www.immobilienscout24.de

IMMOBILIEN SCOUT 24

Methoden zum Befüllen von SCD2

Berlin | 08.12.2010 | Dierk Bachmann

Inhaltsverzeichnis

Kapitel 01

Gewünschte Eigenschaften von SCD2s

- Als Flat Dimension relational in einem Star-Schema gespeichert.
- Verarbeitet mehrere Deltalieferungen je Effektivtag.
- Kann auch Typ1-Spalten beinhalten.
- Performante Befüllung.
- Standard-Spalten der Extraktionsprozesse können verwendet werden.
- Beinhaltet alle gewünschten Standard-Spalten für SCD2s.

Standard-Spalten der Extraktionsprozesse

STG_PRODUKTE		
prid	number	<pk>
enutzer	varchar2	
edatum	date	
gnutzer	varchar2	
gdatum	date	
ods_stcode	varchar2	

- ENUTZER:
  User, der den Datensatz erstellt hat.
- EDATUM:
  Datum der Erstellung des Datensatzes.
- GNUTZER:
  User, der den Datensatz zuletzt geändert hat.
- GDATUM:
  Datum, zu dem der Datensatz zuletzt geändert wurde.
- ODS_STCODE:
  Steuerungscode, der angibt, ob der Datensatz im operativen System aktiv (A) oder gelöscht (L) ist.

Standard-Spalten von SCD2s (1)

REP_DIM_PRODUKT		
dimension_key	number	<pk>
prid	number	
gueltig_von	date	
gueltig_bis	date	
ist_letzte_version	varchar2	
erstellungsdatum	date	
loeschdatum	date	
aenderungsdatum	date	
aenderungsgrund	varchar2	

- DIMENSION_KEY (obligatorisch):
  Surrogate-key, der den PK der Dimension darstellt.
- GUELTIG_VON (obligatorisch):
  Datum, ab dem die Version gültig ist. Dieses Datum beinhaltet keine Uhrzeit.
- GUELTIG_BIS (obligatorisch)
  Datum, bis wann der Datensatz gültig ist. Dieses Datum beinhaltet keine Uhrzeit. Defaultwert: 31.12.9999
- IST_LETZTE_VERSION (obligatorisch)
  „Y“, wenn es sich um die letzte Version eines Datensatzes handelt, ansonsten „N“.
- ERSTELLUNGSDATUM (obligatorisch)
  Datum, zu dem der Datensatz im Quellsystem erstellt wurde.

Standard-Spalten von SCD2s (2)

REP_DIM_PRODUKT		
dimension_key	number	<pk>
prid	number	
gueltig_von	date	
gueltig_bis	date	
ist_letzte_version	varchar2	
erstellungsdatum	date	
loeschdatum	date	
aenderungsdatum	date	
aenderungsgrund	varchar2	

- LOESCHDATUM (nicht obligatorisch)
  Datum, an dem der Datensatz im Quellsystem gelöscht wurde.
- AENDERUNGSDATUM (obligatorisch)
  Datum, an dem der Datensatz geändert wurde. Im Gegensatz zum GUELTIG_VON, beinhaltet dieses Datum auch die Uhrzeit der Löschung.
- AENDERUNGSGRUND (nicht obligatorisch)
  Grund für die Erstellung einer neuen Version. Z.B. „Änderung der Suchregionen“.

Kapitel 02

(1) Gewünschte Eigenschaften von SCD2s
(2) Befüllung per Merge
(3) Befüllung mit Exchange Partition
(4) Vergleich der beiden Methoden

Vorgehensweise

- Bereitstellen der Deltalieferung in einer Tabelle. (Ausgangstabelle)
- Das Befüllen der Dimension geschieht durch ein einzelnes Merge-Statement.
- Beinhaltet die Dimension auch Typ1-Spalten, so ist zusätzlich ein Update-Statement notwendig.
- Das Merge-Statement ist sehr komplex, wenn es alle gewünschten Eigenschaften abdecken soll. Daher wird hier nur das Prinzip erklärt.

