# SAS High-Performance Computing Series

Teil 1 – In-Database Computing

20. SAS Club November 2010

Nicolas Adamek Christoph Wendl





### Im Zeitalter der Datenflut ...

### 1,2 Zettabyte

Weltweite **Datenaufkommen** 2010, entspricht der Menge an digitalen Informationen, die entsteht, wenn jeder Erdenbewohner 100 Jahre lang ununterbrochen *twittern* würde.

### 15 Petabyte

Menge der pro Tag **neu generierten Informationen**. Dies entspricht dem Achtfachen der Informationen in allen US-Bibliotheken.

### +3x

"the size of the largest data warehouse ... triples approximately every two years."

### 18 Months

Alle 18 Monate verdoppelt sich das Datenaufkommen weltweit.

### 1 Million

Kundentransaktionen bei Wal-Mart pro Stunde

### 2 Trillion (10<sup>12</sup>)

Frost & Sullivan's "2004 Healthcare Storage Report" predicts that by 2010, **medical centers** will need to be equipped to hold almost 1 billion terabytes of data, or the equivalent of **2 trillion file cabinets** worth of information.

### **Doppelt**

Laut IDC wächst die Menge an unstrukturierten Daten doppelt so schnell wie in Datenbanken gespeicherte strukturierte Daten





## ... brauchen wir High-Performance Lösungen, die damit umgehen können

- Was ist High-Performance Computing?
  - "Hochleistungsrechnen (englisch: high-performance computing

     HPC) ist ein Bereich des computergestützten Rechnens. Er
     umfasst alle Rechenarbeiten, deren Bearbeitung einer hohen
     Rechenleistung oder Speicherkapazität bedarf."

Quelle: de.wikipedia.org

- Einsatzmöglichkeiten von SAS in einer ganz neuen Dimension
  - Hoch-performante und skalierbare Durchführung analytisch komplexer und/oder Daten-intensiver Verarbeitungsaufträge
  - Extreme Reduzierung von Verarbeitungszeiten bis hin zur Anwendung komplexer Analyseverfahren in nahezu Echtzeit





## SAS High-Performance Computing Initiative

## SAS High-Performance Computing Lösungen

### SAS In-Database Computing

Von SAS
gesteuerte
Analysen und
Verarbeitungen
direkt in den
DatenbankSystemen

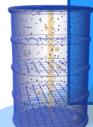
### SAS Grid Manager

Skalierbare
Lastverteilung
von SAS
Prozessen und
SAS Analysen in
einem Rechnerverbund

BURNOUS

### SAS In-Memory Analytics

Verteilung von komplexen SAS Analysen und deren Daten zur In-Memory Verarbeitung in einem Rechnerverbund



# SAS® In-Database Computing Was ist die SAS IDC Initiative?















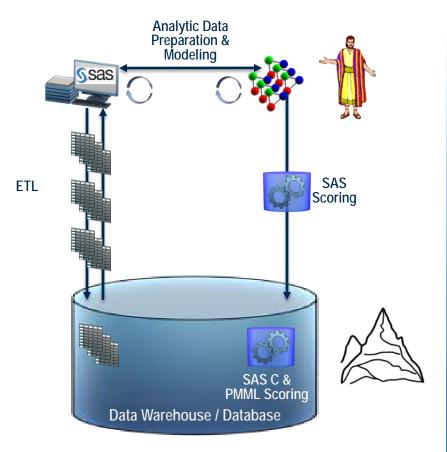
- Ziel der Initiative ist es SAS mit verschiedenen Datenbank-Systemen tiefer zu integrieren
- und damit SAS Verarbeitungen und Analytik im Zusammenspiel noch schneller zu machen, vor allem durch die Vermeidung unnötiger Datentransfers und des Ausnutzen hoch-performanter Datenbanken und Data Warehouse Appliances (MPP)
- Seit dem Beginn der SAS In-Database Computing Initiative 2007 wurde und wird das Spektrum kontinuierlich weiterentwickelt
- Derzeit arbeiten wir mit folgenden Hersteller-Partnern:
  - Teradata, Oracle, IBM, Netezza, AsterData, HP und Greenplum



