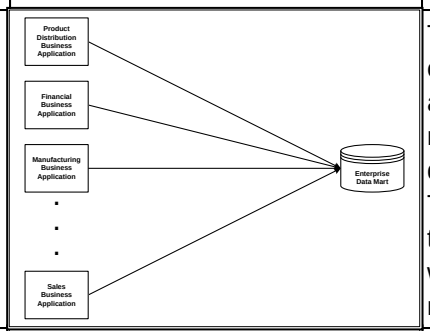
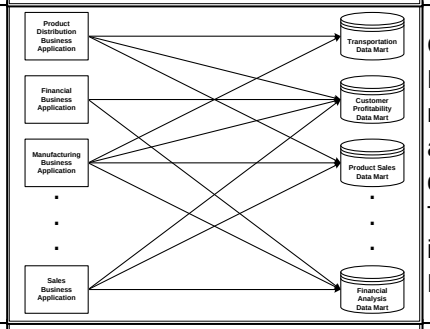
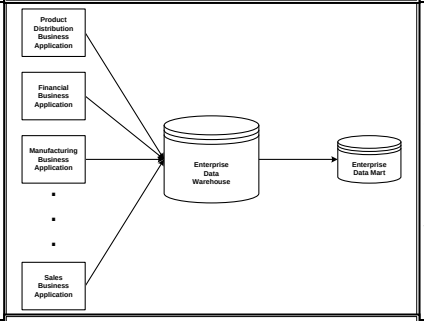
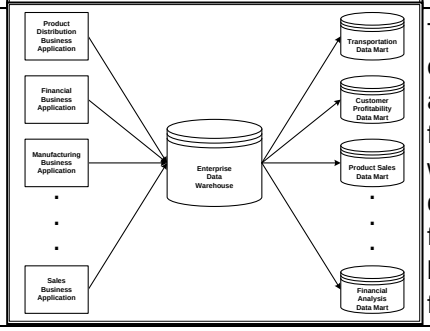
Case	Data Source	Data Warehouse	Data Mart	Architecture Diagram	Strategy Comments
5	many	no	one		This strategy requires conformable dimension analysis to get the requirements for the data mart database. This case confuses the roles of data warehouse and data mart.
6	many	no	many		Cross-functional Reporting. The data mart load processes and interactions are difficult to manage. This is the strategy implemented by SAP BW (Data Lake).
7	many	yes	one		The data mart database duplicates the data warehouse database, This data redundancy conflicts with the strategic goals.
8	many	yes	many		This data warehouse consolidation provides a simplifying construct for the data. The data warehouse database is difficult to generalize from the applications, but it has the most flexibility.

Table 2: Data Mart/Data Warehouse Strategies for Multiple Data Sources

Of the eight cases, cases 1 through 4 do not meet the multiple data sources requirement of the data mart/data warehouse strategy.

Case 1 is the pilot data mart project usually shown in Kimball's examples (Kimball, 1998), but modifying this strategy into a more general strategy is time-consuming, particularly in removing the simplifying assumptions made in the business application. The modifications to the data mart always force changes to all the loading processes.

Case 2 is a project that converts the business application reporting into a series of analytical data marts. These projects are very successful, but they only enhance operational analysis of the application and do not address the requirement of enabling analysis to be performed across multiple data sources. Any modifications to the data mart always force changes to all the loading processes.

Cases 3 and 4 constitute a project that attempts to isolate the data marts from the business application by introducing a database. These projects are difficult because of the data duplication introduced by the isolation database. When the business application has been purchased, this strategy may be necessary to get a data mart from the application data, because the application hides its database from the data marts.

Cases 5 through 8 meet the data mart strategy requirement of enabling analysis across multiple data sources.

The architecture in case 7 involves data duplication and since the data mart is a single first normal form view of the business, this formulation is highly impracticable.

The only acceptable data mart/data warehouse strategies are cases 5, 6, and 8.

