Testing data warehouses with key data indicators
Results with Highspeed

Adalbert Thomalla, Stefan Platz
Agenda

The Problem
  - General Problem
  - Problem within the Project

The Idea

The Solution
  - General Method of Solution
  - Solution within the Project
The Problem
General Problem
General Problem

Test in the project / regression test

- **Non-recurring** assurance of the data quality in a **project** within a specified project plan
- Test and retest multiple deliveries of **mass-data**
- Quality assurance of **historical data** for Basel II, IFRS etc.

Plan

![Plan Diagram]

- Testbegin
- Data-delivery
- Test
- Corrected Data-delivery
- Re-Test
- Scheduled end of project

**Time**
Data verification

- **Recurring** assurance of data quality within production
- Continuous check of the delivery of mass data
- **Additional sources of errors** within recurring data deliveries
Problem
Problem within the Project
Concrete Problem within the Project

Project
- Build and test of a DWH for historical Basel II data

Root-Systems → ETL Processes → Basel II DWH (min. 5 years history) → Subsequent processing

eg. calculation of parameters or regulatory reporting
Concrete Problem within the Project

The original plan

- **Non-recurring** historical data delivery and test of this data set (inclusive Re-Test)
- Handover of the daily data delivery within production
- **No** usage of testing tools intended
Concrete Problem within the Project

Scope of testing

- Around 50 tables – 500 fields – several millions of data records
- Around 500 test cases within 3 levels (Possible value range – Data integrity – End-to-End-Test)

Test execution

- Manual execution and documentation of the tests
- Individual execution of every test case
- Documentation of the test execution within a MS Access testing database
Concrete Problem within the Project

Actual condition

- **Recurring** historical data delivery because of changes and incidents
  - Time- and resources consuming
    (Duration of a complete test cycle around 20 person days)
  - Partial abort of the test because of a new data delivery
  - Concentration on one defined test data
    (One historical month)

Additional requirement

- Recurring verification of the data quality in production
Concrete Problem within the Project

Plan

Testbegin → Data-delivery → Corrected Data-delivery → Re-Test → Scheduled end of project

Test preparation → Test → Re-Test → Time

Current situation

Testbegin → Data-delivery → Corrected Data-delivery → Corrected Data-delivery → New Data-delivery

Test preparation → Test → Test → Re-Test → Test → Re-Test ?! → Time
The Idea
The Idea

Design of a slim test tool using predefined data quality indicators

- Fast test execution!
- No time-consuming repeat of all test cases!
- Automatization!
The Idea

Test: Balance Account 0815 Table A = Balance Account 0815 Table B??
The Idea

Table A

DWH ETL / Transformation

Table B

$Balance_A = \sum Balance$

$Balance_B = \sum Balance$

$Balance_A = Balance_B$?
The Idea

\[ \sum \text{Balance} \quad \sum \text{Creditcards} \quad \sum \text{Defaults} \quad \ldots \sum \ldots \]

\[ \sum \text{Balance} \quad \sum \text{Creditcards} \quad \sum \text{Defaults} \quad \ldots \sum \ldots \]
The Solution
General Method of Solution
The Solution

Administration and configuration of indicators and rules

Calculation engine

Reporting

Access

PDF
Process

System Landscape → Indicators & Rules → Execution → Results → Export Testing-Database
Prerequisite

- SAS and Excel
- Data sources directly in SAS or through links to Oracle or DB2
- Authority to read the data
## CGI Dataquality Indicator Solution

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Server</strong></td>
<td><strong>local</strong></td>
</tr>
<tr>
<td><strong>Export Verzeichnis</strong></td>
<td>D:\CGI Projects\Ratio Tool\</td>
</tr>
<tr>
<td><strong>Datum</strong></td>
<td>04MAR2011</td>
</tr>
<tr>
<td><strong>Ergebnis-Dokument Name</strong></td>
<td>Testing_and_Finance_Demo</td>
</tr>
<tr>
<td><strong>Ergebnis-Dokument Titel</strong></td>
<td>Testing_and_Finance_Demo</td>
</tr>
<tr>
<td><strong>Version SAS Programm</strong></td>
<td>CheckRatios.sas</td>
</tr>
<tr>
<td><strong>Version Excel Master</strong></td>
<td>CheckRatios.xls</td>
</tr>
</tbody>
</table>
Indicators in the context are sums and other aggregate functions like:

