Data Cleansing Strategies

InfoManagement Direct, October 2004

Kuldeep Dongre

The presence of data alone does not ensure that all the management functions and decisions can be smoothly undertaken. There is a compulsive requirement for the data to be meaningful or, in other words, data quality is of utmost importance if management is to take any advantage of the data at their disposal.

Data quality pertains to issues such as:

- Accuracy
- Integrity
- Cleanliness
- Correctness
- Completeness
- Consistency

The quality of data is often evaluated to determine usability and to establish the processes necessary for improving data quality. Data quality may be measured objectively or subjectively. Data quality is a state of completeness, validity, consistency, timeliness and accuracy that makes data appropriate for a specific use.

This article explores the various factors that make data cleansing of the legacy system inevitable and provides strategies that can be adopted. Moreover, it deals with the factors that determine the choice of a particular strategy for data cleansing.

Problem Definition

Practically, any application contains dirty data - data that is meaningless, is not representative of the business it is used in, has some obvious error or becomes meaningless in the new application environment once the legacy system is converted and migrated. I will analyze two scenarios wherein the data is required to be converted and migrated to a new application (target system) from the existing application, which is henceforth referred to as legacy system, and when the legacy system is to be retained.

Data Quality Issues - Root Cause Analysis

Data quality issues in the legacy system arise because of the following factors:
• **Application Errors**: These data errors creep in because of the inability of the legacy system to validate certain user inputs.

• **Human Errors**: This is a major source of dirty data in the legacy system. A major chunk of these errors can be attributed to the legacy system's inability to validate data, but some errors are logical in nature. For instance, consider a date field that refers to the purchase date for a particular piece of equipment. Now, the user may input a date, which can be valid but can be wrong from a business perspective (i.e., the user may enter a date in which the business never existed).

• **Deliberate Manipulations**: A user may be forced by the legacy system to enter data that is prima-facie incorrect but is inevitable because the legacy system would reject the data otherwise. Another source for these deliberate manipulations is the legacy system user who may fudge the data to fulfil certain ends, which may be unethical.

• **Target system model definition**: This can be a factor only when a legacy system data conversion and migration is part of the cleansing project. The target system model may dictate the data to be in a certain format, which cannot be found in the legacy system. Though this is not essentially dirty data, the need for conversion and migration makes it compulsive for the business to cleanse it. Alternatively, depending on the difficulty level of conversion required, it can be incorporated in the conversion process.

The presence of this dirty data poses a serious threat to the management and can affect the decisions that are taken. Data cleansing initiative is a direct consequence of the inability of the management to translate data at hand into effective, winning decisions.

**Data Cleansing Methodology**

In a data conversion project the main objective is to convert and migrate clean data into target system. This calls for a need to cleanse legacy data. Cleansing can be an elaborate process depending on the method chosen and has to be planned carefully to achieve the objective of elimination of dirty data.

Some methods to accomplish the task of data cleansing of legacy system include:

- Automated data cleansing
- Manual data cleansing
- The combined cleansing process

**Automated Data Cleansing**

The generalized method of carrying out an automated data cleansing is detailed below. Refer to Figure 1 for a diagrammatic representation of the process.

1. **Error Identification Process (Data Audit)**: The first and foremost step is to identify and categorise the various errors in the legacy system. This is also called the data audit process. This is generally done after a study of the functionality of the legacy system with the help of business analysts. A data audit would reveal the volume of these error types. A data audit process will provide:
1. The error types that need cleansing. These are called as critical errors types.
2. The error types that can safely be ignored as they are not business critical. These can be classified as non-critical error types.
3. Data volume of each of the critical error types.

Various error types are identified for simplicity of tackling it programmatically in the subsequent data cleansing process.

2. **Error Reporting:** It is necessary to verify the dirty data that is identified and extracted by the error identification process. Business analysts are involved in the verification of this data. This verification is done through user-friendly reports of dirty data based on error type.

![Diagram: Automated Data Cleansing]

**Figure 1: Automated Data Cleansing**

3. **Automated Data Cleansing:** This can typically be a batch process to correct the dirty data based on the error types. All logical error types in data structure can be corrected through this programmed cleansing process. However, one should be
mindful of the fact that there will be some potential error types that cannot be correct by the automated process. Manual intervention is the only way to deal with such error types. This method is dealt in detail later in this particle.

4. **Post Cleanse Data Audit:** Data cleansed automatically by the batch process is required to be verified by re-run of error identification process. This is to ensure the successful completion and flawless functioning of the automatic data cleansing process.

5. **Legacy System Update:** This is a crucial step of the process. The data cleanse process is normally run on a different set of target data sources which are populated at the end of the automated data clean process. This separate data area is called cleansing staging area. This is to avoid any irreversible, accidental, incorrect changes to the production legacy system. It is not a very good idea to update the production legacy system directly from the cleansing process, as in doing so we would not be dealing with the error prone data set but with the entire data set. Instead, the cleansed data in a different data source is used to update the production legacy system after ensuring that the data is clean (in step 4). The end result of this would be a clean legacy system.

**Legacy System Retirement**

When a conversion and migration process is planned process just described varies. The consequence of the migration process would be that the legacy system would eventually be decommissioned and would be replaced by a new target system. Therefore, the updation of legacy system, which constitutes the last step in the automated cleansing process, is generally avoided. Instead, the cleansed data is incorporated in the conversion and migration process, which is directly converted and migrated to the target system. This saves the effort of legacy system update and the risks associated with it.

As a conversion and migration is planned there can potentially be additional error types in legacy system, which arise due to the differences in data models between the source and the target system and the additional data constraints in the target system. These error types are to be categorized for cleansing. In such cases legacy data update is not performed, as it would conflict with the data model of the legacy system.

