A Data Warehouse Case Study

Abstract

Maximizing Decision-making Through Communications, Command and Control of Data from Capture to Presentation of Results.

The essential concept of a data warehouse is to provide the ability to gather data into optimized databases without regard for the generating applications or platforms. Data warehousing can be formally defined as “the coordinated, architected, and periodic copying of data from various sources into an environment optimized for analytical and informational processing”\(^1\).

The Challenge

Meaningful analysis of data requires us to unite information from many sources in many forms, including: images; text; audio/video recordings; databases; forms, etc. The information sources may never have been intended to be used for data analysis purposes. These sources may have different formats, contain inaccurate or outdated information, be of low transcription quality, be mislabeled or be incompatible. New sources of information may be needed periodically and some elements of information may be one time only artifacts.

A data warehouse system designed for analysis must be capable of assimilating these data elements from many disparate sources into a common form. Correctly labeling and describing search keys and transcribing data in a form for analysis is critical. Qualifying the accuracy of the data against its original source of authority is imperative. Any such system must also be able to: apply policy and procedure for comparing information from multiple sources to select the most accurate source for a data element; correct data elements as needed; and check inconsistencies amongst the data. It must accomplish this while maintaining a complete data history of every element before and after every change with attribution of the change to person, time and place. It must be possible to apply policy or procedure within specific periods of time by processing date or event data to assure comparability of data within a calendar or a processing time horizon. When data originates from a source where different policies and procedures are applied, it must be possible to reapply new policies and procedures. Where quality of transcription is low qualifying the data through verification or sampling against original source documents and media is required. Finally, it must be possible to recreate the exact state of all data at any date by processing time horizon or by event horizon.

The analytical system applied to a data warehouse must be applicable to all data and combinations of data. It must take into account whether sufficient data exists at the necessary quality level to make conclusions at the desired significance level. Where possible it must facilitate remediation of data from original primary source(s) of authority.

When new data is acquired from new sources, it must be possible to input and register the data automatically. Processing must be flexible enough to process these new sources according to their own

---

\(^1\) Alan R. Simon, *Data Warehousing For Dummies* ISBN: 0-7645-0170-4
unique requirements and yet consistently apply policy and procedure so that data from new sources is comparable to existing data.

When decisions are made to change the way data is processed, edited, or how policy and procedure is applied, it must be possible to exactly determine the point in time that this change was made. It must be possible to apply old policies and procedures for comparison to old analyses, and new policy and procedure for new analyses.

Defining Data Warehouse Issues

The Lolopop partners served as principals in a data warehouse effort with objectives that are shared by most users of data warehouses. During business analysis and requirements gathering phase, we found that high quality was cited as the number one objective. Many other objectives were actually quality objectives, as well. Based on our experiences, Lolopop defines the generalized objectives in order of importance as:

Quality information to create data and/or combine with other data sources.

In this case, only about one in eight events could be used for analysis across databases. Stakeholders said that reporting of the same data from the same incoming information varied wildly when re-reported at a later date or when it came from another organization's analysis of the same data. Frequently the data in computer databases was demonstrably not contained in the original documents from which they were transcribed. Conflicting applications of policy and procedure by departments with different objectives, prejudices and perspectives were applied inconsistently without recording the changes or their sources, leaving the data for any given event a slave to who last interpreted it.

Timely response to requests for data.

Here, the data was processed in time period batches. In some instances, it could take up to four years to finalize a data period. Organizations requiring data for analysis simply went to the reporting source and got their own copies for analysis, entirely bypassing the official data warehouse and analytical sources.

Consistent relating of information.

An issue as simple as a name – the information that could be used to connect data events to histories for individuals or other uniting objects – had no consistent method to standardize or simplify naming conventions.

Another example, Geographical Information System (GIS) location information had an extravagant infrastructure that was constantly changing. This made comparisons of data from two different time periods extremely difficult.

Easy access to information.

Often data warehouse technologies assume or demand a sophisticated understanding of relational databases and statistical analysis. This prevents ordinary stakeholders from using data effectively and with confidence. In some instances, the personnel responsible for analysis lack the professional and technical skills to develop effective solutions. This issue can stultify reporting to a few kinds of reports and variants that have been programmed over time, and reduces data selection for the analyses to kind of magic applied by clerical personnel responsible for generating reports.
Unleash management to formulate and uniformly apply policy and procedure.

We found that management decisions and mandates could be hindered by an inability to effectively capture, store, retrieve and analyze data.

In this particular instance, no management controls existed to analyze: source of low quality; work rates; work effort to remediate (or even a concept of remediation); effectiveness of procedures; effectiveness of work effort; etc.

Remediation is a good case in point. Management experienced difficulty with the concept of remedying data transcription from past paper forms – even though the forms existed in images that could be automatically routed. The perception was that quantity of data, not quality, was the objective and that no one would ever attempt to fix data by verifying it or comparing it to original documents.

Manage incoming data from non-integrated sources.

Data from multiple, unrelated sources requires a plan to convert electronic data, manage imaging and documents inputs, manage workflow and manage the analysis of data.

In this case, every interface required manual intervention. Since there was no system awareness at the beginning of the capture process as to what was needed for analysis at the end, it was very difficult to make rapid and time effective changes to accommodate changing stakeholder needs.

