

# SAS DI/DM Roadmap 2011

---

Thomas Rocks  
Competence Center Data Integration  
SAS Deutschland



THE  
POWER  
TO KNOW®

# Agenda

---

- Data Integration Studio 4.3
- Data Management Platform 2.1
- Ausblick/Roadmap DI
- Modernisierung: Mehrwerte für SAS Base Kunden



THE  
POWER  
TO KNOW®

**SAS Data Integration Studio  
wird nicht durch  
Dataflux Data Management Studio  
ersetzt!**

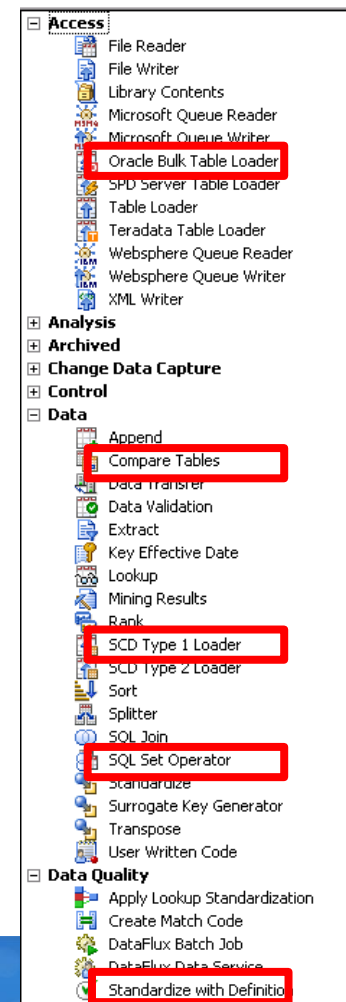
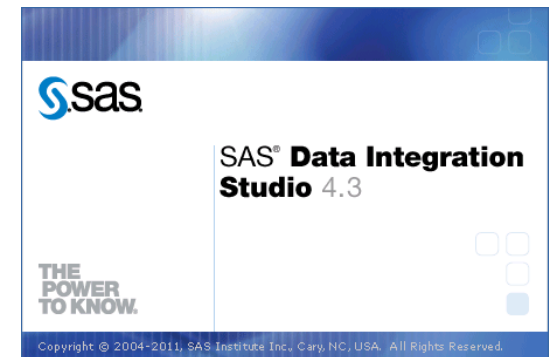
# SAS Data Integration Studio 4.3

## ■ Bedienerfreundlichkeit & Handhabung

- Versionierung
- Jobstatus und Performance Reporting
- Erweiterung des SAS Code Importers
- Erweiterte Metadatensuche
- Source Code Deployment im Batch
- Erweiterter Grid Support

## ■ Transformationen

- ELT & DBMS Pushdown Support
  - SQL Union
  - Expression Builder
    - » Import von User Defined Functions (UDF)
    - » Erweiterte DBMS Unterstützung
  - Oracle Bulk Table Loader
- Tabellen Vergleich um Änderungen zu erkennen
- Type 1 SCD
- Neue Datenquellen:
  - Greenplum
  - Aster Data
  - Sybase IQ



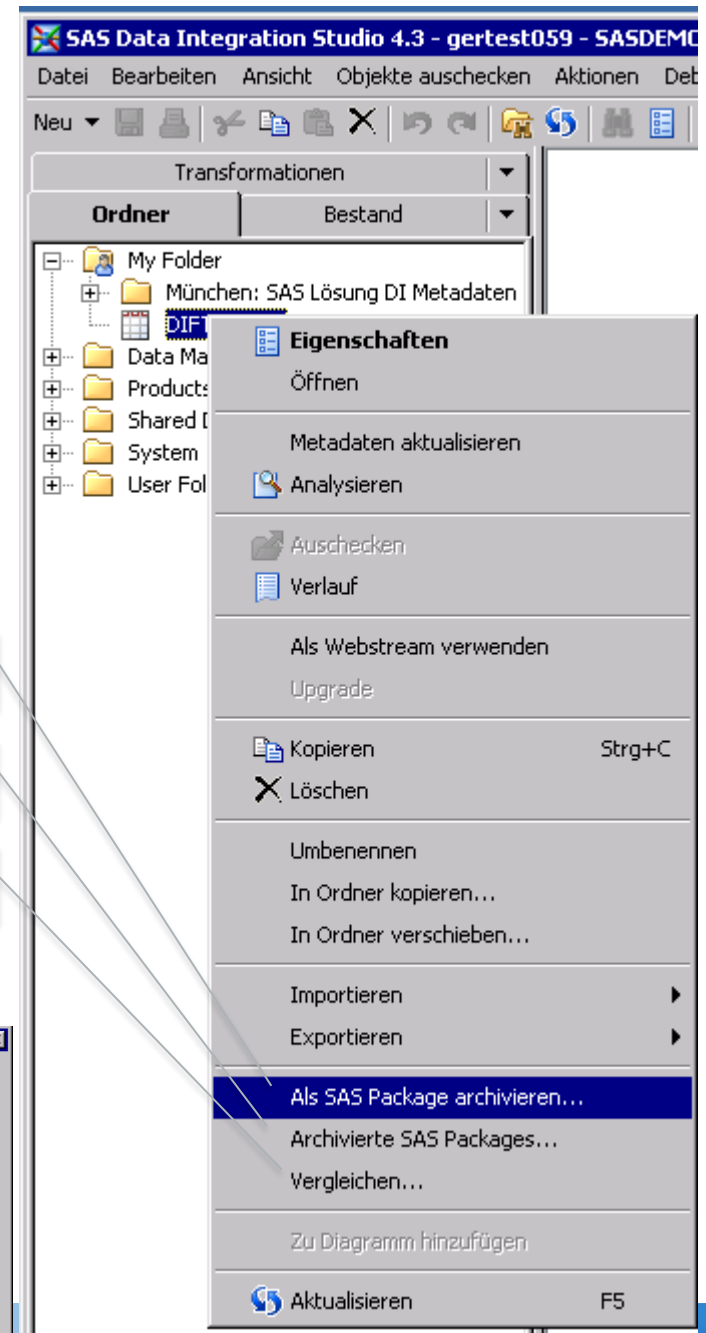
# Versionierung

- 3rd Party Source Code Management Systeme:
  - CVS
  - SubVersion
  - Erweiterbar für weitere Produkte durch offen API

