Predicting tram delays based on weather

Tests to The Transaltlantic Scooters' Project's Architecture

Laura Bakała
Wojciech Kretowicz
Karol Pysiak
Mateusz Szysz

16 January 2022

Object	Test of	Expected result	Actual result
NiFi flow for weather data	whole NiFi flow for weather data	correct flow through all nodes	Figure 1.
table with weather data in HBase	storing weather data in HBase	temperature, air pressure etc. in 3 hour intervals with timestamp of collecting data in HBase table	Figure 2.
weather table in Google Cloud MySQL	adding new records to the speed layer when NiFi flow is executed	processed weather data (with 0s instead of nulls)	Figure 3.
table with trams positions in HBase	storing data in HBase	longitude and latitude of each tram in given moment of time with timestamp of collecting data	Figure 4.
table with timetable in HBase	storing data in HBase	position of the tram stop with lines and directions of trams with timestamp of collecting data	Figure 5.
table with consecutive iterations of rides of each tram in PySpark	creating an order of bus stop (we don't load order from any external source, but we have to extract it from the timetable)	list of consecutive bus stops for a given line for the first iteration at given day	Figure 6.
First Kafka in tram positions NiFi flow	Kafka topic content	unprocessed data with tram positions	Figure 7.
Second Kafka in tram positions NiFi flow	Kafka topic content	processed data with tram positions	Figure 8.
First Kafka in weather flow	Kafka topic content	unprocessed data with weather conditions	Figure 9.
Second Kafka in weather flow	Kafka topic content	processed data with weather conditions	Figure 10.
Kafka with weather aggregates	Kafka topic content	processed data with weather conditions with aggregation	Figure 11.
Kafka with tram lines	Kafka topic content	record with tram stop data	Figure 12.

