

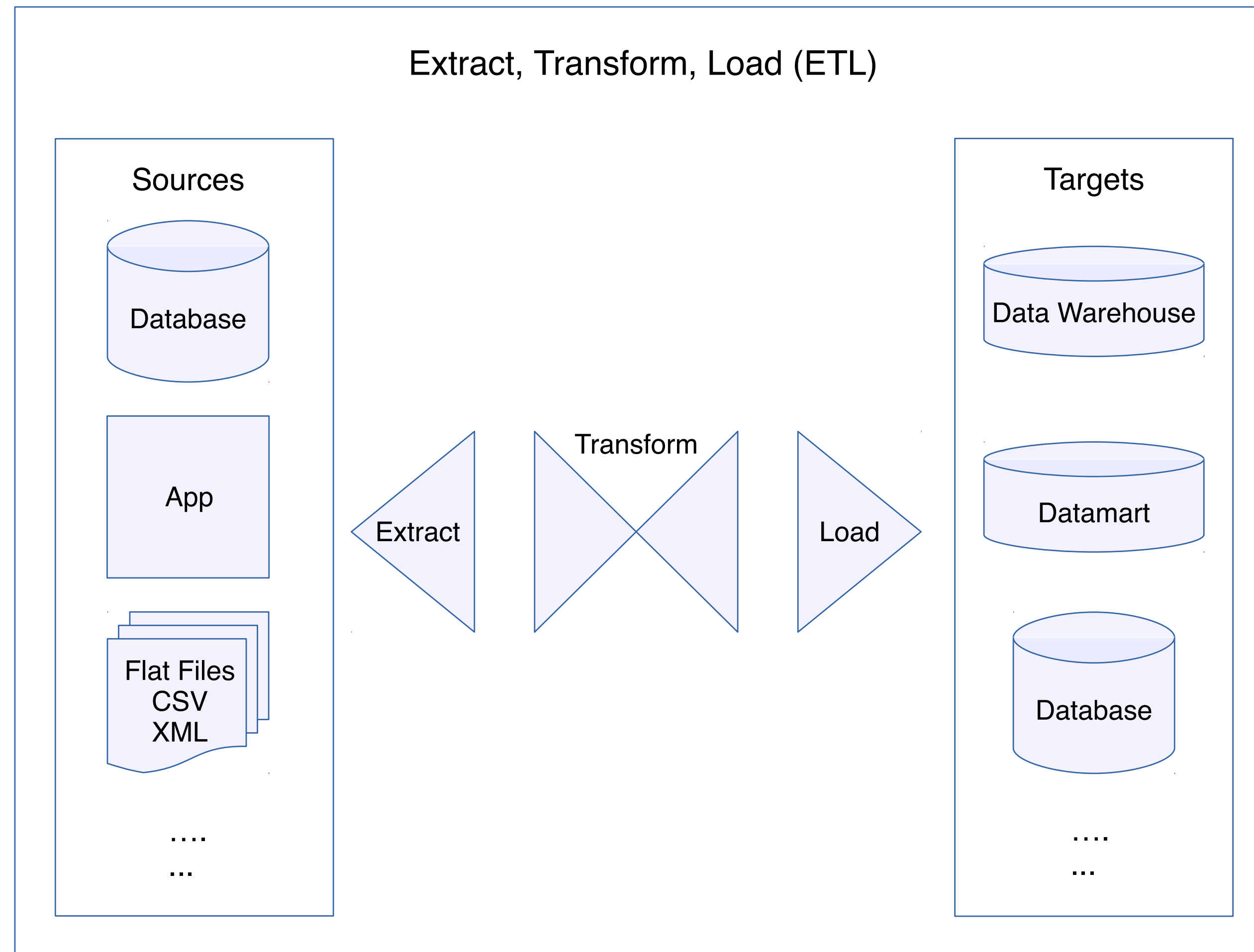
Datenintegration mit Talend DI

Martin v. Löwis, März 2025

Data Integration (DI)

- Informationsintegration: Zusammenführen von Informationen aus verschiedenen Datenbeständen
 - z.B. in data warehouses
 - hier auch: Abgleich (Synchronisation) von Datenbeständen
- führendes System: abhängige Systeme übernehmen Daten
 - z.B. regelmäßig, oder ausgelöst durch Änderungen

