

Merging Of Overlapping GTFS-Feeds

Bachelor Thesis Presentation

Albert-Ludwigs-Universität Freiburg



**UNI
FREIBURG**

03.05.2019

Technische Fakultät, Freiburg im Breisgau

Leo Zeches

Albert-Ludwigs-Universität Freiburg

Overview



1. Introduction
2. Different kinds of overlapping trips
3. Implementation
4. Evaluation and results

1. Introduction: GTFS-Feed



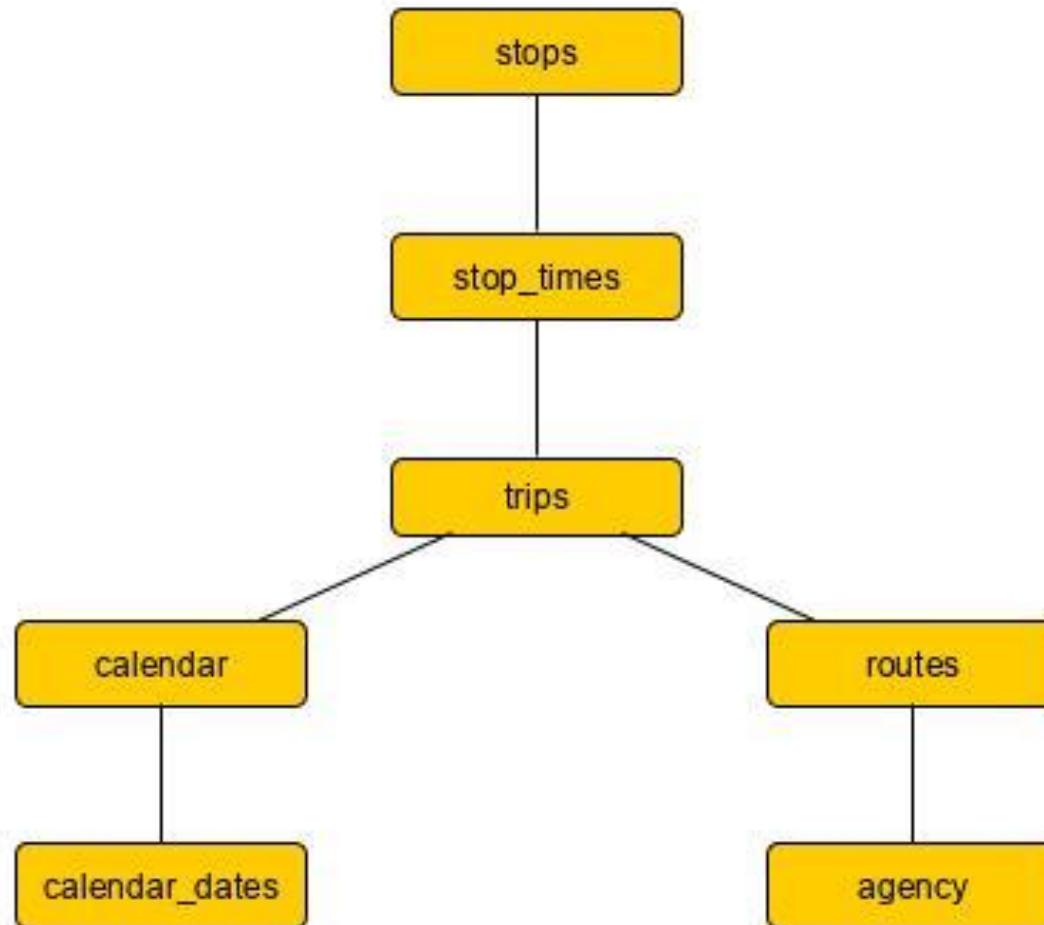
- General Transit Feed Specification
- Zip-file containing csv text files
- 7 required files, 6 optional files
- Entries contain references to other files

Example: stop_times.txt



```
trip_id , arrival_time , departure_time , stop_id , stop_sequence  
STBA,6:00:00 ,6:00:00 ,STAGECOACH,1  
STBA,6:20:00 ,6:20:00 ,BEATTY_AIRPORT,2  
CITY1,6:00:00 ,6:00:00 ,STAGECOACH,1  
CITY1,6:05:00 ,6:07:00 ,NANAA,2  
CITY1,6:12:00 ,6:14:00 ,NADAV,3
```

GTFS Structure



Problem of overlapping feeds



- Different agencies have different GTFS-feeds
- May cover the same area
- Overlapping possible
- Problems when merging overlapping feeds

Example: Two overlapping Feeds



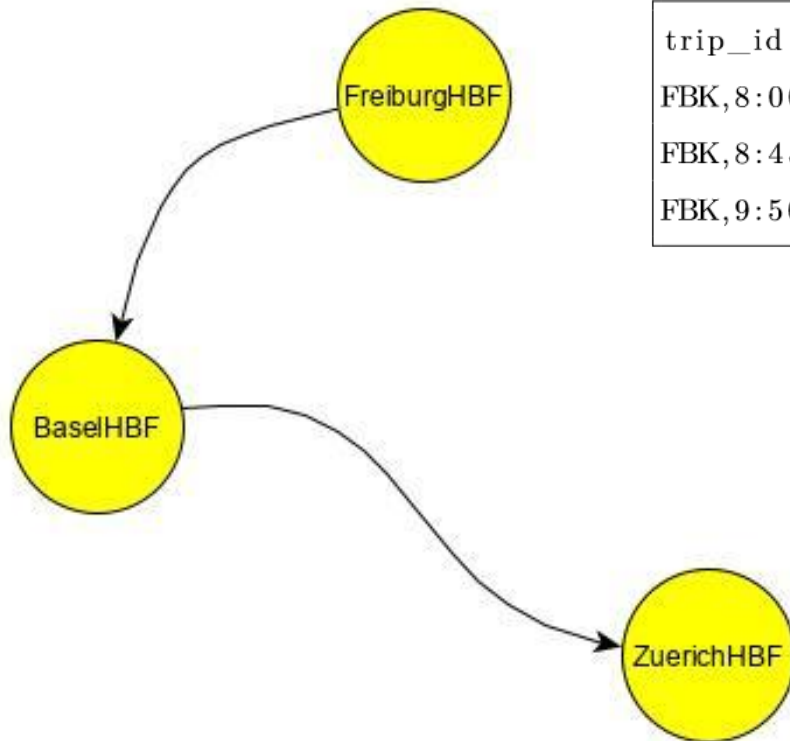
```
trip_id , arrival_time , departure_time , stop_id , stop_sequence
FBK,8:00:00 ,8:05:00 ,FreiburgHBF ,1
FBK,8:45:00 ,8:50:00 ,BaselHBF ,2
FBK,9:50:00 ,10:00:00 ,ZuerichHBF ,3
```

Listing 1.1: Example for *stop_times.txt* for a GTFS feed *A*

```
trip_id , arrival_time , departure_time , stop_id , stop_sequence
BaZue ,8:45:00 ,8:50:00 ,Basel ,1
BaZue ,9:50:00 ,10:00:00 ,Zuerich ,2
```

Listing 1.2: Example for *stop_times.txt* for a GTFS feed *B*

Example: Two overlapping Feeds

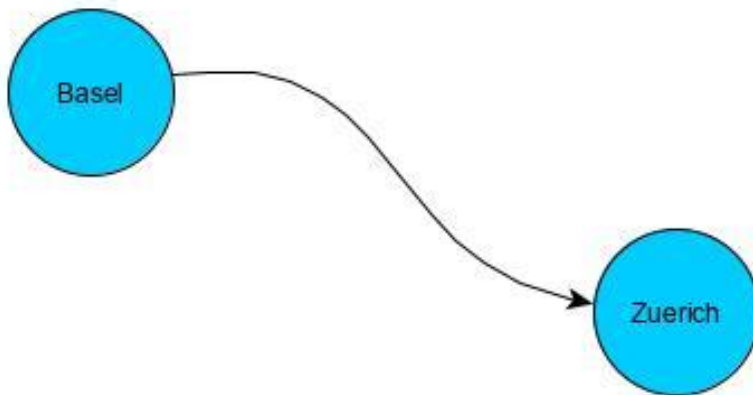


trip_id	arrival_time	departure_time	stop_id	stop_sequence
FBK	8:00:00	8:05:00	FreiburgHBF	1
FBK	8:45:00	8:50:00	BaselHBF	2
FBK	9:50:00	10:00:00	ZuerichHBF	3