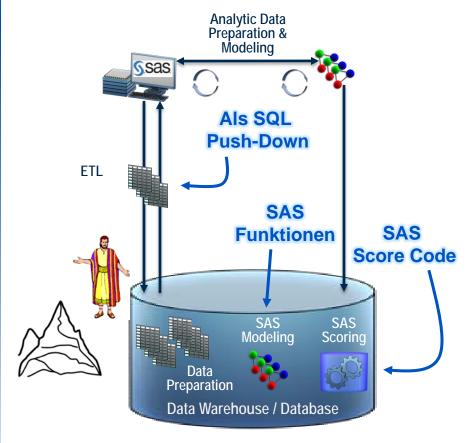


# SAS® In-Database Computing *Architektur-Vergleich – was haben wir implementiert?*

### **Traditional Architecture**



### In-Database Architecture







## SAS® In-Database Computing Technologien

#### **Data Integration**

- 1
- SAS Data Integration: Engere Integration mit Datenbanken durch umfangreiche SQL Push-Down Funktionen und eigenen Transformationen für ausgewählte Datenbank-Systeme
- SAS/ACCESS: Stellt native Verbindungen zwischen SAS und vielen Datenbanken-Systemen zur Verfügung, und bietet damit implizites und explizites SQL Pass-Through



### In-Database Computing in SAS® Data Integration Release Status: SAS® 9.2 M3

- SAS/ACCESS® 9.2 für Relational Databases
  - Optimierung von (PROC) SQL Verarbeitung durch implizites Pass-Through (IPT) (Push-Down) in das datenhaltende Datenbank-System und Reduktion der Datenübertragung:
    - 1. SQL Aggregations-Funktionen (MIN, MAX, AVG, MEAN, FREQ, N, SUM, und COUNT) werden durch das ANSI SQL Aquivalent im Datenbank-System ausgeführt
    - 2. Ausgewählte SAS Funktionen werden in Datenbank-Aquivalente übersetzt. Die Liste der Funktionen ist Datenbank-spezifisch, z.B. für Oracle: ABS, ARCOS (ACOS), ARSIN (ASIN), ATAN, CEIL, COS, COSH, DATEPART, DATETIME (SYSDATE), DTEXTDAY, DTEXTMONTH, TEXTYEAR, EXP, FLOOR, LOG, LOG10, LOG2, LOWCASE (LCASE), SIGN, SIN, SINH, SOUNDEX, SQRT, STRIP (TRIM), TAN, TRANSLATE, TRIMN (RTRIM), UPCASE (UPPER), COALESCE\*, DAY (EXTRACT)\*, MONTH (EXTRACT)\*, YEAR (EXTRACT)\*



# In-Database Computing in SAS® Data Integration

- SAS/ACCESS® 9.2 für Relational Databases
  - Optimierung von (PROC) SQL Verarbeitung durch implizites
     Pass-Through (IPT) (Push-Down) in das datenhaltende
     Datenbank-System und Reduktion der Datenübertragung:
    - 3. IPT Funktions-Katalog (SQL Dictionary) ist erweiterbar
    - 4. Durchreichen von **Table-Joins**, **WHERE-Bedingung** und **UNION/DISTINCT** in das Datenbank-System, sofern möglich (abhängig vom jeweiligen Datenbank-System)
  - Wenn IPT nicht oder nur teilweise möglich, dann werden die Daten für die fehlenden "Operationen" zum SAS Server übertragen
  - Optimierung der Datenbank-Interaktion durch explizites Pass-Through (EPT) von Datenbank-Statements innerhalb von SAS Programmen mittels:

PROC SQL ... CONNECT TO ... EXECUTE



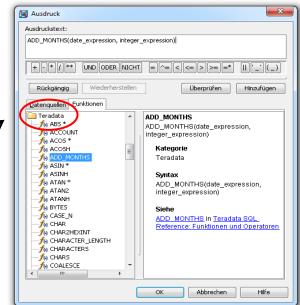


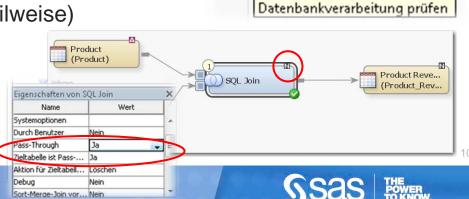
# In-Database Computing in SAS® Data Integration

- Mit SAS® Data Integration Studio:
  - (PROC) SQL basierte Transformationen nutzen IPT Funktionalität der jeweiligen SAS/ACCESS® Engine
  - Teradata Funktionen stehen im Expression Builder zur Verfügung
    - » Expression Builder ist auch erweiterbar!
  - Entwickler-Unterstützung durch visuelle Darstellung von In-Database Computing Schritten vor der Durchführung
  - In-Database fähige Transformationen
    - » Extract, SQL Join, SCD (teilweise)

Copyright © 2010, SAS Institute Inc. All rights reserved.

