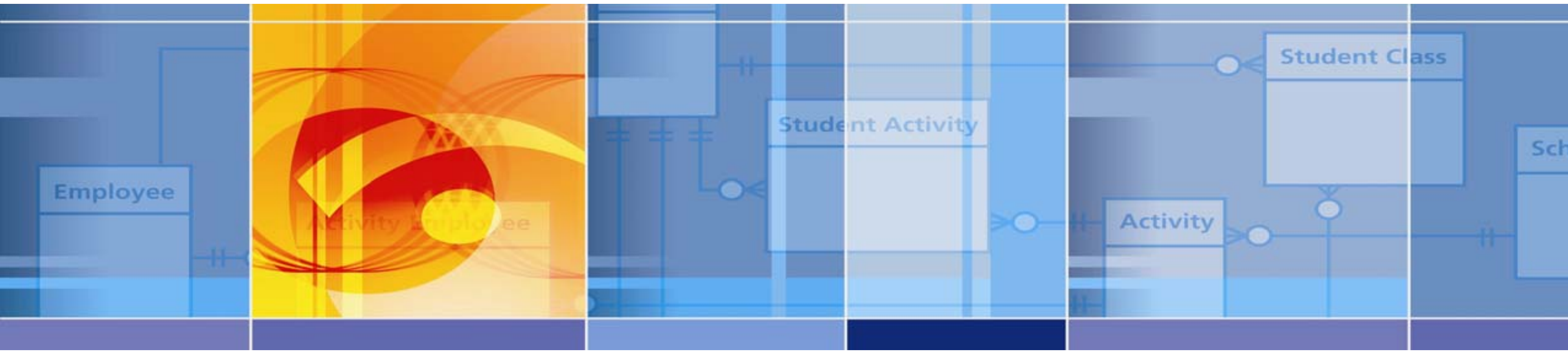


The Optimal Reference Guide:

What's Really "In Store" for Your Data Warehouse?

Data Warehouse Series – Part I

Extraordinary insight into today's education topics

Steven King, ESP Solutions Group
Alexander Jackl, ESP Solutions Group

Foreword by Glynn D. Ligon, Ph.D., ESP Solutions Group

ESP Solutions Group

Table of Contents

Foreword

By Glynn D. Ligon, Ph.D.

Ever seen a warehouse on the front lawn of corporate headquarters? Ever seen a board member or chief executive working in one? No. Even though corporations have them, those warehouses are put in their place and perform their role well outside the visible mainstream of business activity.

So I marvel at the attention that data warehouses are getting in education agencies. The reason is simply that data warehouses are being over sold in education. This is partly to the credit of the IT managers and the data warehouse vendors. This is more the result of our not really understanding what a data warehouse is and is not. This ESP Optimal Reference Guide is the first in a series of reference guides that will define the role of data warehouses in information systems, examine data warehouse design, and define important concepts in education information systems.

My advice as you read this and the other papers in the series is to keep a very open mind about the definition of a data warehouse. Remember that a data warehouse is one component of an information system. Other components such as a metadata dictionary, ETL tools, data cleansing tools, user portal, identity management system, student locator and identifier systems, case management tools, directory management tools, electronic records exchange system, and decision support system are necessary to make the data warehouse a working component of your overall information system. These components all work together accomplish the collection, storage, and analysis/reporting functions of an education information system.

Introduction

Consider this analogy between a furniture store and an information system. As you enter a furniture store, the first things you see are nicely and logically arranged clusters of furniture. Each arrangement has its own particular scheme meant to communicate a particular message to customers.

The storefront of the business is only one part of the overall operation. A furniture store must handle many business processes not apparent to customers. In most stores, the warehouse is in the back of the facility and is often not visible. The table below lists some common characteristic between a furniture store and an information system. Notice how each function requires one or more components to accomplish the requirements of the function.

ESP Insight

Data: noun. 1. plural of datum. 2. individual facts, statistics, or items of information.

