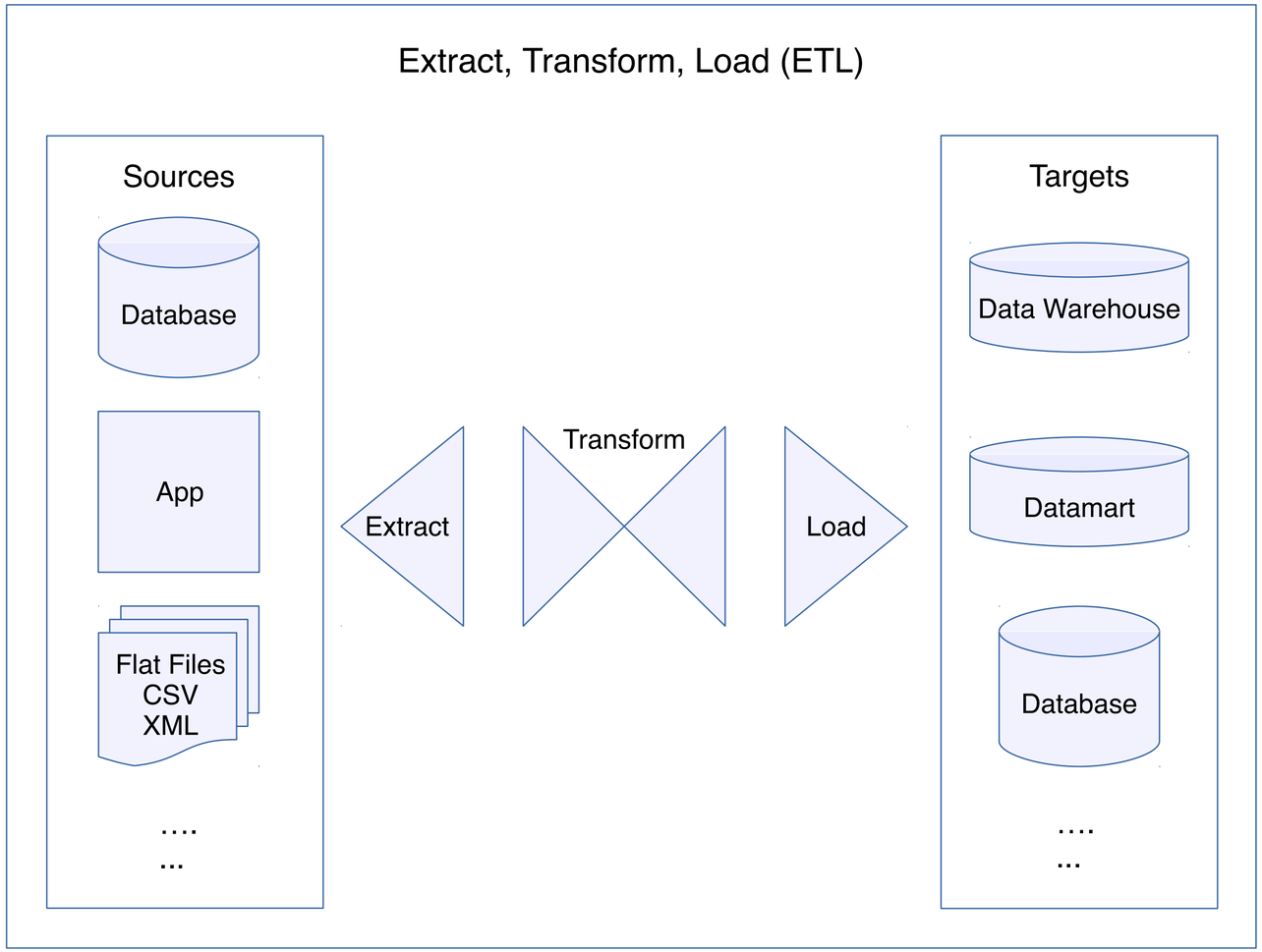
Datenintegration mit Talend DI

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