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Table</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of accounts</td>
<td>Accounts, Scoring</td>
<td>Reconciliation between data sources</td>
</tr>
<tr>
<td>Number of customers</td>
<td>Accounts, Customers</td>
<td>Reconciliation between data sources</td>
</tr>
<tr>
<td>Sum of balance</td>
<td>Accounts, Balance-Sheet</td>
<td>Reconciliation with the Balance-Sheet</td>
</tr>
<tr>
<td>Sum of credit cards balance</td>
<td>Accounts, Balance-Sheet</td>
<td>Reconciliation with the Balance-Sheet</td>
</tr>
<tr>
<td>Number of defaulted accounts without being past due</td>
<td>Accounts</td>
<td>Direct Validation, Integrity</td>
</tr>
<tr>
<td>Number of accounts without scoring</td>
<td>Accounts</td>
<td>Direct Validation, Integrity</td>
</tr>
</tbody>
</table>
From a technical point of view the indicators are summary functions according to the SQL standard (SAS):

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUM</td>
<td>Sum</td>
<td>SUM(Balance)</td>
</tr>
<tr>
<td>AVG</td>
<td>MEAN</td>
<td>Average</td>
</tr>
<tr>
<td>COUNT</td>
<td>FREQ</td>
<td>N</td>
</tr>
<tr>
<td>NMISS</td>
<td>Counting missing values</td>
<td>NMISS(Customer)</td>
</tr>
<tr>
<td>MIN</td>
<td>Smallest value</td>
<td>MIN(Scorevalue)</td>
</tr>
<tr>
<td>MAX</td>
<td>Maximum value</td>
<td>MAX(Scorevalue)</td>
</tr>
<tr>
<td>MIN, PRT, RANGE, STD, STDERR, T, USS, VAR, CSS, CV</td>
<td>Statistics e.g.: STD (standard deviation)</td>
<td>STD(Scorevalue)</td>
</tr>
</tbody>
</table>
• Conditional sum of balance of accounts being in a special product group with CASE WHEN...

\[
\text{SUM(CASE WHEN ARREAR } \leq 0 \text{ AND Default } = 1 \\
\text{ THEN 1} \\
\text{ ELSE 0 END)}
\]

• Complex functions with the SAS Macro engine possible (but knowledge in programming necessary):

\[
\text{SUM(CASE WHEN NPV } \neq ((\text{YEAR1} / (1.1)^1) \% DO i} \\
\text{ = 2 } \% \text{ TO 12; + (YEAR}&i./1.1**&i.) \% \text{ end; )}
\]

\[
\text{THEN 1} \\
\text{ ELSE 0 END)}
\]
The summary functions are placed without a complete SQL function in the MS Excel sheet:

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>VERZEICHNIS</td>
<td>TABELLE</td>
<td>KENNZAHL</td>
<td>KENNZAHL NAME</td>
<td>KENNZAHL BESCHREIBUNG</td>
<td>SOLL ERGEBNIS</td>
</tr>
<tr>
<td>3</td>
<td>D:\Data\</td>
<td>Konten</td>
<td>SUM(Saldo)</td>
<td>Kennzahl_1</td>
<td>Summe Saldo aller Konten</td>
<td>VALUE NE 0</td>
</tr>
<tr>
<td>4</td>
<td>D:\Data\</td>
<td>Konten</td>
<td>SUM(CASE WHEN Produkt = 'Giro' THEN Saldo ELSE 0 END)</td>
<td>Kennzahl_2</td>
<td>Summe Saldo Giro-Konten</td>
<td>VALUE NE 0</td>
</tr>
<tr>
<td>5</td>
<td>D:\Data\</td>
<td>Konten</td>
<td>SUM(CASE WHEN Rueckstand &gt; 0 AND Saldo &lt; 0 THEN 1 ELSE 0 END)</td>
<td>Kennzahl_3</td>
<td>Anzahl Konten mit Rückstand und Habensaldo</td>
<td>VALUE EQ 0</td>
</tr>
<tr>
<td>6</td>
<td>D:\Data\</td>
<td>Konten</td>
<td>SUM(CASE WHEN Rueckstand &lt;= 0 AND Ausfall = 1 THEN 1 ELSE 0 END)</td>
<td>Kennzahl_4</td>
<td>Anzahl ausgefallener Konten ohne Rückstand</td>
<td>VALUE EQ 0</td>
</tr>
</tbody>
</table>
Rules define criteria for combinations of indicators and may be directly assigned to test cases.

