The Data Organization

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

Business Intelligence Architecture

By Rainer Schoenrank

Data Warehouse Consultant

July 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.

Table of Contents

1. Introduction

1.1 Scope of the Document

The scope of the document is the process and data components of the Business Intelligence (BI) reporting process. The architecture diagram shows the relationship between the processing required to generate the BI reports and the databases that hold the business transaction data. The extract, transform and load (ETL) process begins with the OLTP applications (ERP modules) used by the business to record the transactions of the business. The ETL process ends with the data marts that are used to analyze the performance of the business functions.

1.2 Purpose of the Document

An information system architecture is composed of:

- a model for the information system showing the major components and their interactions
- goals to be achieved by the applications which must exist within the architecture
- design criteria for each of the components required in the architecture
- a strategy document prioritizing the implementation of the components
- an implementation plan for each component.

This document outlines the processes and databases necessary to build data marts for the business. The document lists all of the components required and for each component, it gives the necessary requirements and properties.

The document is used as the framework to describe the process and data requirements of the BI reporting process. The high-level properties of the process are described and the properties of the process are used as the QA references to ensure that the process implementation conforms to the strategy selected.

2. Process Architecture

2.1 Process Overview

The Business Intelligence Process Model describes how the data flows through the Enterprise Data Warehouse (EDW) infrastructure with its procedures to the outside world. The diagram below shows the data flows between the data producing (ERP) application processes, the EDW, and the data consuming processes (data marts). As the data is created by the producing processes, it is extracted for the Staging Data Store and transformed into the EDW. The data is consumed by the business modeling processes and is available for display using the OLAP tools on the data marts.

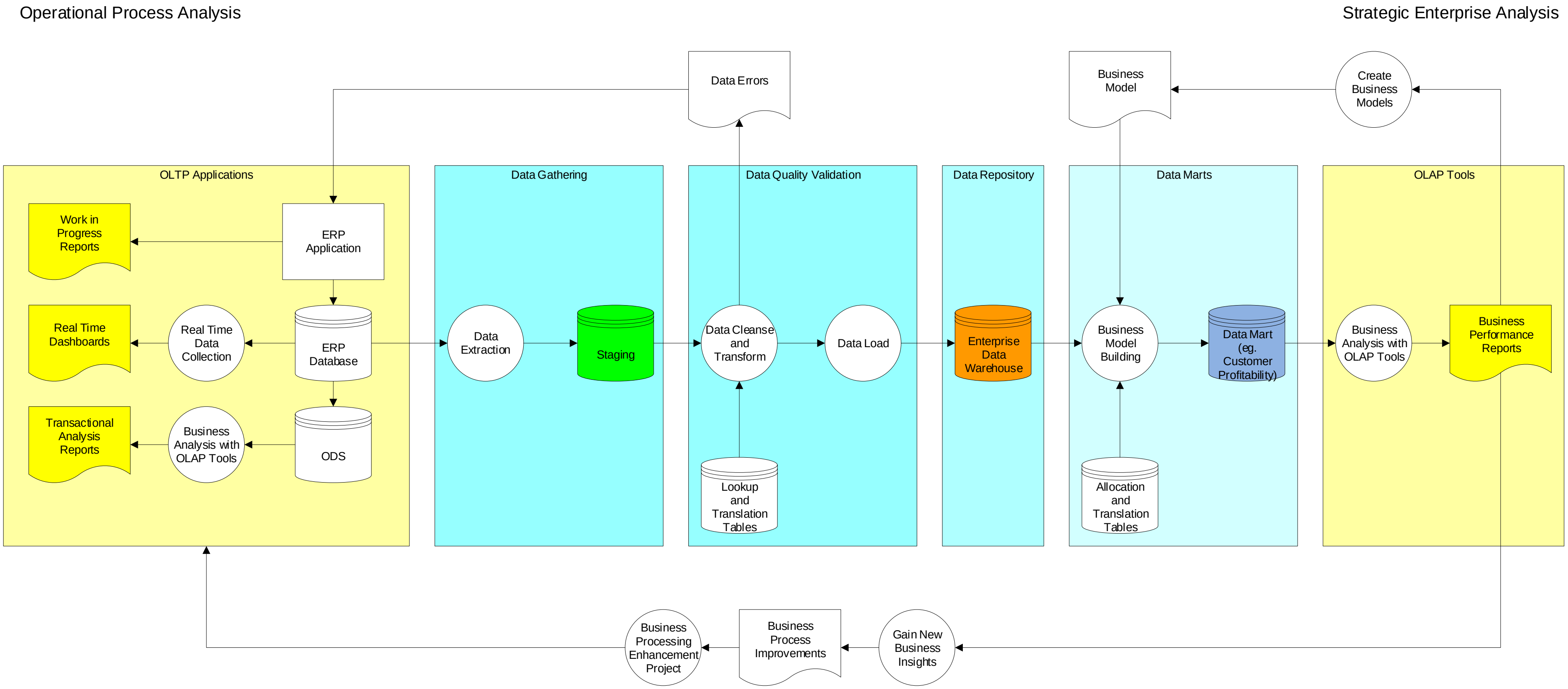
Operational Process Analysis
Strategic Enterprise Analysis
Data Errors
Business Model
Create Business Models
OLTP Applications
Data Gathering
Data Quality Validation
Data Repository
Data Marts
OLAP Tools
Work in Progress Reports
ERP Application
Real Time Dashboards
Real Time Data Collection
ERP Database
Transactional Analysis Reports
Business Analysis with OLAP Tools
ODS
Data Extraction
Staging
Data Cleanse and Transform
Data Load
Lookup and Translation Tables
Enterprise Data Warehouse
Business Model Building
Data Mart (eg. Customer Profitability)
Allocation and Translation Tables
Business Analysis with OLAP Tools
Business Performance Reports
Business Processing Enhancement Project
Business Process Improvements
Gain New Business Insights