Characteristic	Furniture Store	Information System
Public Customer Interface	Customer Showroom	Web Page with High-Level Summaries and Reports
Quick Retrieval Resource	On-Site Warehouse	Data Mart with Most-Needed Data and Reports
Main Storage Facility	Supplier/Corporate Warehouse	Organizational Data Repository or Warehouse
Assembly and Quality Control Facility	Manufacturer Facility	Data Collector and Owner Files (Authoritative Data Source)
Sources	Raw Materials Supplier	Data Providers

The message in this analogy is that for practical reasons, everything is not done in one place in either the furniture business or in the education information business. Despite what purveyors of a comprehensive data warehouse solution may want to us to accept, the Supplier/Corporate/Data Warehouse is not always the most appropriate or efficient location to be manipulating raw data, creating derived statistics, or providing high-level reports to general audiences.

Simply put the overhead and burden to get and maintain all the data in one place is just not necessary. The fundamental reason to have a data warehouse is to achieve efficiencies with, gain access to, and maintain control over your most important and most frequently accessed data. The cumbersome nature of a data warehouse is no problem to a trained user; but to the less sophisticated workers collecting and cleaning data, or policy makers relying upon timely summaries, abiding by all the rules and protocols of an industrial strength data warehouse is simply a waste of their time and energies. Sometimes, it is just more efficient to have some of your data in other places.

This does not minimize the importance of a data warehouse, but it puts the data warehouse back into the context of being only one of the components that make up a comprehensive information system.

Beware of putting too much into your data warehouse. The goal of the data warehouse is to support quick summary and analysis of critical interest, not to store anything and everything in one place.

First-generation student information systems in the 70's scheduled classes, took attendance, reported grades, and were turned off at night. Now schools want special education IEPs, diagnostic assessments linked to lessons plans that are aligned with the state academic standards, discipline incidents differentiated by victims and perpetrators, transportation routes, lunch status, immunization status, special program eligibility status, and Mom's cell phone number in case she's not reading her instant messages. Oh, yes, and overnight updating of district data files for daily reporting.

That seems to be the path data warehouses are on today. They began as a simple concept — consolidate important organizational data into one place. Now the data warehouse is envisioned as holding all data and meeting all demands for access to those data. That is just not practical in most education agencies.

Data Warehouse – A History

Relational Data base theory is relatively new (late 1970's) with database management software (DBMS) not really being developed until the 1980's. Relational data bases offered a lot of advantages and flexibility over the existing flat file and hierarchical data bases of the day, but they tended to be slow. Some systems took a second to update a single transaction.

Research and effort through the 80's was on improving the transaction processing capabilities of these systems. SABRE, The American Airlines reservation system, was viewed as one of the pinnacle systems capable of handling 4,000 transactions a second. (woo hoo).

To get fast data entry required indexing strategies, table tuning, and programming focused on quick location and retrieval of single records by known key data. When these same systems were queried by looser filters (e.g. all records with birth dates after a specified date) they crawled. And all the transaction activities also crawled as CPU time and disk energy was tied up in the open query.

ESP Insight
Data: noun. 1. plural of datum. 2. individual facts, statistics, or items of information.

This resulted in the move to extracting data from the operational transaction system to a separate database to be used for analysis. The analytical requests would then not impact the core business function (making airline reservations, for example).

When analysts develop a query and run it, this often leads to more questions than answers. (Why are assessment results in school A lower than school B?) Consequently, these analysts will want to rerun the query, but with more detail, maybe get School A and B assessment results by gender.

If these queries are run against the operational system, the number of students in each school may be different between the two queries. This isn't good, so the data base supporting analysis must be stable over time. This implies the analysis system is NEVER current and that the queries ALWAYS filter on time.

Most organizations have multiple applications supporting their various core business functions. Each application is built and tuned to handle its particular set of business requirements. Our analysis system however can combine the extracts from these various applications into a single system.