Archivieren

Navigieren

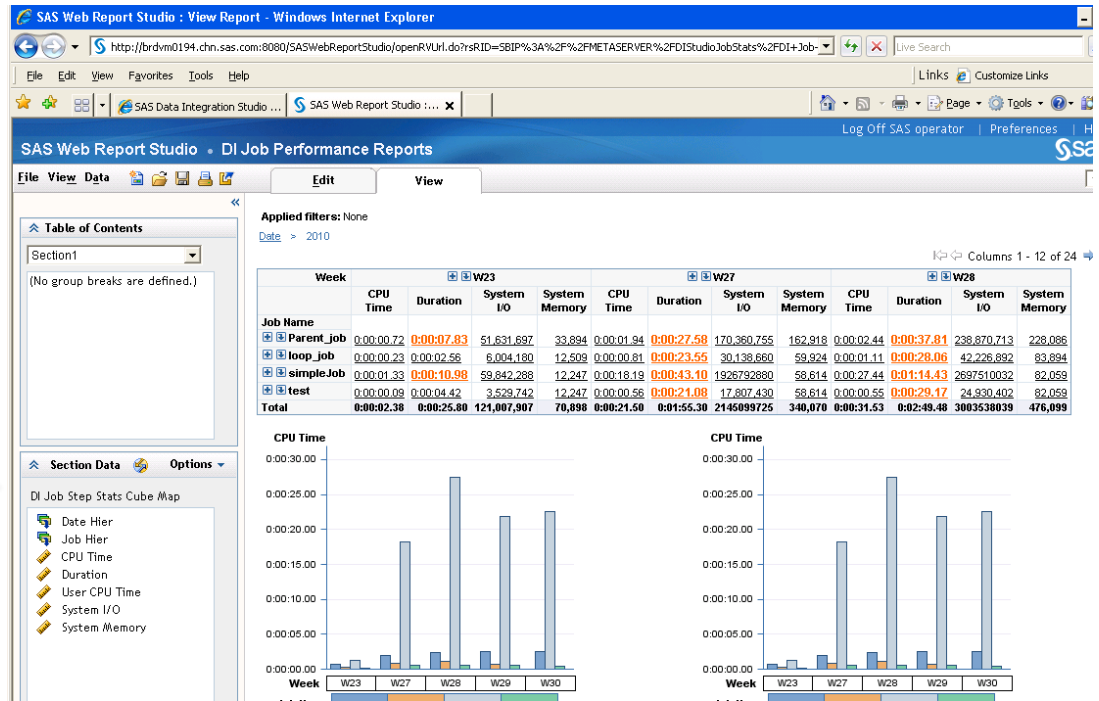
Vergleichen



# Performance Monitoring

- Web Report Studio
- Stored Processes

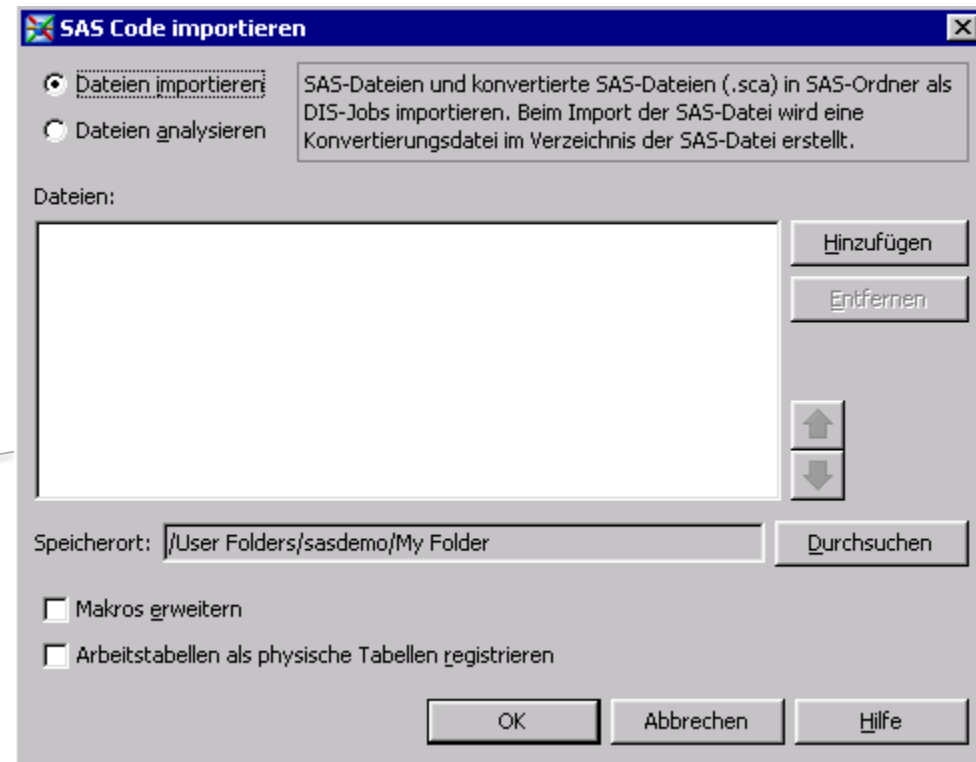
Reports für  
Batch Processing



# SAS Code Importer

- Macro Support
- Analyse Modus
- Kollisionserkennung (Library-Konflikte)
- Erweiterte Protokollierung
- Registrierung von Worktables als Tabellen

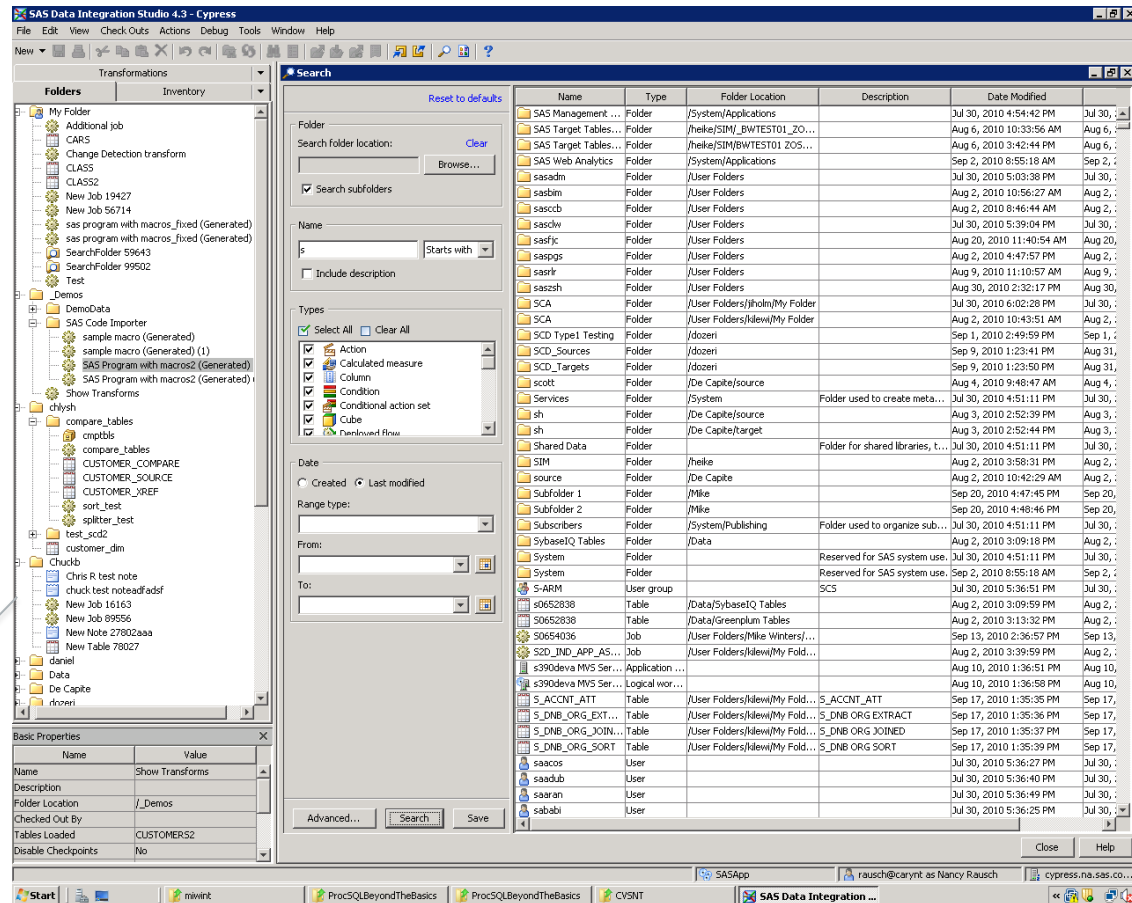
Modernisierung  
Von bestehenden  
BASE SAS  
Programmen.