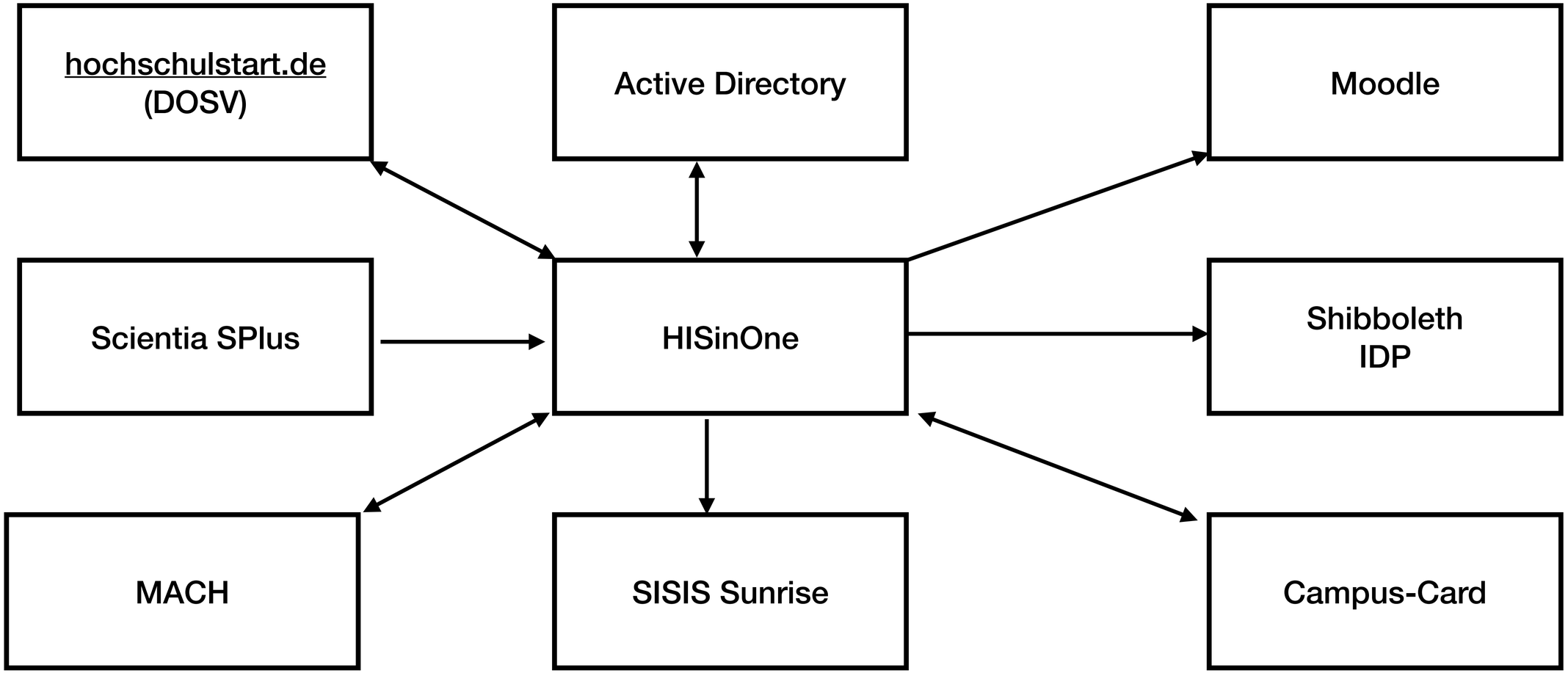


https://en.wikipedia.org/wiki/Extract,_transform,_load#/media/File:Extract,_Transform,_Load_Data_Flow_Diagram.svg