Analysis implies comparison. The comparison can be among entities, among sub groups, or across time. If the comparison is to be "apples to apples" from the various applications, then the data in the system must have consistent data formats, coding, and extract timing. One application may store gender as M or F while another may use 1 and 2. These need to be transformed into a consistent set as the data are loaded into the analysis system.

Business analysis also tends to be around a particular subject, not an operational function. We care about students and their performance, not about the registration or attendance process. Operational systems support the processes; analytical systems study the results across processes.

If we are going to build an analytical system separate from our transactional database applications, and the analytical system has different goals and requirements, then we can rethink and redesign many of the indexing, tuning and other strategies that went into those data base systems.

Thus the Data Warehouse concept was born. The term was coined by Bill Inmon in 1990. He defined a data warehouse as:

"A subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision making process."

He defined the terms in the sentence as follows:

ESP Insight
Bill Inmon is universally recognized as the father of the data warehouse. He has more than 35 years of database technology management experience and data warehouse design expertise.

- **Subject Oriented**: Data that gives information about a particular subject instead of about a company's ongoing operations.
- **Integrated**: Data that is gathered into the data warehouse from a variety of sources and merged into a coherent whole.
- **Time-variant**: All data in the data warehouse is identified with a particular time period.
- **Non-volatile**: Data is stable in a data warehouse. More data is added but data is never removed. This enables management to gain a consistent picture of the business.

Richard Kimball took Inmon's work and extended it in his 1996 book, The Data Warehouse Toolkit. In it, he identified six goals:

1) The data warehouse provides access to organizational data.

2) The data warehouse is consistent.

3) The data in the data warehouse can be separated and combined by means of every possible measure in the business (Slice and Dice).

4) A data warehouse is not just the data, but also a set of tools to query, analyze, and present information.

5) The data warehouse is the place where we publish data.

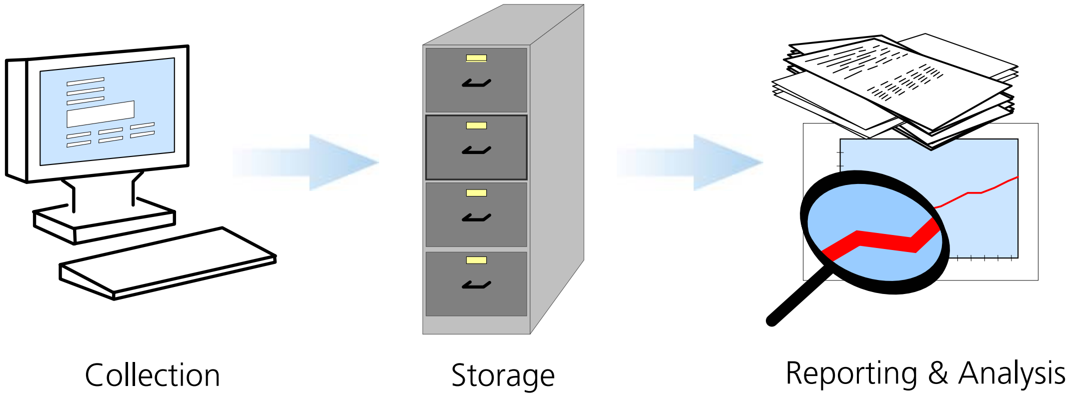
6) The quality of the data in the data warehouse is a driver of business reengineering.

In number 4, Kimball includes the tools to look at and analyze the data as part of the warehouse. This is a critical and necessary function of data management, but is not part of the data warehouse (or at least of the data warehouse data store as discussed later).

When Kimball talks about "publishing" in number 5, he is including a quality assurance and review process. It is here that data are "certified" before they are allowed to be stored in (i.e. "published to") the warehouse.

Education Data Management Model

Education organizations, whether school districts, intermediaries, or state education agencies have 3 main process groups in data management: data collection, data storage, and reporting and analysis.