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TESTFALL</td>
<td>REGEL BESCHREIBUNG</td>
<td>REGEL</td>
</tr>
<tr>
<td></td>
<td>TF_KONTEN_SALDO_GIRO</td>
<td>Summe Saldo Giro-Konten muss der entsprechenden Bilanzposition entsprechen</td>
<td>Kennzahl_2 EQ Kennzahl_5</td>
</tr>
<tr>
<td>2</td>
<td>TF_KONTEN_RUECKSTAND</td>
<td>Plausibilität der Rückstands berechnung</td>
<td>Kennzahl_3 EQ 0 AND Kennzahl_4 EQ 0</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Execution

1. Start
2. Read indicators and rules
3. Indicators calculation
4. Rules checking
5. Export of results
6. Import

Data Sources

Results
Indicators

Results in PDF and Excel

Testing-Database

SAS Program

System Landscape
Indicators & Rules
Execution
Results
Export Testing-Database

Indicators and Rules

Read indicators and rules

Results

Indicators & Rules

Calculation

Rules checking

Export of results

Testing-Database

Import
<table>
<thead>
<tr>
<th>Indicator Name</th>
<th>Kennzahl_4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Number of defaulted accounts without arrear</td>
</tr>
<tr>
<td>Indicator</td>
<td>SUM(CASE WHEN Arrear &lt;= 0 AND Default = 1 THEN 1 ELSE 0 END)</td>
</tr>
<tr>
<td>Expected Result</td>
<td>VALUE EQ 0</td>
</tr>
<tr>
<td>Result</td>
<td>1446</td>
</tr>
<tr>
<td>Compare</td>
<td>Incident</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Testcase</th>
<th>TF_KONTEN_RUECKSTAND</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Plausibility of Arrear</td>
</tr>
<tr>
<td>Rule</td>
<td>Kennzahl_3 EQ 0 AND Kennzahl_4 EQ 0</td>
</tr>
<tr>
<td>Result</td>
<td>Incident</td>
</tr>
</tbody>
</table>
## Kennzahlen Ergebnis Bericht

Verzeichnis=D:\CGI Projects\Ratio Tool\Data\Tabelle=Bilanz

<table>
<thead>
<tr>
<th>Kennzahl Name</th>
<th>Beschreibung</th>
<th>Kennzahl</th>
<th>Soll-Ergebnis</th>
<th>Ergebnis</th>
<th>Soll-Ist-Vergleich</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kennzahl_5</td>
<td>Summe Saldo Giro-Konten</td>
<td>SUM(CASE WHEN Produkt = 'Giro' THEN Saldo ELSE 0 END)</td>
<td>VALUE NE 0</td>
<td>34020000</td>
<td>OK</td>
</tr>
</tbody>
</table>

Verzeichnis=D:\CGI Projects\Ratio Tool\Data\Tabelle=Konten

<table>
<thead>
<tr>
<th>Kennzahl Name</th>
<th>Beschreibung</th>
<th>Kennzahl</th>
<th>Soll-Ergebnis</th>
<th>Ergebnis</th>
<th>Soll-Ist-Vergleich</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kennzahl_1</td>
<td>Summe Saldo aller Konten</td>
<td>SUM(Saldo)</td>
<td>VALUE NE 0</td>
<td>134460000</td>
<td>OK</td>
</tr>
<tr>
<td>Kennzahl_2</td>
<td>Summe Saldo Giro-Konten</td>
<td>SUM(CASE WHEN Produkt = 'Giro' THEN Saldo ELSE 0 END)</td>
<td>VALUE NE 0</td>
<td>34020000</td>
<td>OK</td>
</tr>
<tr>
<td>Kennzahl_3</td>
<td>Anzahl Konten mit Rückstand und Habensaldo</td>
<td>SUM(CASE WHEN Ruckstand &gt; 0 AND Saldo &lt; 0 THEN 1 ELSE 0 END)</td>
<td>VALUE EQ 0</td>
<td>0</td>
<td>OK</td>
</tr>
<tr>
<td>Kennzahl_4</td>
<td>Anzahl ausgefallener Konten ohne Rückstand</td>
<td>SUM(CASE WHEN Ruckstand &lt;= 0 AND Ausfall = 1 THEN 1 ELSE 0 END)</td>
<td>VALUE EQ 0</td>
<td>1446</td>
<td>Fehler</td>
</tr>
</tbody>
</table>
### Regeln Ergebnis Bericht

