Predicting public transport delays

Transatlantic Scooters' BD Project

Laura Bąkała
Wojciech Kretowicz
Karol Pysiak
Mateusz Szysz

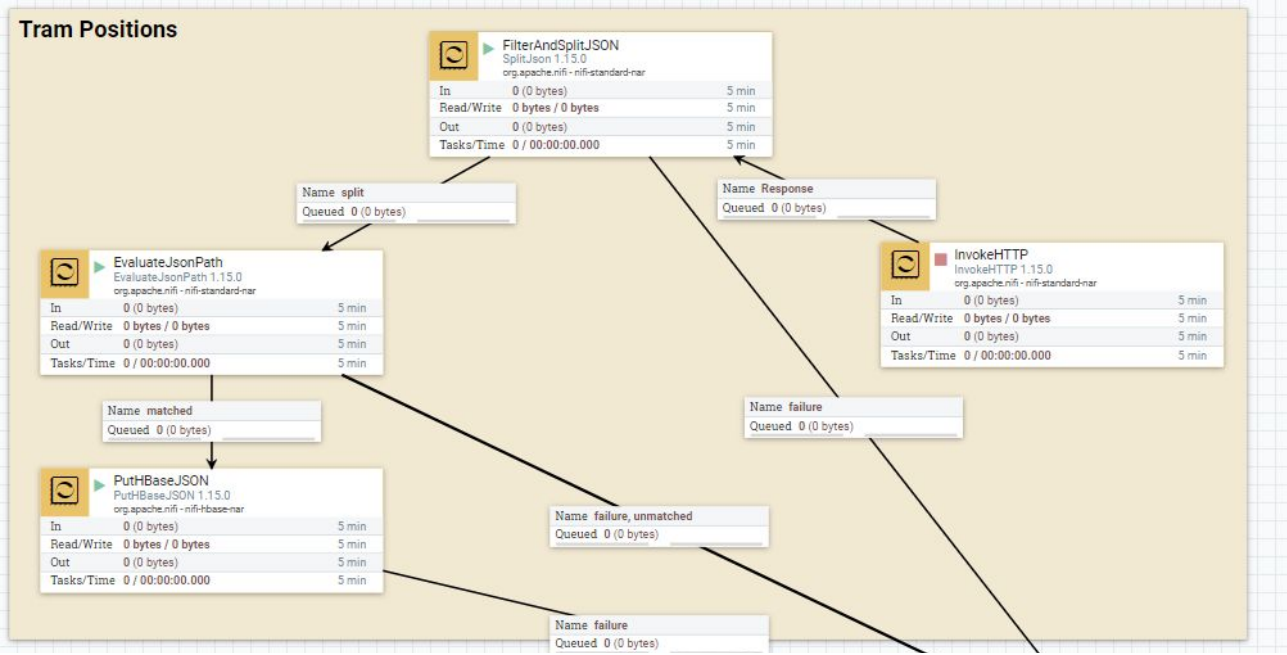
Goals

- Predicting public transport delays based on weather prognoses, traffic and past delays.
- Online monitoring of delays in Warsaw.
- Facilitating choosing appropriate means of transport at given moment and planning travels.

Data sources

- Warszawskie Dane (on-line data about trams and buses including their positions in given moment of time)
- Timetable for buses and trams in Warsaw
- Open-source meteorological data source with prognoses for specific amount of days
- ...*hopefully nothing else*...