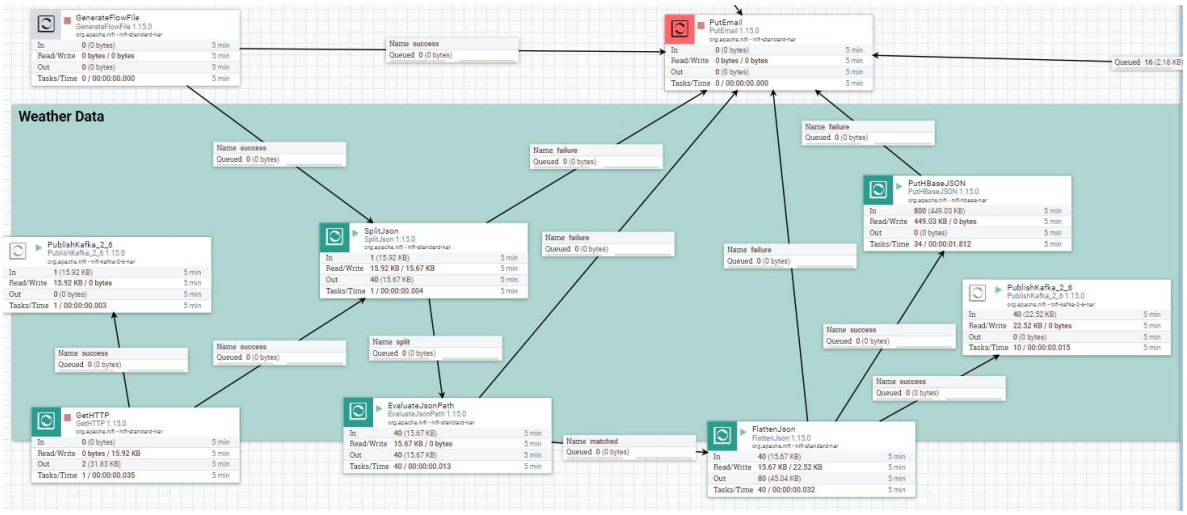

Figure 1: NiFi flow for weather.

```
1642777200                                column=weather_data:main.sea_level, timestamp=1642364300887, value=1011
1642777200                                column=weather_data:main.temp, timestamp=1642364300887, value=270.8
1642777200                                column=weather_data:main.temp_kf, timestamp=1642364300887, value=0
1642777200                                column=weather_data:main.temp_max, timestamp=1642364300887, value=270.8
1642777200                                column=weather_data:main.temp_min, timestamp=1642364300887, value=270.8
1642777200                                column=weather_data:pop, timestamp=1642364300887, value=0.01
1642777200                                column=weather_data:snow.3h, timestamp=1642364172815, value=0.42
1642777200                                column=weather_data:sys.pod, timestamp=1642364300887, value=d
1642777200                                column=weather_data:visibility, timestamp=1642364300887, value=10000
1642777200                                column=weather_data:weather[0].description, timestamp=1642364300887, value=broken clouds
1642777200                                column=weather_data:weather[0].icon, timestamp=1642364300887, value=04d
1642777200                                column=weather_data:weather[0].id, timestamp=1642364300887, value=803
1642777200                                column=weather_data:weather[0].main, timestamp=1642364300887, value=Clouds
1642777200                                column=weather_data:wind.deg, timestamp=1642364300887, value=286
1642777200                                column=weather_data:wind.gust, timestamp=1642364300887, value=12.11
1642777200                                column=weather_data:wind.speed, timestamp=1642364300887, value=6.26
1642788000                                column=weather_data:clouds.all, timestamp=1642364300887, value=92
1642788000                                column=weather_data:dt, timestamp=1642364300887, value=1642788000
1642788000                                column=weather_data:dt_txt, timestamp=1642364300887, value=2022-01-21 18:00:00
1642788000                                column=weather_data:main.feels_like, timestamp=1642364300887, value=264.57
1642788000                                column=weather_data:main.grnd_level, timestamp=1642364300887, value=998
1642788000                                column=weather_data:main.humidity, timestamp=1642364300887, value=75
1642788000                                column=weather_data:main.pressure, timestamp=1642364300887, value=1012
1642788000                                column=weather_data:main.sea_level, timestamp=1642364300887, value=1012
1642788000                                column=weather_data:main.temp, timestamp=1642364300887, value=271.07
1642788000                                column=weather_data:main.temp_kf, timestamp=1642364300887, value=0
1642788000                                column=weather_data:main.temp_max, timestamp=1642364300887, value=271.07
1642788000                                column=weather_data:main.temp_min, timestamp=1642364300887, value=271.07
1642788000                                column=weather_data:pop, timestamp=1642364300887, value=0.21
1642788000                                column=weather_data:snow.3h, timestamp=1642364300887, value=0.2
1642788000                                column=weather_data:sys.pod, timestamp=1642364300887, value=n
1642788000                                column=weather_data:visibility, timestamp=1642364300887, value=10000
1642788000                                column=weather_data:weather[0].description, timestamp=1642364300887, value=light snow
1642788000                                column=weather_data:weather[0].icon, timestamp=1642364300887, value=13n
1642788000                                column=weather_data:weather[0].id, timestamp=1642364300887, value=600
1642788000                                column=weather_data:weather[0].main, timestamp=1642364300887, value=Snow
1642788000                                column=weather_data:wind.deg, timestamp=1642364300887, value=287
1642788000                                column=weather_data:wind.gust, timestamp=1642364300887, value=13.46
1642788000                                column=weather_data:wind.speed, timestamp=1642364300887, value=6.99
186 row(s)
Took 4.5371 seconds
```

Figure 2: Records in HBase's weather table.

1642507200	274.88	1030	66	10000	5.25	20	0.00	0.00	0.00
1642518000	273.32	1031	83	10000	4.05	33	0.00	0.00	0.00
1642528800	272.53	1032	86	10000	3.04	23	0.00	0.00	0.00
1642539600	272.08	1032	85	10000	2.95	20	0.00	0.00	0.00
1642550400	272.10	1030	86	10000	3.38	46	0.00	0.00	0.00
1642561200	272.26	1028	92	10000	4.12	9	0.00	0.00	0.00
1642572000	274.05	1026	90	10000	5.52	42	0.00	0.00	0.00
1642582800	274.79	1024	90	10000	5.52	89	0.00	0.00	0.00
1642593600	275.37	1022	87	10000	4.85	94	0.00	0.00	0.00
1642604400	274.43	1020	90	10000	4.28	100	0.00	0.00	0.00
1642615200	273.59	1018	93	10000	3.89	100	0.00	0.00	0.00
1642626000	273.35	1016	92	10000	3.64	100	0.00	0.00	0.00
1642636800	273.05	1013	84	10000	3.41	100	0.00	0.00	0.00
1642647600	273.21	1009	74	10000	5.06	99	0.00	0.00	0.00
1642658400	273.42	1006	78	10000	6.09	98	0.00	0.00	0.00
1642669200	273.53	1004	95	723	6.07	100	0.00	0.43	0.58
1642680000	273.86	1002	97	24	5.87	100	0.00	1.50	0.86
1642690800	273.40	1003	82	10000	5.24	100	0.00	0.32	0.33
1642701600	272.51	1004	78	10000	7.66	100	0.00	1.38	0.71
1642712400	271.45	1004	94	229	6.97	93	0.00	0.90	0.78
1642723200	271.76	1004	86	10000	6.15	96	0.00	0.21	0.99
1642734000	272.13	1005	93	1670	6.03	100	0.00	1.37	1.00
1642744800	271.08	1007	93	447	5.87	100	0.00	0.93	1.00
1642755600	270.88	1009	92	516	5.44	100	0.00	0.51	0.89
1642766400	271.62	1010	81	10000	5.98	100	0.00	0.76	0.74
1642777200	270.80	1011	77	10000	6.26	84	0.00	0.00	0.01
1642788000	271.07	1012	75	10000	6.99	92	0.00	0.20	0.21