8. Evaluating Acceptable Strategies

The data mart strategies that meet the strategic data mart requirements can be compared on the basis of:

- The number of processes involved.
- The number of databases involved.
- The possibility of using parallelism to speed up the data processing.
- The impact of changes on the processing
- The impact of changes on the databases.

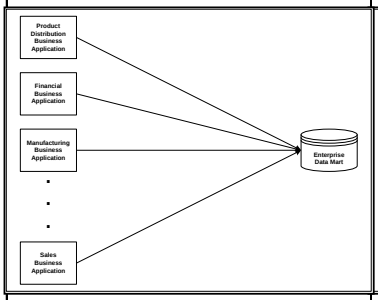
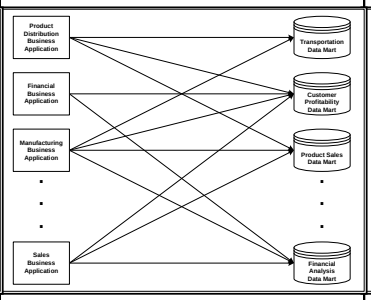
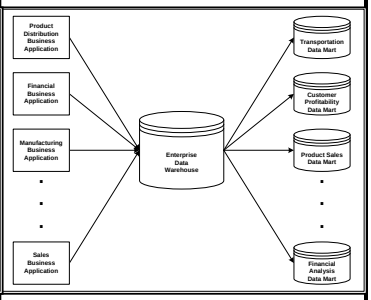
Data Mart Strategy	Case 5	Case 6 Application Strategy	Case 8 Data Strategy
Architecture Diagram of the Data Mart Strategy			
Strategy Summary	many data sources no data warehouse 1 data mart (only a single view)	many data sources no data warehouse many data marts (many views)	many data sources 1 data warehouse many data marts (many views)
Processes			
Number of processes	many	many * many	many + many
Processing complexity	high	high	low
Databases			
Number of databases	1	many	1 + many
Data Mart complexity	high	low	low
Possible Parallelism			
Extraction	no	no	yes
Data Mart	no	possible, but difficult	yes
Operational application changes			
processes impacted	1	many	1
data marts impacted	1	many	0
change complexity	high	high	low
Data Mart changes			
processes impacted	many	many	1
data marts impacted	1	1	1
change complexity	high	high	low
Data Governance Issues			
data lineage	lineage has single branch	lineage has many branches	lineage has single branch
data semantics	meaning is resolved	meaning is not resolved	meaning is resolved
data reconciliation	all reports use the data mart	required to resolve difference	all reports use the data warehouse
Example	Kimball (generalized)	virtual warehouse (SAP BW)	Inmon

Table 3: Comparing Acceptable Strategies

9. Strategy Issues

Regardless of the strategy chosen to build the data marts, three issues must be addressed. These issues are:

- The data marts must be easy to use.
- The data mart's data must have integrity.
- The data used for analysis must be selected, transformed and organized into the data mart.

The ease of use of the data mart and its OLAP analysis tool is how the business will perceive the value of the data mart. A data mart that is slow to deliver results or difficult to understand will not be used by the business. The presentation of the data, the ability to explain the meaning of the data (meta-data) and the availability of data lineage information to explain the source of the data are crucial to how well the users accept the data mart.

The integrity of the data in the data mart is crucial. Users will not trust the data mart if the results don't match the business applications. For example, the financial data mart must be able to recreate the general ledger. This is an issue in the business applications since the data required to operate successfully is less than the data required to record the transactions in the data warehouse. In the warehouse, every business transaction is related to some employee, but for accounts receivable to function properly this relationship is not necessary. To achieve data integrity in the data marts requires more complete and more systematic data collection in the business applications.

The data collection must focus on:

- Accuracy. The values recorded must be correct and translated correctly for the data warehouse. For example, the geographic location can not contain both FR and France as identifiers for the location of France. This would make analysis in the data marts much more difficult.
- Completeness. It is important to collect more timely and complete data (using all master data relationships).
- Consistency. All business sources must use the same master data lists. Keeping the master data list consistent in the data warehouse so that all the business applications can identify their customers and so that customers are not duplicated would be a full-time job.
- Totality. All sources of data must be collected.

