reeeliance

TESTING DATA VAULT 2.0

A REEELIANCE **DELIVER** WHITE PAPER

reeeliance IM GmbH

Budapester Str. 43
10787 Berlin

T +49 30 2693 06 63

info@reeeliance.com
www.reeeliance.com

reeeliance

Testing Data Vault 2.0

The ‘Data Quality Market Survey’ of Gartner found that for the average company in 2017 the direct financial cost of poor data quality is about $15 million per anno. This includes reactive response to data quality issues as well as missed business growth opportunities and an increasing of business risks.

Considering these financial risks of data quality issues, testing has become a necessity in the todays DWH development.

This whitepaper describes how a testing concept for testing Data Vault 2.0 can look like and what to consider in implementing it. The concept is inspired by the agile development process which is an integrated part of the data vault methodology. It proposes a continuous testing to find issues as soon as possible.

Introduction

Data is the new oil

among others Kevin Plank
(CEO of Under Amour)

Quotes like on the left illustrate the outstanding meaning of data in the present economy and explain the strive of companies to become a data-driven company. But as crude is not like crude - data is not like data. Comparable to the high-quality demands on crude, also the quality demands of data is a crucial factor for the value of data.

In the course of data warehouse environments, testing has become a necessity to ensure quality and establish trust in the development. At this, especially the existence of complex business logic, number and diversity of source systems, inconsistencies and redundancies represent typical challenges for the quality of data. Having said this, also the transformation logic within the different modelling methods of data warehouses demanding different requirements on testing the data quality.

One of these methods is the Data Vault 2.0 modeling approach. With its promotion of clearly defined processes, agile workflow including short development cycles and its scalable modelling concept, it already supports high data quality from a modeling perspective. In addition, it enables the design of a comprehensive testing suit based on its distinctive characteristics.

When to test?

In an agile data warehouse development project, testing plays a significant role for the continuous delivery process. Due to the short development cycles and the need to deliver a potentially shippable product, testing must be an integrated part of the development. The continuous testing approach is not exclusively supposed to increase the quality of the development by identifying errors, but also to identify them as soon as possible to reduce the costs of failure. A sequence of different tests sets is recommended:

TYPE OF TEST-SET	TIMING OF THE TEST-SET	PURPOSE OF TEST-SET
SMOKE TEST 1	Between every build and data load	Tests structural components of the vault before the data load starts
SMOKE TEST 2	After every build and data load	Tests basic loading patterns of the data load
COMPONENT TESTS FOR NEWLY ADDED FUNCTIONALITY	After the build and data load of the new functionality (until functionality in implemented correctly)	Tests a specific functionality which is added since the last deployment
REGRESSION TESTS FOR ESSENTIAL COMPONENTS OF THE DATA VAULT MODEL	After the build and the data load on demand	Tests if the addition of the new functionally broke some existing functionality

Considering the high frequency and amount of testing with different test suits, it is mandatory that a certain level of test automation will be implemented. The test automation allows to integrate testing as substantial part of the deployment process and simultaneously minimize manual interaction.

What to test?

Regarding the previous mentioned different test set types, there is a wide range of possible test cases. As indicated, the smoke test between the build and the data load can be used to perform structural tests of data vault entities to verify for example “Are all Data Vault entities are built up and have the right columns and data types?”, “Are the primary key set to the right columns?” or “Are the naming / coding convention applied?”. These kinds of structural tests are very powerful as they give a quick first check on a very early state in the deployment phase and thus ensure a solid foundation of the ongoing process.

To get a quick verification on the loading patterns, a second smoke test set can be implemented. This suite gives the possibility to execute quick checks on the project’s individual patterns like loading orders or checks for the ELT processes., The smoke test set should only cover essential functionality within a short time period since other tests complete the picture.

The regression suit takes an extensive look into the core functionality on the data vault. On this occasion, the time aspect is not that critical, but nevertheless not every test should be part of the regression suit. Test cases must be compiled smartly to achieve a high test-coverage with a small number of test cases.

One of the characteristics of data vault is the massive parallel data processing. Even though this leads to high performance of the data load, it requires the absence of foreign key

constraints within the core. For this reason, one of the essential tests is to verify the referential integrity. However, due to the parallel processing the referential integrity cannot be assured at any point in time. Therefore, the test has to take a temporary discrepancy into account. The time factor is not only relevant for the integrity of the vault but can also be relevant for the completeness checks. In the cases of a near real time data warehouse, the continuous data flow and the data processing from a source or landing zone layer to the model leads inevitable to a temporary discrepancy of these layers. However, this can easily be considered within the setup of the test case. Regarding the actual test for completeness, the tests can be combined with a check for correctness. To do this for the Raw Vault, one of the basic ideas of the modelling approach is followed - recreate the source from the Raw Vault. To check if the Raw Vault is complete and correct, simply recreate the source and compare it against the actual source. Even if the vault was loaded correctly, there might be differences due to some condensing mechanisms within the loading procedures. Therefore, a second reconsolidation with the source can be required to identify if the differences are intentionally or not. At this, it is beneficial that between source layer and Raw Vault layer only hard rules and not the complex business rules are applied.

Testing the complex transformations of the business vault constitutes a great challenge for the testing team. To avoid systemic testing errors, it should be inhibited to simply recreate the logic itself in the tests. An alternative could be to

unite deduced values as aggregations or to check for specific output formats which indicate a correct transformation logic.

TYPE OF TEST-SET	KINDS OF TESTS
SMOKE TEST 1	Structural Tests (e.g. Naming / Coding conventions tests, check for constraints, ….)
SMOKE TEST 2	Data loading Tests
COMPONENT TESTS FOR NEWLY ADDED FUNCTIONALITY	Individual Tests for new functionality
REGRESSION TESTS FOR ESSENTIAL COMPONENTS OF THE DATA VAULT MODEL	Referential Integrity Tests Completeness Check Correctness Check Format checks Aggregation checks Logical testing (manual) Testing against sample data

A complete reconsolidation with a comprehensive completeness and correctness check is not achievable in this case. Nevertheless, there are also other approaches to test the business vault, as seen in the table above.

Conclusion

Data Vault Testing requires careful consideration and planning to master the requirements of complete and consistence testing. In this White Paper, we outlined a procedural frame that allows to comply to these requirements. If you want to learn more about the details, feel free to contact us. We are happy to help.

info@reeeliance.com phone +49 30 2693 06 63