Ausgangstabelle

- Die Ausgangstabelle beinhaltet alle Spalten der Dimension, die keine Standard-Spalten sind.
- Zusätzlich beinhaltet die Ausgangstabelle Standard-Spalten des Extraktionsprozesses:
  - EDATUM:
    Datum der Erstellung des Datensatzes.
  - GDATUM:
    Datum, zu dem der Datensatz zuletzt geändert wurde.
  - ODS_STCODE:
    Steuerungscode, der angibt, ob der Datensatz im operativen System aktiv (A) oder gelöscht (L) ist.

Merge-Statement (1)

- In der USING-Klausel werden die Ausgangstabelle und die Dimension per LEFT OUTER JOIN verbunden.
- Der Join geschieht über den NK. Außerdem wird die Dimension auf die letzte Version der Datensätze eingeschränkt, die nicht gelöscht sind.

```
        .
        .
        .
from    <ausgangstabelle> a
left    join <dimension> d
        on(     d.<nk> = a.<nk>
            and d.gueltig_bis = to_date( '31.12.9999', 'dd.mm.yyyy')
          )
        .
        .
        .
```

Merge-Statement (2)

- Bei Änderungen von Datensätzen in der Dimension muss der alte Datensatz abgeschlossen und ein neuer Datensatz angelegt werden.
- Um das zu erreichen, müssen die ermittelten Datensätze verdoppelt werden. Dabei wird ein Datensatz mit „U“ (Update) und einer mit „I“ (Insert) markiert.

```
        .
        .
        .
from    <ausgangstabelle> a
left    join <dimension> d
        on(     d.<nk> = a.<nk>
            and d.gueltig_bis = to_date( '31.12.9999', 'dd.mm.yyyy')
          )
join    ( select case
                   when level = 1
                   then
                     'I'
                   else
                     'U'
                 end            dml_type
          from   dual
          connect by level <= 2
        ) c
        .
        .
        .
```

Merge-Statement (3)

```
        .
        .
        .
join    ( select case
                    when level = 1
                    then
                      'I'
                    else
                      'U'
                  end              dml_type
          from   dual
          connect by level <= 2
        ) c
        on( -- Änderung
              (     d.<nk> is not null
                and a.ods_stcode <> 'L'
              )
            -- neuer Datensatz
            or(     c.dml_type = 'I'
                and d.<nk> is null
              )
            -- gelöschter Datensatz
            or(     c.dml_type = 'U'
                and a.ods_stcode = 'L'
              )
          )
        .
        .
        .
```

- Da nicht alle Datensätze in der Ausgangstabelle eine Änderung in der Dimension bewirken, muss in der ON-Klausel entsprechende Einschränkungen vorgenommen werden.
- Folgende Fälle müssen beachtet werden:
  - Änderungen
  - Neue Datensätze
  - Löschungen von Datensätzen

Merge-Statement (3)

```
        .
        .
        .
join    ( select case
                    when level = 1
                    then
                      'I'
                    else
                      'U'
                  end              dml_type
          from    dual
          connect by level <= 2
        ) c
        on( -- Änderung
              (     d.<nk> is not null
                and a.ods_stcode <> 'L'
              )
            -- neuer Datensatz
            or(     c.dml_type = 'I'
                and d.<nk> is null
              )
            -- gelöschter Datensatz
            or(     c.dml_type = 'U'
                and a.ods_stcode = 'L'
              )
          )
        .
        .
        .
```

- Da nicht alle Datensätze in der Ausgangstabelle eine Änderung in der Dimension bewirken, muss in der ON-Klausel entsprechende Einschränkungen vorgenommen werden.
- Folgende Fälle müssen beachtet werden:
  - Änderungen
  - Neue Datensätze
  - Löschungen von Datensätzen
- ACHTUNG! Vor dem Löschen können noch Änderungen geschehen sein. Diese würden hier verloren gehen.