Figure 1: Business Intelligence Process Architecture (one source application and one data mart only)

This processing model diagram has been simplified to emphasize the feedback loops by letting all of the source applications be represented by the single element (ERP System). There is actually a fan in of source application data into the STAGE database. The second simplification is the single path to the data mart actually represents the fan out to many subject area data marts.

For this process to provide data required to perform the analysis required for the business key performance indicators (KPI), the data in the EDW must be:

- Complete – contains all of the business transaction and master data
- Comprehensive – all the operational application systems can be mapped on the EDW
- Timely – the data is up to date
- Correct – the data is valid and logically correct

Each symbol on the diagram is a separate entity or process. Each symbol is described below. The lines show how data moves between components of the ETL infrastructure. There are four data flows in the data mart process:

- The main process that loads the data marts
- The data correction feedback for data that does not meet the data quality criteria
- The business model building feedback that describes how the data in the mart is transformed as we understand the business model better.
- The operations improvement feedback during which the knowledge gained by the data marts is used to change the business operations.

2.2 Main Business Intelligence Report Process

2.2.1 Data Sources (ERP System)

These ERP modules automate the business processes and the required record keeping. These applications contain their business semantics, which allows them to make assumptions and simplification within their processing programs and databases. Example – modules of the ERP System are:

Sales Force Automation
Product Design Management
Order Entry
Customer Relationship Management
Production Scheduling
Order Fulfillment
Purchasing
Supplier Management
Receiving
Inventory
Human Resources
Payroll
Fixed Assets
Accounts Receivable
Accounts Payable
General Ledge (the general ledger is the DSS for A/R, A/P, and Payroll)

These applications contain the current state of the business process only. The history of the process is not usually saved past the end of the fiscal year reporting requirement.

