

Checkpoint Charlie – Endstation für schlechte Daten

Jens Behring
its-people

Schlüsselworte

Datenqualität, Oracle, dynamisches SQL, Betriebsführung, Prüfungen

Einleitung

In diesem Vortrag wird ein seit mehreren Jahren im Einsatz befindliches Konstrukt zum Aufspüren von Datenqualitätsproblemen vorgestellt. Dabei wird auch auf die Einbeziehung der beteiligten Fachdienste eingegangen, ohne deren Mitarbeit ein solches Konstrukt zum Scheitern verurteilt ist. Bei der technologischen Umsetzung wurde Wert auf ein einheitliches und zentrales Vorgehen gelegt, um die Administration möglichst einfach zu halten. Geplante Erweiterungen und Verbesserungen werden ebenso erwähnt, wie Probleme hinsichtlich einer ausgelagerten Betriebsführung.

Zentrales Prüfungssystem

Datenqualität ist in allen Datenbanken, in die Daten aus anderen Systemen importiert werden, ein Thema. Dabei ist es unerheblich, ob es sich hierbei um ein Datawarehouse-System handelt oder um ein klassisches Datenbanksystem. Ungeprüfte Übernahme fehlerhafter oder widersprüchlicher Daten ist in vielen Fällen sogar eher schlimmer als ein pauschales Ablehnen von Datenlieferungen, sollte ein Fehler bemerkt werden. Eine mögliche Lösung des Problems könnte ein Verschieben der erhaltenen Daten nach /dev/null sein, allerdings führt dies langfristig nicht zu dem gewünschten Ergebnis.

Im Rahmen des hier beschriebenen Prüfungskonzeptes wurde ein anderer Weg gewählt. Zunächst wurden einige Grundvoraussetzungen definiert:

- Alle Prüfungen, egal wie komplex auch immer, lassen sich innerhalb einer Datenbank als Select-Statement formulieren
- Eine Veränderung der Prüfungen soll ohne DBA-Rechte möglich sein
- Keine Mischung verschiedener technologischer Ansätze (z.B. Datenbank-Check-Constraints auf der einen Seite und Prüf-Skripte auf der anderen)
- Prüfungen sollen temporär deaktiviert werden können
- Eine Ausnahmenliste soll für einzelne Objekt-Prüfungskombinationen gepflegt werden können
- Ergebnisempfänger sollen sinnvolle und verwertbare Meldungen erhalten

Im Rahmen der weiteren Planungen entstand die Idee zu einem Konstrukt aus Tabellen und einer API in PL/SQL. In den Tabellen werden die abnehmenden Fachdienste und Prüfungen verwaltet, mit Hilfe der API ist ein Ausführen der Prüfungen möglich. Die Hinterlegung der Prüfungen kann Abbildung 1 entnommen werden.