Merge-Statement (4)

- Ermitteln von Standard-Spalten.
- Es ist wichtig, dass für Updates das ermittelte GUELTIG_VON gleich dem GUELTIG_VON in der Dimension ist, da es als Match-Kriterium verwendet wird.

```
merge /*+ append */ into <dimension> m
using( select case
                when c.dml_type = 'U'
                then
                  d.gueltig_von
                else
                  trunc( nvl( a.gdatum, a.edatum))
              end                                   gueltig_von
       ,      case
                when c.dml_type = 'I'
                then
                  to_date( '31.12.9999', 'dd.mm.yyyy')
                when a.ods_stcode = 'L'
                then
                  trunc( a.gdatum)
                when c.dml_type = 'U'
                then
                  trunc( a.gdatum) -1
              end                                   gueltig_bis
       ,      a.gdatum                             aenderungsdatum
       ,      a.edatum                             erstellungsdatum
              .
              .
              .
```

Merge-Statement (5)

- Match-Kriterium ist der NK und das GUELTIG_VON.
- Beim Update wird das GUELTIG_BIS gesetzt und somit der Datensatz abgeschlossen.
- Außerdem wird das Flag IST_LETZTE_VERSION auf „N“ gesetzt, wenn der Datensatz nicht gelöscht wurde.

```
                .
                .
                .
      ) u
      on(     m.<nk>          = u.<nk>
          and m.gueltig_von = u.gueltig_von
        )
when matched
then
  update
  set     m.gueltig_bis         = u.gueltig_bis
  ,       m.ist_letzte_version = case
                                   when u.ods_stcode = 'A'
                                   then
                                     'N'
                                   else
                                     'Y'
                                 end
                .
                .
                .
```

Merge-Statement (6)

```
                .
                .
                .
when not matched
then
  insert
  ( dimension_key
  , gueltig_von
  , gueltig_bis
  , aenderungsdatum
  , erstellungsdatum
  , <nk>
  .
  .
  .
  , ist_letzte_version
  )
  values
  ( <sequenz>.nextval
  , u.gueltig_von
  , u.gueltig_bis
  , u.aenderungsdatum
  , u.erstellungsdatum
  , u.<nk>
  .
  .
  .
  , 'Y'
  )
;
```

- Beim Insert wird der DIMENSION_KEY aus einer Sequenz befüllt.
- Das Flag IST_LETZTE_VERSION wird auf „Y“ gesetzt.

Was nicht berücksichtigt wurde

- Änderungen am Tag der Löschung.
- Löschen eines Datensatzes an dem Tag, an dem er erstellt wurde.
- Mehrfache Deltalieferung für den gleichen Effektivtag.
- Typ1-Änderungen.

Kapitel 03

(1) Gewünschte Eigenschaften von SCD2s
(2) Befüllung per Merge
(3) Befüllung mit Exchange Partition
(4) Vergleich der beiden Methoden

Vorgehensweise

- Partitionierung der Dimension nach GUELTIG_BIS.
- Bereitstellen der Deltalieferung in einer Tabelle. (Ausgangstabelle)
- Aufteilen der Befüllung in zwei Schritten:
  - Vorbereiten der Daten.
  - Daten in die Dimension per exchange partition und Typ1-Änderungen per Update laden.
- Die Logik befindet sich in einem PL/SQL-Package. Es gibt für jeden der beiden Schritte eine Prozedur.
- Diese Prozeduren können in einem OWB-Mapping aufgerufen werden.