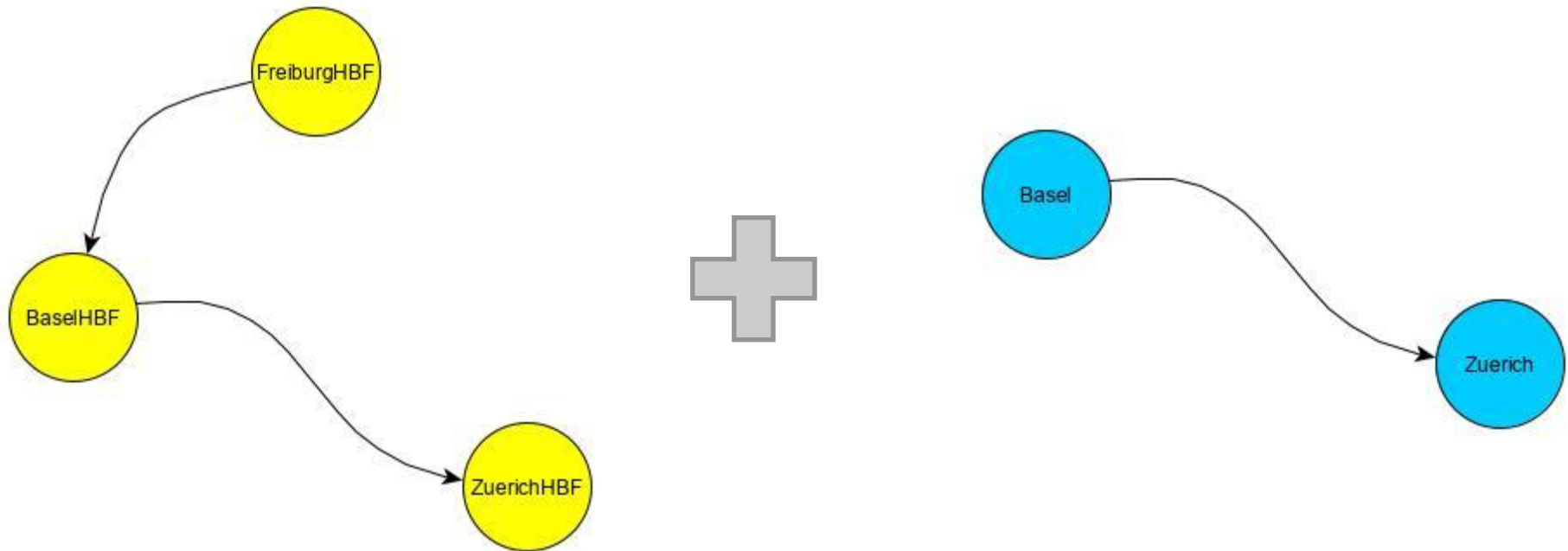
Example: Two overlapping Feeds



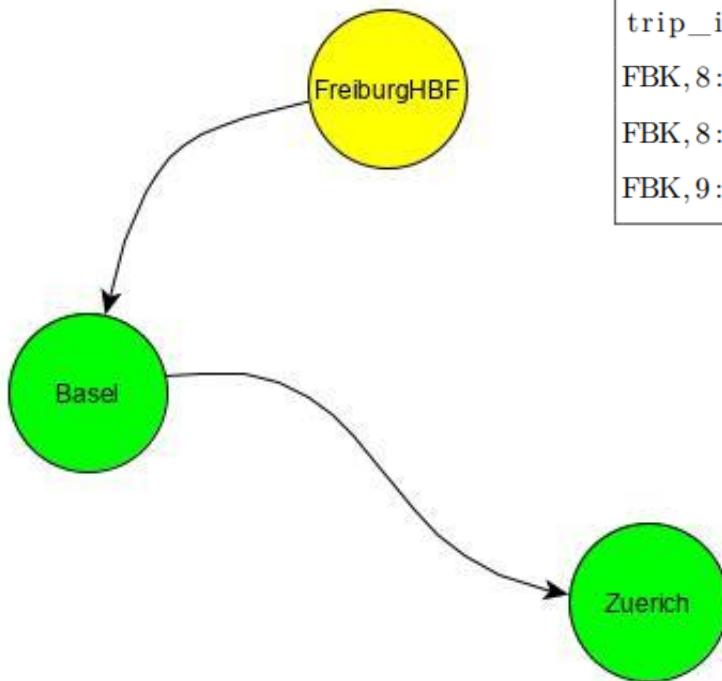
```
trip_id , arrival_time , departure_time , stop_id , stop_sequence  
BaZue , 8:45:00 , 8:50:00 , Basel , 1  
BaZue , 9:50:00 , 10:00:00 , Zuerich , 2
```



Example: Two overlapping Feeds



Example: Two overlapping Feeds

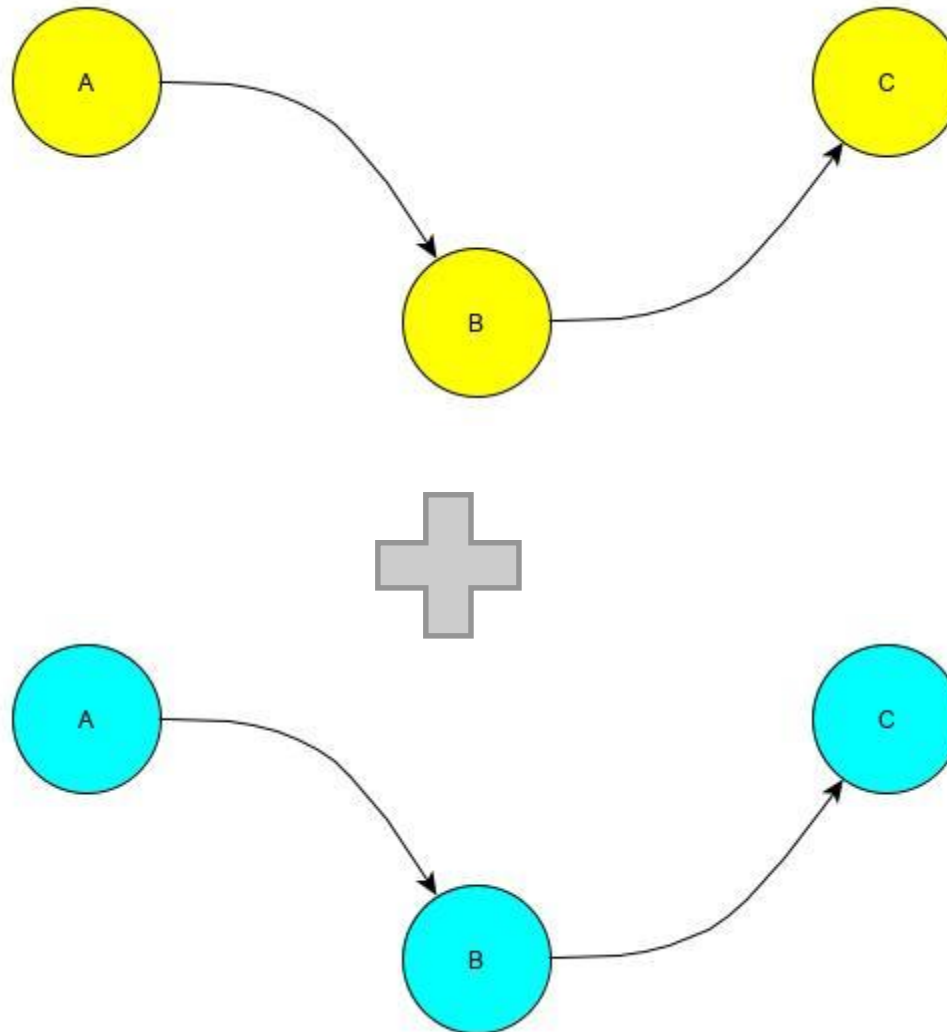


```
trip_id , arrival_time , departure_time , stop_id , stop_sequence  
FBK,8:00:00 ,8:05:00 ,FreiburgHBF ,1  
FBK,8:45:00 ,8:50:00 ,Basel ,2  
FBK,9:50:00 ,10:00:00 ,Zuerich ,3
```

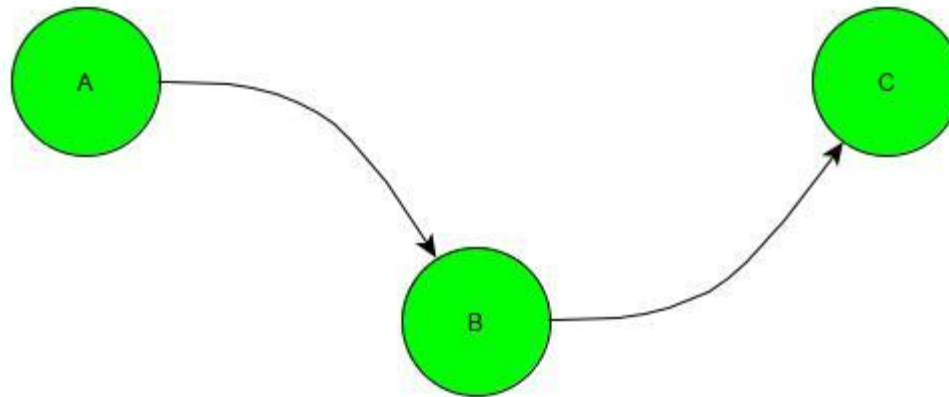
2. Different Kinds of Trips:



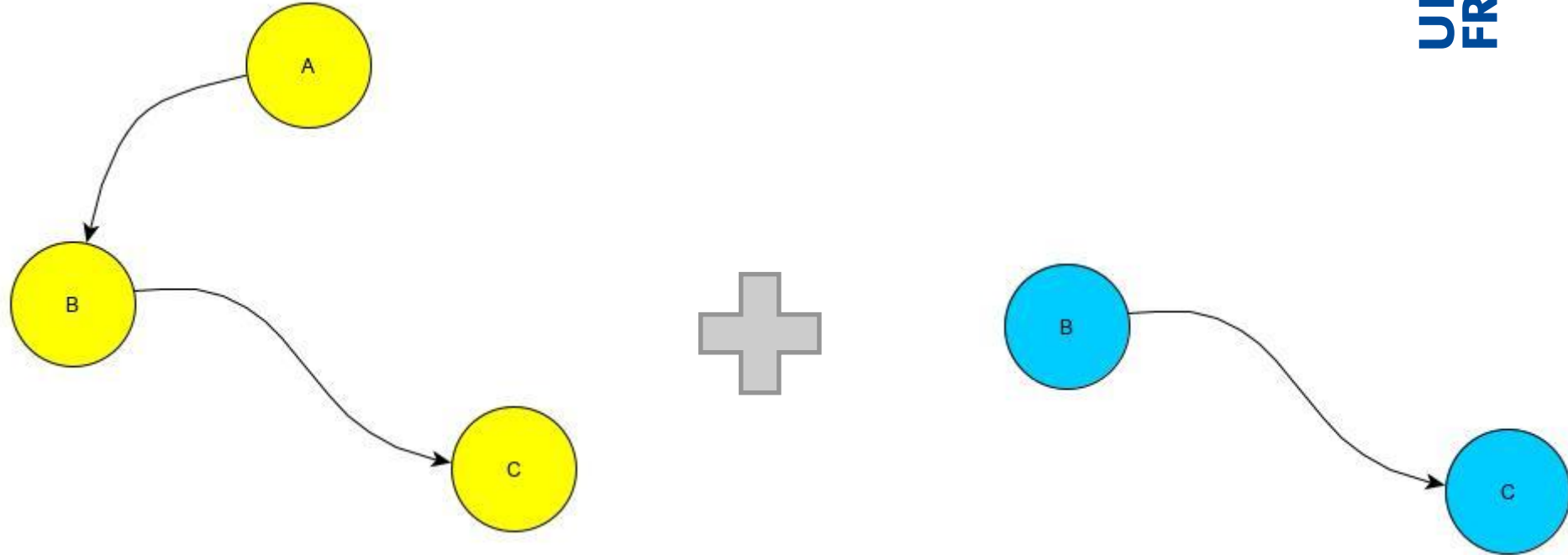
2.A Equivalent Trips



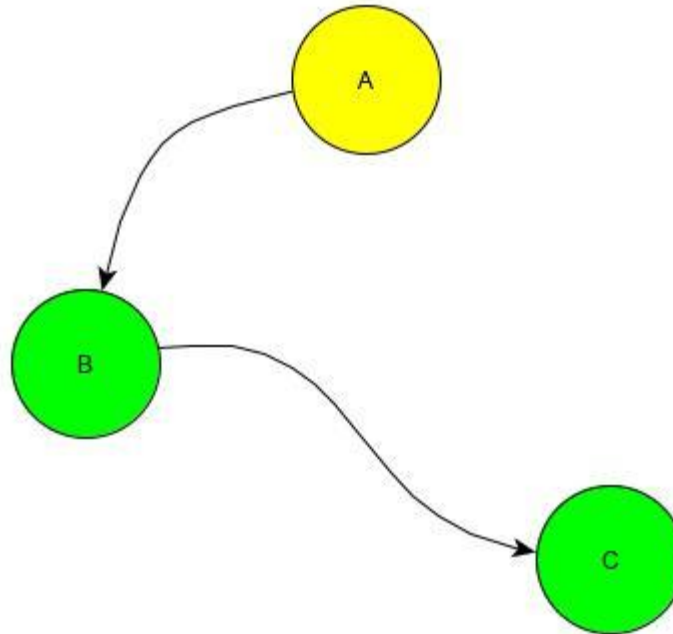
2.A Equivalent Trips



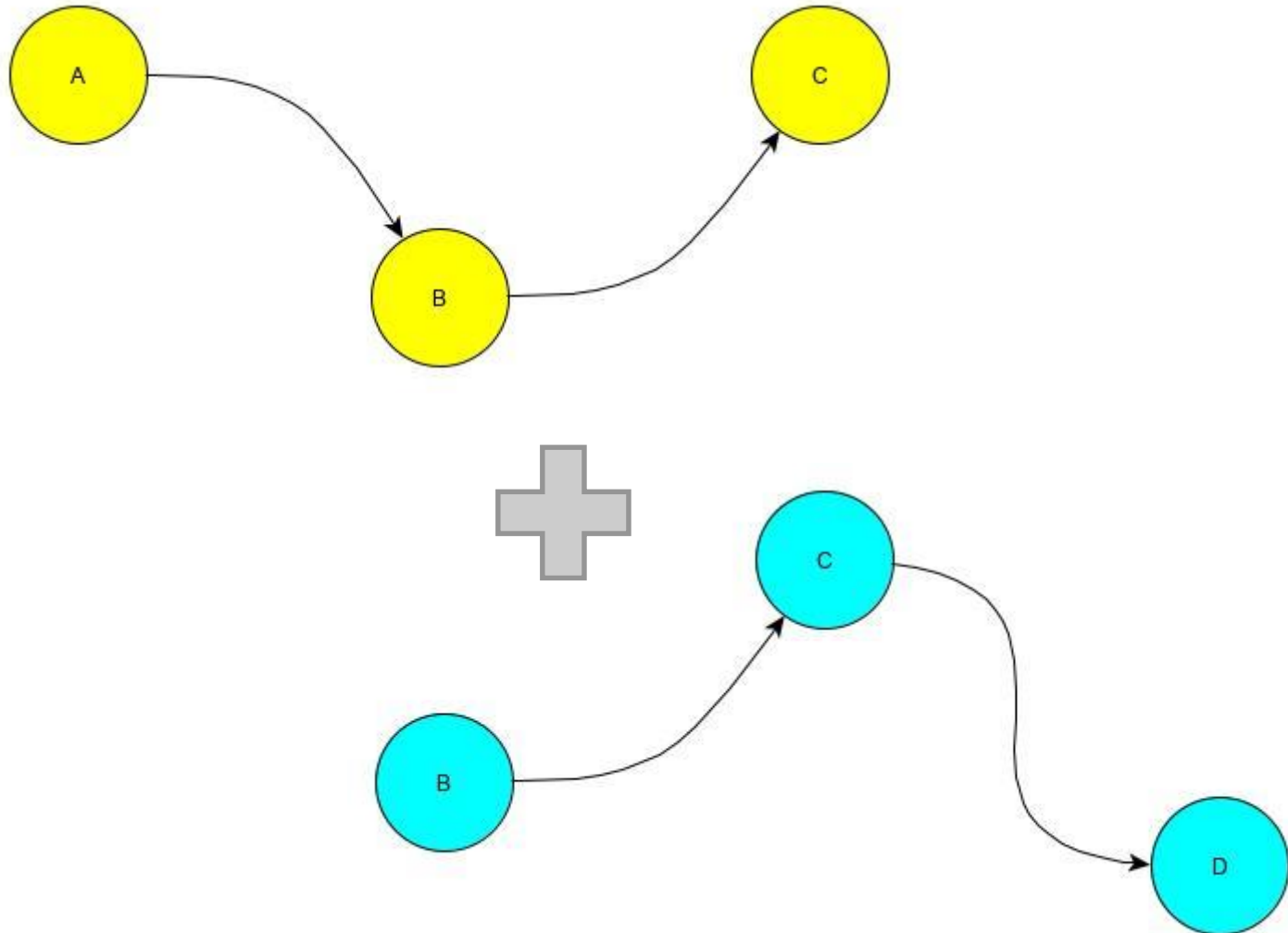
2.B Included Trips



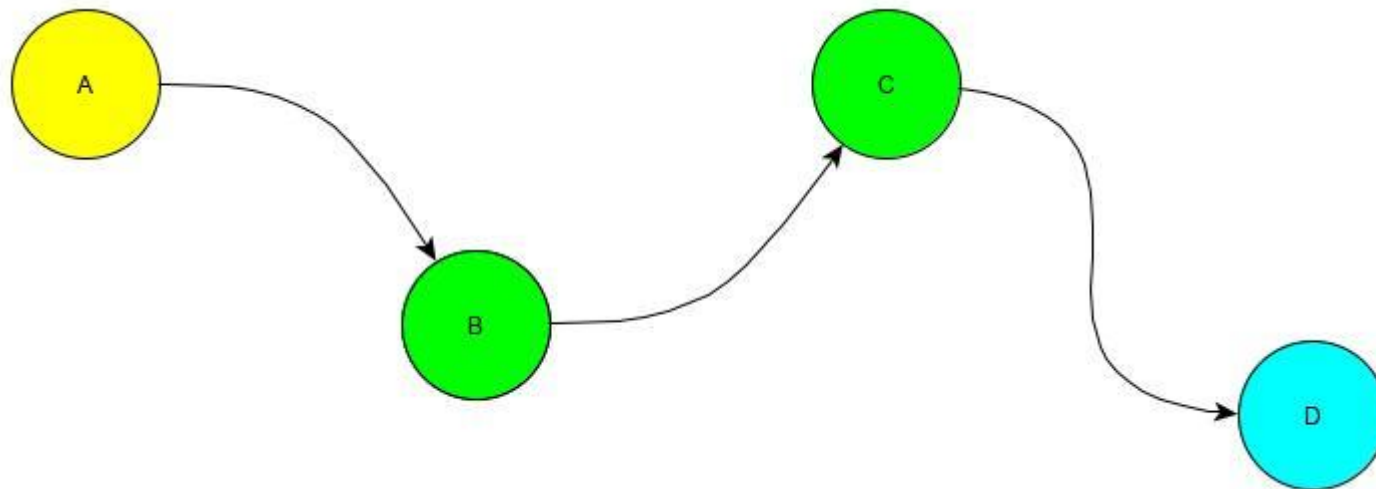
2.B Included Trips



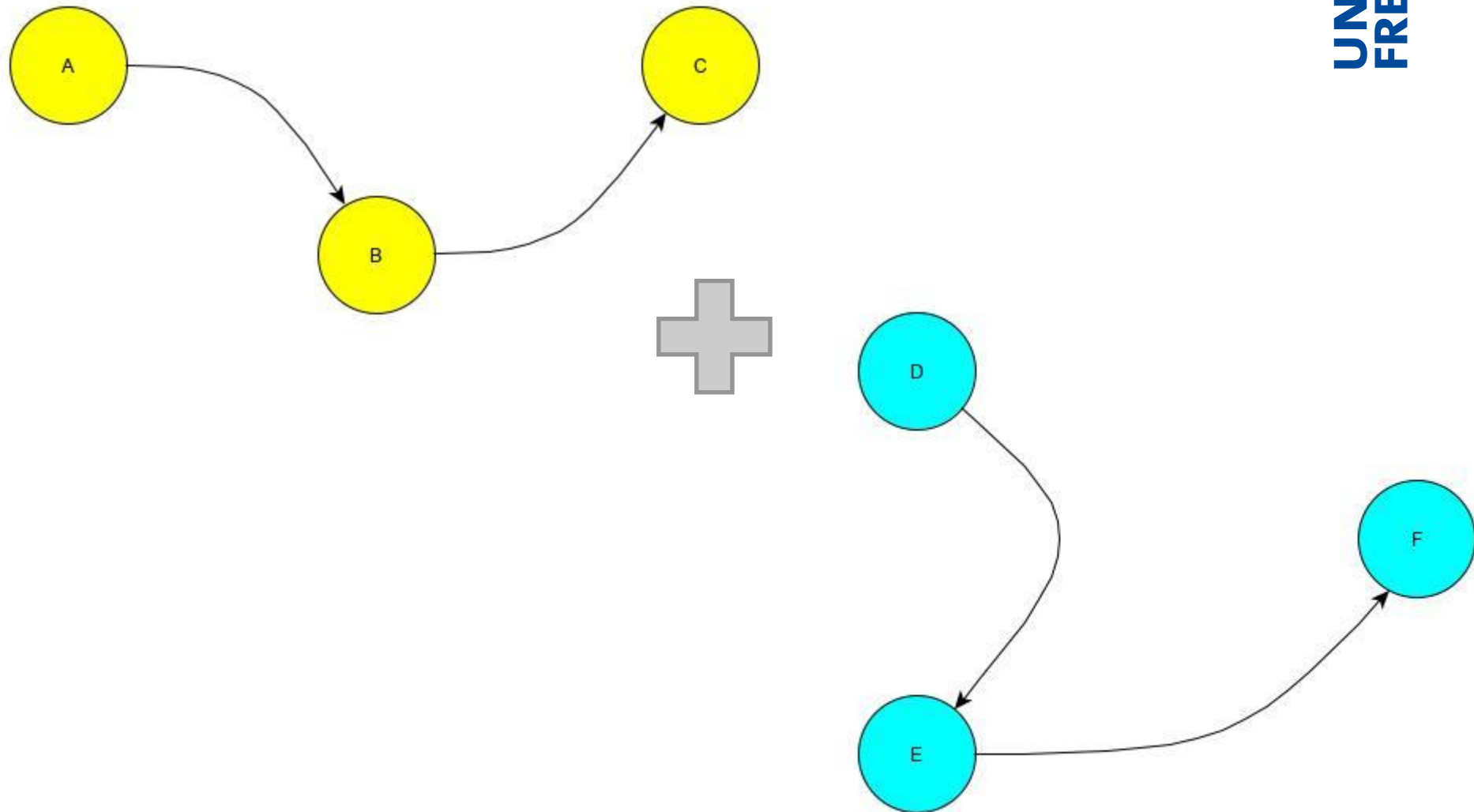
2.C Partially Included Trips