56 rows in set (0.01 sec)

Figure 3: Records in the serving layer concerning weather.

```
3202_2022-01-16 00:15:37                column=position:Brigade, timestamp=1642365054501, value=36
3202_2022-01-16 00:15:37                column=position:Lat, timestamp=1642365054501, value=52.299297
3202_2022-01-16 00:15:37                column=position:Lines, timestamp=1642365054501, value=1
3202_2022-01-16 00:15:37                column=position:Lon, timestamp=1642365054501, value=20.935259
3202_2022-01-16 00:15:37                column=position:Time, timestamp=1642365054501, value=2022-01-16 00:15:37
3202_2022-01-16 00:15:37                column=position:VehicleNumber, timestamp=1642365054501, value=3202
3202_2022-01-16 00:15:53                column=position:Brigade, timestamp=1642365060487, value=36
3202_2022-01-16 00:15:53                column=position:Lat, timestamp=1642365060487, value=52.29922
3202_2022-01-16 00:15:53                column=position:Lines, timestamp=1642365060487, value=1
3202_2022-01-16 00:15:53                column=position:Lon, timestamp=1642365060487, value=20.935135
3202_2022-01-16 00:15:53                column=position:Time, timestamp=1642365060487, value=2022-01-16 00:15:53
3202_2022-01-16 00:15:53                column=position:VehicleNumber, timestamp=1642365060487, value=3202
3202_2022-01-16 00:16:03                column=position:Brigade, timestamp=1642365067490, value=36
3202_2022-01-16 00:16:03                column=position:Lat, timestamp=1642365067490, value=52.299137
3202_2022-01-16 00:16:03                column=position:Lines, timestamp=1642365067490, value=1
3202_2022-01-16 00:16:03                column=position:Lon, timestamp=1642365067490, value=20.934391
3202_2022-01-16 00:16:03                column=position:Time, timestamp=1642365067490, value=2022-01-16 00:16:03
3202_2022-01-16 00:16:03                column=position:VehicleNumber, timestamp=1642365067490, value=3202
3202_2022-01-16 00:16:14                column=position:Brigade, timestamp=1642365073498, value=36
3202_2022-01-16 00:16:14                column=position:Lat, timestamp=1642365073498, value=52.299232
3202_2022-01-16 00:16:14                column=position:Lines, timestamp=1642365073498, value=1
3202_2022-01-16 00:16:14                column=position:Lon, timestamp=1642365073498, value=20.934195
3202_2022-01-16 00:16:14                column=position:Time, timestamp=1642365073498, value=2022-01-16 00:16:14
3202_2022-01-16 00:16:14                column=position:VehicleNumber, timestamp=1642365073498, value=3202
3202_2022-01-16 00:16:25                column=position:Brigade, timestamp=1642365083538, value=36
3202_2022-01-16 00:16:25                column=position:Lat, timestamp=1642365083538, value=52.299232
3202_2022-01-16 00:16:25                column=position:Lines, timestamp=1642365083538, value=1
```