Ausgangstabelle

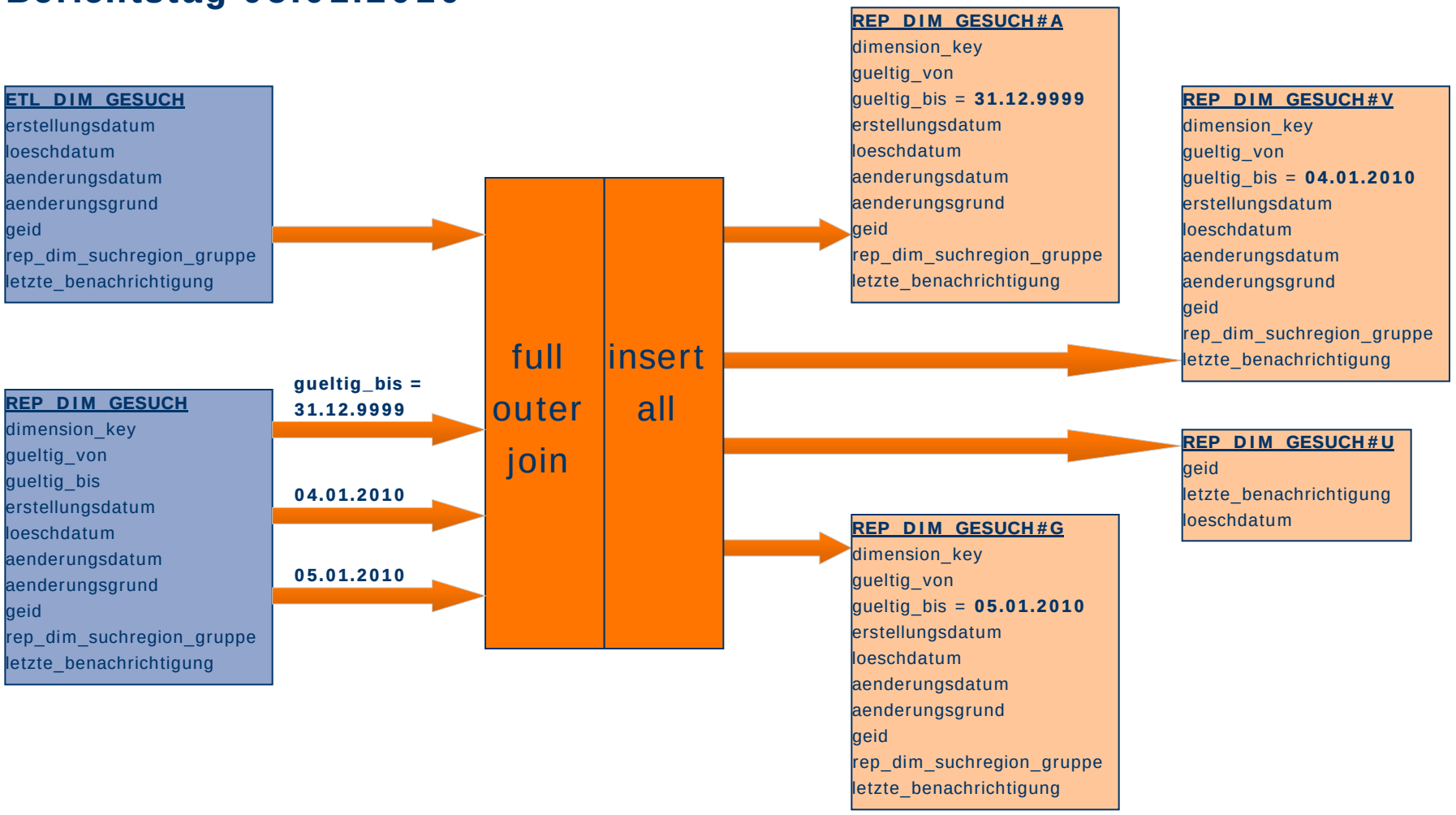
- Die Ausgangstabelle beinhaltet alle Spalten der Dimension, die keine Standard-Spalten sind.
- Zusätzlich gibt es Standard-Spalten für die Ausgangstabelle:
  - AENDERUNGSGRUND (nicht obligatorisch)
    Diese Spalte muss vorhanden sein, wenn es auch in der Dimension vorhanden ist.
  - ERSTELLUNGSDATUM (obligatorisch)
  - LOESCHDATUM (obligatorisch)
    Dieses Feld muss auch dann in der Ausgangstabelle vorhanden sein, wenn es sich nicht in der Dimension befindet.
  - AENDERUNGSDATUM (obligatorisch)

Vorbereiten der Daten (1)

- Erstellen von drei Tabellen (Austauschtabellen), die beim Exchange Partition mit der Dimension verwendet werden:
  - Tabelle, die alle aktuellen Datensätze beinhaltet. (GUELTIG_BIS = 31.12.9999)
  - Tabelle, die alle neu gelöschten Datensätze beinhaltet. (z.B. GUELTIG_BIS = 05.01.2010)
  - Tabelle, die alle Datensätze beinhaltet, die durch eine Änderung beendet werden. (z.B. GUELTIG_BIS = 04.01.2010)
- Erstellen einer Tabelle, die die NKs der Datensätze beinhaltet, die durch eine Typ1-Änderung betroffen sind. Diese Tabelle beinhaltet zusätzlich die zu ändernden Spalten.
- Befüllen der erstellten Tabellen.
- Hinzufügen von lokalen Indizes und Constraints zu den Austauschtabellen.
- Erstellen von Statistiken zu den Austauschtabellen.