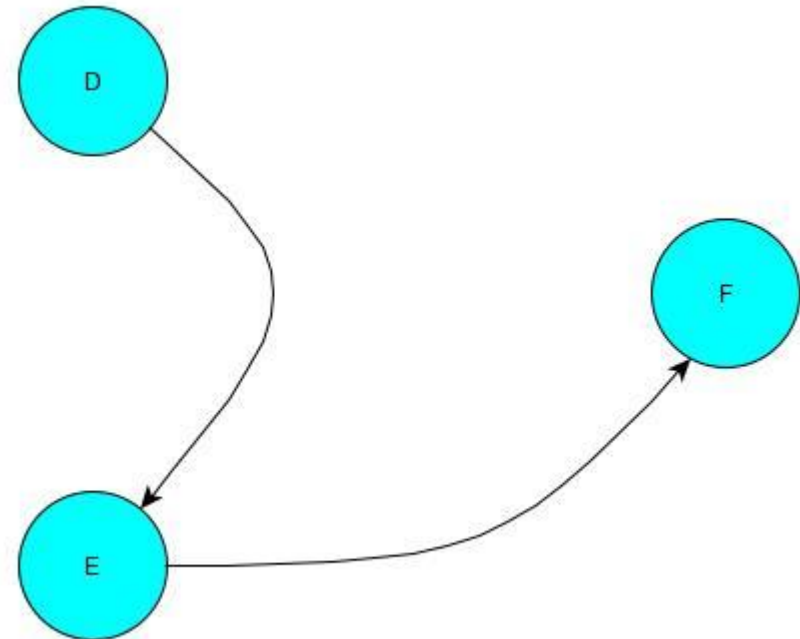
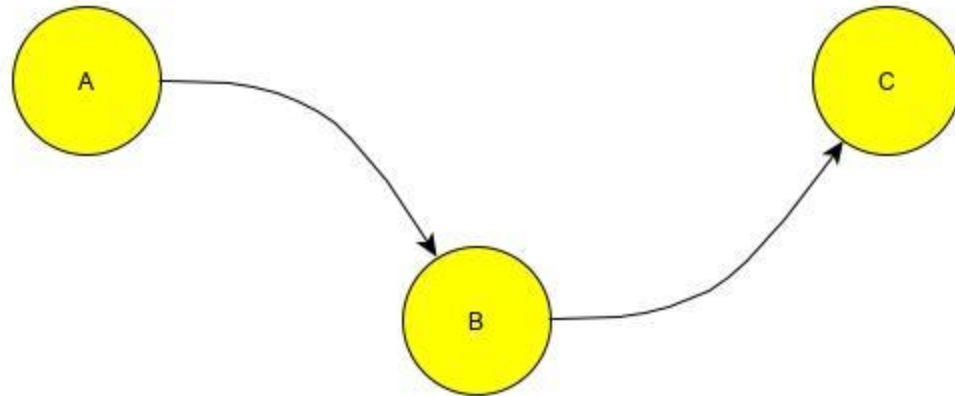
2.C Partially Included Trips



2.D No overlap



2.D No overlap

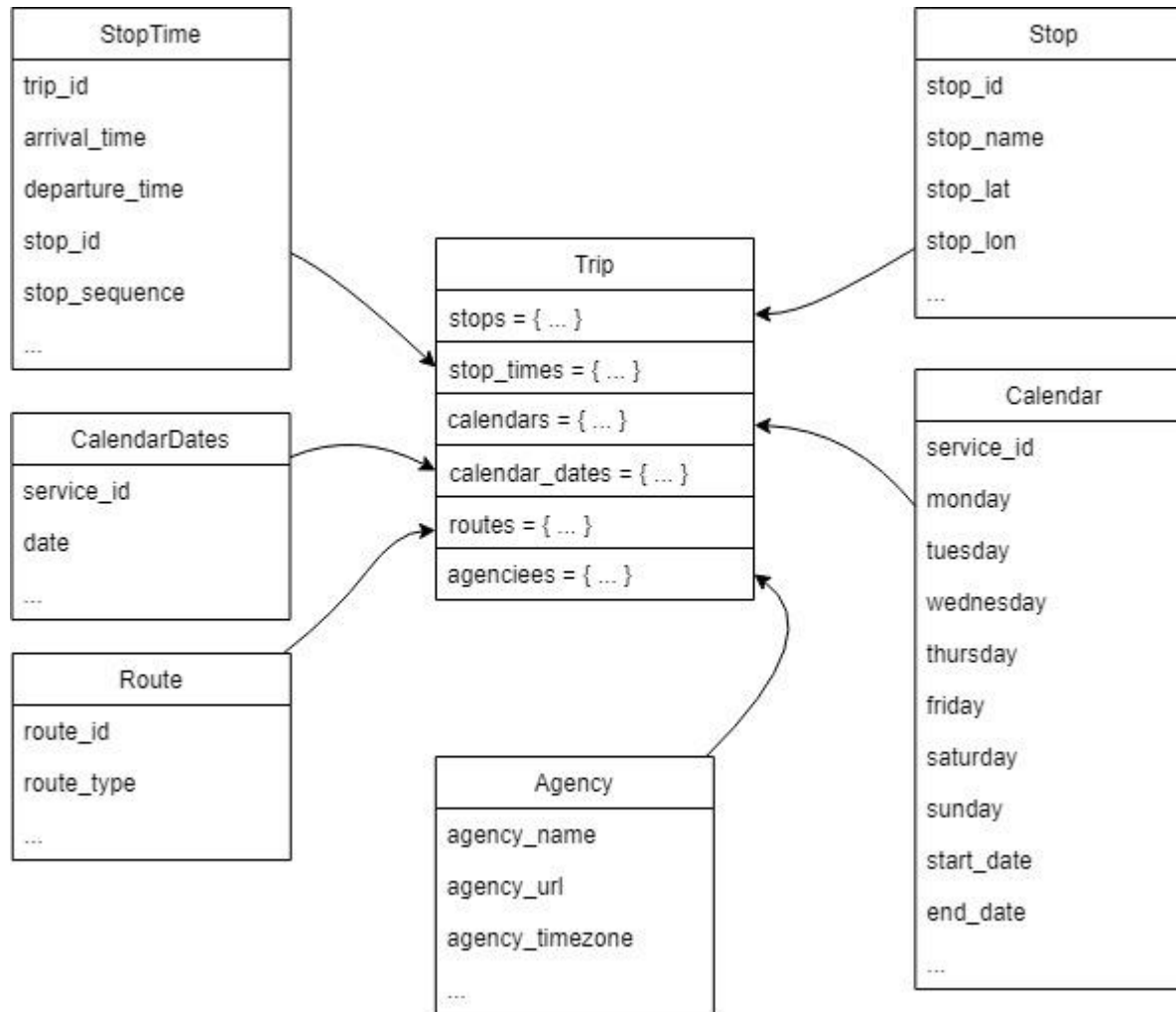


3.1 Data Structures



- Specific objects for each file
- Trip objects contain dictionaries with other objects
- Comparisons of trips between calendars and stop_data tuples