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



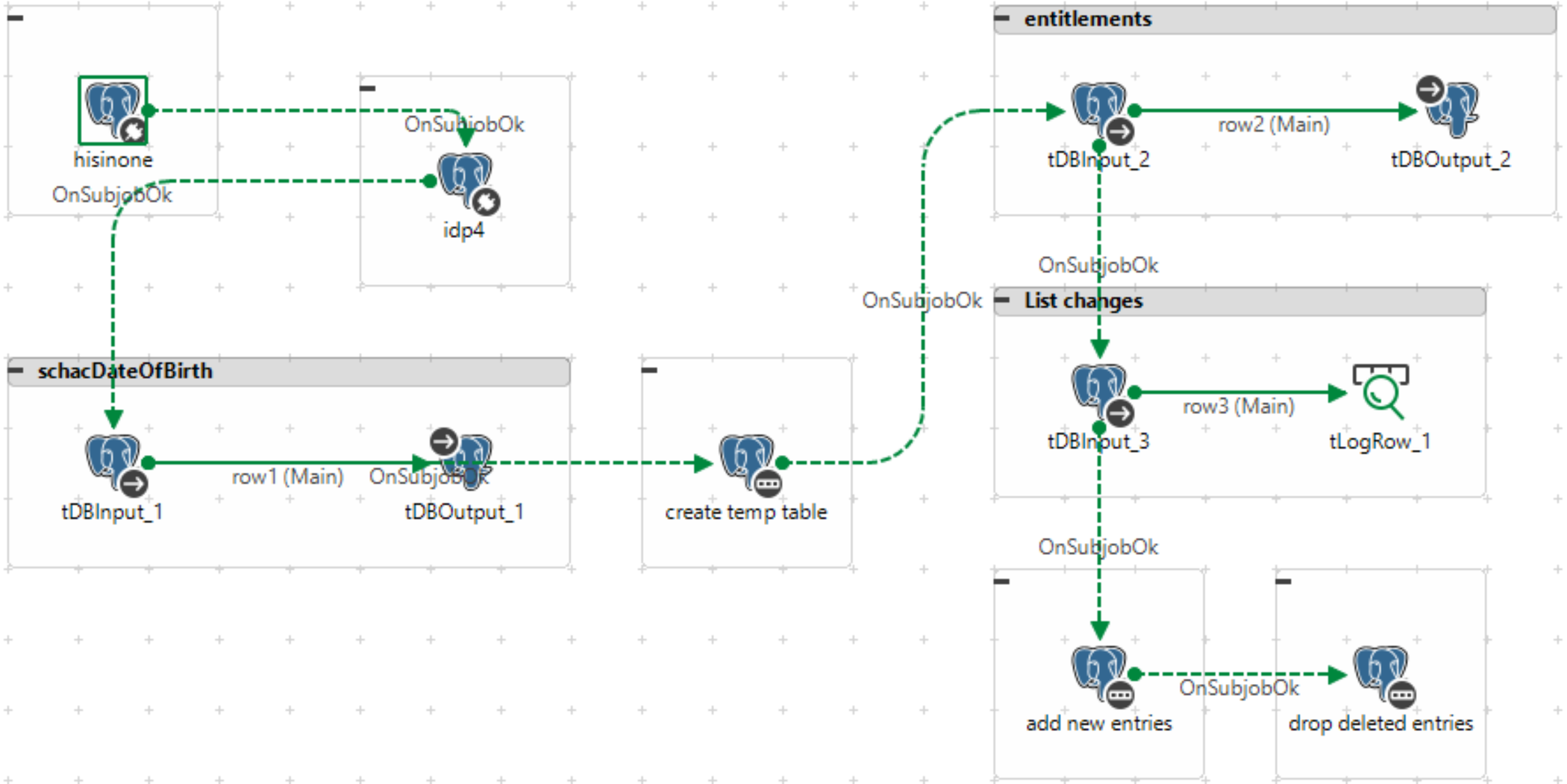
Talend

- 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