Figure 4: Records in HBase's the table for positions of trams.

```
hbase(main):009:0> scan 'timetable', {LIMIT => 10}
ROW                                  COLUMN+CELL
 00000969-3cd9-4bed-b042-ddae238998af    column=trams:busstopId, timestamp=1642365269496, value=1078
 00000969-3cd9-4bed-b042-ddae238998af    column=trams:busstopNr, timestamp=1642365269496, value=03
 00000969-3cd9-4bed-b042-ddae238998af    column=trams:direction, timestamp=1642365269496, value=rondo \xC5\xBBaba
 00000969-3cd9-4bed-b042-ddae238998af    column=trams:lat, timestamp=1642365269496, value=21.031907
 00000969-3cd9-4bed-b042-ddae238998af    column=trams:line, timestamp=1642365269496, value=1
 00000969-3cd9-4bed-b042-ddae238998af    column=trams:long, timestamp=1642365269496, value=52.273599
 00000969-3cd9-4bed-b042-ddae238998af    column=trams:time, timestamp=1642365269496, value=08:36:00
 00002a17-813f-49be-8b57-7067619de650    column=trams:busstopId, timestamp=1642177632179, value=3007
 00002a17-813f-49be-8b57-7067619de650    column=trams:busstopNr, timestamp=1642177632179, value=06
 00002a17-813f-49be-8b57-7067619de650    column=trams:direction, timestamp=1642177632179, value=Metro Wierzbno
 00002a17-813f-49be-8b57-7067619de650    column=trams:lat, timestamp=1642177632179, value=21.023100
 00002a17-813f-49be-8b57-7067619de650    column=trams:line, timestamp=1642177632179, value=18
 00002a17-813f-49be-8b57-7067619de650    column=trams:long, timestamp=1642177632179, value=52.189770
 00002a17-813f-49be-8b57-7067619de650    column=trams:time, timestamp=1642177632179, value=19:40:00
 000030b7-867a-4e27-ab39-c73b0ccfccac    column=trams:busstopId, timestamp=1642180376217, value=2011
 000030b7-867a-4e27-ab39-c73b0ccfccac    column=trams:busstopNr, timestamp=1642180376217, value=03
 000030b7-867a-4e27-ab39-c73b0ccfccac    column=trams:direction, timestamp=1642180376217, value=Wsp\xC3\xB3lna Droga
 000030b7-867a-4e27-ab39-c73b0ccfccac    column=trams:lat, timestamp=1642180376217, value=21.102623
 000030b7-867a-4e27-ab39-c73b0ccfccac    column=trams:line, timestamp=1642180376217, value=24
 000030b7-867a-4e27-ab39-c73b0ccfccac    column=trams:long, timestamp=1642180376217, value=52.242414
 000030b7-867a-4e27-ab39-c73b0ccfccac    column=trams:time, timestamp=1642180376217, value=16:20:00
 000093d0-8c88-4859-8fbd-f660e9b52421    column=trams:busstopId, timestamp=1641551810657, value=4121
 000093d0-8c88-4859-8fbd-f660e9b52421    column=trams:busstopNr, timestamp=1641551810657, value=04
 000093d0-8c88-4859-8fbd-f660e9b52421    column=trams:direction, timestamp=1641551810657, value=Och-Teatr
 000093d0-8c88-4859-8fbd-f660e9b52421    column=trams:lat, timestamp=1641551810657, value=20.981497
 000093d0-8c88-4859-8fbd-f660e9b52421    column=trams:line, timestamp=1641551810657, value=1
 000093d0-8c88-4859-8fbd-f660e9b52421    column=trams:long, timestamp=1641551810657, value=52.216757
 000093d0-8c88-4859-8fbd-f660e9b52421    column=trams:time, timestamp=1641551810657, value=15:25:00
 00009688-9269-433e-bb28-5e4e59ecd06d    column=trams:busstopId, timestamp=1641485068873, value=1086
 00009688-9269-433e-bb28-5e4e59ecd06d    column=trams:busstopNr, timestamp=1641485068873, value=03
 00009688-9269-433e-bb28-5e4e59ecd06d    column=trams:direction, timestamp=1641485068873, value=Kondratowicza
 00009688-9269-433e-bb28-5e4e59ecd06d    column=trams:lat, timestamp=1641485068873, value=21.024676
 00009688-9269-433e-bb28-5e4e59ecd06d    column=trams:line, timestamp=1641485068873, value=1
 00009688-9269-433e-bb28-5e4e59ecd06d    column=trams:long, timestamp=1641485068873, value=52.296618
 00009688-9269-433e-bb28-5e4e59ecd06d    column=trams:time, timestamp=1641485068873, value=10:07:00
 0000a5ac-fcf0-412f-95fa-d91e8c8b6422    column=trams:busstopId, timestamp=1638105724083, value=4005
 0000a5ac-fcf0-412f-95fa-d91e8c8b6422    column=trams:busstopNr, timestamp=1638105724083, value=04
```