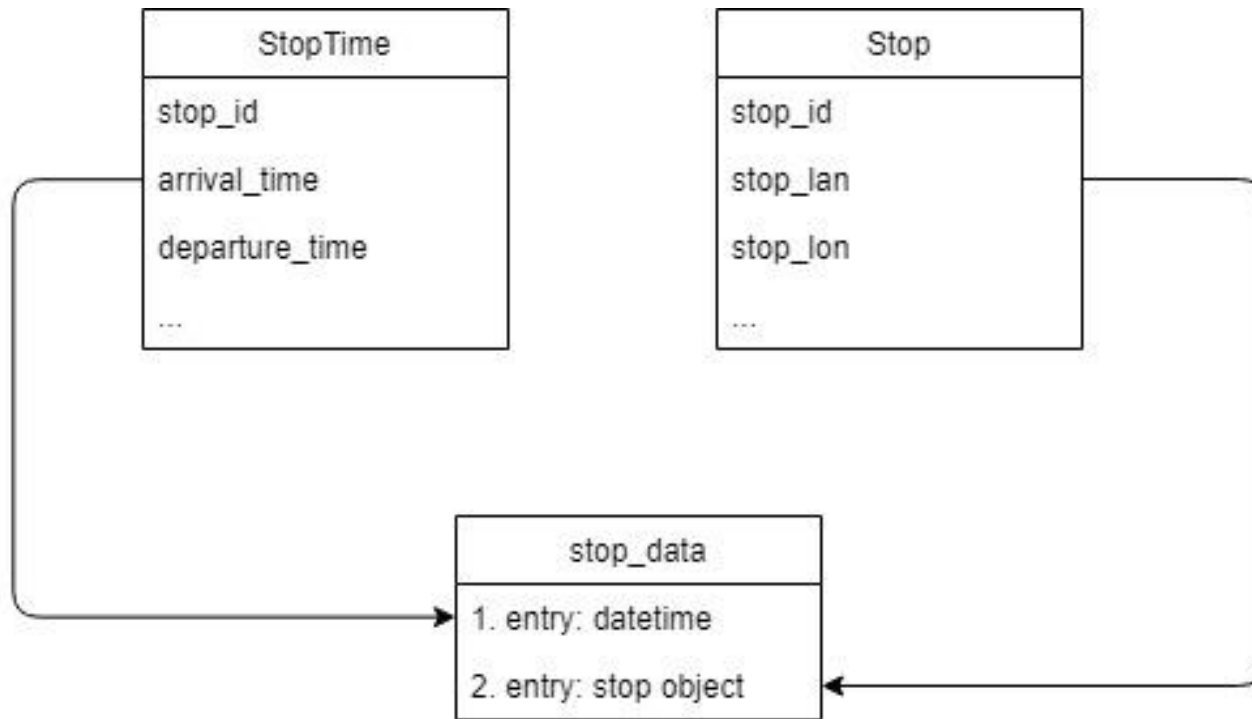
3.1 Data Structures



Stop Data Tuples



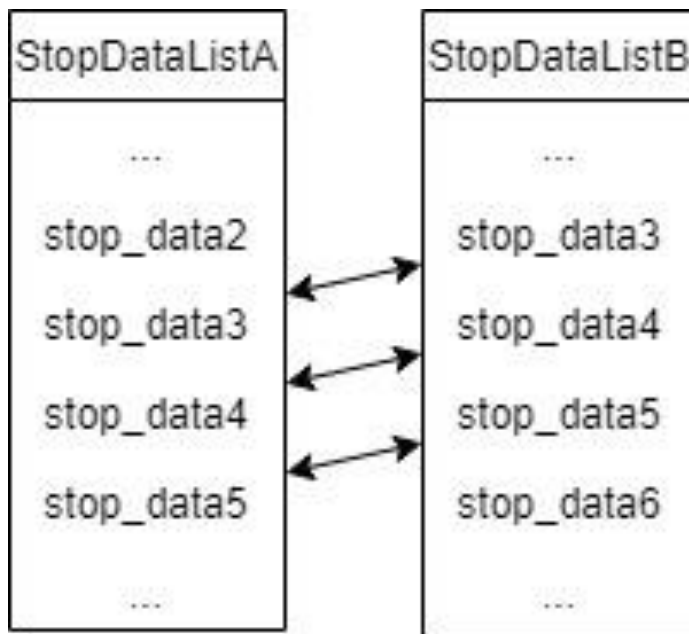
- Containing datetime object and stop object
- Used to compare the course of two trips



3.2 Comparisons between trips



- Seperate comparisons between stop data and calendars



	Mo	Tu	We	Th	Fr	Sa	Su		
calendarA	1	1	1	1	1	0	0	start_date	end_date
	↑	↑	↑	↑	↑	↑	↑	20170101	20200101
calendarB	0	0	0	0	1	1	1		
	↓	↓	↓	↓	↓	↓	↓	20170101	20180101