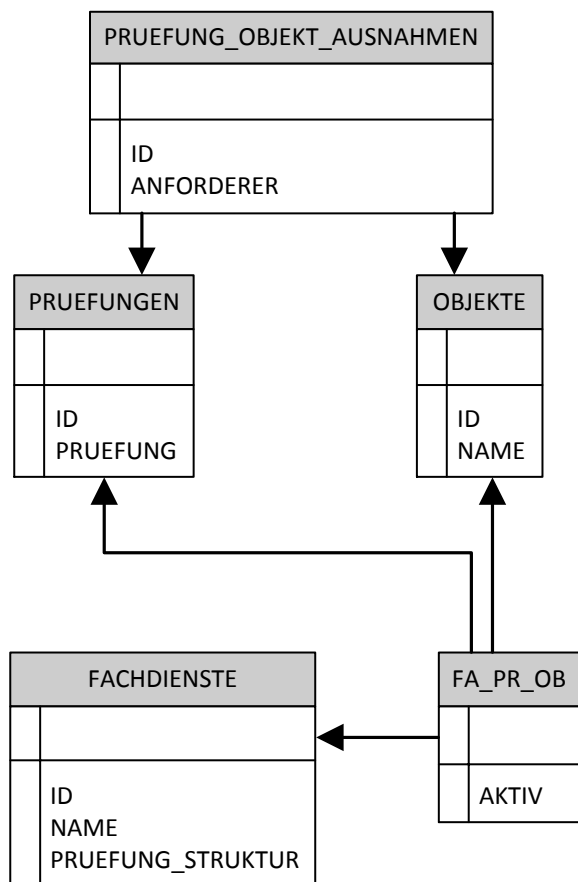


Abbildung 1: ER-Diagramm

Für jeden abnehmenden Fachdienst wurde definiert, welche Felder aus den gelieferten Dateien in Kombination mit bestimmten Fehlertexten für die Bearbeitung der Fehler oder der Widersprüche nötig sind. Alle diesem Fachdienst zugeordneten Prüfungen müssen so gebaut sein, dass sie diese Strukturvorgaben erfüllen. Durch diese Abstimmung konnte bei den Fachdiensten eine recht hohe Akzeptanz erreicht werden.

Durch das Aktiv-Flag können einzelne Prüfungen aktiviert oder deaktiviert werden. Dies ist z.B. zu bestimmten Zeitpunkten sinnvoll, in denen entweder eine Bearbeitung der Meldungen nicht möglich ist, oder aber auf Grund von bestimmten Zuständen im liefernden System mit einer erhöhten, fachlich aber nicht zu begründeten Anzahl an Meldungen zu rechnen ist.

In der Ausnahmenliste können einzelne Datensätze mit ihrem Primärschlüssel und einer bestimmten Prüfung eingetragen werden. Die geschieht nur nach Rücksprache mit dem Fachdienst und einer schriftlichen Bestätigung dieser Ausnahme.

Die Prüfungen wurden in Zusammenarbeit mit dem abnehmenden System und dem liefernden Fachdienst erstellt, auch um beide Seiten hinsichtlich der Belange der jeweils anderen Seite ein wenig zu sensibilisieren.

In obigem Datenmodell ist noch eine Zuordnung von Prüfungen zu div. Fachdiensten möglich. Dies hat sich in der Praxis allerdings als untauglich erwiesen und ist in dieser Form nicht zum Einsatz gekommen.

Dieses beschriebene Konstrukt ist seit 2004 produktiv im Einsatz. Zurzeit finden sich im System ca. 500 Prüfungen für 4 Fachdienste. Diese Prüfungen werden je nach Fachdienst wöchentlich, monatlich oder nur wenige Male pro Jahr ausgeführt. Die Information der Anwender erfolgt zum einen per Mail inkl. eines CSV-Anhangs, zum anderen ist es möglich, sich die Prüfungsergebnisse in einer APEX-Anwendung anzuschauen. In einem Fall werden die Ergebnisse sogar direkt ins liefernde SAP-System geladen und stehen dort dem Anwender wöchentlich aktualisiert zur Verfügung.

Geplante Erweiterungen

Im Laufe der Jahre hat sich natürlich gezeigt, dass im Jahr 2004 nicht alle Eventualitäten berücksichtigt wurden. So sind zurzeit die Einträge in der Ausnahmenliste zeitlich unbefristet, so dass diese Möglichkeit inzwischen ab und an dazu genutzt wurde, um sich lästiger regelmäßig wiederkehrender Meldungen zu entledigen. Dies soll in der aktuell geplanten Version erweitert werden.

Aktuell ist es nur möglich alle Prüfungen eines Fachdienstes aufzurufen. Einzelne Prüfungsaufrufe sind nur möglich, in dem alle anderen Prüfungen deaktiviert werden. Erfolgt ein solcher Aufruf, wird automatisch auch, sofern vorgesehen, eine Mail verschickt. In Zukunft soll zum einen der regelmäßige Batchbetrieb möglich sein (analog zu heute), alternativ soll es aber auch eine Möglichkeit geben, Prüfungen online im Rahmen einer Anwendung aufzurufen und sich das Ergebnis nur in der Oberfläche präsentieren zu lassen. Dies ist z.B. für Live-Auswertungen im Rahmen einer Besprechung sinnvoll aber auch für den Entwickler, um eine Prüfung „mal eben“ testen zu können.