The organizations required for the business applications are not the only organizations needed for analysis. For example, sales, production and logistics might all use customer to organize sales, but sales organize customers by account hierarchy, production organize customers by volume categories and logistics organize customers by distance from the distribution center. So, there are at least three different organizations of customer. In general, how many organizations of data are necessary? This depends on business operational and analytical requirements but probably between five and nine organizations per data mart.

10. The Recommended Solution

Table 4 compares the data mart/data warehouse strategy requirements to the strategy cases from Tables 1 and 2. The requirement column shows whether a particular strategy case meets that requirement. The requirements are:

1. Disallow any wholesale redundancy or duplication of a database's data.
2. Enable tracing the detailed data back to the source application data.
3. Enable analysis across multiple sources of data.
4. Reduce data processing complexity.
5. Allow parallel data processing.

STRATEGY				REQUIREMENT				
Data Source	Data Warehouse	Data Mart		reduce data duplication	trace data source	across source analysis	reduce processing complexity	enable parallel processing
one	none	one	**Case 1**	no	yes	no	yes	no
one	none	many	**Case 2**	yes	yes	no	yes	yes
one	one	one	**Case 3**	no	yes	no	yes	no
one	one	many	**Case 4**	no	yes	no	yes	yes
many	none	one	**Case 5**	yes	no	yes	no	no
many	none	many	**Case 6**	yes	no	yes	no	no
many	one	one	**Case 7**	no	yes	yes	yes	no
many	one	many	**Case 8**	yes	yes	yes	yes	yes

Table 4: The Requirements Met by Each Strategy

The recommended data mart strategy is Case 8 as shown in Figure 3. The recommended strategy meets all the requirements and it has simplification and generalization advantages.