Vorbereiten der Daten (2)

Berichtstag 05.01.2010



Prozedur zum Vorbereiten der Daten (1)

```
prepare_large_scd2( i_dim_table                 => 'rep_dim_gesuch'
                  , i_dim_table_owner           => 'isis_report'
                  , i_src_table                 => 'etl_dim_gesuch'
                  , i_berichtstag               => 20100105
                  , i_col_nk_list               => 'geid'
                  , i_seq_dim_key               => 'dim_gesuch_seq'
                  , i_scd1_col_list             => 'loeschdatum,letzte_benachrichtigung'
                  , i_estimate_tab_stat_percent => 5
                  , i_aenderungsdatum_is_scd2   => 'N'
                  )
```

- i_dim_table: Name der Dimension.
- i_dim_table_owner: Schema, in dem sich die Dimension befindet.
- i_src_table: Name der Ausgangstabelle.
- i_berichtstag: Tag, auf den sich die Deltalieferung bezieht.
- i_col_nk_list: Liste aller Spalten, die zum Natual Key gehören.
- i_seq_dim_key: Name der Sequenz, die zum Befüllen des Dimension_Key‘s verwendet wird.

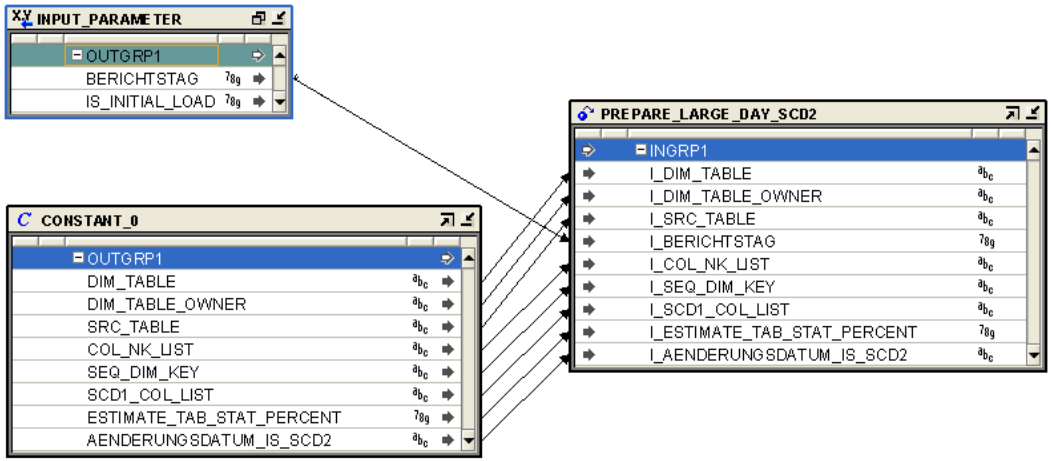
Prozedur zum Vorbereiten der Daten (2)

```
prepare_large_scd2( i_dim_table                 => 'rep_dim_gesuch'
                  , i_dim_table_owner           => 'isis_report'
                  , i_src_table                 => 'etl_dim_gesuch'
                  , i_berichtstag               => 20100105
                  , i_col_nk_list               => 'geid'
                  , i_seq_dim_key               => 'dim_gesuch_seq'
                  , i_scd1_col_list             => 'loeschdatum,letzte_benachrichtigung'
                  , i_estimate_tab_stat_percent => 5
                  , i_aenderungsdatum_is_scd2   => 'N'
                  )
```

- i_scd1_col_list: Liste der Spalten der Dimension, die vom Typ 1 sind.
- i_estimate_tab_stat_percent: Prozentualer Anteil der Zeilen, die Oracle zum Analysieren verwenden soll.
- i_aenderungsdatum_is_scd2: Hier wird festgelegt, ob das Änderungsdatum auch eine Typ2-Spalte ist.