Weiterhin soll den verschiedenen Schritten beim Laden der Daten Rechnung getragen werden und eine weitere logische Gruppierung geschaffen werden. Damit wird es möglich, SAP-Daten vor dem Laden in die Staging-Area zu prüfen oder aber die SAP-Daten nach dem ersten Verarbeitungsschritt – es ist durchaus denkbar, dass einige Prüfungen erst nach einer gewissen Verarbeitung sinnvoll sind. Dies ist vor allem dann der Fall, wenn Daten aus mehreren liefernden Systemen gegeneinander geprüft werden, da eine Zusammenführung dieser Daten nicht unbedingt in der Staging-Area erfolgt.

Ein nicht zu unterschätzender Punkt ist die Absprache mit der Betriebsführung. Im hier beschriebenen Fall erfolgte die Betriebsführung der Datenbank durch die Mitarbeiter des Endkunden – insofern waren Zugriffsrechte und andere Restriktionen kein großes Thema. Aktuell erfolgt jedoch die Überführung zu einem IT-Dienstleister ins Rechenzentrum. In diesem Zusammenhang hat sich das verwendete dynamische SQL als ein großes Problem erwiesen. Zur Zeit erfolgt die Prüfung der Daten mit Hilfe von „execute immediate“, was in der Betriebsführung ausnahmslos verboten ist. Dadurch ist eine Umstellung auf Sysref-Cursor nötig, welche in diesem Fall erlaubt sind.

Ein weiteres Betriebsführungs-Problem ist die bisherige Möglichkeit, zur Laufzeit Select-Statements in die Datenbank zu laden, welche kurz darauf als Prüfung ausgeführt werden können. Hier kann es natürlich in bestimmten Fällen zu Performanceeinbußen kommen, wenn ein extrem unperformantes Statement abgesetzt wird.

Diskussionen mit der Mail-Server-Betriebsführung hinsichtlich Mailspamming im großen Stil konnten nach einigen Telefonaten und Meetings aus der Welt geschafft werden – immerhin werden aktuell pro

Woche ca. 150 Mails mit einem Datenvolumen von ca. 500MB verschickt..

Im Rahmen der Erweiterung wird es auch möglich sein, Prüfungen der Kategorie „kritisch“ zuzuordnen. Wird im Rahmen eines Ladeprozesses mind. ein Datensatz gefunden, der gegen eine kritische Prüfung verstößt, so wird dies dem aufrufenden Programm als Ergebnis mitgeteilt. Hier kann dann ein Abbruch der weiteren Verarbeitung bis zur Klärung des Sachverhaltes vorgenommen werden.

Zusammenfassung

Die hier vorgestellte Möglichkeit der Prüfung von gelieferten Daten hat Ihre Praxistauglichkeit im Rahmen von 7 Jahren produktiven Einsatz gezeigt. Datenprüfungen werden an einer Stelle zentral verwaltet und sind so auch über Urlaubs- und Krankheitsphasen einzelner Mitarbeiter verfügbar.

Sollten Widersprüche nicht korrigierbar sein, ist es möglich, einzelne Datensätze von Prüfungen auszuschließen, um Anwender nicht über lange Zeiträume mit wiederkehrenden Meldungen zu nerven und. Damit würde die Akzeptanz eines solchen Systems unweigerlich sinken.

Die reichhaltigen Möglichkeiten der Bereitstellung der Prüfungsergebnisse sorgen dafür, dass der Aufwand für die Ergebnisempfänger in überschaubaren Grenzen bleibt. Dies ist im Sinne einer möglichst zeitnahen und vollständigen Abarbeitung der Widersprüche ein wichtiger Punkt.

Die Qualität der gelieferten Daten ist in vielen, leider nicht in allen Bereichen deutlich verbessert worden. In einigen Fällen war dies auch darauf zurückzuführen, dass sich alle betroffenen Fachdienste – liefernder, verarbeitender als auch empfangender Fachdienst –gemeinsam mit dem Thema Datenqualität und vor allem auch mit den Erwartungen und Problemen der Gegenseite auseinandergesetzt haben.

Kontaktadresse:

Jens Behring
its-people Hochtaunus GmbH
Lyoner Straße 44-48
D-60528 Frankfurt

Telefon: +49 (0) 69-2475210-0
Fax: +49 (0) 69-2475210-21
E-Mail: jens.behring@its-people.de
Internet: www.its-people.de