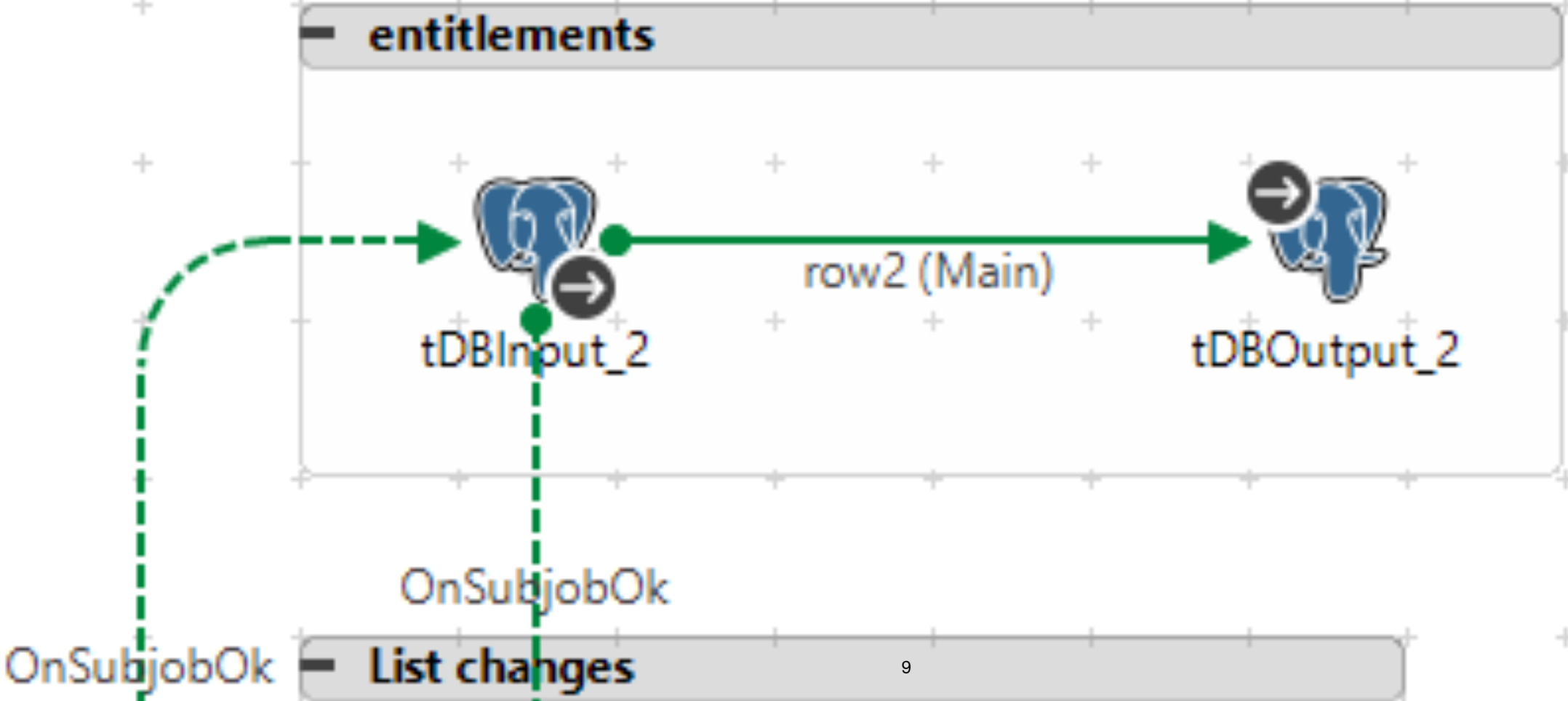
Beispiel: Deutschlandticket

- 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

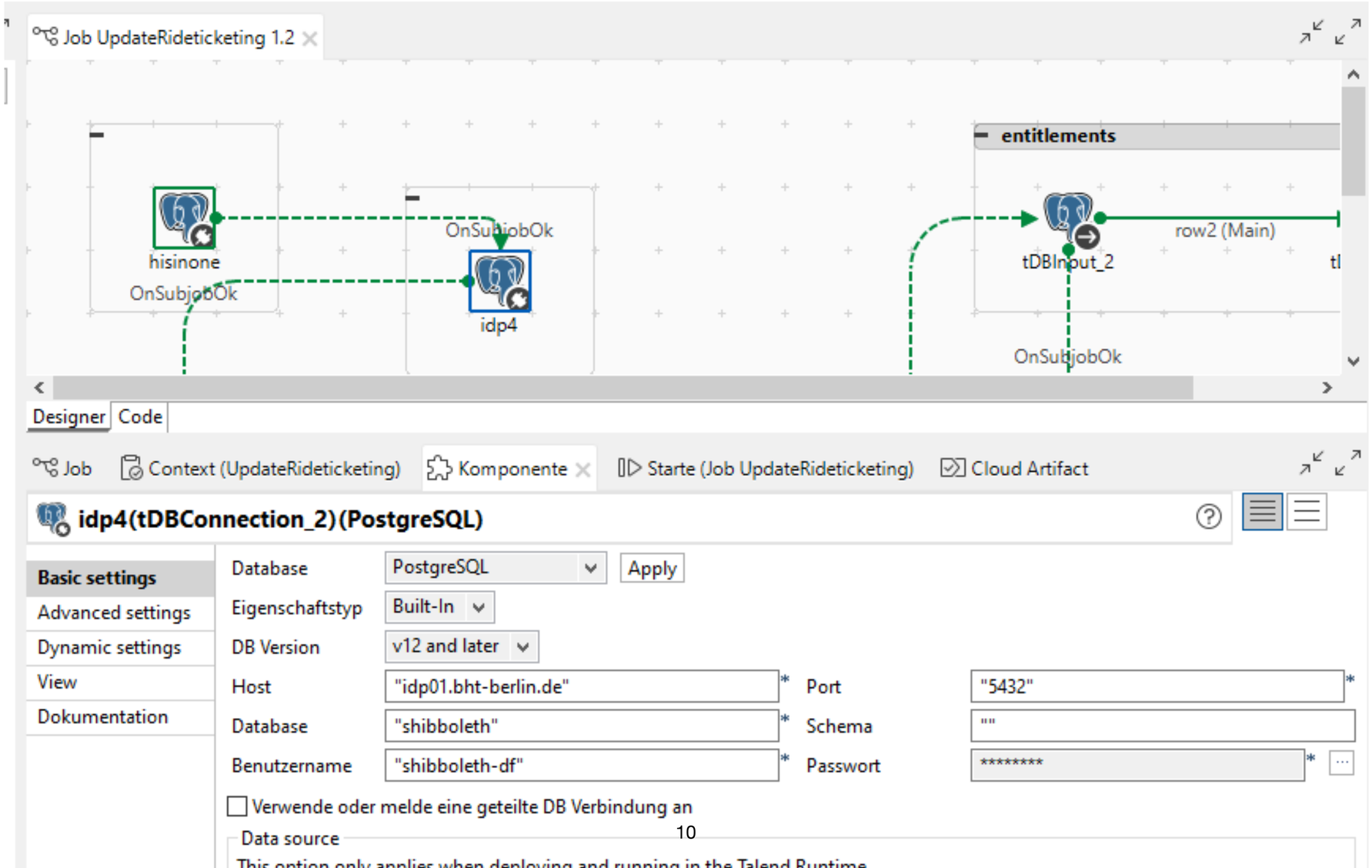