# Integrierte Metadatenuche

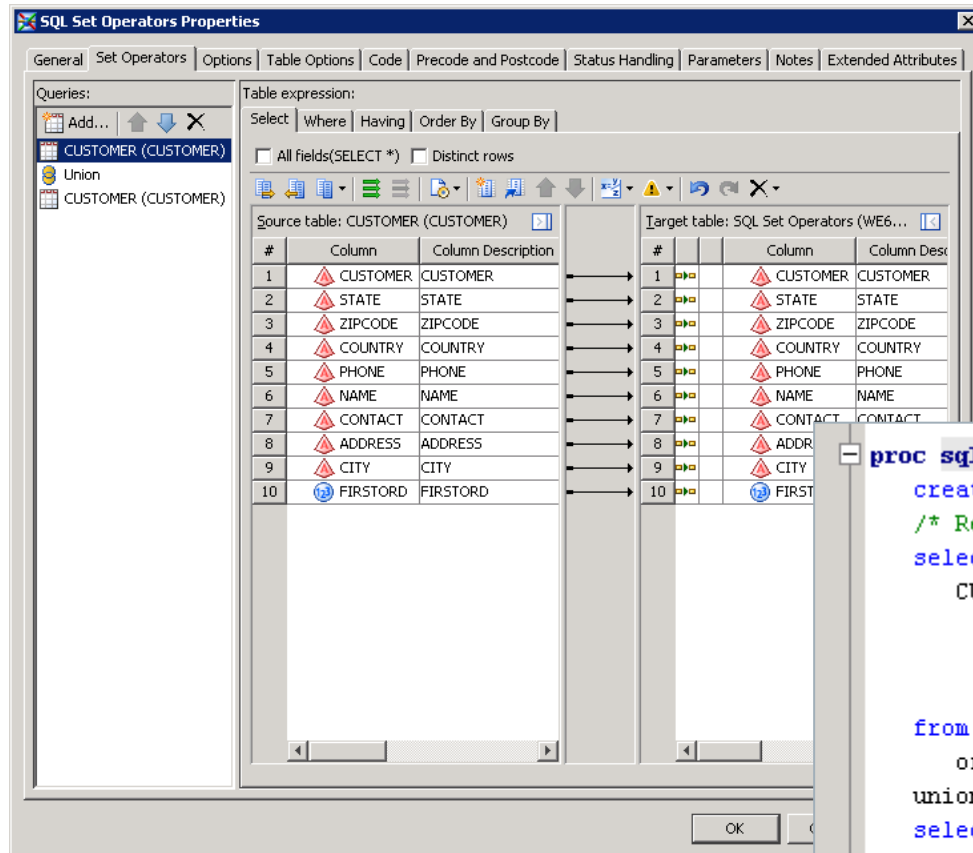
- Suchen von Metadatenobjekten
- Komplexe Filtermöglichkeiten
- Wiederverwendbarkeit durch Speicherungsmöglichkeit

Metadaten-Analyse  
Nicht nur für  
Experten!





# Erweiterte ELT Unterstützung

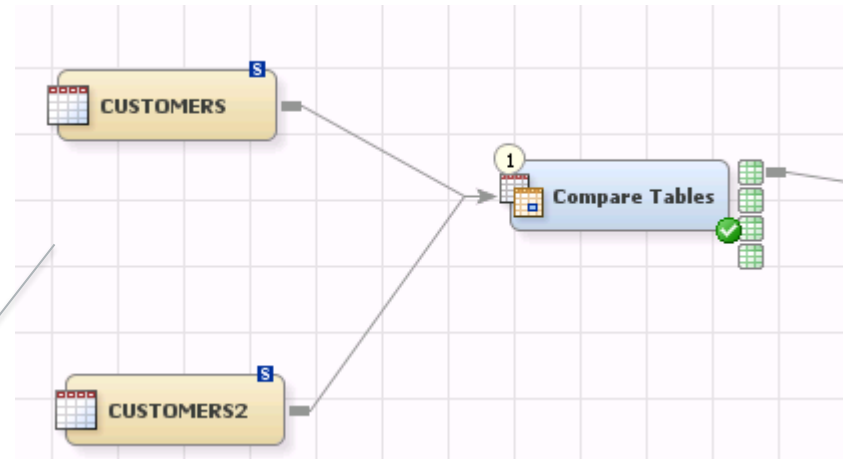


- SQL Union Transformation:
  - UNION
  - UNION ALL
  - INTERSECT
- Optimiert für DBMS

```
= proc sql;  
  create table orabk.CUSTOMER1 as  
  /* Returns all unique rows from the two query results. */  
  select  
    CUSTOMER.CUSTOMER length = 8  
    format = $8.  
    informat = $8.  
    label = 'CUSTOMER'  
  from  
    orabk.CUSTOMER  
  union all  
  select  
    "" as CUSTOMER length = 8  
    format = $8.  
    informat = $8.  
    label = 'CUSTOMER'  
  from  
    orabk.CUSTOMER  
;  
quit;
```

# Compare Tables Transformation

- Vergleichen von 2 Tabellen auf Unterschiede
- Performance-Optimierung durch Nutzen von Hashing



Effiziente Ermittlung  
von Änderungen in  
Tabellen

# SCD Type 1 Transformation

- Aktualisierung ohne Historisierung

Effiziente Übernahme  
von Änderungen

**SCD Type 1 Properties**

General | **Keys** | Mappings | Options | Table Options | Code | Precode and Postcode | Parameters | Notes | Extended Attributes

**Natural key**

Select the match keys from the source table and map them to the comparison table columns if needed.

Available source columns:

Source Column
CUSTNUM
CUSTNAME
CUSTCITY

Selected source column mappings:

Source Column	Target Column
CUSTNUM	CUSTNUM
CUSTNAME	CUSTNAME
CUSTCITY	CUSTCITY

**Surrogate key**

Select the target table column where surrogate keys will be generated. The new record expression is used to generate values for new records. The macro variable "NewMatchKey" is used to hold the maximum key value.

☐ Generate a Surrogate Key

Surrogate key column:

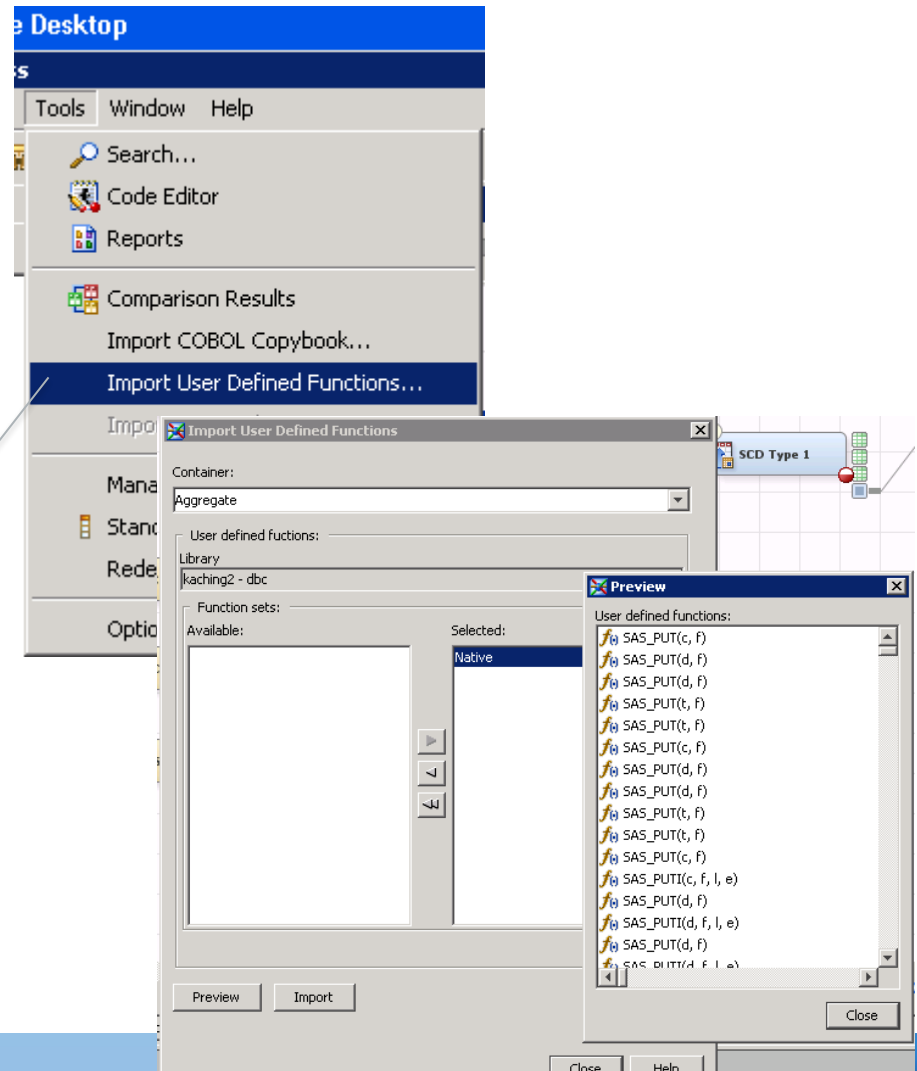
New record expression:

Define Max Key

# Importieren von DBMS User Defined Functions

- Importieren und Nutzen im Expression Builder
- Unterstützt Standard DBMS UDFs und Enterprise Miner publishing Mining Analytics UDFs
- Parametererkennung

Erweiterte  
Unterstützung von  
DBMS Pushdown



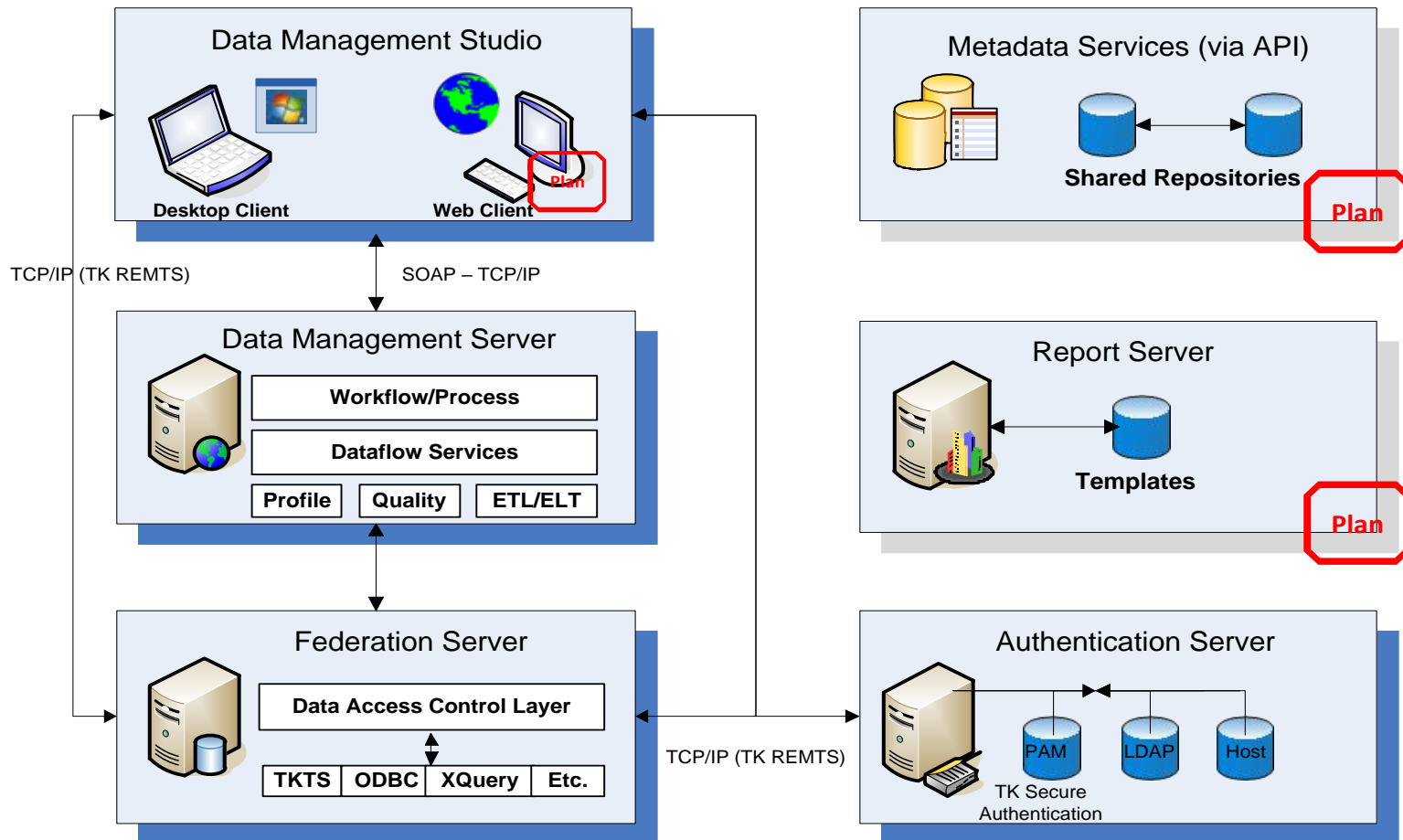
# Was ist Data Management?

Data Management ist die von Analysten geforderte Zusammenführung von

- Data Quality
- Data Integration
- Master Data Management
- Metadata Management
- Enterprise Data Connectivity & Federation

# Die Data Management Platform - Architektur

Data Management Platform - Architektur



# Data Management Studio – Die Steuerzentrale

Informationen in einer Oberfläche

The screenshot shows the Data Management Studio 2.1 interface. The left sidebar contains a navigation pane with the following items: Information, Data Connections, Collections, Folders, Data Management Servers, and Administration. The main window is divided into several panes. The top pane is the 'Overview' pane, which displays a list of recent files and a section for the 'DataFlux Community of Experts' with various blog posts. The bottom pane is the 'Default Settings' pane, which displays configuration information for the host, license, authentication, and application. Overlaid on the screenshot are several grey boxes with white text, each pointing to a specific feature or section of the interface. These boxes are labeled: 'Zuletzt bearbeitet' (Last edited), 'Zugriff auf Dataflux Blog' (Access to Dataflux Blog), 'Einstellungen' (Settings), 'Navigation', 'Data Management Methodology', 'Dokumentation' (Documentation), and 'Web-Links'. The 'Data Management Methodology' box also contains a circular diagram with the following steps: MONITOR, CONTROL, DEFINE, PLAN, DISCOVER, DESIGN, EXECUTE, and EVALUATE.