- » Table Loader, Teradata
   Table Loader
- EPT mit User-Written Code und Transformationen







# In-Database Computing in SAS® Data Integration - Anwendungsfall

Damit können DWH-Beladungsstrecken komplett als **ELT** Prozesse **mit SAS** direkt im Datenbank-System ablaufen.

Extract

- Nur mehr benötigte Daten extrahieren und weiterverarbeiten
- Sehr oft befinden sich diese im selben Datenbank-System

Load

- Rohdaten direkt vom Vorsystem in der Ziel-Datenbank ablegen und direkt dort transformieren
- Nutzen High-Speed Bulk-Loader der Datenbank um diese Daten zu laden

**Transform** 

- Transformieren der Daten mit optimierten und hochperformanten Datenbanken
- Verwenden von Datenbank-spezifischem SQL und Benutzerfunktionen



## SAS In-Database Computing Technologien

#### **Data Integration**

1

- SAS Data Integration: Engere Integration mit Datenbanken durch umfangreiche SQL Push-Down Funktionen und eigenen Transformationen für ausgewählte Datenbank-Systeme
- SAS/ACCESS: Stellt native Verbindungen zwischen SAS und vielen Datenbanken-Systemen zur Verfügung, und bietet damit implizites und explizites SQL Pass-Through

#### **Foundation & Analytics**

2

 Base SAS + Module: Ausgewählte SAS Prozeduren (PROC) werden in SQL "umgeschrieben" und der Datenbank zur Verarbeitung übergeben



### 1

# In-Database Computing in der SAS® Foundation Release

- Release Status: SAS® 9.2 M3
- In der aktuellen Version von SAS® 9.2 stehen folgende Base SAS Prozeduren als In-Database PROCEDURES zur Verfügung
  - PROC FREQ, PROC MEANS, PROC SUMMARY, PROC TABULATE
  - PROC REPORT, PROC SORT, PROC RANK
- Generieren zur Laufzeit Datenbank-SQL für die gesamte oder Teile der Prozedur-Verarbeitung
- Voraussetzung ist die jeweilige SAS/ACCESS® Engine als Verbindungsmechanismus zwischen dem SAS Server und der Datenbank
- Als Datenbank-System werden derzeit Teradata, DB2 (UNIX und PC Hosts) und Oracle unterstützt
- Aktiviert werden die In-Database PROCEDURES durch die System (oder LIBNAME) Option SQLGENERATION





## In-Database Computing in der SAS® Foundation Was passiert unter der Motorhaube?

- 1 OPTIONS **SQLGENERATION**=none;
- 2 LIBNAME ORA ORACLE path=xe USER='sasdemo' PASSWORD='{sas001}U0FTcHcx';
- 3 PROC FREQ DATA=ora.cars;

TABLE Origin;

RUN;

Die Prozedur FREQ

		ORIGIN			
ORACLE_2: Prepared: on connection 0 SELECT "ORIGIN" FROM CARS		Häufigkeit	Prozent		Kumulativer Prozentwert
NOTE: Es wurden <b>428</b> Beobachtungen gelesen aus Datei ORA.CARS.		158	36.92	158	36.92
		123	28.74	281	65.65
	USA	147	34.35	428	100.00



# In-Database Computing in der SAS® Foundation Was passiert unter der Motorhaube?

- 1 OPTIONS **SQLGENERATION=DBMS**;
- 2 LIBNAME ORA ORACLE path=xe USER='sasdemo' PASSWORD='{sas001}U0FTcHcx';
- 3 PROC FREQ DATA=ora.cars; TABLE Origin;

RUN;

Die Prozedur FREQ

ORACLE\_3: Prepared: on connection 0
select COUNT(\*) as ZSQL1, case when COUNT(\*) >
COUNT(TXT\_1."ORIGIN") then ' ' else
MIN(TXT\_1."ORIGIN") end as ZSQL2 from CARS
TXT 1 group by TXT 1."ORIGIN"

ACCESS ENGINE: SQL-Anweisung wurde an das DBMS zum Abrufen von Daten übergeben.