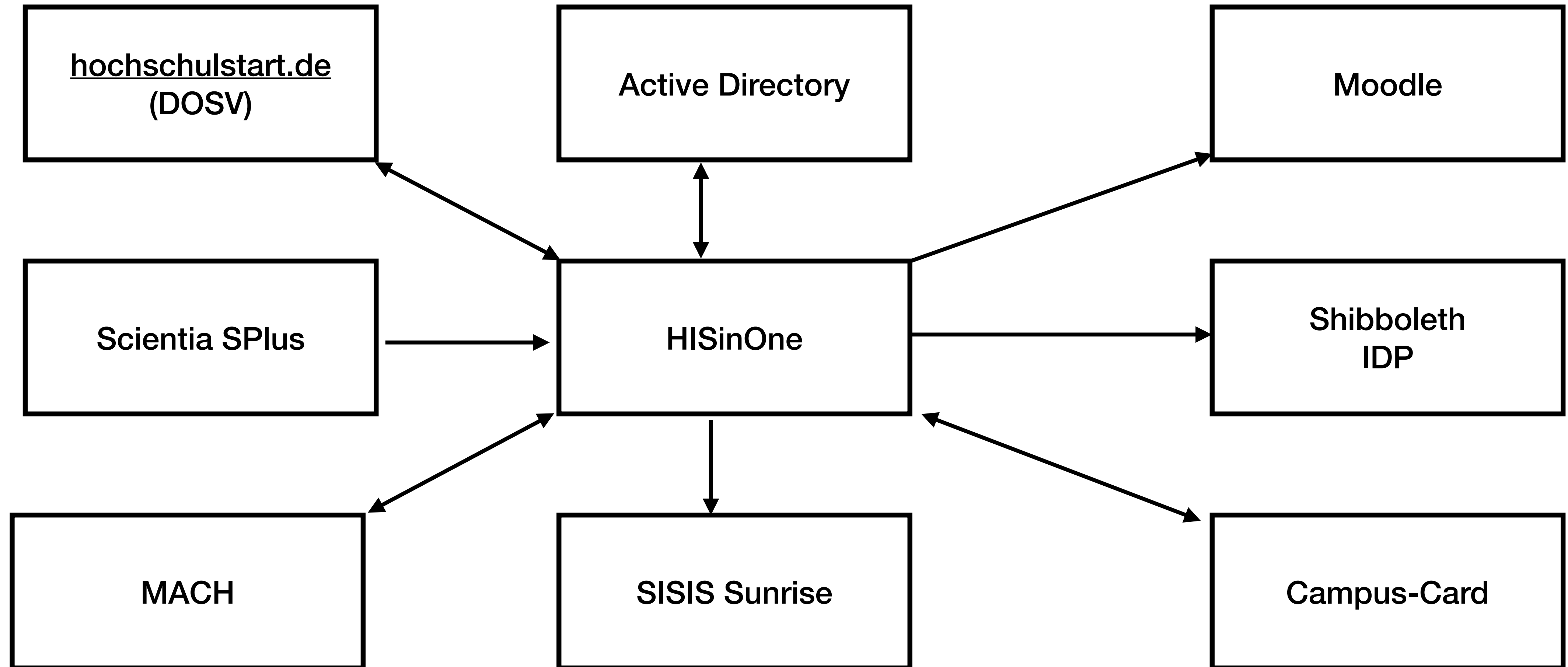
ETL: Extract, Transform, Load



Anwendungsfall: HISinOne

- Campus-Management-System
- Verwaltung von Studierendendaten
 - Stammdaten: Name, Postanschrift, Krankenkasse, Nutzername, Matrikelnummer
 - Semesterweise Daten: Bewerbungen, Immatrikulation / Rückmeldung, Rechnungen / Zahlungen
 - Leistungsdaten: Belegungen, Noten
- Stundenplanung: Räume, Zeiten, Lehrkräfte
- Erstellung von Bescheiden (Zeugnisse, Immatrikulationsbescheinigungen)