**Zuletzt bearbeitet**

**Zugriff auf Dataflux Blog**

**Einstellungen**

**Navigation**

**Data Management Methodology**

**Dokumentation**

**Web-Links**

**Data Management Methodology Diagram:**

```
graph TD; MONITOR --> CONTROL; CONTROL --> DEFINE; DEFINE --> PLAN; PLAN --> DISCOVER; DISCOVER --> DESIGN; DESIGN --> EXECUTE; EXECUTE --> EVALUATE; EVALUATE --> MONITOR
```

**Recent Files:**

File Name	Path	Modified
First_Job	dfr://DataFlux Sample/Jobs/First_Job(DataJob)	10/29/2010 4:54:12 PM
First_Job	dfr://DataFlux Sample/Jobs/First_Job(XMLSQLQuery)	10/29/2010 1:24:49 PM
DB2_ORDER_FACT Query	dfr://DataFlux Sample/Jobs/DB2_ORDER_FACT Query(XMLSQLQuery)	10/29/2010 1:54:59 PM

**DataFlux Community of Experts:**

- Thu, 04 Nov 2010 13:00:13 +0000  
[Returning to the Scene of the Crime](#)  
Join data expert Phil Simon as he returns to...
- Wed, 03 Nov 2010 13:00:25 +0000  
[Shared Responsibility](#)  
Should everyone share responsibility for data...
- Mon, 01 Nov 2010 13:00:08 +0000  
[When the Contract is Completed](#)  
What happens when a data consultant finishes her contract? Joyce Norris-Montanari gives an insider's view.
- Fri, 29 Oct 2010 13:00:37 +0000  
[What are You Taking Off the Plate?](#)  
Why aren't managers always on board with data management initiatives? Dylan Jones explores the issue.
- Thu, 28 Oct 2010 13:00:36 +0000  
[Lessons and Observations from IDEAS, Part II](#)

**Default Settings:**

Host ID:	"78e4009e8878 00a0c6000000 0026b9d7c358"
Active QKB:	C:\Program Files\DataFlux\QKB\QKB2010A
License Primary Method:	SAS
License Primary Location:	C:\Program Files (x86)\DataFlux\DMStudio\2.1
Authentication Server:	gertro-1:21030
User Directory:	C:\Users\gertro\AppData\Roaming
Application Root:	C:\Program Files (x86)\DataFlux\DMStudio\2.1

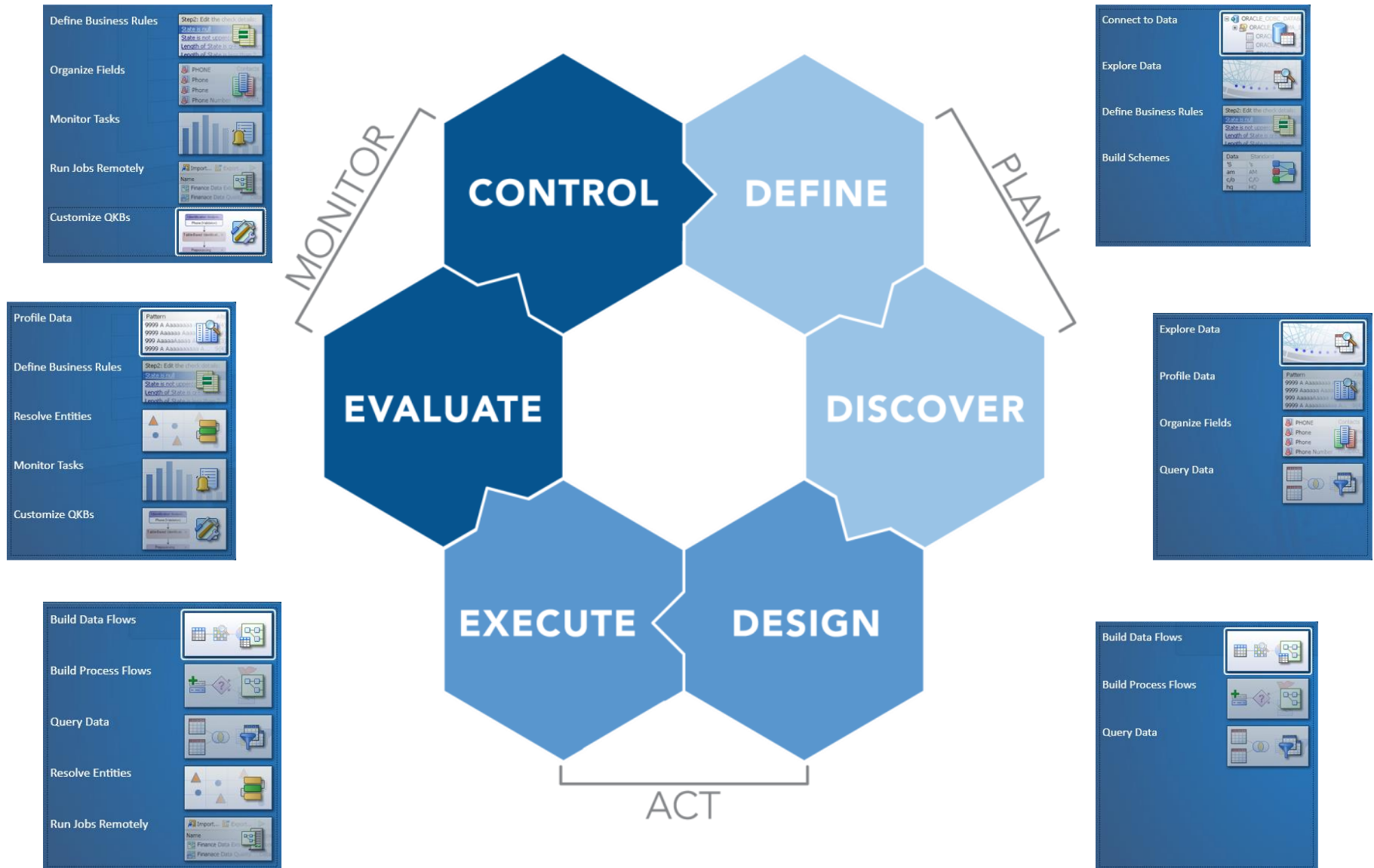
**Documentation Links:**

- Expression Language Reference Guide
- ODBC Driver Online Help
- ODBC Driver Reference Guide
- Federation Server TSSQL Reference Guide
- Data Management Server User's Guide
- Federation Server User's Guide
- Authentication Server User's Guide
- Data Management Server Administrator's Guide
- Federation Server Administrator's Guide
- Authentication Server Administrator's Guide

**Web-Links:**

- DataFlux.com
- News and Events
- MyDataFlux Portal
- Technical Support
- Downloads
- Documentation

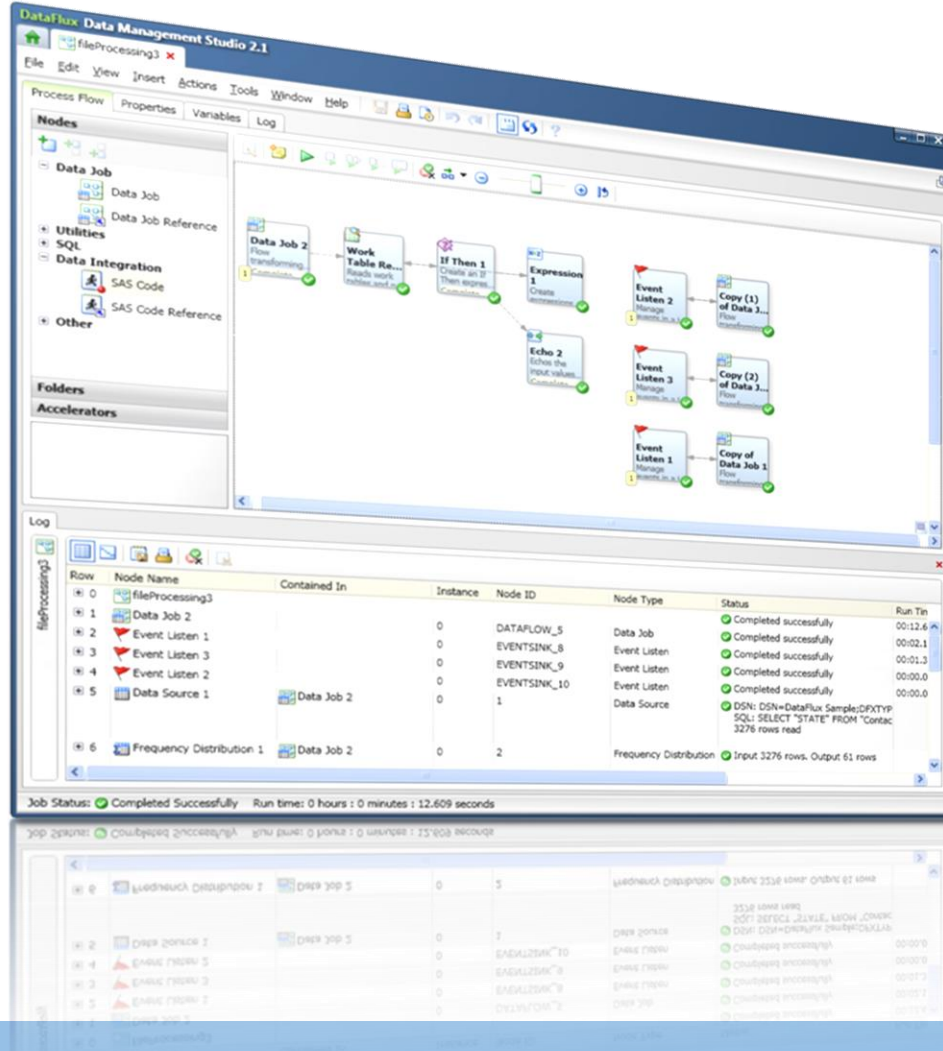
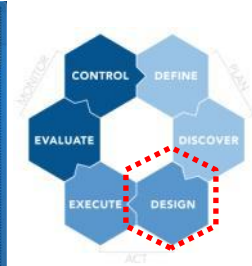
# Data Management Lifecycle Methodology





