The Six Principles of BW Data Validation

The Problem

Users do not trust the data in your BW system.

The Cause

By their nature, data warehouses store large volumes of data. For analytical purposes, the final dataset is frequently aggregated and the transactional detail is lost. When users look at the analytical reports generated out of the BW system, the results they see can not be easily verified. As a result, creating user confidence in data quality becomes difficult and user trust is eroded, which can ultimately undermine the success of the EDW initiative.

The Response

Don’t let data quality be random – establish an organizational mindset, and an IT process to validate data, and ensure optimal data quality for end users, before the end users find the problems themselves.

The Six Principles of Data Validation

First principle: validate across all data - a ‘representative sample’ is insufficient. All it takes is a few missing or duplicated records, or garbled key figures, to ruin the accuracy of many reports.

Steps:
1) Determine what application areas to validate
2) Determine validation level of granularity
3) Determine inclusion time frame for data validation (days, months, quarters or years)
4) Determine validation frequency (daily, weekly, monthly)

Second principle: validate at high level of granularity - primary data validation should not be performed at the ‘low’ level of data granularity. Document or line item-level validation is not necessary, until a possible data quality problem has been identified, and it explodes the effort involved in validation.

Data validation is intended primary to provide broad analysis of data quality, with the ability to narrow down investigation toward smaller subsets of the data where quality problems exist. Designed properly, the validation process helps to quickly refine the investigation and thereby reduce the amount of data that must be analyzed at low granularity levels to find quality problems.
Third principle: put business meaning behind the validation elements – validate data elements that are meaningful to the business, such as Item Net Value, Sales Quantity, etc… In technical terms this is unnecessary. Validation can be accomplished by simply using the numbers behind key figure values. We can not underestimate though, that this data validation is done for the end user, with the aim of improving the user confidence in the data. As such, the ‘numbers’ being validated must be meaningful to the user.

Fourth principle: automate – make data validation a continuous and automated process. It is preferable to execute validation functions in carefully timed sequence to the data loading process. Integration of Automated Data Validation is a critical part of the design, quality assurance and production phases of the project lifecycle. Successful data quality process integration will offer substantial benefits for the BI implementation.

Fifth principle: know the source of your data - it is very important to know and understand the data of each application area in the source system, and how it is entered and stored in the source system database. Data validation is not a simple process. It is not purely technical, but also requires a quality management mindset, and a commitment to data validation process. Data validation requires understanding of the associated business processes, data entry modes, and data warehousing theory. Without this, the validation functions cannot achieve their full potential.

Note:
Validation does not mean that you must check everything about the data. Rather, it should ensure that critical elements like the number of records and/or sums of certain value fields (key figures) arrive and are updated properly into the data warehouse. With this knowledge, setting up your validation rules and processes becomes more ‘mechanical’.

The key things to know are:

1) Which fields are populated by the application?
2) Which of these fields are utilized in data warehouse reporting?
3) What is the frequency with which this information is loaded into the data warehouse?
4) When is data archived out of the source system?

Sixth principle: customize your validation approach - not all BW objects are created the same, nor are the validated in the same way. Validation frequency, selection parameters, and comparison object definition differ depending upon BW object type – an operational data store is fundamentally different in its record types, lengths, and retention periods than a highly summarized cube will have.
BW allows almost limitless data transformation and staging within the architecture, and therefore it is important not to stop at the enterprise data warehouse (EDW) layer. Validation should occur across all layers of the data warehouse architecture, from EDW up through all analytical decision support system (DSS) objects.

Some BW objects may allow validation by using the inherent redundancy of data in the BW system. In such cases, use the redundancy to cross-validate data.

**Options for Data Validation**

Past experience with client BW implementation has shown us that clients most often choose one of the following approaches to BW data validation and quality control:

**Ignore the issue**
This is the most commonly selected option among SAP BW clients. In this scenario, the IT organization lets the BW users find out whether data is right or not. If they don’t hear any complaints, they assume the data is fine. If an issue is raised by a user, then ad-hoc validation testing is performed, but no ongoing process is implemented. This approach is risky because it allows the possibility that errors reside in the BW system undetected for long periods of time, after which the impact of the errors is much greater. Further, it undermines the usability and reputation of the BW system with users. The effect of frequent data reload requests is most visible here.