NOTE: SQL-Generierung wird zur Erstellung von Häufigkeits- und Kreuztabellen verwendet.

#### ORIGIN

Häufigkeit	Prozent		Kumulativer Prozentwert
158	36.92	158	36.92
123	28.74	281	65.65
147	34.35	428	100.00

ZSQL1	ZSQL2		
147	USA		
123	Europe		
158	Asia		







# In-Database Computing in der SAS® Release Status: 1.2 mit SAS® 9.2 M3

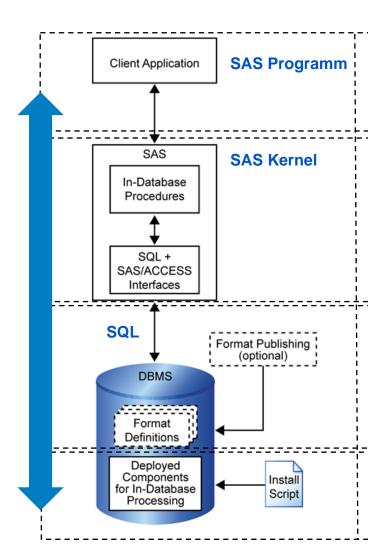
- SAS<sup>®</sup> Analytics Accelerator for Teradata
  - Folgende Base SAS®, SAS/STAT® und SAS/ETS® Prozeduren stehen zur teilweisen Verarbeitung innerhalb eines Teradata Systems zur Verfügung:
    - » PROC CORR (Base SAS), PROC CANCORR, PROC FACTOR, PROC PRINCOMP, PROC REG, PROC SCORE, PROC VARCLUS, PROC TIMESERIES (SAS/ETS)
  - Keine eigene PROC Implementierung (selbe Syntax), sondern die bestehenden PROCs wurden erweitert:
    - » Generieren dynamisch SQL Code, der bei der Ausführung in Teradata, die …
    - » ... von SAS bereitgestellten UDFs (User-Defined Functions) durchführt, z.B. SAS\_ZACORR zur Ermittlung einer Sum-of-Squaresand-CrossProducts (SSCP) Matrix
  - Voraussetzungen:
    - » SAS® 9.2: Base SAS, SAS/ACCESS für Teradata, SAS/STAT® und SAS/ETS®
    - » Teradata V13 (Linux)





# In-Database Computing in der SAS® Foundation Release

Release Status: SAS® 9.2 M3



- Ihr SAS Programm ... in <u>ihrem</u> **SAS Client**(z.B. SAS® Display Manager, SAS® Enterprise Guide, SAS® Data Integration Studio, SAS® BI Clients)
- SQLGENERATION (Libname | System) aktiviert In-Database Prozeduren
- Base SAS, SAS/STAT® und SAS/ETS® PROC Kernel-Statements erzeugen SQL
- SAS/ACCESS Interface für DB-Verbindung und Ausführung
- Verwenden (implizit oder explizit) eines SAS
   Format-Katalogs in der Datenbank mit
   SAS\_PUT()
- Derzeit verfügbar für die SAS/ACCESS<sup>®</sup>
   Engines Teradata & Netezza
- SAS® Analytics Accelerator for Teradata User-Defined Function (z.B. SAS\_ZACORR)



## SAS In-Database Computing Technologien

#### **Data Integration**

- 1
- SAS Data Integration: Engere Integration mit Datenbanken durch umfangreiche SQL Push-Down Funktionen und eigenen Transformationen für ausgewählte Datenbank-Systeme
- SAS/ACCESS: Stellt native Verbindungen zwischen SAS und vielen Datenbanken-Systemen zur Verfügung, und bietet damit implizites und explizites SQL Pass-Through

#### **Foundation & Analytics**

2

 Base SAS + Module: Ausgewählte SAS Prozeduren (PROC) werden in SQL "umgeschrieben" und der Datenbank zur Verarbeitung übergeben

#### **Data Mining**

3

• SAS Scoring Accelerator: Erweiterung zu SAS Enterprise Miner, um Scoring Modelle innerhalb von Datenbank-Systemen zu registrieren und durchzuführen