Figure 5: Records in HBase's timetable table.

```
df = spark \
  .read \
  .format("kafka") \
  .option("kafka.bootstrap.servers", "instance-tram-1:9092") \
  .option("subscribe", "lines") \
  .load()

table = df.select(col("value").cast("string")) .alias("csv").select("csv.*")
table = table.selectExpr("split(value,',')[0] as busstopId" \
                ,"split(value,',')[1] as busstopNr" \
                ,"split(value,',')[2] as line" \
                ,"split(value,',')[3] as direction" \
                ,"split(value,',')[4] as lon" \
                ,"split(value,',')[5] as lat" \
                        ,"split(value,',')[6] as time" \
                        ,"split(value,',')[7] as iter")

table = table.withColumn('busstopNr', table.busstopNr.cast(IntegerType()))\
.withColumn('line', table.line.cast(IntegerType()))\
.withColumn('lon', table.lon.cast(FloatType()))\
.withColumn('lat', table.lat.cast(FloatType()))\
.withColumn('time', table.time.cast(TimestampType()))\
.withColumn('iter', table.iter.cast(IntegerType()))
table = table.withColumn('time', 60*hour(table.time) + minute(table.time))
table = table.withColumn('coords', array(table.lon, table.lat))
```

```
table.show(20)
```

```
[Stage 0:>                                                          (0 + 1) / 1]
```

busstopId	busstopNr	line	direction	lon	lat	time	iter	coords
R-03	0	15	Wołoska	52.1885	20.999907	223	0	[52.1885, 20.999907]
3240	7	15	Samochodowa	52.18876	21.00332	225	0	[52.18876, 21.00332]
R-04	0	4	"Zgrupowania AK "...	52.299137	20.934156	225	0	[52.299137, 20.93...]
3116	3	15	Telewizja Polska	52.188824	21.007105	226	0	[52.188824, 21.00...]
R-04	0	1	"Zgrupowania AK "...	52.299137	20.934156	227	0	[52.299137, 20.93...]
3115	3	15	Metro Wierzbno	52.18887	21.0114	227	0	[52.18887, 21.0114]
6061	5	1	Marymoncka	52.299557	20.935863	228	0	[52.299557, 20.93...]
6014	3	4	Przy Agorze	52.297985	20.942766	228	0	[52.297985, 20.94...]
3114	3	15	Królikarnia	52.18895	21.016396	228	0	[52.18895, 21.016...]
6013	3	4	UKSW	52.294888	20.94566	229	0	[52.294888, 20.94...]
6011	3	4	Szpital Bielański	52.29012	20.950191	230	0	[52.29012, 20.950...]
3007	5	15	Puławska	52.189762	21.02384	230	0	[52.189762, 21.02...]
6014	3	1	Przy Agorze	52.297985	20.942766	230	0	[52.297985, 20.94...]
6013	3	1	UKSW	52.294888	20.94566	231	0	[52.294888, 20.94...]
3007	4	15	Malczewskiego	52.190685	21.024548	231	0	[52.190685, 21.02...]
6010	3	4	AWF	52.28773	20.952522	231	0	[52.28773, 20.952...]
6011	3	1	Szpital Bielański	52.29012	20.950191	232	0	[52.29012, 20.950...]
3006	6	15	Park Dreszera	52.19437	21.024326	232	0	[52.19437, 21.024...]
6008	3	4	Podleśna-IMiGW	52.28333	20.956453	232	0	[52.28333, 20.956...]
3005	6	15	Morskie Oko	52.197018	21.024158	233	0	[52.197018, 21.02...]

only showing top 20 rows

Figure 6: Exemplary table with order of tram stops.