Prozedur zum Vorbereiten der Daten (3)

- Die Prozedur wird im OWB als Transformation in einem Mapping aufgerufen.

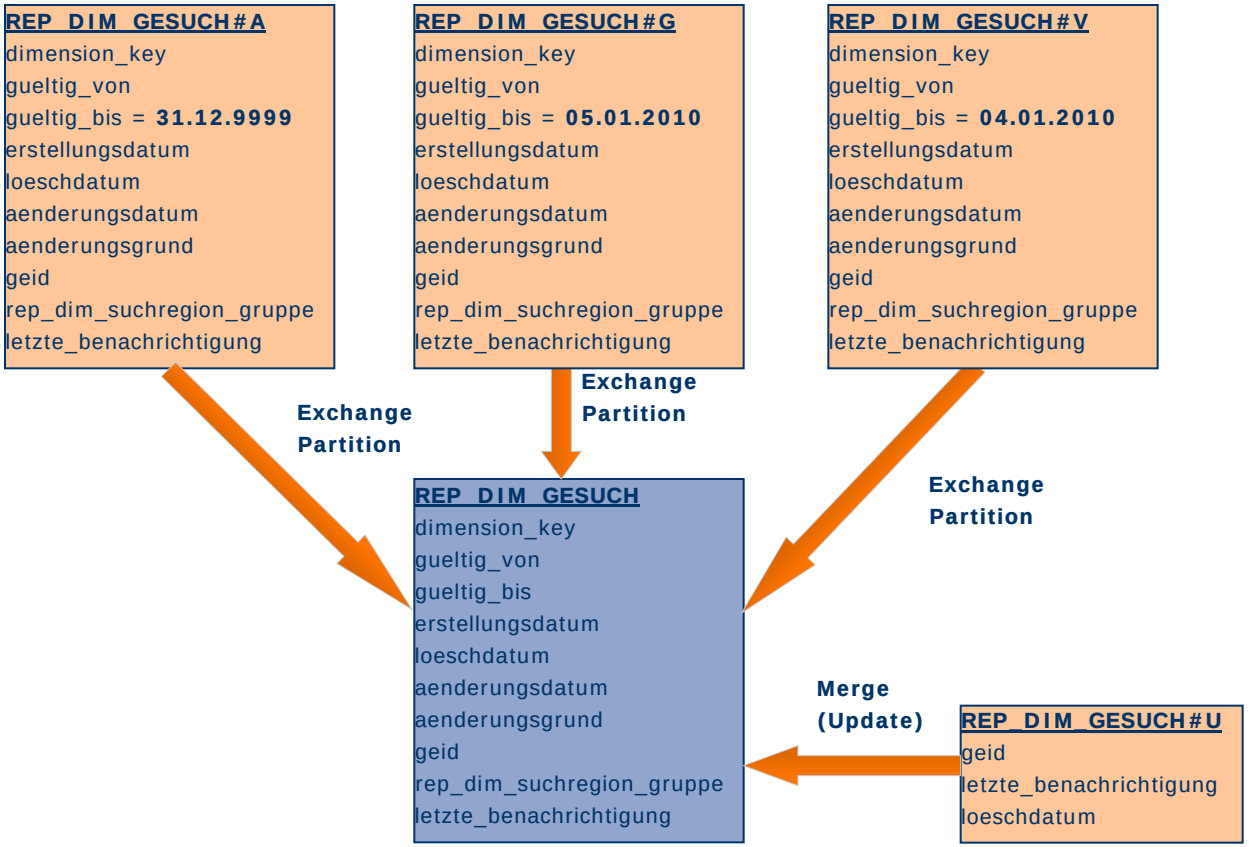


Laden der Daten (1)

- Der Status aller FKs, die die Dimension referenzieren und der Status aller globalen Constraints der Dimension wird auf DISABLED gesetzt, wenn globalen Constraints nicht beim Exchange Partition mitgepflegt werden. (update global indexes)
- Truncate der Partition für den Effektivtag. (zum Effektivtag gelöschte Datensätze)
- Truncate der Partition für den Effektivtag -1. (zum Vortag beendete Datensätze)
- Exchange Partition zwischen den Austauschtabellen und der Dimension.
- Update der SCD1-Felder, die sich geändert haben.
- Globale Indizes werden neu aufgebaut.
- Der Status aller FKs, die die Dimension referenzieren und der Status aller globalen Constraints der Dimension wird auf ENABLED gesetzt, wenn globalen Constraints nicht beim Exchange Partition mitgepflegt werden.