3.3 Optimization

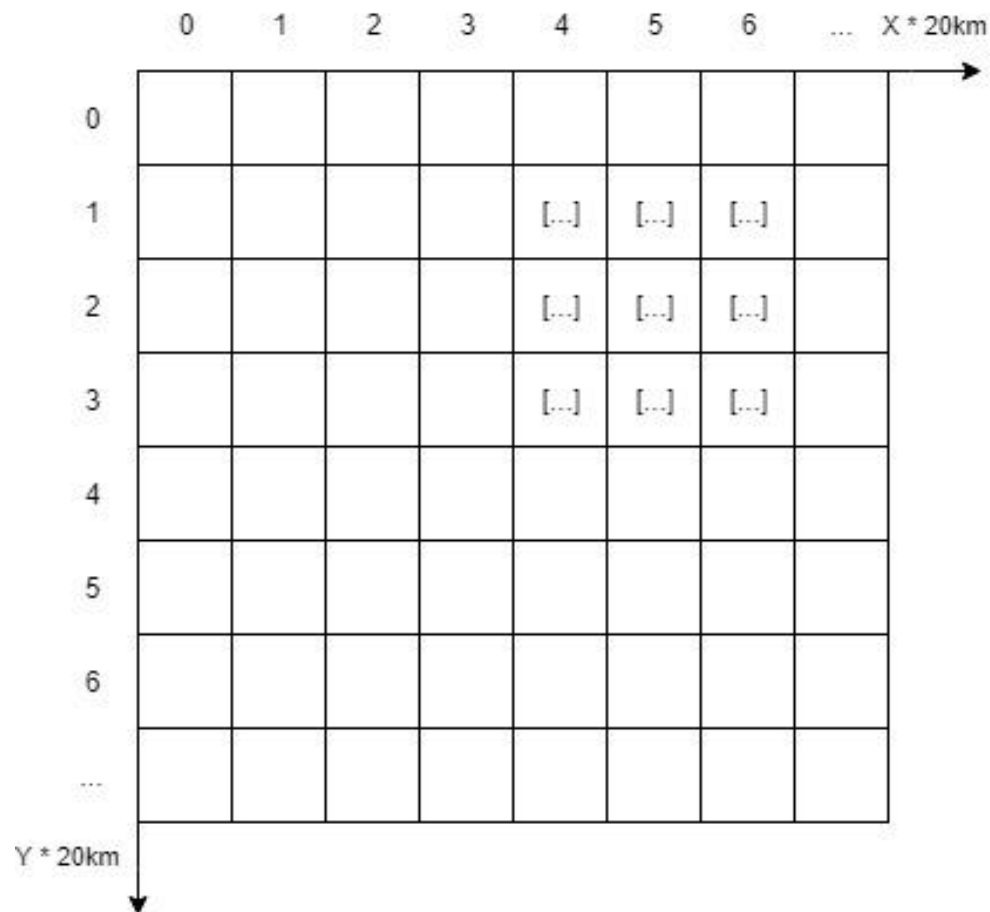


3.3 Optimization: 2D-Grid structure

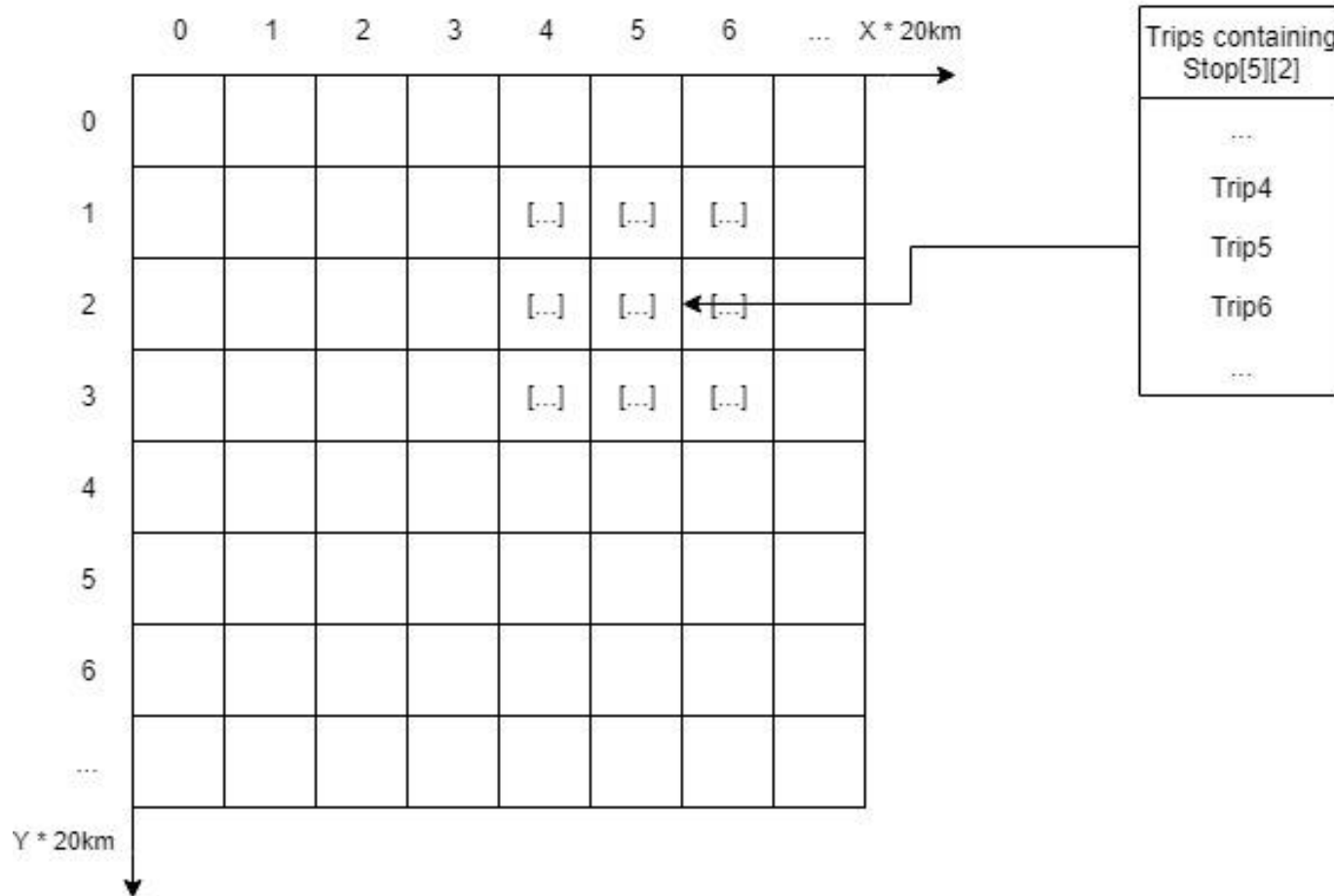


- <https://freegeographytools.com/2009/google-earth-coordinate-system-grids>

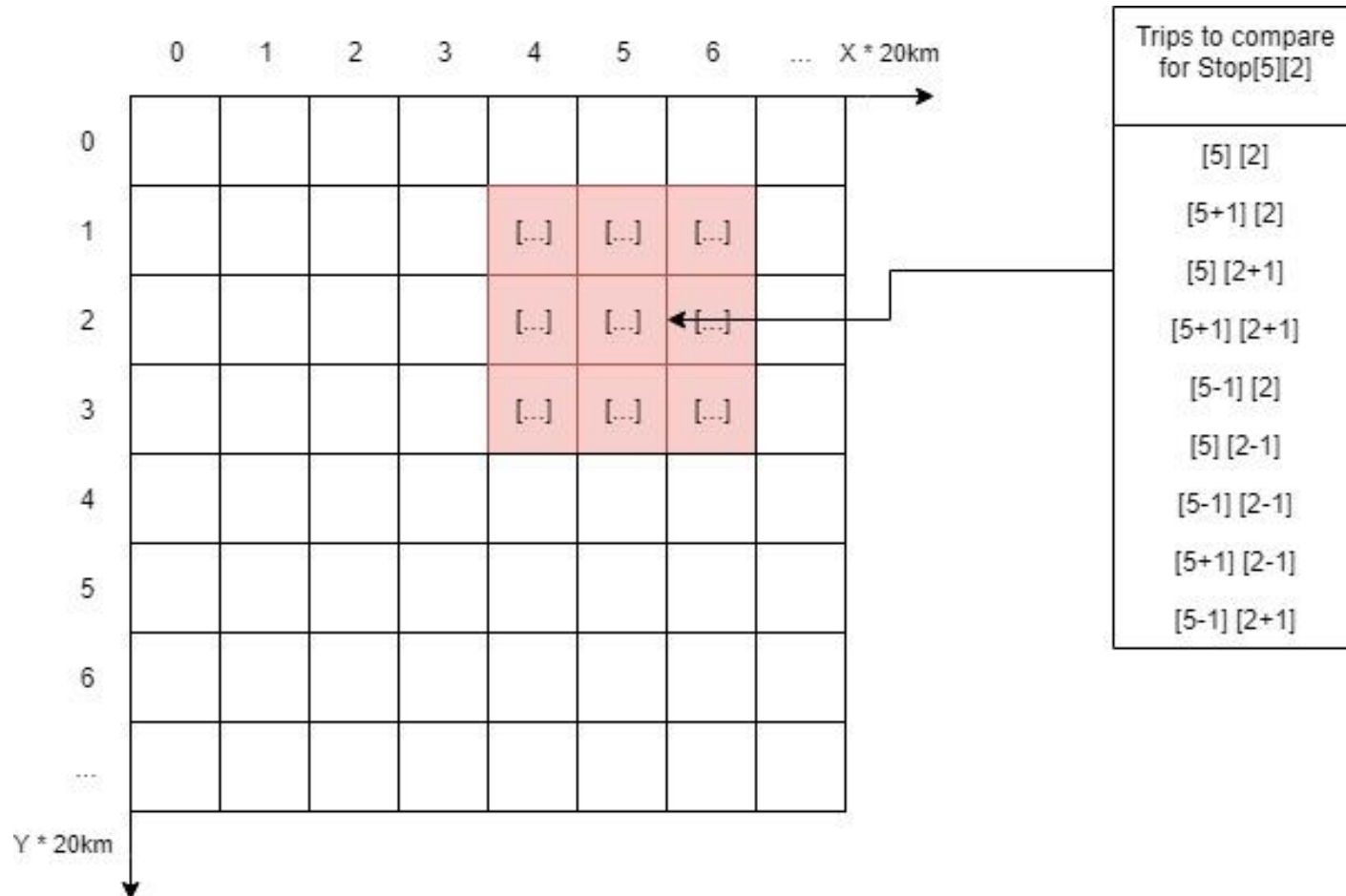
3.3 Optimization: 2D-Grid structure



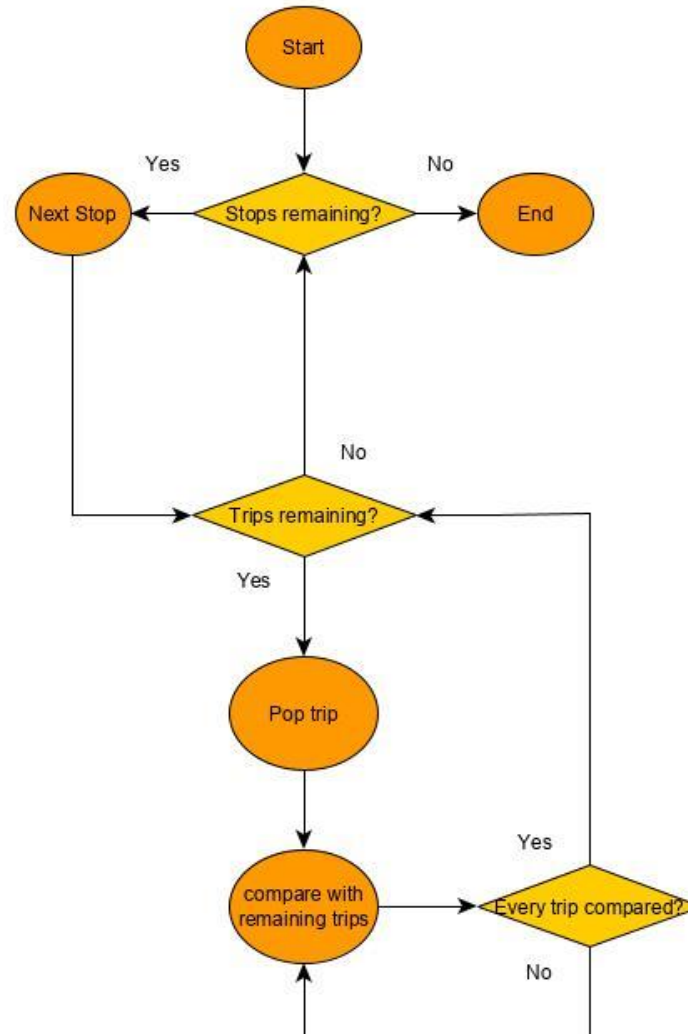
3.3 Optimization 2D-Grid structure



3.3 Optimization: 2D-Grid structure



Flowchart

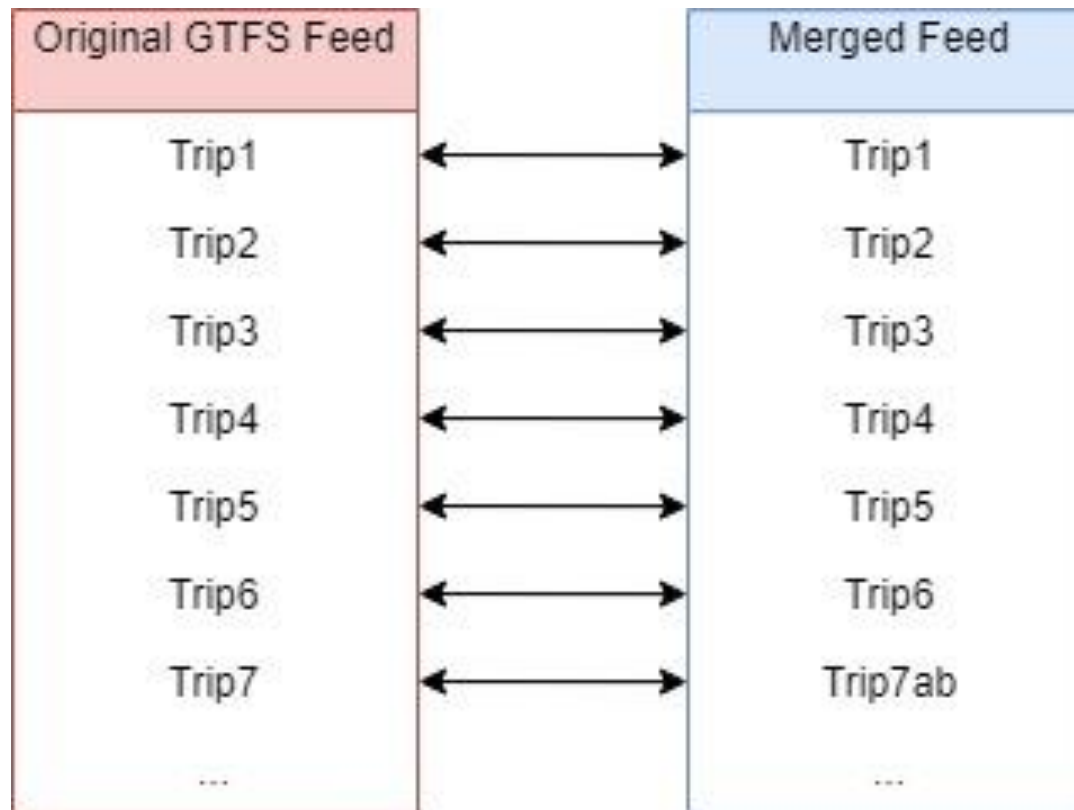


4. Evaluation



- Evaluation using a ground truth
- No „real“ ground truth possible
- Creation of a „fake“ ground truth
- Splitting and merging of an existing feed
- Possibility to add „random“ noise to trips