The operational applications implicitly contain a model and organization of the business for which they were developed, for example ORACLE Financials is based on 1980’s American accounting practice while SAP R/3 is based on batch job manufacturing and 1970’s European accounting practice. **The major assumption that is made is the record keeping is done on paper.**

The operational application also makes assumptions and simplifications (shortcuts) in its data collection since some ‘nice to have’ data is not necessary for efficient operations. For example, invoices are aggregations of line items sold and the invoice assumes all the line items start from the same place and get delivered to the same place. This allows the tax and freight to be calculated for the aggregated invoice rather than assigned to individual line items.

Also, the payroll system assumes that salaried employees work 40 hours a week, so no time card records are created for salaried employees to save on space and compensated time off is treated as an exception. Having the timecards in the EDW explicitly would make the creation of a personnel data mart much simpler.

These applications contain multiple groups of data, not all of which are required in the data warehouse. To function correctly, the operational applications contain the following types of data:

- Application configuration data
- User security data
- Workflow definition data
- Batch processing data
- Business transaction data and master data.

As an operational history database, the EDW is only interested in the completed business transaction and master data.

2.2.2 Data Extraction Process

For this step the operational business data is not transformed or aggregated. The data is checked for conformance to the field definition, cross field validations and the data is moved to the Staging database.

The objectives of this process are:

- Minimize the impact on the OLTP applications
- Extract only business data (no processing data, no configuration data, etc.)
- Extract only the newly modified data from the OLTP application
- Obtain a consistent snapshot of the business data
- Operate on static data when cleansing and transforming for the data repository

There are two approaches to this process:

- Extract a snapshot of the operational application data
- Publish the interface for the data that the operational application will have to move to staging.

The difference in the approaches is where and when the errors in the data required for the data warehouse are made public. The first approach requires no effort on the part of the operational application. Usually the data warehouse team programs the data extraction process and deficiencies in the data are not known until the attempt to load the data into the data warehouse. At that point, the errors are part of the data warehouse process instead of deficiencies in the data source.

By publishing a data interface to the staging data and having the operational application move the data, the process and data deficiencies become part of the application and not the data warehouse process. This will make the operational application take ownership of the data correction process. The difficulty in this approach is that the data warehouse requirements need to be known in detail in order to publish the data interface making this approach more time consuming.

The Data Extraction process uses an ETL Tool (Informatica, DataStage, etc.) to move the data from the operational application modules to the Staging database. The process will keep track of its execution history so that all the data modified in the ERP system since that last extraction will be moved into the Staging database.

2.2.3 Staging Data Store

The Staging database contains tables that are images of the operational source data. The purpose of the database is to allow the processes to operate on static data when cleansing and transforming the data for the data repository. The tables in the database do not have explicit primary keys so that all operational data is collected. Also, the database management system (DBMS) does not enforce any referential integrity.

The Staging database contains the latest data extraction values only. Normally, the tables are truncated before the data extraction process is begun.

The Staging Data Store has the following properties:

Staging Repository Property	Value
purpose	holds images of the data source tables
scope	the source data tables for all the OLTP applications
number of data sources	a single OLTP application for each table
conceptual data model	none
logical data model	none
implemented database	isolated tables without keys or relationships
referential integrity	none
primary dimensions	none
secondary dimensions	none
facts	• all master data changes (i.e., customer state change) • sales order process measurements (dollar and quantity measurement) • purchase order process measurements (dollar and quantity measurement) • sales process event measurements
measurement granularity	atomic (i.e., sales order/invoice line or equivalent)
time granularity	day
update processing frequency	daily

Table 1: Properties of the Staging Data Store

2.2.4 Data Cleanse and Transform Process

The Data Cleanse and Transform Process transform the data in the staging database (ERP data model) into the EDW database (Enterprise data model).