Integrationsschnittstellen



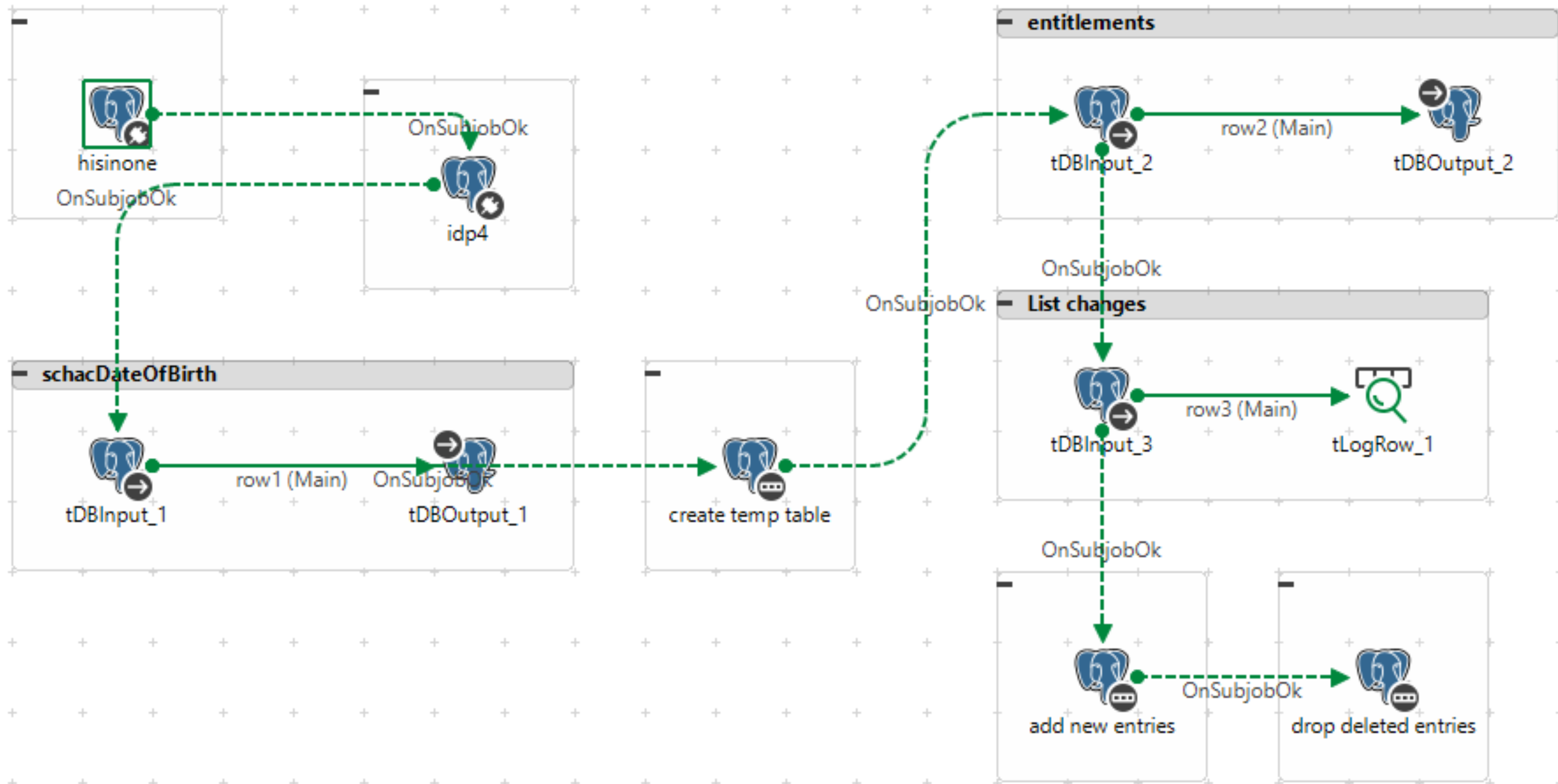
Talend

- Grafische Programmiersprache zur Datenintegration
- Entwicklungsumgebung: Talend Studio
 - basiert auf Eclipse
- Vordefinierte Komponenten zum Zugriff auf Systeme
 - Datenbanken, Dateien, Webservices, E-Mail-Versand, ...
- Laufzeitumgebung: Administration Center, Job Server
 - Jobs werden in ein Java-Programm übersetzt, was geplant ausgeführt wird