4. Evaluating



4. Results



GTFS-Feed 1: sample-feed.zip

file	lines
trips.txt	11
stops.txt	9
stop_times.txt	28
calendar.txt	2
calendar_dates.txt	1
routes.txt	5
agency.txt	1

Mode	0	1	2	3
Noise 0	5.714%	5.714%	0.000%	0.000%
Noise 1	5.714%	5.714%	0.000%	0.000%
Noise 2	78.571%	78.571%	45.946%	45.946%

running time (milliseconds)	38.86914
memory usage (MiB)	36.1797

Table 1: Evaluation of sample-feed.zip with grid optimization

4. Results



- Running time for sample-feed.zip:
 - With grid optimization: 38.86914 ms
 - Without grid optimization: 47.87111 ms

4. Results



GTFS-Feed 2: chilliwack_premerged.zip

file	lines
trips.txt	462
stops.txt	291
stop_times.txt	11512
calendar.txt	4
calendar_dates.txt	1
routes.txt	10
agency.txt	1

Mode	0	1	2	3
Noise 0	2.414%	2.385%	0.488%	0.454%
Noise 1	2.539%	2.491%	0.548%	0.470%
Noise 2	37.766%	37.696%	16.088%	15.934%

running time (milliseconds)	16633.26716
memory usage (MiB)	47.6719

Table 2: Evaluation of chilliwack_premerged.zip
with grid optimization

4. Results



- Running time for sample-feed.zip:
 - With grid optimization: 16633.26 ms
 - Without grid optimization: 77048.06 ms



Thank you

Questions?

- Different Comparison Modes:
 - Mode 0: only equivalent trips
 - Mode 1: only equivalent + included trips
 - Mode 2: only equivalent + partially included trips
 - Mode 3: every kind of trip

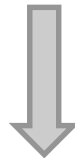
- Different Noise Levels:
 - Level 0: No noise at all
 - Level 1: Small amount of noise
 - Level 2: Large amount of noise

- Noise creation:
 - Change names of stop objects
 - Change coordinates of stop objects
 - Change arrival_time and departure_time of stop time objects
 - Change of trip_id

- Noise Level 1:
 - 30% chance of noise
 - stop_name will be mixed up
 - arrival_time and departure_time will be changed (max 2 min. difference)

- Example Noise Level 1:

```
stop_id ,stop_name ,stop_lat ,stop_lon  
BEATTY_AIRPORT,Nye County Airport ,36.868446,-116.784582  
BULLFROG,Bullfrog ,36.88108,-116.81797  
STAGECOACH,Stagecoach Hotel & Casino ,36.915682,-116.75167  
NANAA,North Ave / N A Ave,36.914944,-116.761472
```

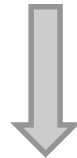


```
stop_id ,stop_name ,stop_lat ,stop_lon  
BEATTY_AIRPORT,yN eoCnuytA rioptr ,36.868446,-116.784582  
BULLFROG,uBllrfgo ,36.88108,-116.81797  
STAGECOACH,tSgaceaohcH tole& C sani oD,36.915682,-116.751677  
NANAA,oNtr hvA e / N AvA e,36.914944,-116.761472
```

- Noise Level 2:
 - 80% chance of noise
 - Difference in arrival_time and departure_time of up to 10 min
 - Changes to trip_id
 - Changes to stop coordinates

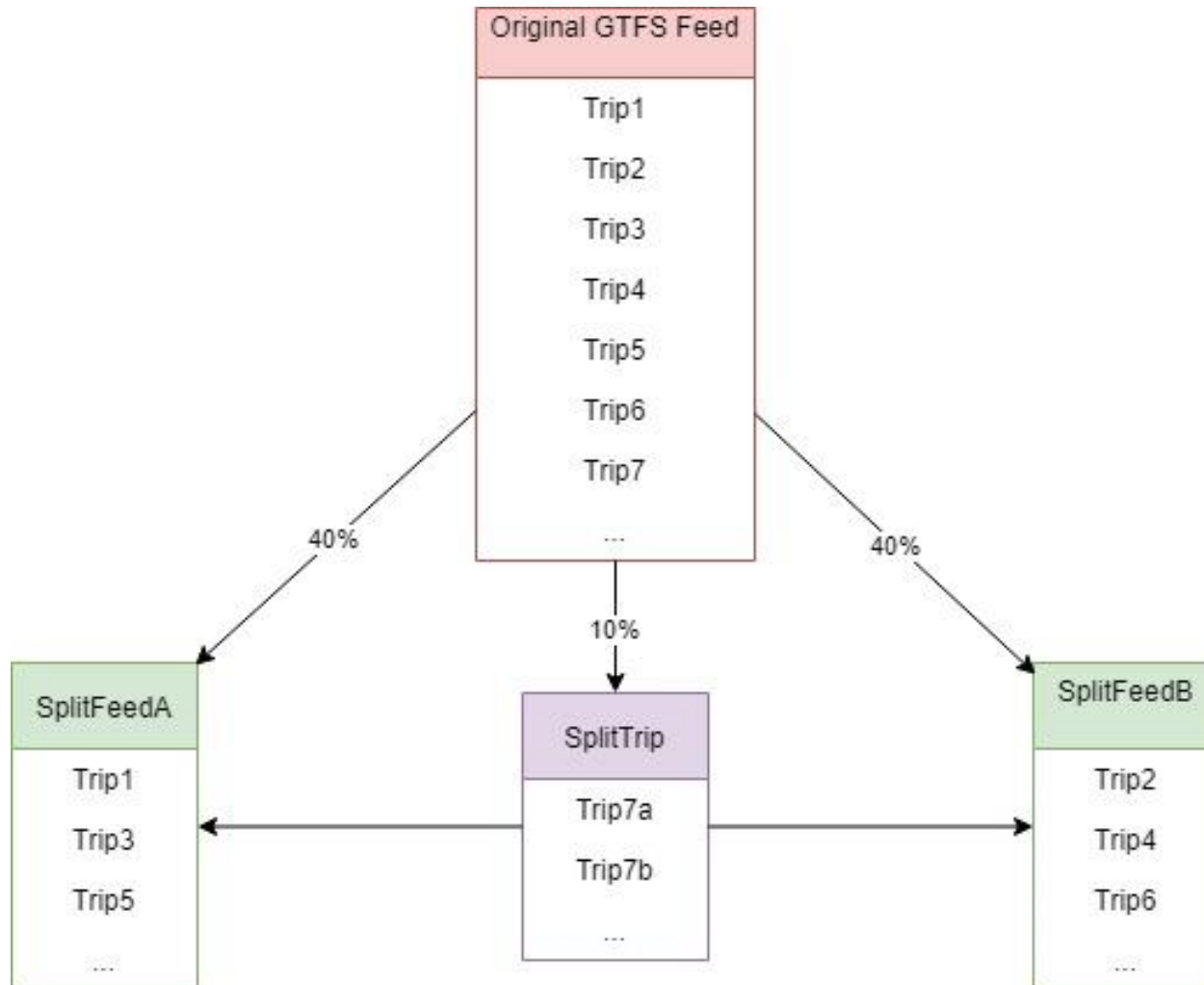
- Example Noise Level 2:

```
stop_id,stop_name,stop_lat,stop_lon  
BEATTY_AIRPORT,Nye County Airport,36.868446,-116.784582  
BULLFROG,Bullfrog,36.88108,-116.81797  
STAGECOACH,Stagecoach Hotel & Casino,36.915682,-116.75167  
NANAA,North Ave / N A Ave,36.914944,-116.761472
```