Talend-Job-Übersicht



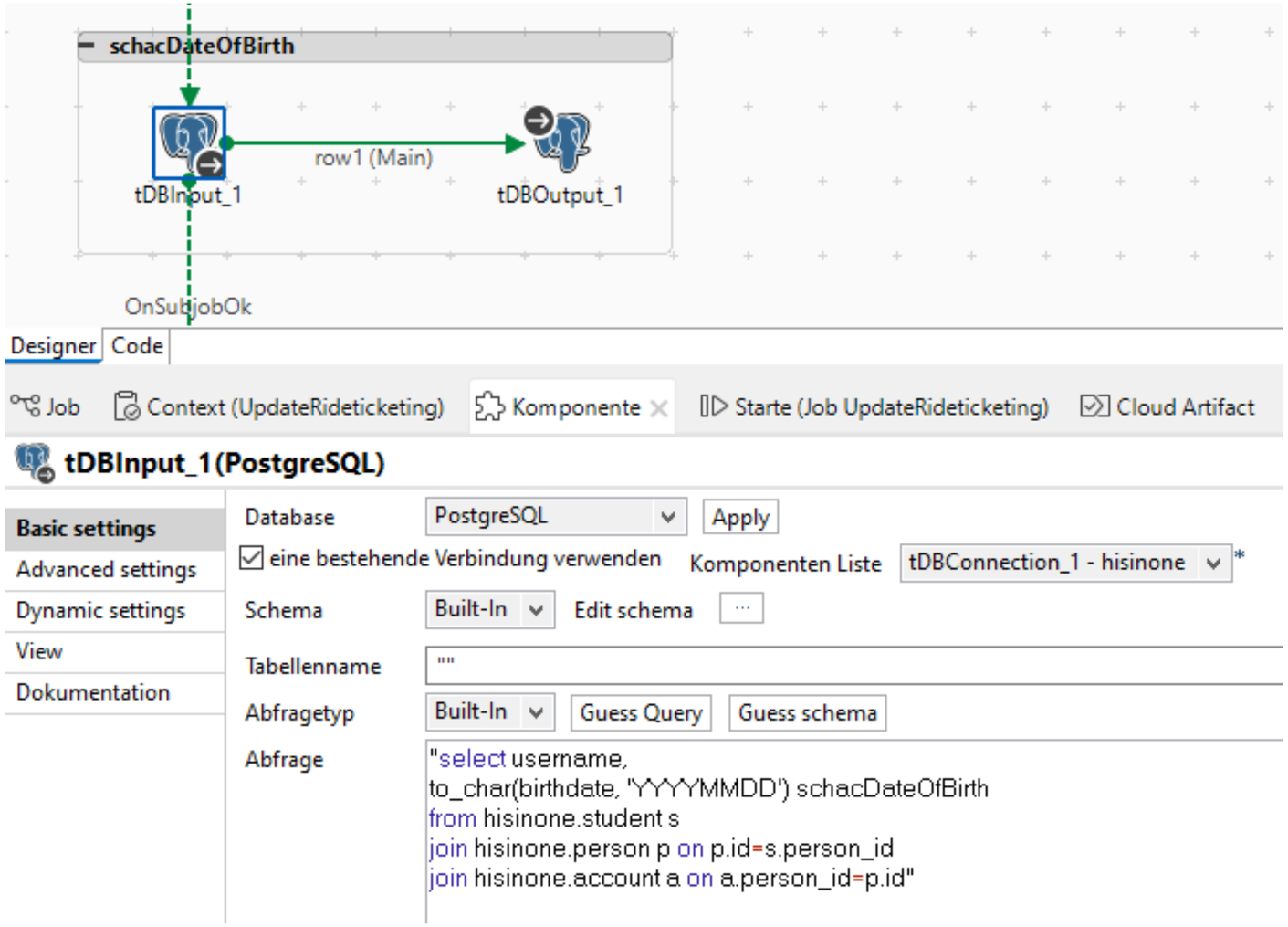
Subjobs


entitlements
row2 (Main)
tDBInput_2
tDBOutput_2
OnSubjobOk
OnSubjobOk
List changes