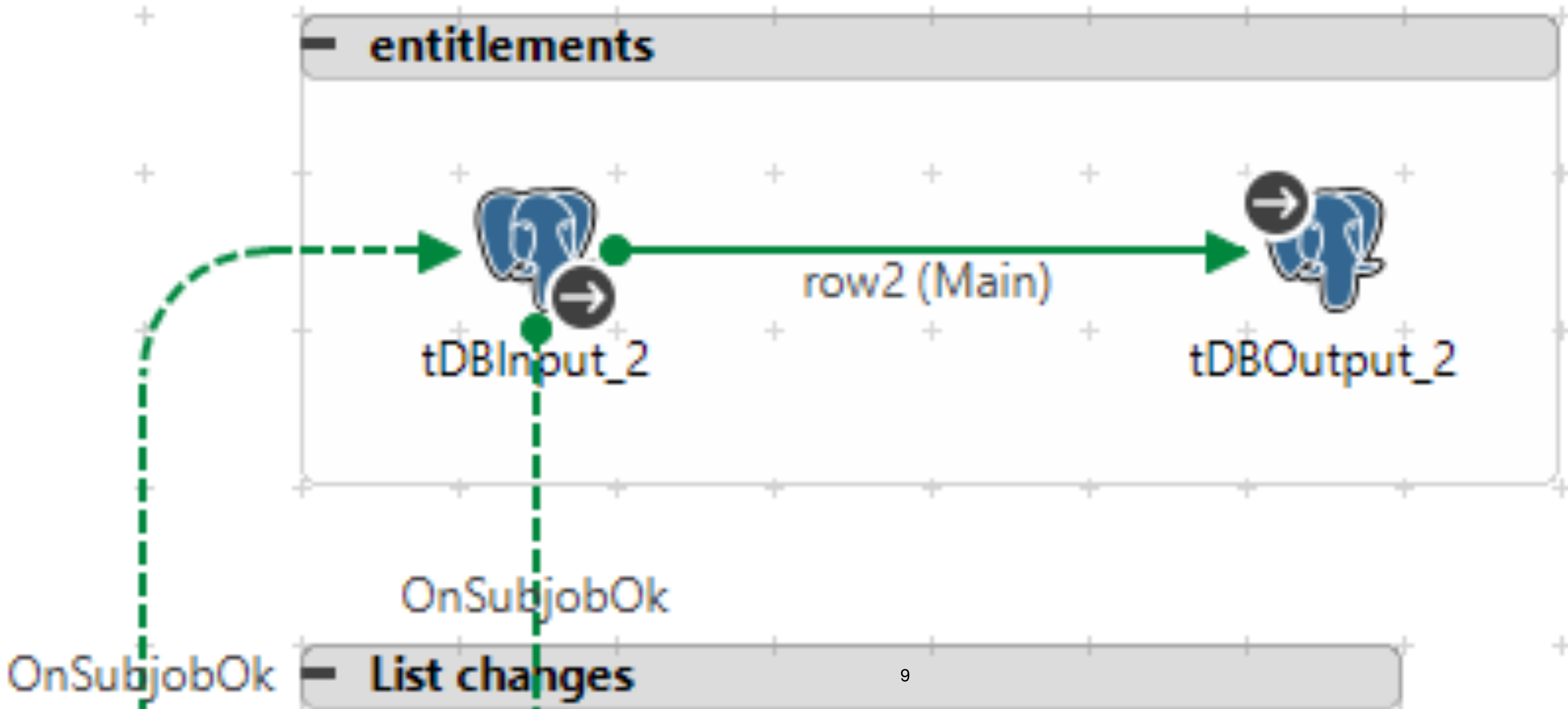
Beispiel: Deutschlandticket

- Externer Dienstleister benötigt Personenattribute, um Anspruch auf Semesterticket zu validieren
 - Stammdaten: Name, Geburtsdatum
 - Semesterdaten: wurde das Semesterticket bezahlt?
- Umsetzung mit Shibboleth IDP
- IDP-Datenbank benötigt zusätzliche Attribute

Talend-Job-Übersicht



Subjobs



Komponenten: DB-Connection

The screenshot displays the Talend Studio interface for a job named "Job UpdateRideticketing 1.2". The job design shows three components: "hisinone", "idp4", and "tDBInput_2". The "hisinone" component is connected to "idp4" via a dashed green arrow labeled "OnSubjobOk". The "idp4" component is connected to "tDBInput_2" via a dashed green arrow labeled "OnSubjobOk". The "tDBInput_2" component is connected to a table named "entitlements" via a solid green arrow labeled "row2 (Main)".