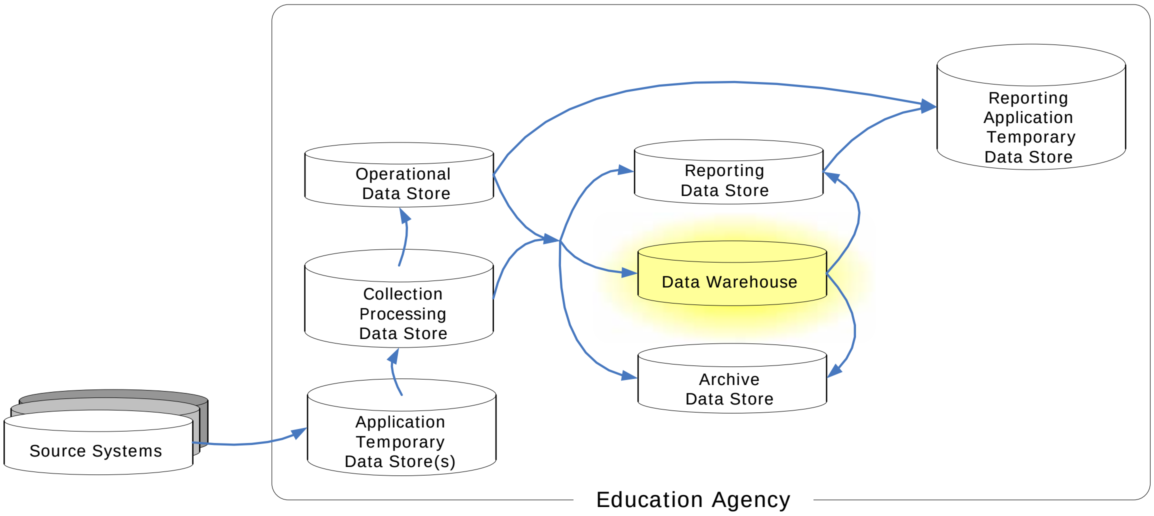
The "Data Warehouse", in its narrowest definition is only part of number 2. Kimball generously expands the definition to include Reporting and Analysis and frankly most “Data Warehouse” companies in this space have their hands in all three of the core process groups with particular emphasis on reporting and analysis.

This paper will explain this context of the comprehensive educational data system from the perspective of an agency data manager or CIO and provide a clear overarching framework for dealing with the specific sets of student, staff, and organization data for which they are stewards.

Characteristics of the Data Stores

Most organizations have multiple applications supporting their various core business functions. These various functions and processes also have supporting data stores.

There are many different ways to talk about storing data. This white paper will not delve into the details of schema types, relational vs. XML vs. OLAP issues, and other database specific issues. Rather, this paper will take a functional look at the types of data stores that an educational enterprise will need and what is noteworthy about each type.


Reporting Application Temporary Data Store
Operational Data Store
Reporting Data Store
Collection Processing Data Store
Data Warehouse
Archive Data Store
Source Systems
Application Temporary Data Store(s)
Education Agency

Data tends to flow from the bottom left in this diagram. Local education agency data comes from school district sources. State or intermediate education agency data comes from LEAs. There are often multiple sources – program offices, student information systems, staff information systems, etc.

Application/Temporary Data Store

When an education agency automates a data collection, it may build an application which may need temporary storage. It is this temporary application data store that holds a user's interim work before the collection is finished and the data are submitted. The application uses the Temporary Data Store (TDS) to manage certification of data, run preliminary checks on the data and in preparation to send the data onward. The data is only housed so long as needed to move the data to an operational data store, thus there is usually no permanence and no record though this can change from application to application. The structure of a TDS is defined by optimizing access speed in moving the data back and forth.

Collection Processing Data Store

The agency processes the data in the collection processing data store. The collection processing data store contains the current data from a particular collection or source. It is interactive and dynamic and is not a storage bin where data is stored for the long term. Data is frequently created, updated, and deleted. It is usually granular data – or at least at the same level as the data that was collected.