<table>
<thead>
<tr>
<th>Testfall</th>
<th>Beschreibung</th>
<th>Regel</th>
<th>Ergebnis</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF_KONTEN_SALDO_CIRO</td>
<td>Summe Saldo Ciro-Konten muss der entsprechenden Bilanzposition entsprechen</td>
<td>Kennzahl_2 EQ Kennzahl_5</td>
<td>OK</td>
</tr>
<tr>
<td>TF_KONTEN_RUECKSTAND</td>
<td>Plausibilität der Rückstandsberechnung</td>
<td>Kennzahl_3 EQ 0 AND Kennzahl_4 EQ 0</td>
<td>Fehler</td>
</tr>
</tbody>
</table>

[Results]

E-MAIL
Testing Database

- Documentation of the results within self-developed testing database
- Documentation of test cases and test executions
- Incident-Reporting
- Status-Tracking
- Import of Rules results
Constraints of the Solution

- Indicators are partially „just“ indicators for an incident
- Not all test cases possible, e.g. End-to-End-Test
- Explanatory power of indicators compared to test of individual records
Benefits of the Solution

Situation before

Situation after

Test cases

Test cases
Benefits of the Solution

Before

- Analysis: 25%
- Conception: 15%
- Execution: 50%
- Documentation: 10%

After

- Analysis: 50%
- Conception: 25%
- Execution: 15%
- Documentation: 10%
Benefits of the Solution

- Light weigh realization of the Idea with MS Excel and SAS
- Standardized checking logic through summary functions
- Performant realization of the indicators with SAS
- Editing of indicators through business department possible
- Fast reports with results in MS Excel and PDF
- Integration in testing database possible (e.g. MS Access)
Capabilities

- **Project**
  - Reduce testing effort
  - Regression tests and Ad Hoc Retests

- **Continuous data verification**
  - Daily usage to assure the quality of input data
  - Complete Data Warehouse
The Solution Implementation in the Project
Implementation in the Project

Before

- Around 500 test cases in 3 levels
  (Possible values – Integrity – End-to-End-Test)
- Focusing on one selected data set (one historical month)
- Duration of one complete cycle around 20 person days

After

- Around 400 test cases of level 1 and 2
  (Possible values – Integrity)
- Test of every historical month possible
- Duration working with the tool: around 5 hours
  ➔ Duration of one complete cycle around 8 person days
Implementation in the Project

Project Success
✓ Automated and fast execution of the test cases
✓ Complete test of the data possible
✓ Assured data quality within the scheduled project deadline

Production
✓ Verification of daily and monthly data possible
Implementation in the Project

Current situation

Testbegin → Data-delivery → Test → Corrected Data-delivery → Corrected Data-delivery → New Data-delivery → Re-Test → Test

With the tool

Testbegin → Test preparation → Test → Test → Test → Test → Test → Test → Production

Scheduled end of project

Time
Discussion
Contact

Adalbert Thomalla
Lead Consultant
+49 (0)211 5355 0
adalbert.thomalla@cgi.com

Stefan Platz
Senior Consultant
+49 (0)211 5355 0
stefan.platz@cgi.com
Our commitment to you
We're trusted advisors committed to delivering solutions that control spending and improve productivity.
PROPRIETARY AND CONFIDENTIALITY NOTICE

Confidentiality
The material contained within this document is proprietary and confidential. It is the sole property of CGI Inc. (CGI), and may not be disclosed to anyone except for the purpose of bidding for work for or on behalf of CGI. Based upon the extent of the information provided, individuals with access to this information may be required to sign a nondisclosure confidentiality agreement.

Intellectual Property Rights
The contents of this document remain the intellectual property of CGI at all times. Unauthorised reproduction, photocopying or disclosure to any other party, in whole or in part, without the expressed written consent of CGI is prohibited. All authorised reproductions must be returned to CGI immediately upon request.

Trademarks
All trademarks mentioned herein, marked and not marked, are the property of their respective owners.