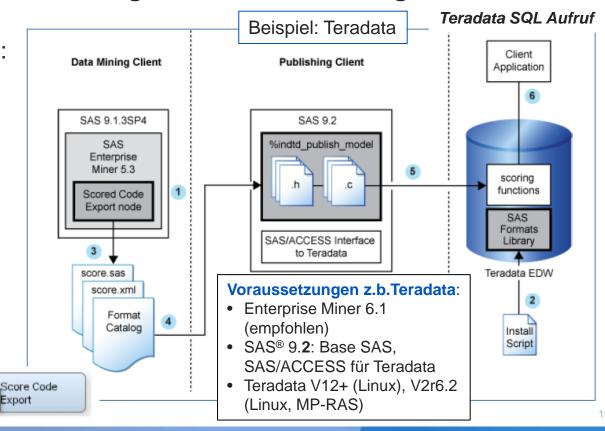




## In-Database Computing in SAS Analytics

### SAS® Scoring Accelerator:

- Erweiterung für SAS® Enterprise Miner®, um Scoring-Modelle direkt im Datenbank-System auszuführen – Score Code Export Knoten
- Integriert mit SAS® Model Manager mittels Model Packages
- Verfügbar für die Datenbank-Systeme:
  - » DB2
  - » Netezza
  - » Teradata
- Unterstützt die Übergabe von Modellen an DBA
- Reduziert Fehler-Anfälligkeit in der Übergabe
- Beschleunigt den Prozess





## SAS In-Database Computing Technologien

#### **Data Integration**

- 1
- SAS Data Integration: Engere Integration mit Datenbanken durch umfangreiche SQL Push-Down Funktionen und eigenen Transformationen für ausgewählte Datenbank-Systeme
- SAS/ACCESS: Stellt native Verbindungen zwischen SAS und vielen Datenbanken-Systemen zur Verfügung, und bietet damit implizites und explizites SQL Pass-Through

#### **Foundation & Analytics**

2

 Base SAS + Module: Ausgewählte SAS Prozeduren (PROC) werden in SQL "umgeschrieben" und der Datenbank zur Verarbeitung übergeben

#### **Data Mining**

3

• SAS Scoring Accelerator: Erweiterung zu SAS Enterprise Miner, um Scoring Modelle innerhalb von Datenbank-Systemen zu registrieren und durchzuführen

#### **Solutions**

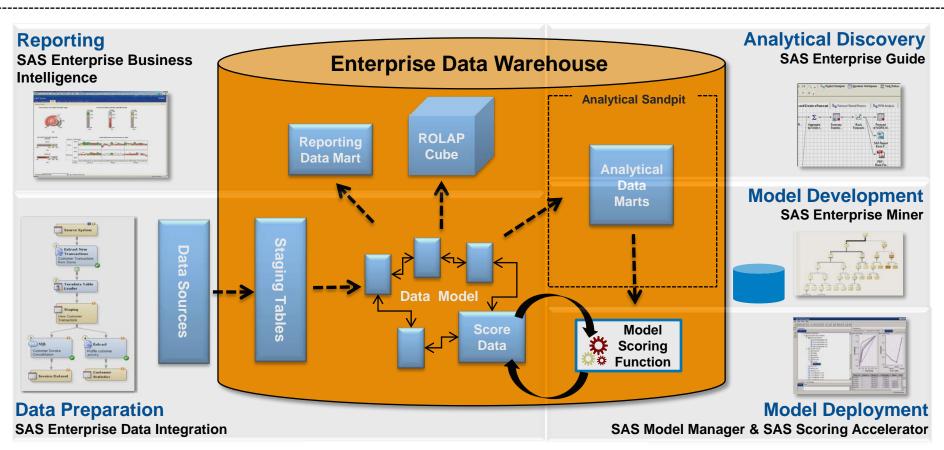


- z.B. SAS Anti-Money Laundering: Durchführen von Regeln (Szenarios) direkt in einem Teradata System
- SAS Industry Solutions: Die DDS Datenmodelle der Lösungen können in verschiedenen Datenbank-Systemen eingesetzt werden