This data cleansing and transform process uses six steps to clean and transform the data in the staging database:

1. Clean – remove empty records, replace nulls with default values.
2. Reversing the application processing assumptions (i.e., removing the shortcuts taken by the OLTP applications).
3. Validation – ensuring that the values of the data are correct. (e.g., null values, right data type, range of values, etc.)
4. Translation – changing code values from the OLTP application values to the data repository values. (e.g., 1 to Male, 2 to Female)
5. Referential Integrity Checking – ensuring that the relationships embedded in the data exist. (e.g., sales process measurement record points to a valid customer, etc.)
6. Checking the data for conformance to the business rules. Business rule validations are logical errors in the value of the data, for example, a sales order that is in ‘filled’ or ‘completed’ status, cannot have quantity field with a value of zero.

The process generates data error reports that allow the data to be corrected in the source OLTP application.

2.2.5 Lookup and Translation Tables Database

The Lookup and Translation Tables database is used to store the tables that contain the translations for the data as it moves from Staging to the data repository. The tables allow the application coded data to be translated into the code values used in the data repository. A simple example is the translation of the values in the person Gender field. The values 1 and 2 used in an OLTP application are translated into the values M and F used in the data repository.

2.2.6 Data Load Process

The Load Process moves the validated data in the staging database (OOP data model) into the data repository database (Relational data model).

The stage tables are processed in sequence by table type:

1. First all of the validation table changes are loaded,
2. Then the new master data records are loaded
3. Then the master data changes are loaded
4. Finally, the process measurement (fact) tables are loaded into the data repository.

This processing order is necessary so that no data is rejected because of data integrity issues. No data that fails the declared referential integrity is allowed into the data repository. The requirement of this processing is that the interface into the data repository is constant regardless of the nature of source applications.

The data is not aggregated into the data repository. Data aggregation is done when the data is loaded into the data marts.

2.2.7 Enterprise Data Warehouse Database

The EDW database contains only a database and no processes. The database contains the history of the business measurements and master data changes consolidated from many data sources.

The objective of the EDW database is to collect all of the business data so that the data is of high quality and consistent. In this way reports of the company's status from different points of view can be compared without having to resolve inconsistencies in the data.

For the EDW database to provide data required to perform the business intelligence analysis, the data in the data repository must be:

- Complete – contains all of the business process measurement and master data, including
    - The history of all the business measurements
    - The history of all the master data changes
- Comprehensive enough to include expected, actual, and forecast for sales, purchasing and production events
- General enough to contain similar data (e.g., invoices) from multiple sources (SAP R/3, JD Edwards, PeopleSoft, etc.) without design modifications
- Timely – the data is up to date
- Correct – the data is valid and logically correct

To meet the goals of data quality, the EDW database is designed using the following data analysis principles:

- non-overlapping business process measurements
- non-overlapping master data (data dimensions)
- separate the master data from its organizations (e.g., customer from market segmentation)

As a 5th normal form (5NF) relational database, the data repository allows us to specify the data integrity outside of and across the OLTP applications.

The properties of the EDW database are shown in Table 2.

Data Warehouse Property	**Value**
purpose	• consolidates OLTP application data into an enterprise wide structure • accumulates the history of the business process measurements • accumulates the history of the master data changes • implements and ensures data quality • acts as the isolation interface between data sources and data marts
scope	enterprise (at least 4 master data areas, probably more depending on organization chart and chart of accounts)
number of data sources	many, one source for each OLTP application whose data will be consolidated in the data warehouse
conceptual data model	star
logical data model	snowflake
implemented database	normalized relational database (5th normal form)
referential integrity	declared and enforced by the database management system (DBMS)
primary dimensions	all master data
secondary dimensions	any column in the database
facts	• buy transactions (dollar and quantity measurement) • sell transactions (dollar and quantity measurement) • process event transactions (time measurement)
measurement granularity	atomic (i.e., invoice line or equivalent)
time granularity	day
update processing frequency	daily

Table 2: Properties for a Data Warehouse

The EDW design must be:

- Comprehensive enough to include expected, actual, and forecast for sales, purchasing and production events
- general enough to contain similar data (e.g., invoices) from multiple sources (SAP R/3, JD Edwards, PeopleSoft, etc.) without design modifications

The EDW design assumes that every data object has a unique name.