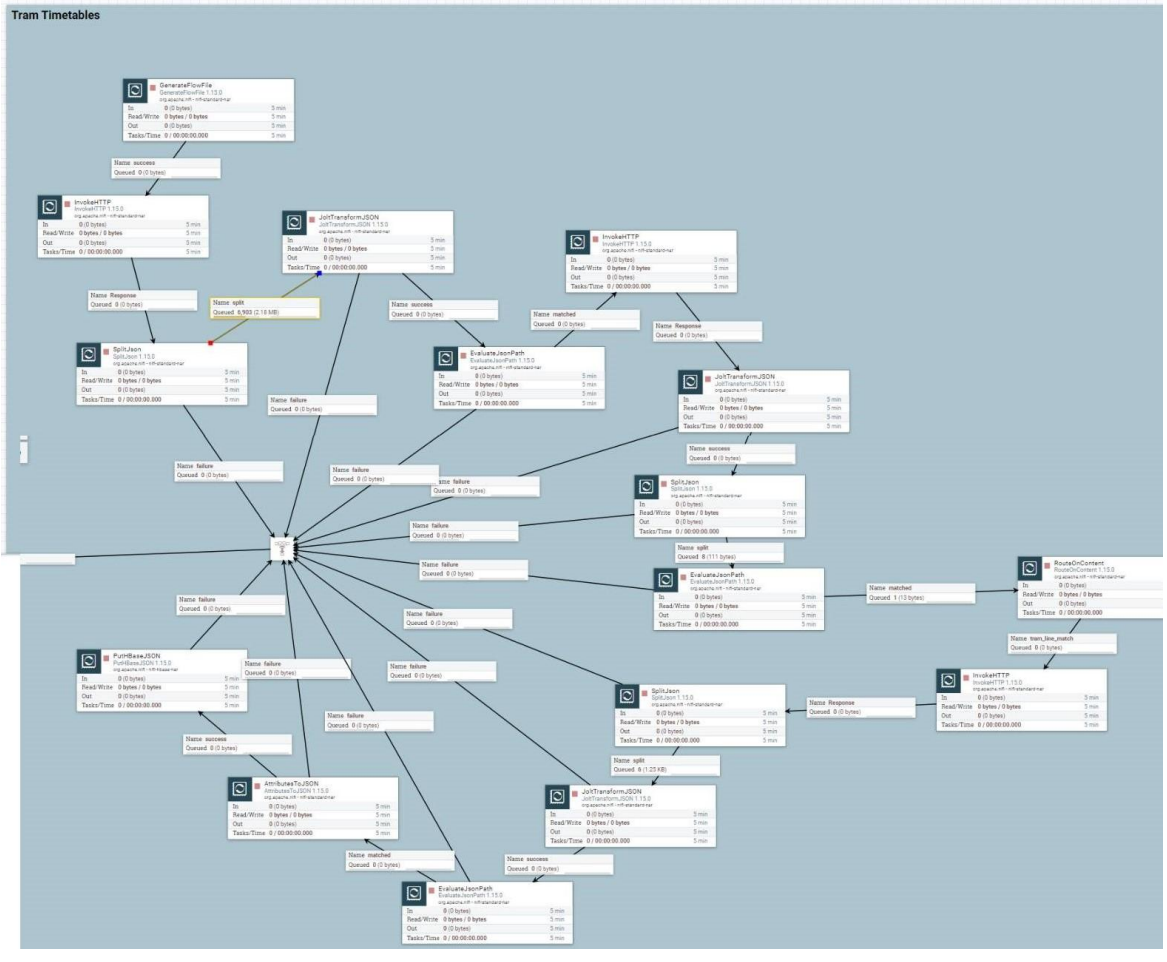
Warszawskie Dane



Warszawskie Dane: trams' locations

1. Retrieve data about trams' positions
2. Filter out positions (roughly) outside the boundaries of Warsaw
3. Save data to HBase