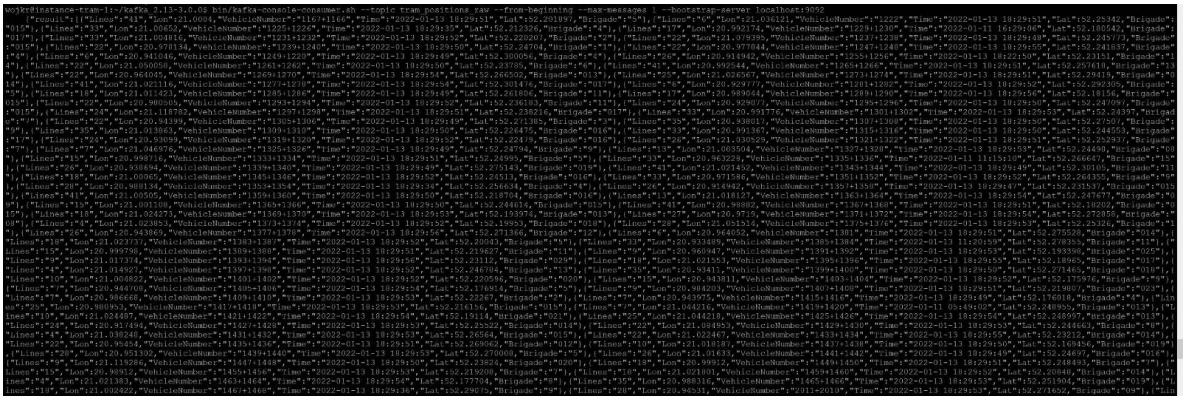

Figure 7: Unprocessed data with tram positions.

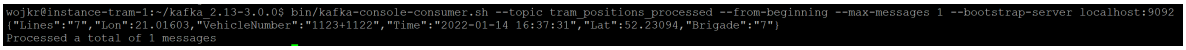

Figure 8: Processed data with tram positions.

Figure 9: Unprocessed data with weather conditions.

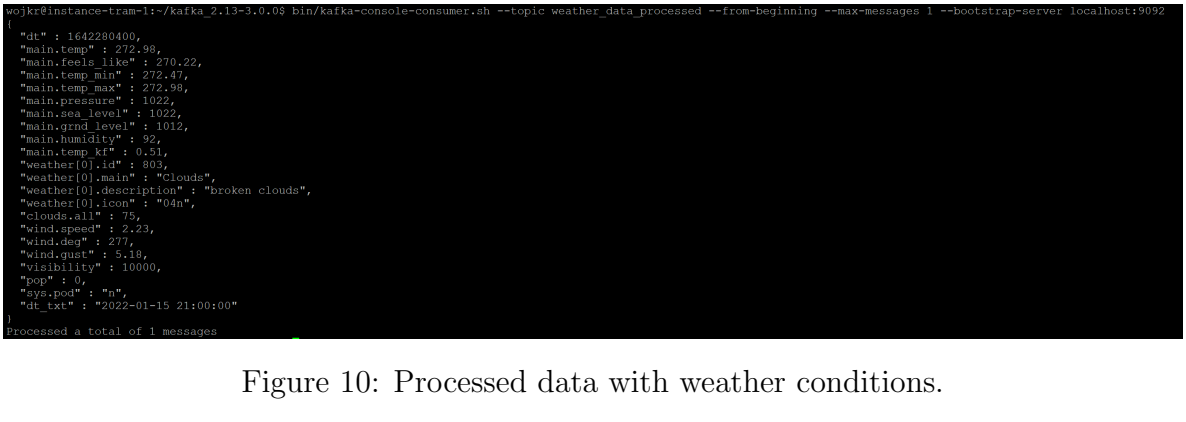

Figure 10: Processed data with weather conditions.

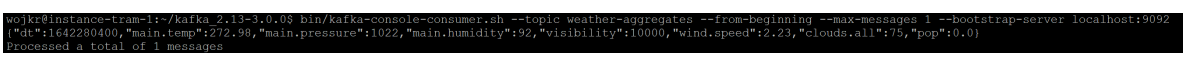

Figure 11: Processed data with weather conditions with aggregation.

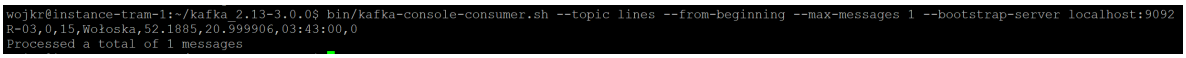

Figure 12: Exemplary record with tram stops data.