**Other Variations**

In projects wherein data cleansing is the only motive and the legacy system is to be used, a legacy system update is performed. Moreover, in such cases an effort is to be made with respect to the origin of these error types. The investigations would almost always lead to a change to the legacy system. It calls for a decision to be taken on the legacy system rectification for its identified vulnerabilities. Decisive factors which affect the decision include questions:

- Is the legacy system change impact with respect to cost and time scales too high to be feasible?
- Does the change eliminate each and every error type and eliminate the need to rerun the data cleansing process time and again?
• Is it practical to ignore the defects in the legacy system and continue with a standardized automated data cleansing process at timely intervals?
• Is it feasible to migrate the data in existing legacy system to another system?

An analysis based on these factors will result in a set of answers that could decide the future of the legacy system.

**Manual Data Cleansing**

The need for manual data cleansing arises from the fact that not all the errors can be automatically cleansed. There exist certain error types wherein neither a logical conclusion can be drawn nor rules can be formulated about the value that a particular field will take. The only way to cleanse data is to do it manually. A generalised process can be formulated for such a data cleansing procedure in the same lines as we did for the automated process. Figure 2 depicts the manual data cleanse process.

![Figure 2: Manual Data Cleansing Process](http://www.information-management.com/infodirect/20041029/1012952-1.html?zkPrint...19/01/2010)
• **Error Identification Process:** The error identification process would give rise to error reports categorised on the basis of error types.

• **Error Reporting - Data Cleansing Spreadsheets:** The error reports here serve a different purpose unlike the ones in automated data cleansing process. The error reports are sent for manual correction. The business experts would manually correct the error reports and send it back.

• **Post-Cleanse Data Audit:** In the manual process this cycle comprising of data cleansing spreadsheets and post-cleanse data audit is iterative. Manual intervention will have its own share of errors which implies that more than one set of error reports are to be produced to cleanse the residual errors after a rerun of data error identification process has been run.

Projects in which data cleansing is a part of the data conversion and migration process the manually cleansed spreadsheets are incorporated into the conversion process. The target system is populated with the cleansed data from the spreadsheets.

**Legacy System Continuance**

Unlike automated data cleansing process, the effort, cost and time involved in the manual process is huge; so, the manual cleansing process cannot be repeated. Consequently, in projects where no migration is planned, the reason for these errors should be thoroughly investigated. Once the reason for the dirty data is found efforts should be made in the direction of avoiding the reoccurrence of these errors. In dealing with such errors we do not have the privilege of doing a manual data cleanse process all over again.

**The Combined Data Cleansing Process**

Largely, the dirty data present in legacy systems are in both categories:

• Data error types that can be automatically cleansed without manual intervention and

• Data error types that require manual intervention to be cleansed.

This nature of data dictates the use of a combined process. The data errors are categorized into those which can be resolved by the automatic process and those which require manual corrections. The result of this is the employment of a combined data cleansing process.

**Choice of Cleansing Process**

Various factors drive the final decision by the management on the type of cleansing process that is used.

**Automated Cleansing Process:** Automated cleansing process is adopted when:

• Data to be cleansed is too huge to accomplish it manually.

• All or majority of the data errors can be fixed programmatically by applying logical rules.
• The cost involved in manual cleansing is high when compared to the time-scales in which it can be done with an automated process in place.
• The automated process is planned to be re-used at timely intervals owing to the fact that the legacy system change is not feasible.
• The legacy system is going to be replaced by a target system and all the data in the legacy system is subjected to a conversion process.

Manual Cleansing Process: This is best suited in the following circumstances:
• Erroneous data cannot be fixed programmatically.
• Data volumes to be cleansed are very less making the automation process laborious in comparison to the manual process.
• The legacy system is immediately planned to be replaced by a different application and cleansing is a part of the conversion process.

Combined Cleansing Process: The combined process can be employed in the following scenarios:
• Erroneous data are distributed equally between the ones that can be addressed automatically and the ones that are to be handled manually.
• Use of a single process does not produce appreciable data cleansing in the system.

Management Information
An important feature of a data cleansing exercise is the production of data quality statements that reflect the percentage increase in data quality obtained. It should address each and every category of error that is cleansed with the percentage achievement obtained.

Residual Dirty Data
Achieving 100 percent data cleansing in reality is a bit difficult. Despite all the efforts, there still exists a certain percentage of dirty data. This residual dirty data should be reported, stating the reasons for the failure in data cleansing for the same.

Data cleansing is a tedious and time taking exercise which requires a sound methodical strategy, wise choice in the process to be used and a basic understanding of the applicable legacy system and the target system.

Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cleansing Staging Area</td>
<td>A separate area wherein only the error prone records are dumped and the cleansing process is executed.</td>
</tr>
<tr>
<td>Data Audit</td>
<td>An in-depth analysis of the legacy system which reveals the data error types that are prevalent in the system.</td>
</tr>
<tr>
<td>Dirty Data</td>
<td>Data that is required to be cleansed.</td>
</tr>
<tr>
<td>Legacy System</td>
<td>System currently in use by the business.</td>
</tr>
<tr>
<td>Target System</td>
<td>System which would replace the existing legacy system.</td>
</tr>
</tbody>
</table>
**Information Sources:**

*Kuldeep Dongre* is a data conversion and migration analyst with Tata Consultancy Services Ltd. He has more than four years of experience in data conversion and migration, data cleansing, ETL, government accounting and data structure rationalization. You can reach him at kul.deep@corporg.net.

For more information on related topics, visit the following channels:

- Data Quality