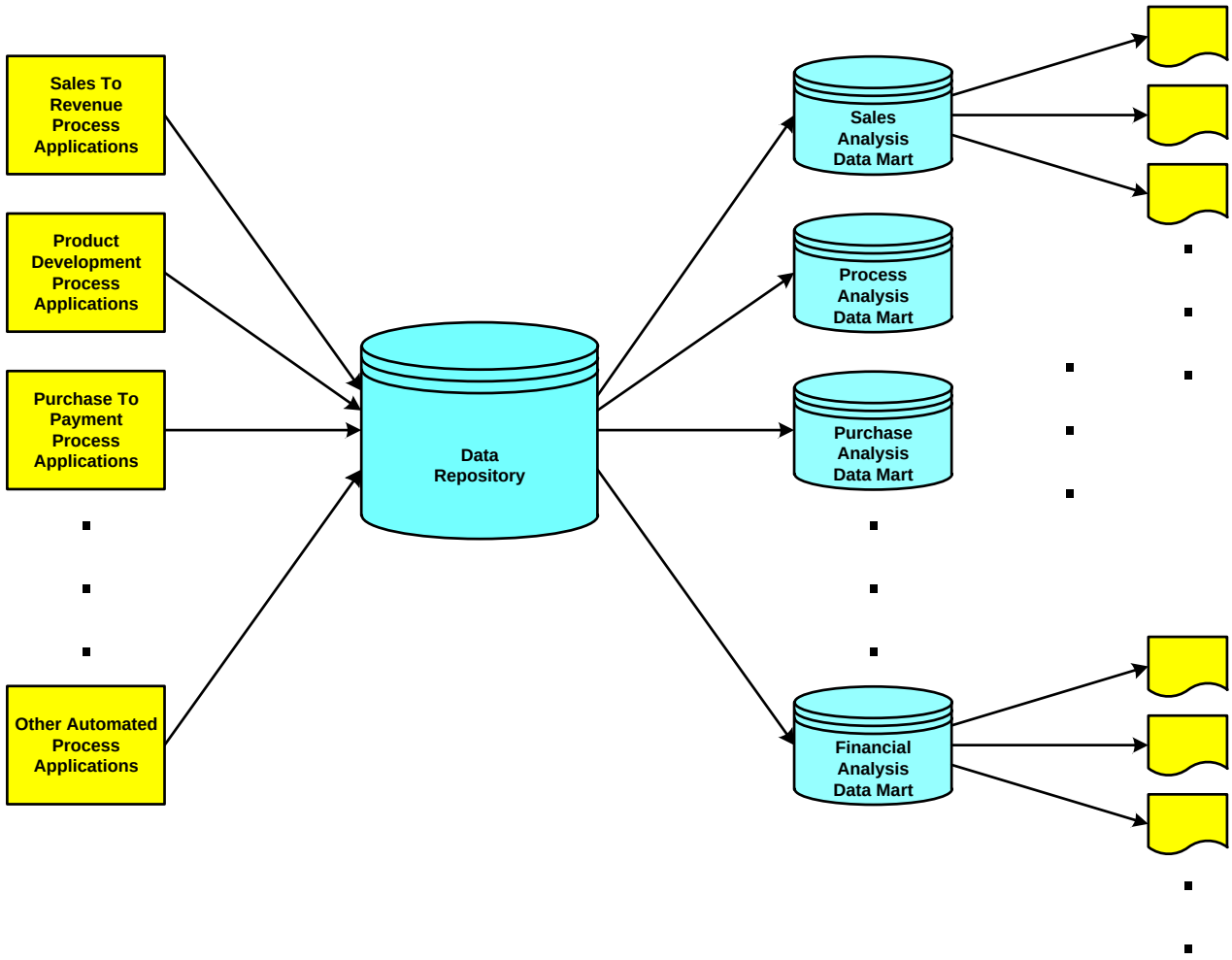


Figure 3: The Recommended Data Mart Strategy

The advantages of a single data warehouse are that it:

- Minimizes the effects of data source changes on data marts.
- Minimizes the effects of data mart reporting changes on data sources.
- Provides a consolidation database for the source data and becomes a single source of truth for the data.

The advantages of multiple data marts are:

- Processing simplicity. Using many data marts trades more storage space for simplicity of processing.
- More available processing parallelism.
- Data localization, which focuses each data mart on a particular business problem and its solution.

11. Next Steps

Having chosen the data mart strategy, the next step is to document the data warehouse architecture. The architecture document will describe the scope of the proposed data warehouse solution. The conceptual level of the data warehouse architecture is shown in Figure 4.