## **SAS In-Database Computing**

### End-to-End In-Database Verarbeitung mit der SAS Plattform



**SAS Platform for Business Analytics** 

SAS Metadata Server





# Motivation für SAS In-Database Computing Wieso ist In-Database *Ihr Thema*?

### Einsparungen

- Es "wandern" weniger Daten zwischen der Datenbank & SAS:
  - Reduktion der Datenmengen, die über ein Netzwerk übertragen werden muss
  - Weniger Daten-Redundanzen zwischen DWH und Analyse-System
- Geringerer Storage-Bedarf für das SAS System
- Bessere Ausnutzung bereits vorhandener High-Performance Datenbank- und Data Warehouse Technologien

### Akzeptanz

- Datenbank-Systeme sind in der Regel bereits vorhanden, etabliert und sehr oft quasi schon "bezahlt"
- Sie sind bereits Bestandteil von IT Management Prozessen (Operations)
- Es kann oft auf umfangreiches Know How zurückgegriffen werden
- Fachlogik wird dabei weiterhin sicher und zentral verwaltet (Compliance & Governance)
- Geringeres Sicherheitsrisiko (Daten "wandern" weniger)





## Wie starte ich mit SAS In-Database Computing in meinem Unternehmen?

- Klären Sie das Umfeld ...
  - Was für eine SAS Anwendung haben Sie? Wie nutzen Sie SAS?
  - Welche SAS Software ist im Einsatz? Wo befinden sich ihre Daten?
  - Welche Datenbank-Systeme sind im Einsatz? Gibt es bereits einen Zugriff auf diese Datenbank-Systeme?
  - Wer ist dafür zuständig (DBA)?
- Identifizieren Sie mögliche In-Database Verarbeitungen
  - Welche SAS Prozesse lassen sich in das Datenbank-System verlagern?
  - Welche Voraussetzungen müssen erfüllt sein?
- ... und dann "Probieren Sie es aus"! 3.
- Nutzen Sie unser Informationsangebot:
  - SAS Support: <u>SAS In-Database Technology</u>, SAS/ACCESS Produktdokumentation
  - Hilfe? sprechen Sie mit SAS Professional Services ...





## SAS In-Database Computing in der Praxis

 Christoph Wendl – SAS Consultant Professional Services – zeigt Ihnen...



- Beispiele:
  - ELT mit SAS® Data Integration Studio und SAS/ACCESS® Interface für Oracle
    - » Transformieren und Verknüpfen von Daten direkt in einer Oracle Datenbank
  - In-Database Prozedur PROC MEANS in SAS® Enterprise Guide und SAS/ACCESS® für Oracle
    - » Datenanalyse aus der Sicht eines Endbenutzers mit SAS Enterprise Guide und direkter Verarbeitung (intransparent) im Oracle Datenbank-System



# Beispiel 1 – ELT - SQL Join direkt in der Datenbank (Oracle)

```
proc sql;
   connect to ORACLE
       PATH=xe USER=sasdemo
                              PASSWORD="{SAS002}97ECE44B0F19E0DE0F1E0782"
   execute
      insert into sasdemo.ORDER FACT QUANTITY COUNT
      select
         count (BI ORDER STAGING.ITEM QTY) as QUANTITY COUNT,
         BI STORE DIM.STORE ID,
         BI STORE DIM.RETAIL OUTLET NAME
      from
         sasdemo.BI ORDER STAGING BI ORDER STAGING,
         sasdemo.BI STORE DIM BI STORE DIM
      where
         BI_ORDER_STAGING.STORE_ID = BI_STORE_DIM.STORE_ID
      group by
         BI STORE DIM.STORE ID,
         BI STORE DIM.RETAIL OUTLET NAME
   ) by ORACLE;
   %rcSet(&sqlrc);
   disconnect from ORACLE;
quit;
```