Although there are exceptions, there is usually one operational data store per collection. How long data is maintained in the Collection Processing Data Store is governed by the business needs of the application it is serving and the system as a whole. When the data from all the respondents is in, cleaned, and final, then the data should be loaded to the Operational Data Store, the Reporting Data Store, the Data Warehouse, and the Archive Data Store.

Because certain collections; the October 1st collection, the End of Year collection, ADA/ADM can have strong financial and accountability implications, these data stores are sometimes kept active for months and even years after the collection window is closed. Once the data is considered "final" then often the whole collection will be archived so that the question: "What was the 'official' accepted October 1st report in 2006?" can be answered.

Operational Data Store

An operational data store contains current data. It is interactive and dynamic but is usually the source of this year's current data. Data is frequently created, updated, and deleted. It is usually granular data, or at least at the same level as the data that was collected.

Changes are recorded in the Audit Data Store along with the time and the source of the change (direct user change, update from a source collection, etc.)

There is some reporting done from an Operational Data Store but it is usually focused on either the current state of the data (How many kids are enrolled in Shady Hill School this fall?) or data collection activity itself (When did the LEA submit the report? How many error messages were there?).

Reporting Data Store

A reporting data store is the set of tables optimized for queries, retrieval or building particular graphical representations of the data. The Reporting Data Store is solely optimized for data visualization and indexed for the highest speed of retrieval of the most critical and prioritized data. Like a Data Warehouse, data in the Reporting Data Store is time stamped and contains significant history.

A particular Reporting Data Store may be focused on one or selected areas of analysis or reporting based either on business unit or by a particular need (publishing standard test scores, for instance). There usually is a user-friendly front end to the Reporting Data Store – possibly integrated with Decision Support tools. In most cases, the questions to be analyzed or the reports to be generated are known. Data in the Reporting Data Store are organized to support those needs.

ESP Insight

Data: noun. 1. plural of datum. 2. individual facts, statistics, or items of information.

These types of data stores are functional descriptions. In reality the lines, as with everything, get blurry and hard to make out. Sometimes an Operational Data Store has some longitudinal data from multiple sources and sometimes a Decision Support tools are built on top of an Operational Data Store. Each of these data storage functions though have a different set of needs and requirements and it is worth it for an education agency to work out what its requirements are and which of these storage types best suit those needs.

Data Warehouse

Agencies build Data Warehouses to support their analysis and ad hoc query needs. Data are selected for inclusion based on how well they can be analyzed or contribute to analysis. As such, one cannot buy a data warehouse. The data sources and particular topics of potential interest are unique to each agency and therefore, the tables to be loaded are unique to each situation. A data warehouse vendor may have models that have been successful in the past, but the final designs are all local.

Data from multiple sources, programs, and agencies are combined. Like the Reporting Data Store, data in the Data Warehouse are historical and time stamped. Significant amounts of energy go into coordinating and making the data consistent and complete. It is critical that agency staff NOT underestimate the amount of data cleaning that will be necessary to integrate the data in the system when it has not been compared before. It always surprises agency personnel how much data cleansing is necessary.

Each Data Warehouse should have a data dictionary that all elements conform to. Since the point of data mining is to look at data in new ways, to combine data from disparate systems, a critical component is a complete data dictionary that describes exactly what data are collected, the time period covered, the population about which the data were collected, and any other information an analyst might need. For the analysis to be an appropriate one, the analyst must have this information.

Archive Data Store

This is often not a separate store but is functionally decomposed here because it has very different requirements than the other data stores. It is not necessarily designed for high-speed access as much as it is designed to maintain a strict ability to store change data accurately. It is more important that it can absorb data quickly than it is that it releases data quickly.

Reporting Application/Temporary Data Store

This is an instance of the application/temporary store except it is focused on serving the data needs of the retrieval, analysis, and presentation applications rather than the data collection ones. It draws its data from the Reporting Data Store.

Conclusion

While playing a pivotal role for analysis and program performance monitoring, the data warehouse is nonetheless just one component in a complete education information management system. Too much emphasis has been placed on the data warehouse at the expense of the other components. In many cases, users are looking for a well structured and functioning reporting data store and system.