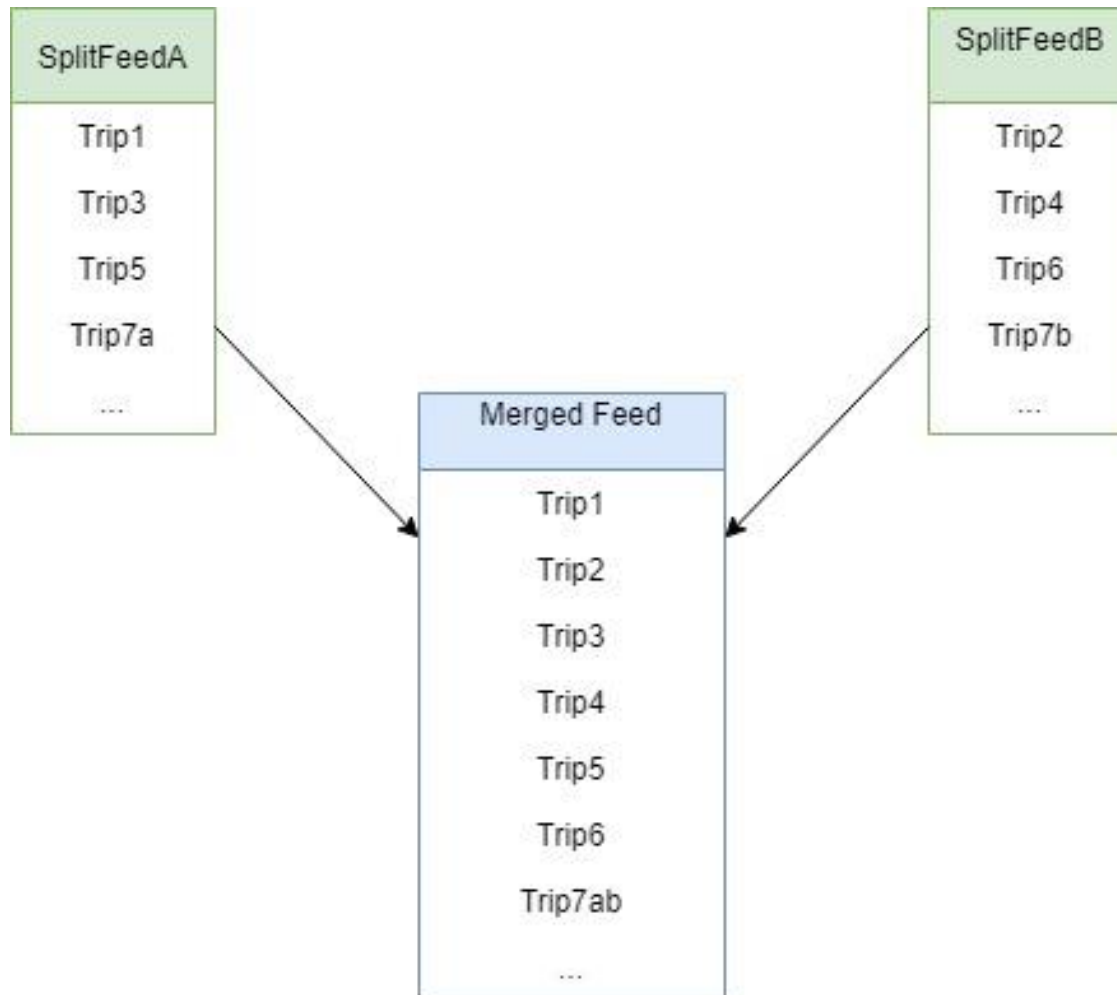


```
stop_id,stop_name,stop_lat,stop_lo  
BEATTY_AIRPORT,yN eoCnuytA rioptr,36.870497335,-116.782659944  
BULLFROG,uBllrfgo,36.881965733,-116.816864765  
STAGECOACH,tSgaceaohcH tole& C sani oD,36.91588751,-116.75167  
NANAA,oNtr hvA e / N AvA e,36.914944,-116.761472
```

Backup Slides



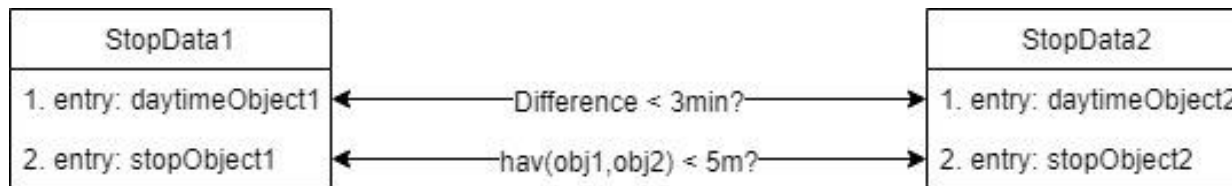
Backup Slides



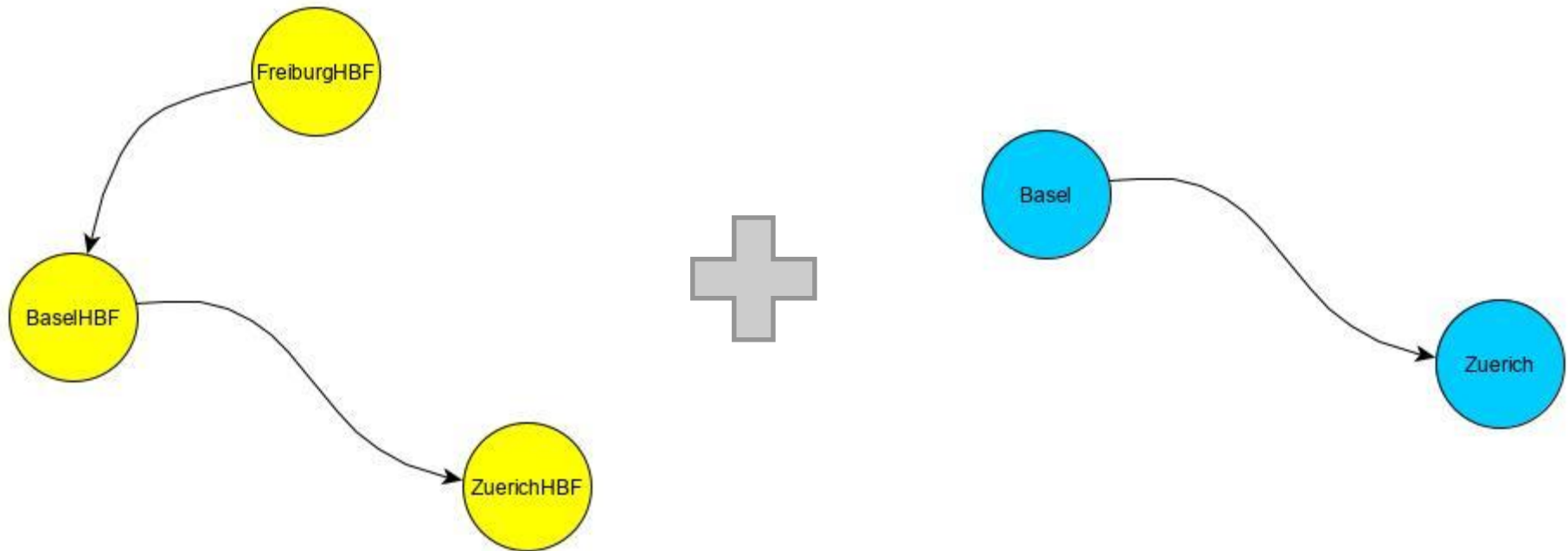
Compare Stop Data:

- Calculate distance between stops using haversine formula
- Calculate difference of arrival time
- StopData tuples are equal if:
 - Distance between stops $< 5\text{m}$
 - Difference in arrival time $< 3\text{min}$

- Comparisons between Stop Data Tuples



Backup Slides



Comparison Example:

- StopData1: BaselHBF
 - arrival_time1 = 8:45:00
 - stop_lat1 = 47.559601
 - stop_lon1 = 7.588576
- StopData2: Basel
 - arrival_time2 = 8:46:25
 - stop_lat2 = 47.559611
 - stop_lon2 = 7.588576

Comparison Example:

- $\text{abs}(\text{daytime}(8,45,0) - \text{daytime}(8,46,25))$
 $= 1,25 \text{ min} < 3 \text{ min}$
- $\text{haversine}(\text{stop_lat1}, \text{stop_lon1}, \text{stop_lat2}, \text{stop_lon2})$
 $= 1.112 \text{ m} < 5 \text{ m}$

Backup Slides



Feed db_fv_premerged.zip

file	lines
trips.txt	6925
stops.txt	714
stop_times.txt	102100
calendar.txt	2557
calendar_dates.txt	1
routes.txt	1485
agency.txt	14
Size zipped:	0.672 MB
Size unzipped:	4,28 MB

Feed ch_fv.zip

file	lines
trips.txt	17625
stops.txt	5897
stop_times.txt	161471
calendar.txt	12455
calendar_dates.txt	12346
routes.txt	1544
agency.txt	62
Size zipped:	1,21 MB
Size unzipped:	8,28 MB

Backup Slides



Merged Feed: "ch_fv + db_fv_premerged.zip"

file	lines
trips.txt	11448
stops.txt	1862
stop_times.txt	142879
calendar.txt	3454
calendar_dates.txt	1178
routes.txt	2722
agency.txt	76
running time (milliseconds)	38'640'449.180
running time (minutes)	644.007
memory usage (MiB)	4445.705
Size zipped:	1.7 MB
Size unzipped:	6.43 MB

Table 3: Merging of ch_fv.zip and
db_fv_premerged.zip

Backup Slides



Feed fraser_valley_feed.zip

file	lines
trips.txt	483
stops.txt	243
stop_times.txt	10461
calendar.txt	5
calendar_dates.txt	0
routes.txt	15
agency.txt	1
Size zipped:	0.118 MB
Size unzipped:	0.845 MB

Feed comox_valley_feed.zip

file	lines
trips.txt	2927
stops.txt	638
stop_times.txt	102075
calendar.txt	7
calendar_dates.txt	2
routes.txt	23
agency.txt	1
Size zipped:	0.872 MB
Size unzipped:	6,03 MB

Backup Slides



Merged Feed: 'commox_valley_feed + fraser_valley_feed.zip'

file	lines
trips.txt	1576
stops.txt	878
stop_times.txt	51350
calendar.txt	12
calendar_dates.txt	1
routes.txt	38
agency.txt	2
running time (milliseconds)	2'555'623.13175
running time (minutes)	42.59
memory usage (MiB)	285.4
Size zipped:	0.393 MB
Size unzipped:	3.54 MB

Table 4: Merging of commox_valley_feed.zip and fraser_valley_feed.zip