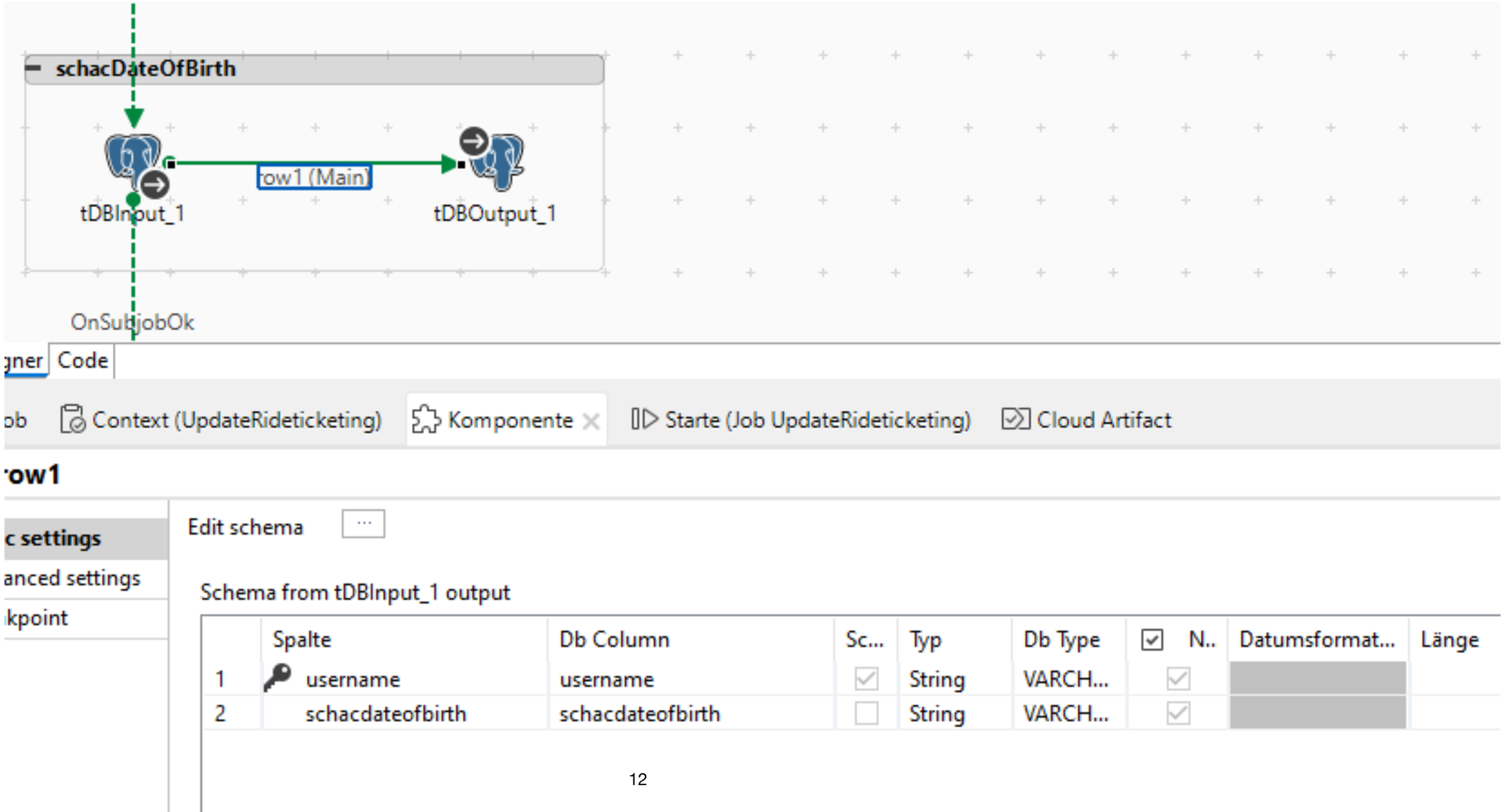
Komponenten: DB-Connection



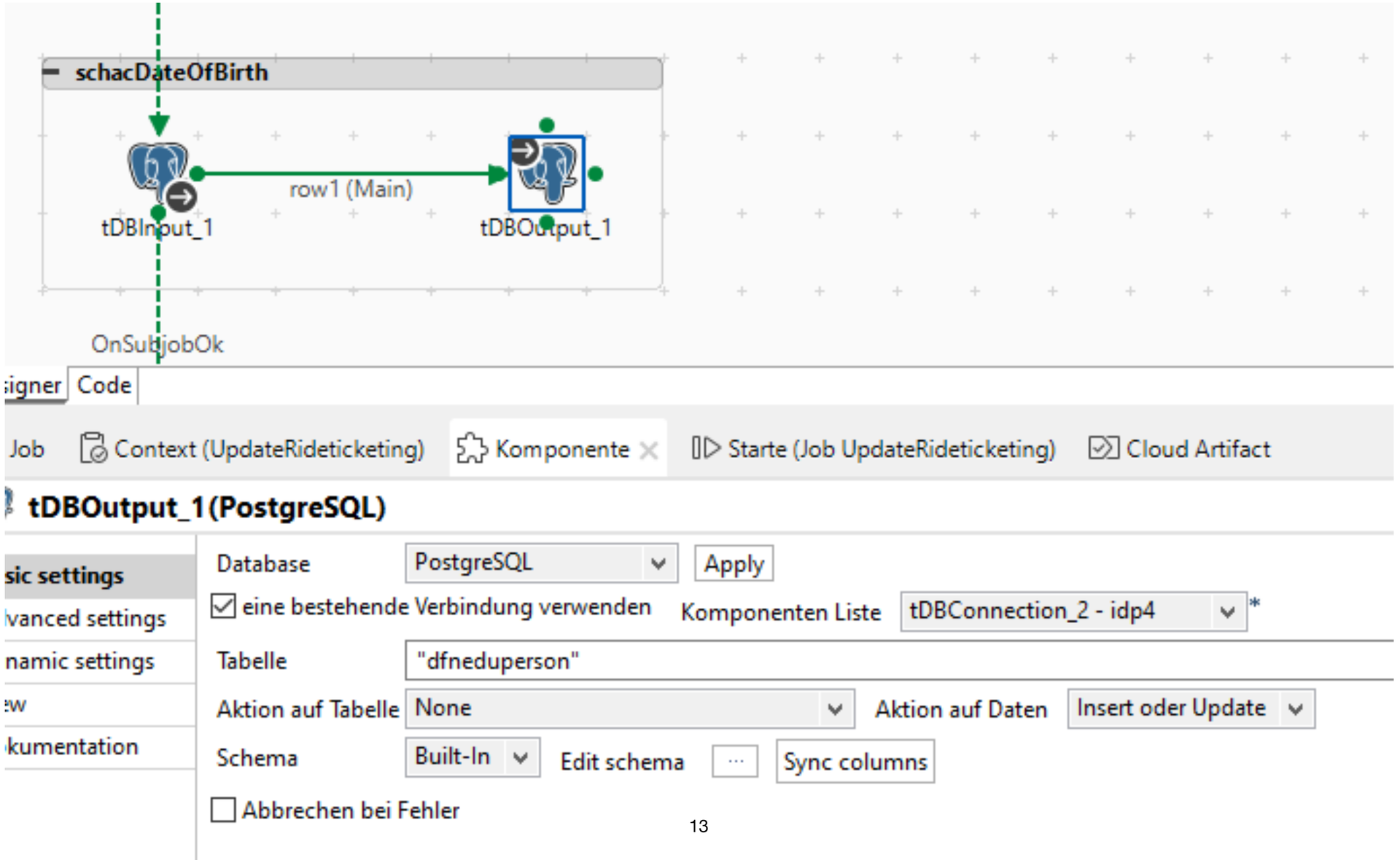
Komponenten: Datenbankabfrage



Rows


schacDateOfBirth
ow1 (Main)
tDBInput_1
tDBOutput_1
OnSubjobOk
gner
Code
ob
Context (UpdateRideticketing)
Komponente
Starte (Job UpdateRideticketing)
Cloud Artifact
ow1
c settings
anced settings
kpoint
Edit schema
Schema from tDBInput_1 output
Spalte
Db Column
Sc...
Typ
Db Type
N..
Datumsformat...
Länge
1
username
username
String
VARCH...
2
schacdateofbirth
schacdateofbirth
String
VARCH...