A well designed and implemented data warehouse can be quite powerful. But just like any tool, it needs to be used for the right purpose. If an agency administrator is sold a data warehouse just to satisfy simple reporting needs then resources have probably been wasted. If a data warehouse is overburdened with archive and reporting uses, then it will be unable to truly function as needed.

The next paper in this series will delve deeper into the characteristics of a good data warehouse and describe the design features that can allow it to shine for analysis.

Key Concepts, Constructs, and Definitions

Aggregation – The transformation of granular data into "aggregated data"- that is to say the summation of the granular data using business rules defining which data is to be included, from what sources, and what transforms to enact on the data.

Archive Data Store – This data store (often a set of tables in the Data Warehouse instance) maintains the list of changes that have happened in the Operational Data Store and the Warehouse and archives of the transactional data stores.

Collation – The grouping of data together without aggregation, or transforms of any kind.

Collection Processing Data Store – Typically the staging platform for data moving from Source Systems to an Operational Data Store. The data arrives in its most detailed state reflecting the most granular transactions.

Data Element – A discrete category of data, e.g. "age," "ethnicity," "test score."

Data Mart – A subset of Data Warehouse data spun off to serve the specific data analysis needs of a subgroup of end users, such as a particular agency program or operating unit, executive management, and so forth.

Data Warehouse – A centralized source of key data drawn from various Systems of Record and including longitudinal and source data brought together for the purposes of data integration in line with the agency's analysis and reporting requirements.

Decision Support System – An IT-enabled system that facilitates the integration of critical agency information so that management may employ that information to inform planning and decision making.

Extract, Transform, and Load (ETL) – The process and IT tools employed to draw out (extract) data from Source Systems, to systematically alter the data (transform) to conform with the database structure of the Data Warehouse, and to deposit (load) that data into the warehouse.

Metadata repository – This type of repository stores data about the data, including: descriptions of what kind of information is stored where, how it is encoded, how it is related to other information, where it comes from and how it is related to our business.

Online Analytical Processing (OLAP) – a method of interactive data selection and analysis employed in conjunction with data warehousing and decision support systems.

Operational Data Store – this is the data store that has the current operational data of the enterprise. Current questions are directed primarily against this store.

Reporting Data Store – the Data Store(s) that serves Data Mart or Marts. It is optimized for high-speed search and retrieval.

Source System – Typically a transactional IT system, such as a financial, human resources, student information, or assessment management system, that feeds the agency's Decision Support System and Data Warehouse.

System of Record – See Source System.

Temporary Data Store – A temporary storage space for data, usually data in mid-transaction. There is no reporting or persistence of this data.   Instances can be found behind the data collection applications and serving the reporting and presentation applications (these are sometimes known as display stores).

About ESP Solutions Group

ESP Solutions Group provides its clients with *Extraordinary Insight*™ into K-12 education data systems and psychometrics. Our team is comprised of industry experts who pioneered the concept of "data driven decision making" and now help optimize the management of our clients' state and local education agencies.

ESP personnel have advised school districts, all 52 state education agencies, and the U.S. Department of Education on the practice of K-12 school data management. We are regarded as leading experts in understanding the data and technology implications of the **No Child Left Behind Act (NCLB)**, **Education Data Exchange Network (EDEN)**, and the Schools **Interoperability Framework (SIF).**

Dozens of education agencies have hired ESP to design and build their student record collection systems, federal reporting systems, student identifier systems, data dictionaries, evaluation/assessment programs and data management/analysis systems.

To learn how ESP can give your agency *Extraordinary Insight*™ into your K-12 education data, contact Greg Nadeau at (781) 370-1017 or gnadeau@espsg.com.

This document is part of *The Optimal Reference Guide* Series, designed to help education data decision makers analyze, manage, and share data in the 21st Century.

ESP Solutions Group

(512) 458-8364
www.espsolutionsgroup.com
Austin • Boston • Washington DC