The configuration panel for the "idp4(tDBConnection_2)(PostgreSQL)" component is shown below the job design. The configuration is as follows:

Property	Value
Database	PostgreSQL
Eigenschaftstyp	Built-In
DB Version	v12 and later
Host	"idp01.bht-berlin.de"
Port	"5432"
Database	"shibboleth"
Schema	""
Benutzername	"shibboleth-df"
Passwort	*****

Additional options include:

- Verwende oder melde eine geteilte DB Verbindung an
- Data source: 10
- This option only applies when deploying and running in the Talend Runtime

Komponenten: Datenbankabfrage

The image shows a software development environment with a component diagram and a configuration panel.

Component Diagram:

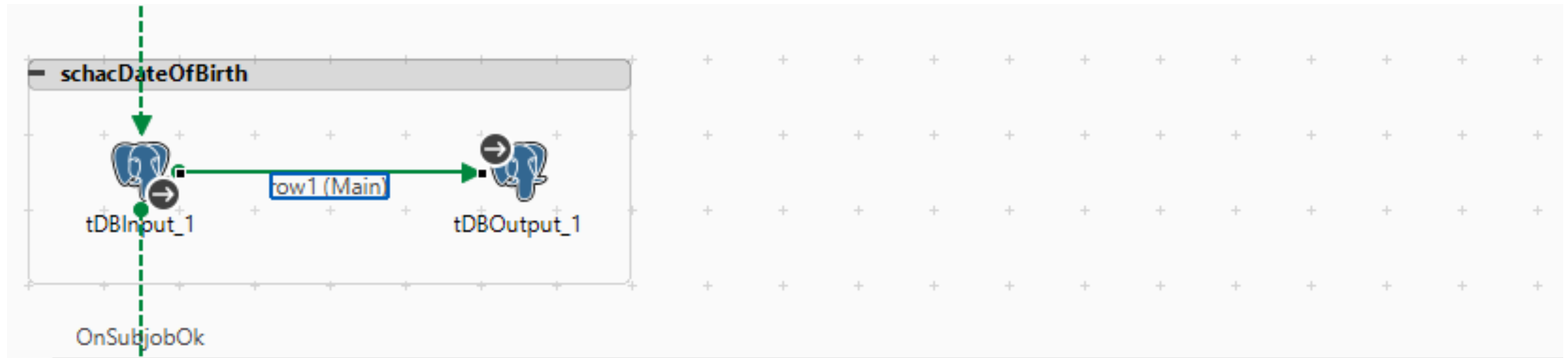
- Component: **schacDateOfBirth**
- Input Component: **tDBInput_1** (PostgreSQL icon)
- Output Component: **tDBOutput_1** (PostgreSQL icon)
- Flow: **row1 (Main)** connects tDBInput_1 to tDBOutput_1.
- Trigger: **OnSubjobOk** is connected to tDBInput_1.

Configuration Panel: tDBInput_1 (PostgreSQL)

- Basic settings**
- Database: PostgreSQL (Apply)
- eine bestehende Verbindung verwenden
- Komponenten Liste: tDBConnection_1 - hisinone *
- Schema: Built-In (Edit schema ...)
- Tabellenname: ""
- Abfragetyp: Built-In (Guess Query, Guess schema)
- Abfrage:

```
"select username,
to_char(birthdate, 'YYYYMMDD') schacDateOfBirth
from hisinone.student s
join hisinone.person p on p.id=s.person_id
join hisinone.account a on a.person_id=p.id"
```