2.2.8 Business Model Building Process

The Business Model Building (data mart load) process moves the data from the data repository into the data mart. For this step the data is transformed by:

- resolving all EDW database codes with their descriptions (translation)
- discarding irrelevant Master Data organizations
- flattening the Master Data organizational hierarchies (classification of data)
- selecting all of the required master data attributes (business dimensions)
- aggregating the facts for the business model
- allocating unmeasured facts such as demographics, costs, etc. according the required algorithms
- calculating the metrics for the business model

2.2.9 Translation and Allocation Tables Database

The Translation and Allocation Tables database is used to store the tables that contain the translations for the data as it moved from EDW to the data marts. The tables allow the EDW data to be translated, aggregated, and allocated into the data used in the data marts. Also, the database management system (DBMS) does not enforce any referential integrity.

2.2.10 Data Mart

A data mart is the business reporting database for a business processing area. The data model is a cube with the dimensions required for this business subject area. Each data mart has the properties shown in Table 3.

Data Mart Property	**Value**
purpose	An extract of the data repository for reporting and analysis
scope	A single business subject area or problem area
number of data sources	One, the data repository
conceptual data model	star
logical data model	snowflake
implemented database	denormalized database table (first normal form)
referential integrity	not required for reporting
primary dimensions	all master data required by the business model
secondary dimensions	any column in the data mart
facts	value metrics (dollars) quantity metrics time metrics performance ratios
measurement granularity	as required by the business model
time granularity	as required by the business model
update processing frequency	as required by the business model

Table 3: Properties for a Data Mart

2.2.11 Business Analysis Process

The BI/Reporting Tool (Hyperion, Brio, Business Objects, Cognos, etc.) has the functionality that allows the business users to access the Data Mart to generate the business performance indicators and metrics.

2.2.12 Business Performance Indicators Reports

The business performance indicators are presented as:

- Reports – a snapshot of the data mart at a point in time
- Inquiries (queries) – displays value of a metric now

2.3 Data Correction Feedback Loop

2.3.1 Data Errors Reports

The Data Errors Reports process produces lists of data records that do not meet the data quality requirements of the EDW. These reports are sent to the data source applications so that the data errors will be corrected and then extracted again.

2.4 Business Model Feedback Loop

2.4.1 Create Business Models

The process for creating business models examines the data required for the business performance indicators, reverse engineers the data to the structures in the EDW and creates the structure required for the data mart that is used to produce the reports.

2.4.2 Business Model Document

The Business Model Document specifies how the data in the EDW is selected, aggregated and transformed as it is loaded into the data mart.

2.5 Operations Improvement Feedback Loop

These components are not part of the project scope.

2.5.1 Gain New Business Insights

The analysis of the business performance data allows you to gain new insights into the operation of the business and how the operations can be improved.

2.5.2 Business Process Improvement Document

The Business Process Improvement Document contains the new business insights and details the benefits and costs of changing the business operations.

2.5.3 Business Processing Enhancement Project

The prioritization and funding of projects to change the business operations goes through the project development process.

3. Other Components

3.1 Source Application Components

These components are not part of the project scope.

3.1.1 Real Time Data Collection

In order to create dashboards that display the real-time condition of the business, a real-time server process is required that extracts selected data from several operational applications. This process would collect the required data at fixed time intervals and organize it into the dashboard repository.

3.1.2 Real Time Dashboards

The Real Time Dashboards are updated on a regular basis by the real-time data collection server after the dashboard repository has been updated.

3.1.3 Work in Progress Reports

The Work in Progress reports are part of the operational applications. They display the current state of the data in the operational application to show the work in progress.