# Data Management Studio

## Highlights



### ■ Business Rules

- Erstellen und Verwalten von Regeln  
⇒ **Business Rule Manager**

### ■ Data Management

- Erstellen und Verwalten von Schemas
- Nativer Zugriff auf DBMS Typen
- “ELT” basierte Daten Manipulation (z.B. In-Database Processing)  
⇒ **Data Job Editor**

### ■ Process Modelling

- Parallel Processing Orchestration
- Bedingte Ausführung
- Event Triggering
- Erstellen und Ausführen von Workflows  
⇒ **Process Job Editor**



## Data Connections



- Data Connections
  - DataFlux Sample
  - DB2ODBC
  - DB2TRO
    - gertro-1
  - dBASE Files
  - Excel Files
  - FEDTRO
    - DB2TRO
      - gertro-1
    - SASTRO
      - local
  - FoxPro Files
  - MS Access 97 Database
  - MS Access Database
  - PRISM
  - SAMPLE
  - SASTRO
    - local
  - Text Files
  - Visio Database Samples

DB2TRO

Total count: 1

ODBC Verbindungen

Federation Server  
VerbindungenLokale DSN für Federation  
Server Verbindungen

Information

Data Connections

Collections

Folders

Data Management Servers

Administration

Locked by gertro

gertro-1:21030 EUROPE\gertro as gertro



## Collections



DataFlux Sample

New Collection

 New Collection  
COLLECTION

Total count: 106

Insert Fields...

Name	Table	Database	Catalog	Schema	Type	Length	Description
NAME	BLCLASS	DB2TRO	gertro-1	SASDEMO	Character	8	
SEX	BLCLASS2	DB2TRO	gertro-1	SASDEMO	Character	1	
AGE	BLCLASS2	DB2TRO	gertro-1	SASDEMO	Numeric	53	
HEIGHT	BLCLASS2	DB2TRO	gertro-1	SASDEMO	Numeric	53	
WEIGHT	BLCLASS2	DB2TRO	gertro-1	SASDEMO	Numeric	53	
NAME	DB2CLASS	DB2TRO	gertro-1	SASDEMO	Character	8	
SEX	DB2CLASS	DB2TRO	gertro-1	SASDEMO	Character	1	
AGE	DB2CLASS	DB2TRO	gertro-1	SASDEMO	Numeric	53	
HEIGHT	DB2CLASS	DB2TRO	gertro-1	SASDEMO	Numeric	53	
WEIGHT	DB2CLASS	DB2TRO	gertro-1	SASDEMO	Numeric	53	
NAME	DB2CLASS2	DB2TRO	gertro-1	SASDEMO	Character	8	
SEX	DB2CLASS2	DB2TRO	gertro-1	SASDEMO	Character	1	
AGE	DB2CLASS2	DB2TRO	gertro-1	SASDEMO	Numeric	53	
HEIGHT	DB2CLASS2	DB2TRO	gertro-1	SASDEMO	Numeric	53	
WEIGHT	DB2CLASS2	DB2TRO	gertro-1	SASDEMO	Numeric	53	
CUSTOMER_ID	DB2_ORDER_FACT	DB2TRO	gertro-1	SASDEMO	Numeric	12	
EMPLOYEE_ID	DB2_ORDER_FACT	DB2TRO	gertro-1	SASDEMO	Numeric	12	
STREET_ID	DB2_ORDER_FACT	DB2TRO	gertro-1	SASDEMO	Numeric	12	
ORDER_DATE	DB2_ORDER_FACT	DB2TRO	gertro-1	SASDEMO	Temporal	10	
DELIVERY_DATE	DB2_ORDER_FACT	DB2TRO	gertro-1	SASDEMO	Temporal	10	
ORDER_ID	DB2_ORDER_FACT	DB2TRO	gertro-1	SASDEMO	Numeric	12	
ORDER_TYPE	DB2_ORDER_FACT	DB2TRO	gertro-1	SASDEMO	Numeric	53	
PRODUCT_ID	DB2_ORDER_FACT	DB2TRO	gertro-1	SASDEMO	Numeric	12	
QUANTITY	DB2_ORDER_FACT	DB2TRO	gertro-1	SASDEMO	Numeric	53	
TOTAL_RETAIL_PRICE	DB2_ORDER_FACT	DB2TRO	gertro-1	SASDEMO	Numeric	53	
COSTPRICE_PER_UNIT	DB2_ORDER_FACT	DB2TRO	gertro-1	SASDEMO	Numeric	53	
DISCOUNT	DB2_ORDER_FACT	DB2TRO	gertro-1	SASDEMO	Numeric	53	
CUSTOMER_ID	DB2_ORDER_STAR	DB2TRO	gertro-1	SASDEMO	Numeric	12	
EMPLOYEE_ID	DB2_ORDER_STAR	DB2TRO	gertro-1	SASDEMO	Numeric	12	
STREET_ID	DB2_ORDER_STAR	DB2TRO	gertro-1	SASDEMO	Numeric	12	
ORDER_DATE	DB2_ORDER_STAR	DB2TRO	gertro-1	SASDEMO	Temporal	10	
DELIVERY_DATE	DB2_ORDER_STAR	DB2TRO	gertro-1	SASDEMO	Temporal	10	
ORDER_ID	DB2_ORDER_STAR	DB2TRO	gertro-1	SASDEMO	Numeric	12	

Selbst erstelltes Dictionary  
von Tabellen und Spalten

Information

Data Connections

Collections

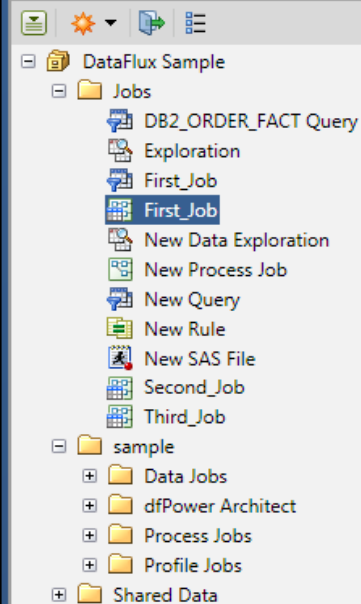
Folders

Data Management Servers

Administration



## Folders

**First\_Job**  
DATA JOB

## Identification

Name: First\_Job  
Description:  
Note:  
Type: Data Job  
ID: dfr://DataFlux Sample/Jobs/First\_Job(DataJob)  
Date Created: 29.10.2010 12:10  
Last Modified: 04.11.2010 15:41  
Lock Status: Locked by gertro