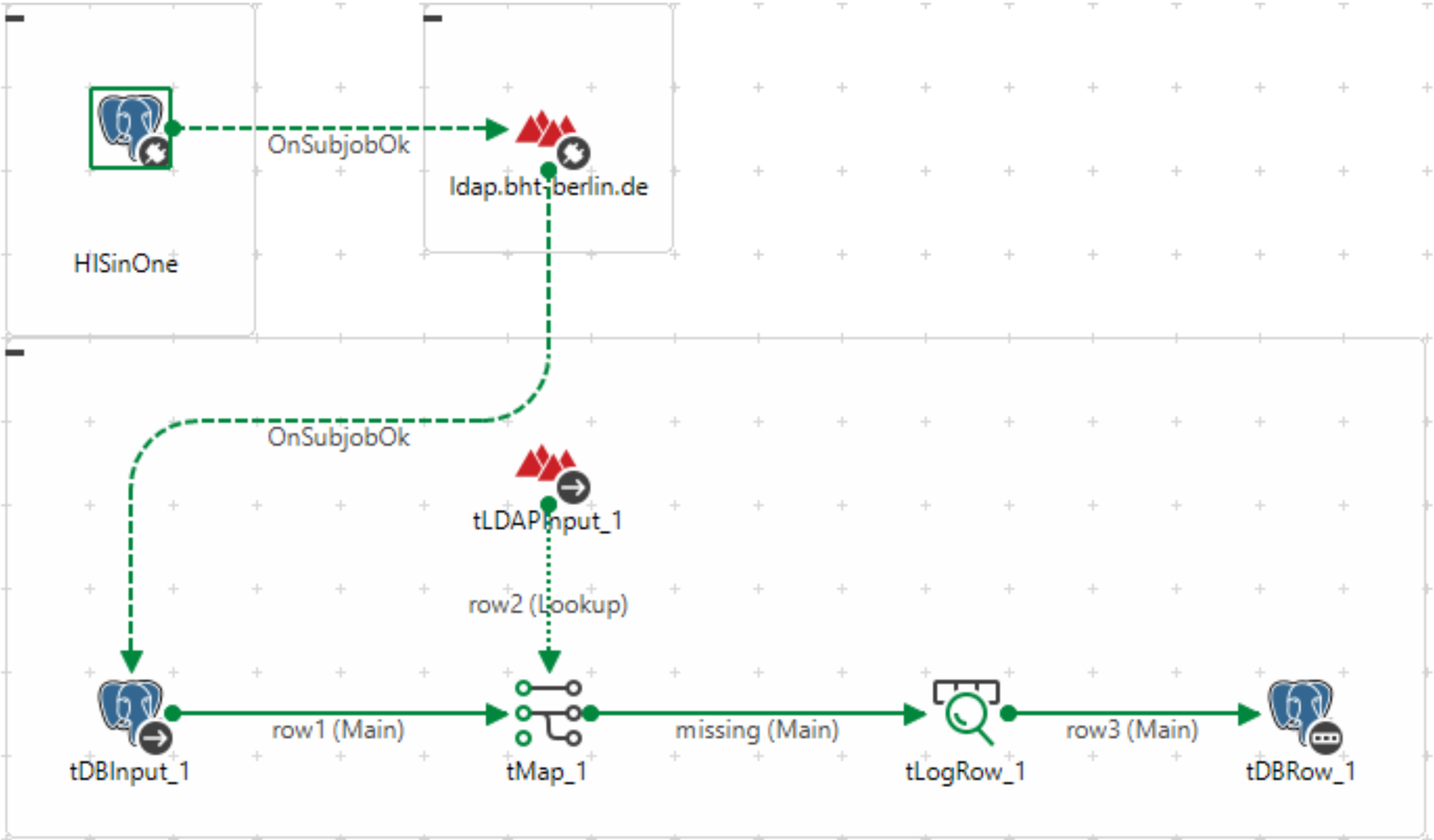
Komponente: DB-Ausgabe



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