Rows



Designer Code

Job Context (UpdateRideticketing) Komponente x Starte (Job UpdateRideticketing) Cloud Artifact

row1

Basic settings

Advanced settings

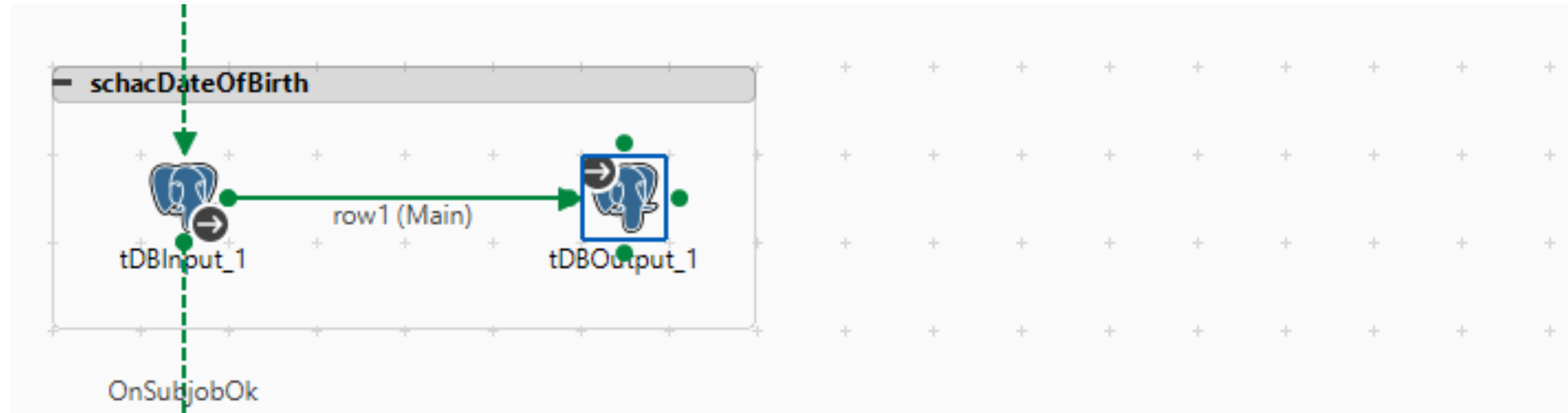
Component

Edit schema

Schema from tDBInput_1 output

	Spalte	Db Column	Sc...	Typ	Db Type	<input checked="" type="checkbox"/> N..	Datumsformat...	Länge
1	username	username	<input checked="" type="checkbox"/>	String	VARCH...	<input checked="" type="checkbox"/>		
2	schacdateofbirth	schacdateofbirth	<input type="checkbox"/>	String	VARCH...	<input checked="" type="checkbox"/>		

Komponente: DB-Ausgabe



Designer Code

Job Context (UpdateRideticketing) Komponente x Starte (Job UpdateRideticketing) Cloud Artifact

tDBOutput_1 (PostgreSQL)

Basic settings

Database PostgreSQL Apply

Advanced settings

eine bestehende Verbindung verwenden Komponenten Liste tDBConnection_2 - idp4 *

Dynamic settings

Tabelle "dfneduperson"