# Beispiel 2 – SQL Join: Upload von SAS Dataset in die Datenbank (Oracle)

```
LIBNAME ORA ORACLE PATH=xe SCHEMA=sasdemo USER=sasdemo PASSWORD="{SAS002}97ECE44B0F19E0DE0F1E0782";
proc datasets lib = ORA nolist nowarn memtype = (data view);
     delete BI STORE DIM SAS;
  quit;
  /*--- Create a new table ----*/
  data ORA.BI STORE DIM SAS
           (dbnull = (store id = YES
                     retail outlet name = YES
                     retail outlet format cd = YES
                     retail outlet type cd = YES
                     store size = YES
                      effective from dttm = YES
                     effective to dttm = YES
                     processed dttm = YES));
     attrib store id length = 8 format = 12. informat = 12. label = 'Retail Outlet Location Key';
     attrib retail outlet name length = $40 label = 'Retail Outlet Name';
     attrib retail outlet format cd length = $32 label = 'Retail Outlet Format Code';
     attrib retail outlet type cd length = $32 label = 'Retail Outlet Type Code';
     attrib store size length = 8 format = NLNUM10.2 informat = NLNUM10.2 label = 'Total Size';
     attrib effective from dttm length = 8 format = DATETIME20. informat = DATETIME20.
                label = 'Effective From Datetime':
      attrib effective to dttm length = 8 format = DATETIME20. informat = DATETIME20.
               label = 'Effective To Datetime';
     attrib processed dttm length = 8 format = DATETIME20. informat = DATETIME20.
               label = 'Processed Datetime';
run:
```

```
proc append base = ORA.BI_STORE_DIM_SAS
  data = mysas.BI_STORE_DIM_SAS force;
run;
```



# Beispiel 2 – SQL Join: Ausführen des Joins nach Upload in die Datenbank

```
proc sql;
   connect to ORACLE
       PATH=xe USER=sasdemo PASSWORD="{SAS002}97ECE44B0F19E0DE0F1E0782"
   );
   execute
      insert into sasdemo.ORDER FACT QUANTITY COUNT
      select
         count (BI ORDER STAGING.ITEM QTY) as QUANTITY COUNT,
         BI STORE DIM SAS.store id,
         BI STORE DIM SAS.retail outlet name
         sasdemo.BI STORE DIM SAS BI STORE DIM SAS,
         sasdemo.BI ORDER STAGING BI ORDER STAGING
      where
         BI STORE DIM SAS.store id = BI ORDER STAGING.STORE ID
      group by
         BI STORE DIM SAS.store id,
         BI STORE DIM SAS.retail outlet name
   ) by ORACLE;
   %rcSet(&sqlrc);
   disconnect from ORACLE;
quit;
```





### Beispiel 3 – PROC MEANS

- Project Log
- LIBNAME Option: SQLGENERATION=DBMS





## Tipps & Tricks (1)

#### Allgemeine Tips:

- SASTRACE statement um den SQL Code, der an die Datenbank geschickt wird zu sehen: options sastrace=',,,d' sastraceloc=saslog nostsuffix;
- > SASTRACE statement um die Durchführungszeit der einzelnen Schritte zu sehen: options options sastrace=',,,ts' sastraceloc=saslog nostsuffix;
- Funktionen verwenden, die die Datenbank verarbeiten kann (z. Bsp.: TRIM kann nicht in Oracle verarbeitet werden, TRIMN schon), SAS/ACCESS Dokumentation überprüfen!
- Primary Index: SAS-Datasets brauchen keinen, manche DBMS aber schon (Beispiel Teradata zwecks Verteilung über die Knoten)
- System Option DBIDIRECTEXEC, leitet alle Statements die möglich sind an die Datenbank weiter
- > SAS Makro statt einer SAS Funktion, um IPT zu gewährleisten:

```
proc sql;
    select account, ship_date from dbms.neworders where ship_date < today() + 7;
    quit;

$let target_date=%eval(%sysfunc(today()) + 7);

proc sql;
    select account, ship_date from dbms.neworders where ship_date < &target_date;
    quit;</pre>
```





## Tipps & Tricks (2)

#### SAS Data Integration Studio:

- ➤ ELT: Output Tabellen direkt in die Datenbank umleiten, Property kann auf der Tabelle, der Transformation, dem Job oder als globale Option gesetzt werden
- Check Database Processing

#### Mögliche Ursachen, wenn kein impliziter Pass-Through gelingt:

- Zu einem Datenbankfeld fehlt im DBMS das SAS-Format
- ➤ Die gewählte SAS Funktion gibt es nicht (in der Form) im DBMS
- ➤ Es wird eine Transformation genutzt, die kein SQL produziert (z.B. PROC Transpose)
- ➤ Es wird auf eine Tabelle referenziert, die nicht im DBMS liegt (gerne auch eine WORK-Tabelle, hinter den Kulissen…)







Vielen Dank für Ihre Aufmerksamkeit

Haben Sie Fragen?