Start time:

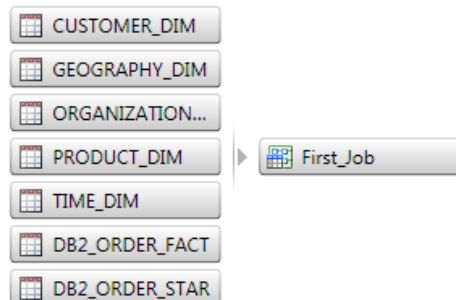
End time:

Run time:

[View Log](#)

## Lineage

&gt; First\_Job



# Objekteigenschaften & Ausführungsstatus

## Ordnersicht der erstellten Objekte

## Impact Analyse

Information

Data Connections

Collections

Folders

Data Management Servers

Administration

Locked by gertro

gertro-1:21030

EUROPE\gertro as gertro

**Administration**

- Authentication Servers
  - gertro-1
- Federation Servers
  - gertro-1
- Macro Files
  - macros
- Quality Knowledge Bases
  - 2010A
- Reference Sources
- Repository Definitions
  - DataFlux Sample

Administration  
Authentication Server

Administration  
Federation Server

Dataflux Macros

Administration  
QKB

Administration von Fremddaten  
(USPS & Geocode)

Administration  
Dataflux Repositories

Information

Data Connections

Collections

Folders

Data Management Servers

Administration

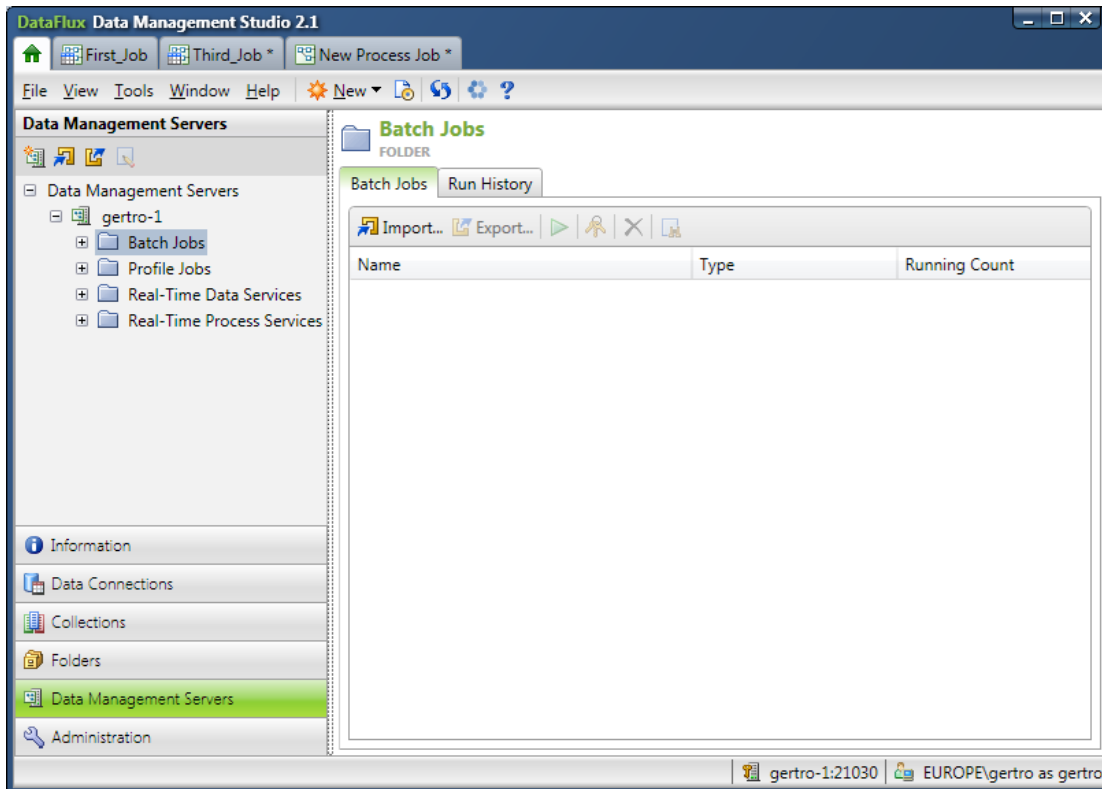
Locked by gertro

gertro-1:21030

EUROPE\gertro as gertro

# Data Management Server

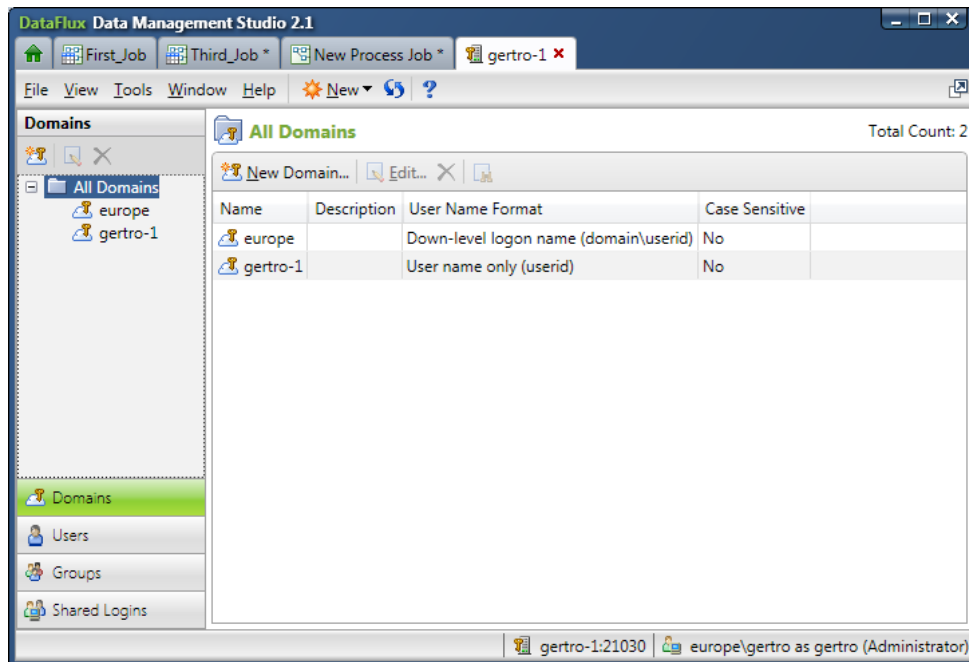
Der „Dienstleister“



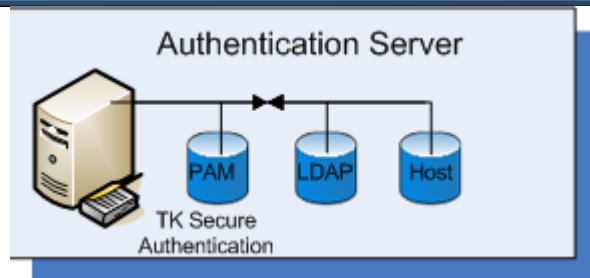
- Registrierung von
  - Jobs
  - Services (SOA)
- Vergleichbar Dataflux Integration Server
- Plattformen:
  - Windows 32 bit
  - Windows 64 bit
  - Solaris x86 64-bit
  - Solaris Sparc 64-bit
  - Linux x86 32-bit
  - Linux x86 64-bit
  - HP-UX Itanium 64-bit
  - AIX 64-bit
  - *HP-UX Pa-Risc geplant*

