Philipps Universität Marburg

T Square

Project Step 2: Integration

Timo Scheerer, Timo Büchert

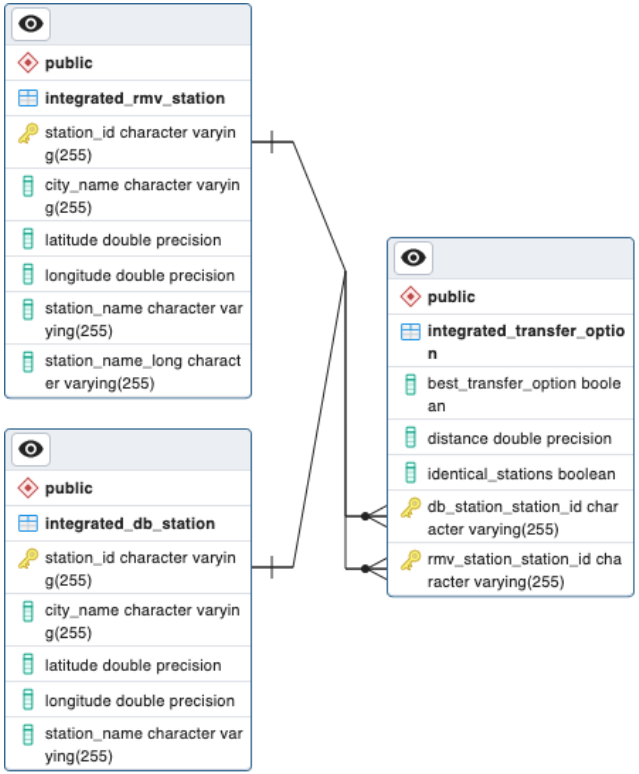
Integration Pipeline

Philipps Universität Marburg

1. Integrated Schema (ER Model)
2. Integration Pipeline Architecture
3. Filling the DB and RMV tables
4. Cluster relevant transfer option candidates using the city associated to the RMV and DB stations
5. Calculate distances within clusters and fill the table if distance is below threshold
6. Identification of identical stations
    1. Filter by distance (lower threshold)
    2. String similarity of the station names
7. Identification of best options

Integrated Schema (ER Model)

Philipps Universität Marburg



Integration Pipeline Architecture

Philipps Universität Marburg


```java
@Bean
public Job process(final JobRepository jobRepository,
                   final JobCompletionNotificationListener listener) {
    return new JobBuilder( name: "importDataJob", jobRepository)
            .incrementer(new RunIdIncrementer()) JobBuilder
            .listener(listener)
            .start(splitImportFlow()) JobFlowBuilder
            .next(splitStageFlow()) FlowBuilder<FlowJobBuilder>
            .next(transferOptionFlow())
            .end() FlowJobBuilder
            .build();
}
```

application.properties

```
spring.batch.jdbc.initialize-schema=always
spring.datasource.url=jdbc:postgresql://docker.local:5432/postgres
spring.datasource.username=postgres
spring.datasource.password=postgres
spring.datasource.driver-class-name=org.postgresql.Driver
spring.jpa.database-platform=org.hibernate.dialect.PostgreSQLDialect
spring.jpa.hibernate.ddl-auto=update
chunk.size=1000
import.db.file.name=DB_Bahnhof_alle.csv
import.rmv.file.name=RMV_Haltestellen.csv
integration.distance.threshold.meters=300
integration.equality.threshold.meters=200
integration.equality.threshold.levenshtein=5
integration.equality.threshold.levenshtein.cityname=6
integration.equality.prefixlength.cityname=4
```

Filling the DB and RMV tables through mapping / calculating the relevant attributes

Philipps Universität Marburg

RMV Station →	Integrated RMV Station
hafasId	
rmvId	
dhid	stationId
hstName	stationName
nameFahrplan	
xIplWert, yIplWert	longitude
xWgs84, yWgs85	longitude, latitude
lno	
gueltigAb, gueltigBis	
verbund1IstgleichRmv	
gemeindename	cityName
land, rp, ortsteilname	
ags values	

DB Station →	Integrated DB Station
evaNr	stationId
ds100	
ifopt	
name	stationName, cityName (extracted from name)
verkehr	
laenge	longitude
breite	latitude
betreiberName	
betreiberNr	
status	

Clustering by city names

Philipps Universität Marburg

- For each DB station we search for RMV stations in the same city
- One word city names like 'Marburg' are easy wins
- Cities like 'Frankfurt a.M.' have different string representations
- If no exact match is found candidates are searched and best is used
  - Candidates...
    - … have common 4 char prefix
    - … have a Levenshtein distance below 6

Identification of similar stations example

Philipps Universität Marburg

RMV	DB
Marburg Hauptbahnhof	Marburg(Lahn)
Marburg Südbahnhof	Marburg Süd
Frankfurt (Main) Hauptwache	Frankfurt(M)Hauptwache

Identification of similar stations

- Take all stations in a specific range as equality candidates
- Take the station names and remove parts between brackets
- Compute three Levenshtein distances and take the minimum
  - Take RMV station name and append 'Bahnhof' to DB station name
  - Take RMV station name and append 'Hauptbahnhof' to DB station name
  - Take RMV station name and DB station name
- If a pair of stations has a Levenshtein distance below a certain threshold assume its equal

Best Options

Philipps Universität Marburg

- All identical station transfers are best options
- The transfer option with the smallest distance which is no identical station transfer is best option
- There may be two best options for one station (one identical and one non-identical)