**Manual validation**
Some BW clients choose to assign a team that will select and manually compare datasets in the source and BW systems for comparison and validation of a ‘statistically significant’ representative data sample according to a pre-written validation test script. This can lead still to undetected data errors and omissions, due to exclusion and human error and it is costly.

If you are going to validate data manually, you will have to determine which table to view in the R/3 system (via SE16n), and then use the Manage function in BW to view the data records in the ODS or Cube (note: to sum records, or to capture record counts in BW, you will have to export the data into a spreadsheet for external manual validation).

If you are validating data that requires information from multiple R/3 tables, you will have to build a join in R/3 to allow comparison of data from both tables to the data that resides in BW.

In sum, this is complex and tedious work, and will require dedication of work hours from knowledgeable resources to perform this validation accurately. Regardless of resource skill, however, you will likely be forced to compare ‘representative sample’ sets of data, since it is impossible to validate hundreds of thousands or millions of records by hand.
Find an application – or start coding
Some BW clients will decide to write their own validation programs for ‘critical’ data areas. This can be done, but is often expensive and time-consuming, and adds an additional maintenance burden to the Support team. Alternatively, you can search on the market for software applications that will validate your R/3 and BW data in flexible, generic and automated fashion.

Effects of the Poor BW Data Quality

Apparent Effects

Certainly, the most visible effect of data quality problems in your BW system is the end user ‘defection rate’ – your users don’t trust the data, and therefore cease to use the system.

A secondary, but visible effect is that the longer a data problem exists in the system, the more widespread its impact is on the system, and organizational decision making. This is most critically noticed in Finance, where a single wrong number can invalidate an entire series of reports and the decisions based upon them.

Frequent data reloads is the next visible effect. When a data warehouse is small, and the data has not yet been archived out of the source systems, this is not a serious issue. But as volume grows and time passes, its impact can be more pronounced.

Hidden Effects

There is an unexpected, yet significant effect: users will require greater record detail to be retained in the BW system, in order to allow them to validate the analytical report results themselves. This leads to data record size and volume explosion, which impacts load and query performance, increases system maintenance costs, and blurs the real purpose of the data warehouse – to provide analytical support for business/organizational decisions. It is also worth mentioning that the idea of validating the BW data by comparing to the source system(R/3) data on transactional record level is not feasible due to the number of records created in the R/3 system (which only increases over time).

How to Build Organizational Commitment to BW Data Quality

All successful SAP BI implementations eventually achieve some level of user confidence. The question is – at what cost? Including an Automated Data Validation Process to the BI implementation pays a quick and significant return on investment.
Most organizations perform ‘some’ form of validation, but it is rarely part of a formalized process, and is part of a commitment to data quality management. Organizations often get caught up with tracking data load statuses, packet statuses, and so on, as an indicator of BW data quality. While that is certainly one part of ensuring proper data quality, it is not the whole story – loads can complete, but data errors arise because of flaws in data extraction, transformation and update.

Data validation should not be a ‘one-time’ or ad-hoc process. There must be a thoughtful, well-planned and continuous validation process that is supported by organizational commitment to ensuring error-free data quality. If the best way an organization can plan for its future is to review the results of past decisions and actions, the data used in this decision making process must be correct.

The organization must:

a) communicate the importance of data quality to the end users and IT staff
b) establish and publish data quality standards for the SAP BW system
c) define and initiate data validation policy and procedures – involve process and data experts in the definition of specific validation rules and logic
d) incorporate data validation into system process automation
e) promote a culture of predictability, reliability and accountability for BW data quality

**When to Validate**

We suggest that data validation in SAP BW is included in both stages of system lifecycle and must become an automated process in:

- project design & development
- production support