Laden der Daten (2)

Berichtstag 05.01.2010



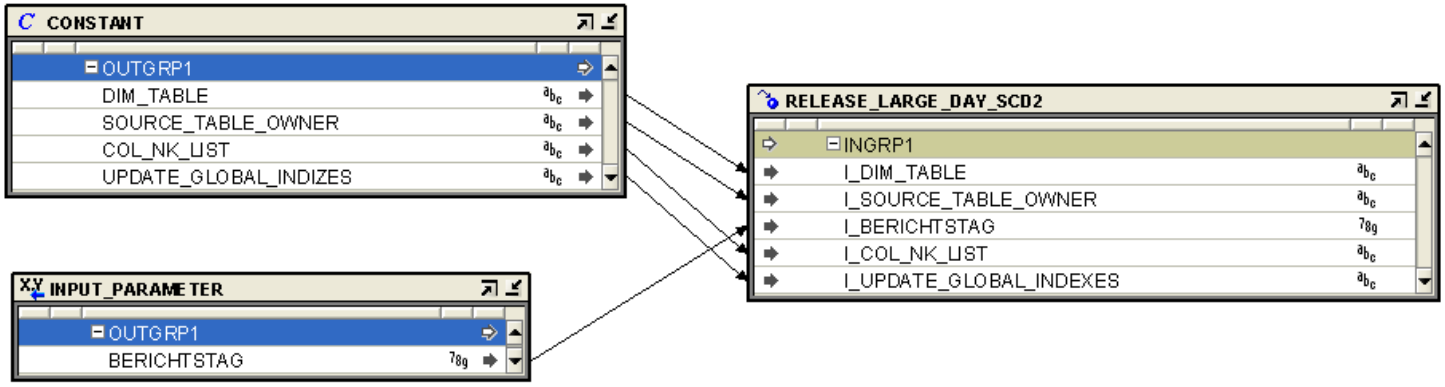
Prozedur zum Laden der Daten (1)

```
release_large_scd2( i_dim_table              => 'rep_dim_gesuch'
                  , i_source_table_owner     => 'isis_report'
                  , i_berichtstag            => 20100105
                  , i_col_nk_list            => 'geid'
                  , i_update_global_indexes  => 'N'
                  )
```

- i_dim_table: Name der Dimension.
- i_source_table_owner: Schema, in dem sich die Austauschtabelle befindet.
- i_berichtstag: Tag, auf den sich die Deltalieferung bezieht.
- i_col_nk_list: Liste aller Spalten, die zum Natual Key gehören.
- i_update_global_indexes: Beim exchange partition werden globale Indizes von Oracle mitgepflegt. Ansonsten von der Prozedur.

Prozedur zum Laden der Daten (2)

- Die Prozedur wird im OWB als Transformation in einem Mapping aufgerufen.



Kapitel 04

(1) Gewünschter Eigenschaften von SCD2s
(2) Befüllung per Merge
(3) Befüllung mit Exchange Partition
(4) Vergleich der beiden Methoden

Vergleich der beiden Methoden

- Befüllung per Merge:
  - Partitionierung nach GUELTIG_BIS ist nicht sinnvoll, da dazu ROW MOVEMENT eingeschaltet werden müsste.
  - Für kleine und mittelgroße Dimensionen geeignet.
- Befüllung mit Exchange Partition:
  - Overhead durch Abfrage des Data Dictionary.
  - Einfache Handhabung durch Standardisierung und kapseln der Logik in einem Package.
  - Für mittelgroße und große Dimensionen geeignet.

Fragen ?

Vielen Dank für Ihre Aufmerksamkeit!

Kontakt:
Immobilien Scout GmbH
Andreasstraße 10
10243 Berlin

Fon: +49 (0)30 243 01-1593
Email: dierk.bachmann@immobilienscout24.de
URL: www.immobilienscout24.de