Reproducible Reporting Results

We found that reporting of data was not reproducible and the reasons for differences in reporting were not retrievable, undermining confidence in the data, analysis and reporting.

One may essentially summarize these objectives as quality challenges that require a basic systems engineering approach for resolution.

Our Findings

We determined that existing data warehousing systems have evolved as a bridge to data rather than a new method to make effective use of an enterprise’s information. Out of this experience came the Lolopop automated data warehouse solution. Lolopop presents new concepts supporting a complete data communications, command and control capability, enhancing the ability to assemble and analyze data using quality and analytical standards.

First, the Lolopop approach requires a foundation for establishing a definition of quality
Quality Concepts

Foremost among these is the concept of Source of Authority (SOA) as a starting point for tracking and measuring quality. A data warehouse must accurately and precisely reflect the truth if it is to provide usable analysis and accurate decision making. To the extent it is truthful, analyses and decisions are accurate and usable. If data element values cannot be accountably traced to their sources, and the truth of the sources assessed, one never knows whether information is reliable or not.

Lolopop defines SOA as an “information source designated by the organization as unquestionably correct”. SOA enforces the rule for attribution of accepted value or any change and offers the ability to apply new policy to prefer another source as more authoritative. The SOA concept originated from our discovery that data entry clerks were transcribing what they felt should have been reported versus what was reported; with no consistent or documented basis for making their “corrections”. Example of SOA use:

Suppose an organization assigns (or accepts) the authority for a data element to be an on-site reporter’s written verification. If there’s an error in the data, the organization applies a data correction policy (or rule) that accepts higher authority from an alternate source (SOA #2), which now becomes the SOA. Next year, effective January 1st, the organization equips all its on-site reporters with scanning devices to record the data (SOA #3). Now, the organization applies a new policy essentially saying “prior to 12/ 31 the highest SOA for this data element is source two and beginning 1/1 it’s source three”.

Resolution of indistinct or conflicting naming conventions is required to prevent imprecise or erroneous comparisons and ensure the ability to relate data from different sources. Lolopop naming conventions establishes the ability to standardize object identification and comparison fields to distinguish what is being compared ... regardless of source name, object instantiation, or variable.

Lolopop quality components ensure the accuracy and precision of data acquisition, transcription, processing and remediation as measured against the authoritative origin (SOA) of the data element information.

Remediation is required to correct erroneous capture or transcription of data elements, and allow transcription of additional data elements not originally captured. Lolopop remediation components include quality verification, quality assurance and quality measurement. Quality and confidence become part of the provenance for a dataset indicating what quality levels it represents and what confidence can be expected in statistical analyses.

The Lolopop Data Warehouse Communications, Command and Control Center (DWC) solution is comprised of many components as illustrated below. Future papers will discuss Lolopop’s: architecture; rules engine; workbench utilities; data acquisition and coordination; and, compliance with the Federal Data Quality Act of 2002.

---

2 The Rules Engine contains a number of proprietary algorithms.
Data Acquisition

Data provenance and reproducible results require unique data storage. The Lolopop DataStack™ ensures that subsequent analyses: 1) exactly reproduce the content of the previous dataset analysis; or 2) reports the changes recorded based on dates of occurrence that explain absolutely the difference from the earlier analysis. Datasets have fixed content that has computed or measured quality and confidence levels, and change logging included in the dataset.

Routing and Scheduling

Lolopop includes a proprietary state seeking routing and scheduling component. State seeking means that routing initiates with a set of unpopulated data elements. User management specifies rules for data quality and confidence objectives which are converted into a set of unsatisfied data elements – the Dataset Definition.

Lolopop's routing and scheduling system processes incoming data according to rules until: objectives are met; objectives cannot be met; or a remediation plan is presented.

Analytics

One Lolopop component populates the Dataset Instantiation with data elements meeting the specified quality traceable to SOA while supporting the required result confidence level. Another builds the analysis plan, selects the appropriate analysis tool and defines all calculations that will be performed against the Dataset Instantiation: including data selection and interpretation; aggregate calculation; and testing. One may choose among: comparison by different times, locations or conditions; trend by time period; significant counts, central values and range values; statistical display; topographical display; and, alarm or trigger. A non-parametric approach is selected when the data is abnormally distributed or the sample is too small.

Conclusion

In today's aggressively competitive environment, fast, accurate and focused decisions are required for both survival and success. Conditions are constantly changing in more complex contexts than ever before. Decisions depend on data that is timely, accurate, focused, and responsive to changes in sources and form.

As decision makers are pressed to make faster and better decisions, it has become clear that decision support systems are not keeping up. Lolopop represents new and better thinking about the challenge of providing decision support with sound concepts for: acquiring and presenting data; measuring its quality; determining when enough data is available to make a decision; qualifying sources; applying policy uniformly; selecting the right method of analysis; and, providing stable explainable reporting.

Better decisions facilitate survival and success. Lolopop provides the infrastructure to make better decisions.

Authors: David P. Wesenberg & Joanne E. Peterson
For more information contact:
Joanne Peterson (800)-544-1210
or Joanne@abator.com

Lolopop's DataStack is a proprietary data storage technique.