# Authentication Server

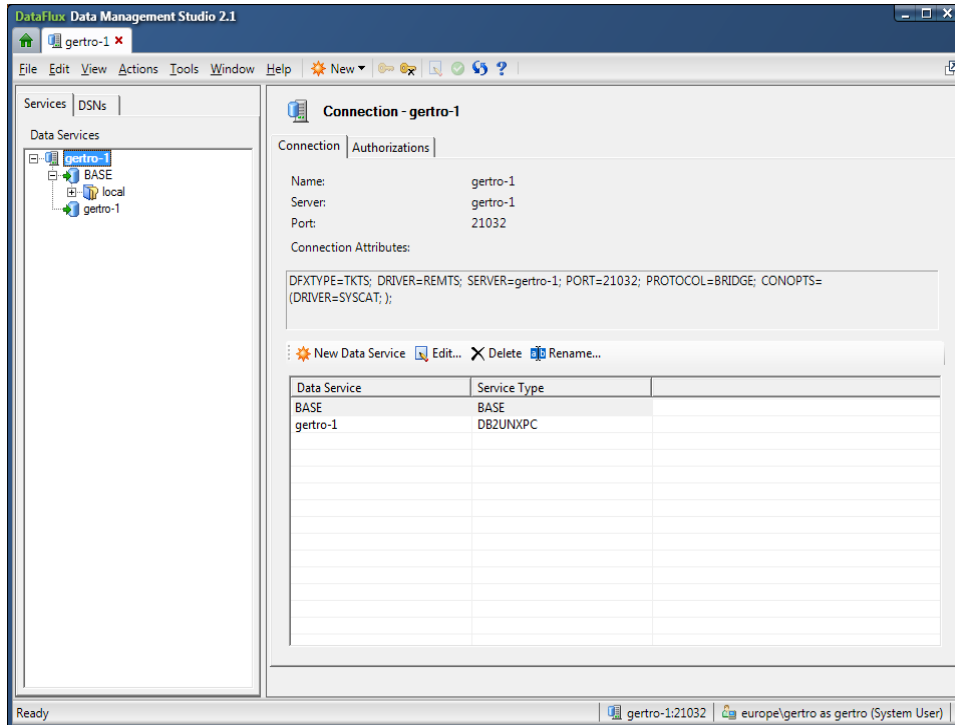
Der „Türsteher“



- Verwaltung von:
  - Domänen
  - Benutzern
  - Gruppen
  - Gemeinsam genutzten Logins
- Supporte Authentication Provider:
  - Host
  - LDAP
  - Active Directory
  - PAM



# Die „Datenzentrale“



- Technologie (SAS Table Server):

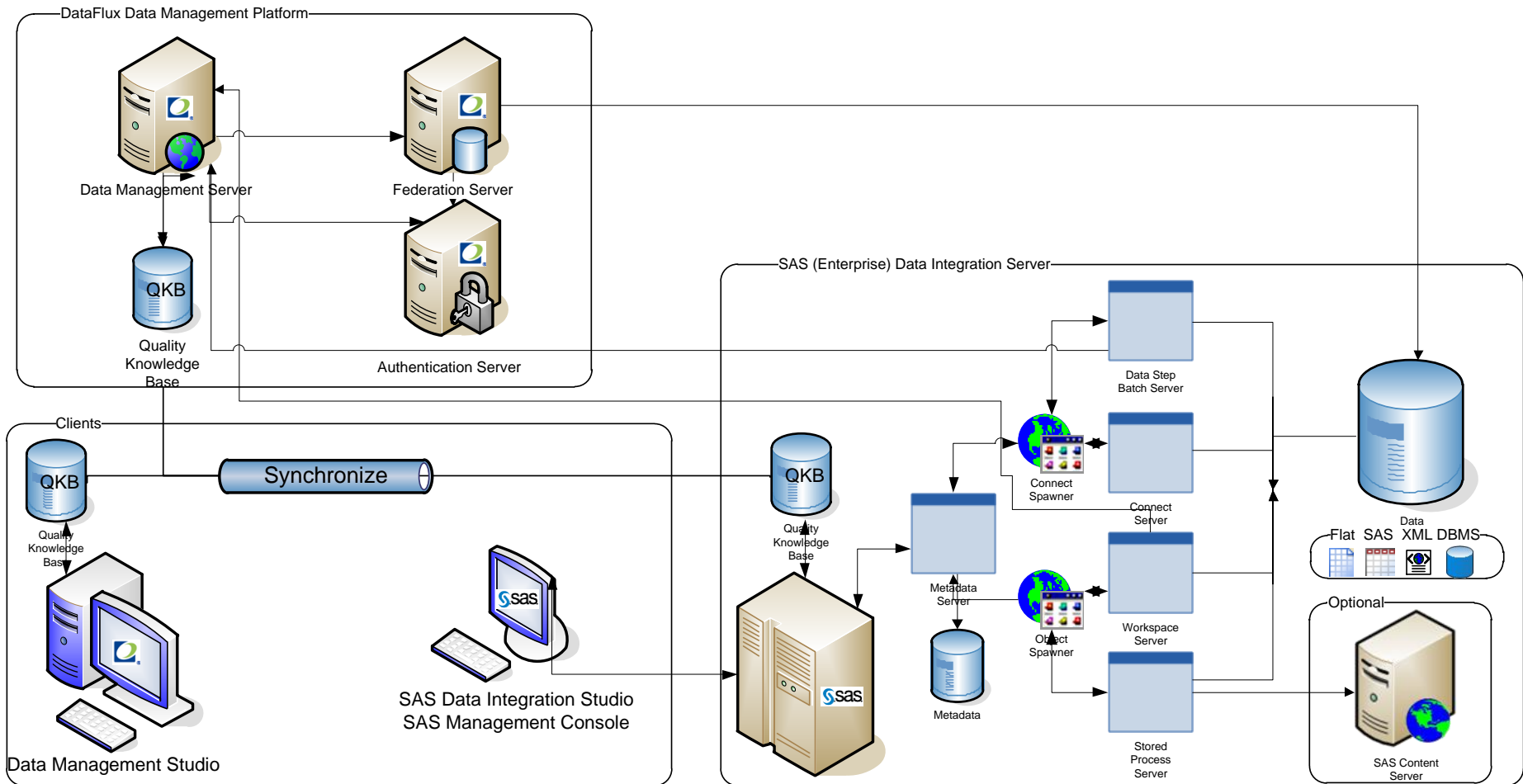
- Skalierbar
- Threaded
- Multi-user
- Einheitlich
  - Federation Server TSSQL
  - DBMS native SQL
- Sicher (über Authentication Server)

- Datenquellen:

- SAS
- DBMS:
  - DataFlux Driver for Base SAS
  - DataFlux Driver for Oracle
  - DataFlux Driver for Teradata
  - DataFlux Drive for DB2 CS
  - DataFlux Driver for DB2 z/OS
  - DataFlux Driver for MySQL
  - DataFlux Driver for ODBC
  - DataFlux Driver for SAP



# SAS DI/DMP Architektur



# Zusammenfassung Data Management Platform

Die Data Management Platform Version 2.1 ist der Nachfolger von dfPower Studio:

- Geführter Einstieg über Methodology
- Integrierte Oberfläche mit einheitlichem Look & Feel
- Einsteigerfreundlich

Die Data Management Platform 2.1 (**GA seit 11/2010**) oder aktueller wird dfPower Studio 8.x unter SAS 9.3 ersetzen.

Folgende Erweiterungen sind für SAS DI/DM geplant:

- Short-Term:
  - Erweiterter ELT (Pushdown) Support
  - JMS Support
  - Hadoop Support
  - Weitere DBMS-spezifische Loader
- Long-Term:
  - Erweiterte Möglichkeiten zur gegenseitigen Nutzung von DI Studio und der DMP Platform:
    - Business Rules
    - DI Prozess Modellierung
    - Scheduling
  - Event Stream Processing
  - Solution Integration
  - Performance Optimierung
    - Pipelining
    - Grid Support