View

Aktion auf Tabelle None Aktion auf Daten Insert oder Update

Documentation

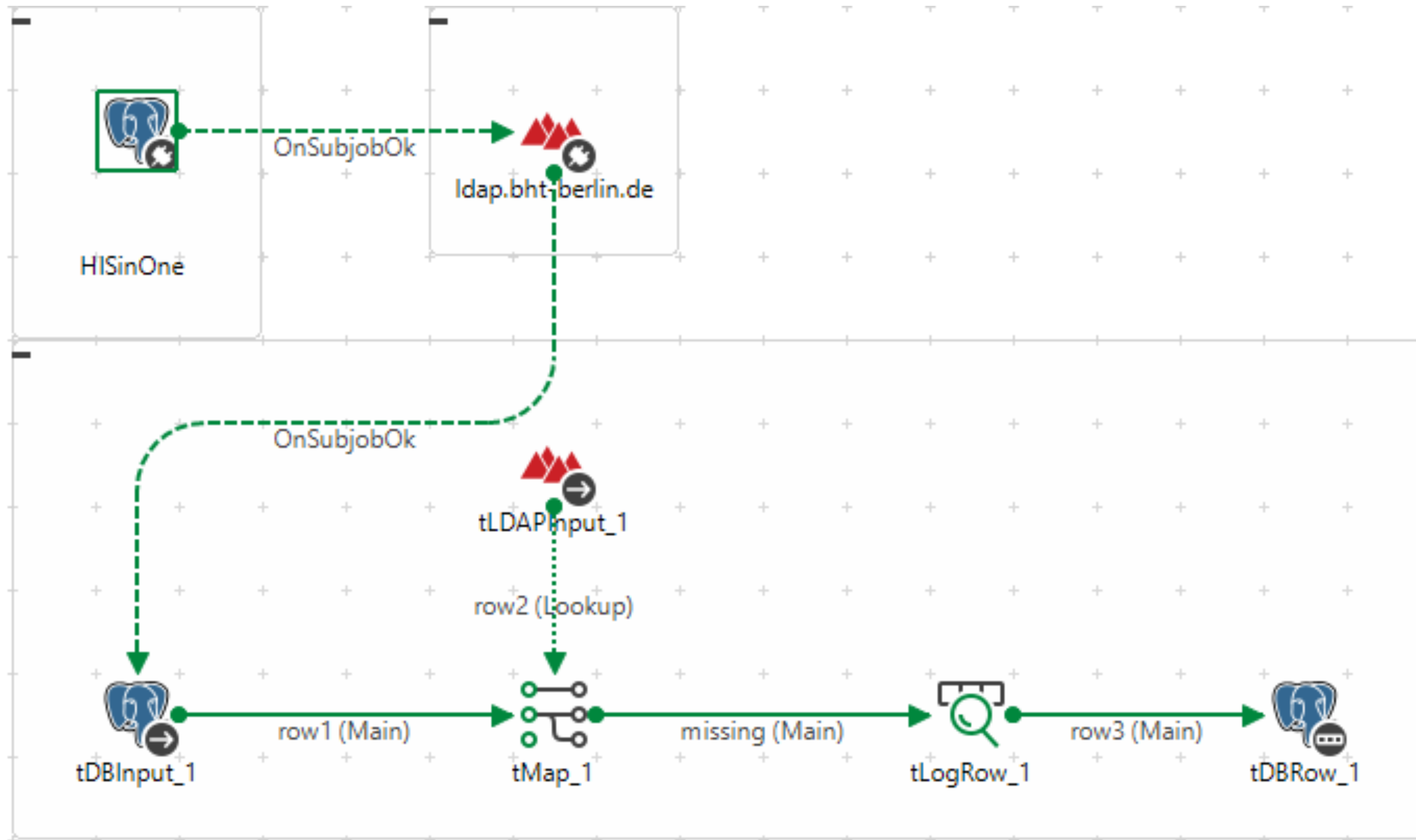
Schema Built-In Edit schema Sync columns

Abbrechen bei Fehler

Beispiel: Löschen abgelaufener Nutzer

- Nutzer, die im Active Directory gelöscht wurden, sollen aus HISinOne ebenfalls gelöscht werden.
- Konzeptionell Join-Operation über zwei Systeme

Löschen von Account-Einträgen



Security

- Talend-Jobs haben oft weitreichende Schreibrechte auf zahlreiche Systeme
- Credentials: Talend speichert Passwörter verschlüsselt mit einem Master-Key
 - im git-Repository werden nur die verschlüsselten Passwörter abgelegt
 - Editor (Talend Studio) und Job Server kennen den Master-Key
- Windows: Job Server kann als Service Account laufen, mit den Rechten eines Domänennutzers
 - Zugriffe z.B. auf SMB-Shares dann ohne Passwort möglich
- Strategie: Principle of Least Privilege

Administration Center

- web-basierte Verwaltungsoberfläche
- Einrichtung von neuen Projekten
- Berechtigungen
- zeitgesteuerte Ausführung ("cron")
- Monitoring erfolgreicher Job-Ausführung

Praktische Erfahrungen

- steile Lernkurve
- hohe Produktivität bei der Umsetzung neuer Integrationsprozesse
- Entwicklung von Lizenzkosten:
 - ursprünglich Open Source, mit kostenfreier Version
 - inzwischen von Qlik übernommen, nur noch gegen Lizenzkosten erhältlich