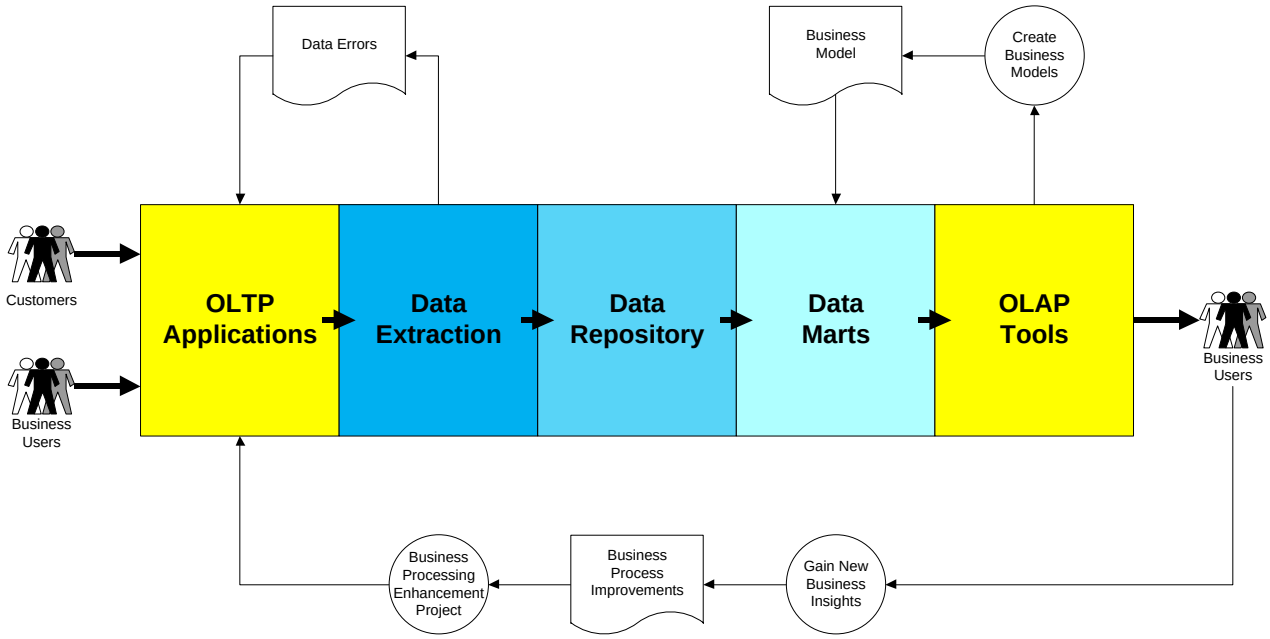


Figure 4: The Data Warehouse Architecture

The architecture has five major components and three feedback loops.

The components are:

1. The OLTP application – this is the collection of all the data sources and describes how the processing applications interact.
2. The data extraction – this is the collection of the data extraction processes, the data validation processes and the data warehouse loading processes. A well-designed extraction process (ETL process) will:
    - Remove the data assumptions and simplifications (shortcuts) made by the data source applications.
    - Save the data translation rules.
    - Simplify the ETL programming.
3. The data repository – this is the database that consolidates the data into a single relational model. The well-designed data warehouse database will:
    - Enforce data integrity.
    - Enable extensibility.
    - Store the master data organizations.
    - Contain the resolution for business data issues (i.e., different identifiers for the same customer, vendor, employee or product).
    - Resolve organizational complexity.

4. The data marts – these are the processes that extract, transform, allocate and load data into the data marts. Each data mart represents a unique facet of the business. The well-designed data mart will:
    - Simplify the data mart type. The data mart should be a cube, not a star schema. A cube resolves the issue of slowly changing dimensions and does not require programming assistance to create reports.
    - Limit the number of dimensions per data mart. There should only be a few dimensions to simplify the data presentation and enhance user understanding of the meaning of the data
5. The OLAP (Online Analysis Processing) tools – these are the tools that the business uses to interact with the data mart. These tools vary from reports and spreadsheets to BI packages such as Cognos and Hyperion.

The users create data in the processing architecture and consume the data in business intelligence architecture. The data only move from left to right through the architectures. The feedback mechanisms are external to the architectures. The architecture will have three feedback processes:

1. The data correction feedback process – this process sends any errors found during data extraction and data warehouse loading back to the source application for correction.
2. The business model feedback process – this process sends any errors and omissions in the data mart data back to the data mart architecture for correcting the extraction process.
3. The operations improvement feedback process – this process converts the insights gained from the data analysis into improvements on how to conduct business.

The architecture document will describe these components in detail and give the properties of these components required to ensure a successful implementation. The architecture document is the basis for the data warehouse project estimation and organization.

12. Bibliography

Inmon, William [1989]. *Data Architecture: The Information Paradigm*, Wellesley: QED Information Sciences, Inc.

Kimball, Ralph [1998]. *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses*, New York: John Wiley and Sons, Inc.

Pacioli, Luca Bartolomes [1494]. *Summa de artithmetica, geometria, proportioni et proportionalita*, Venice.

ten Have, O. [1986]. *The History of Accountancy*, 2nd ed., Palo Alto: Bay Books.