Warszawskie Dane



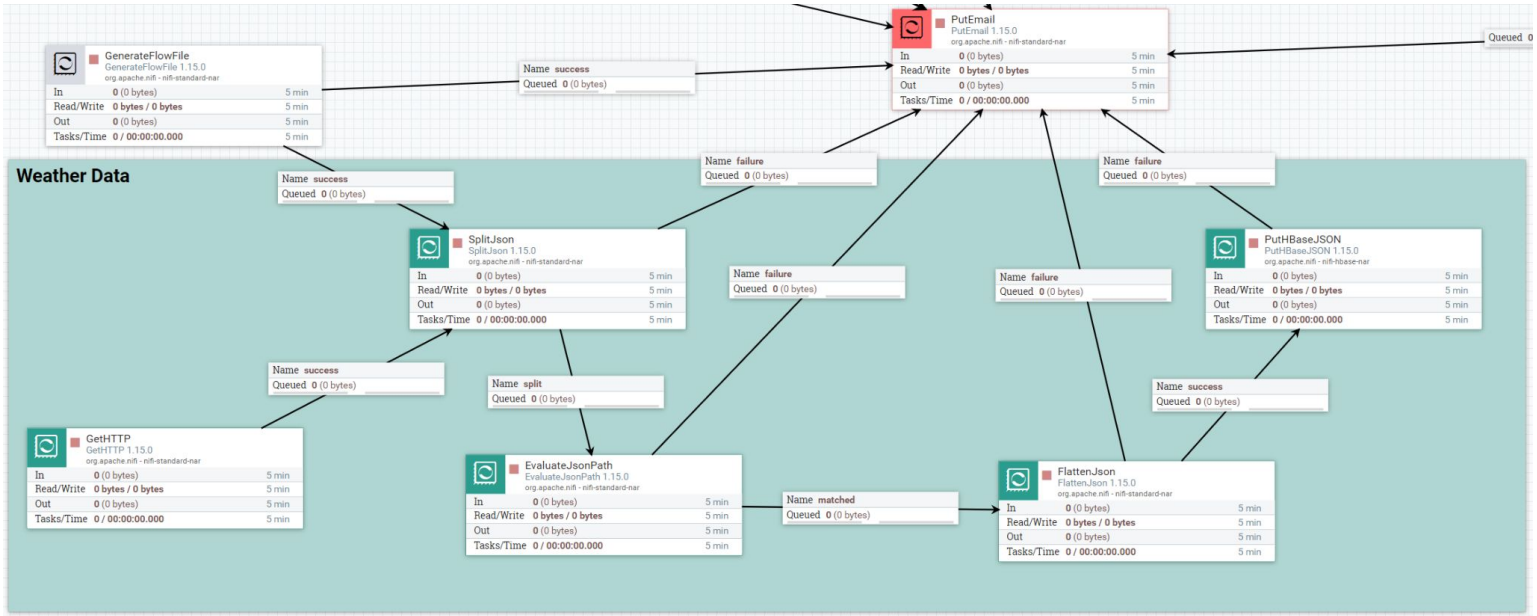
Warszawskie Dane: schedules

1. Retrieve the list of bus/tram stops
2. For each stop:
    a. Retrieve the list of lines stopping there
    b. Filter tram lines (1- & 2-digit codes)
    c. For each tram line:
        i. Retrieve timetable
        ii. Split timetable into separate departure times
        iii. Add relevant attributes from previous queries
        iv. Save to data HBase

OpenWeather


```
{
  "cod": "200",
  "message": 0,
  "cnt": 40,
  "list": [
    {
      "dt": 1637506800,
      "main": {…},
      "weather": […],
      "clouds": {…},
      "wind": {…},
      "visibility": 10000,
      "pop": 0,
      "sys": {…},
      "dt_txt": "2021-11-21 15:00:00"
    },
    {
      "dt": 1637517600,
      "main": {…},
      "weather": […],
      "clouds": {…},
      "wind": {…},
      "visibility": 10000,
      "pop": 0,
      "sys": {…},
      "dt_txt": "2021-11-21 18:00:00"
    },
    {
      "dt": 1637528400,
      "main": {…},
      "weather": […],
      "clouds": {…},
      "wind": {…},
      "visibility": 10000,
      "pop": 0,
      "sys": {…},
      "dt_txt": "2021-11-21 21:00:00"
    },
    {
      "dt": 1637506800,
      "main": {
        "temp": 281.4,
        "feels_like": 278.73,
        "temp_min": 281.4,
        "temp_max": 281.84,
        "pressure": 1004,
        "sea_level": 1004,
        "grnd_level": 993,
        "humidity": 85,
        "temp_kf": -0.44
      },
      "weather": [
        {
          "id": 803,
          "main": "Clouds",
          "description": "broken clouds",
          "icon": "04n"
        }
      ],
      "clouds": {
        "all": 75
      },
      "wind": {
        "speed": 4.59,
        "deg": 270,
        "gust": 8.64
      },
      "visibility": 10000,
      "pop": 0,
      "sys": {
        "pod": "n"
      },
      "dt_txt": "2021-11-21 15:00:00"
    },
```

OpenWeather



OpenWeather

```
1638435600  column=weather_data:clouds.all, timestamp=1638019987294, value=100
1638435600  column=weather_data:dt, timestamp=1638019987294, value=1638435600
1638435600  column=weather_data:dt_txt, timestamp=1638019987294, value=2021-12-02 09:00:00
1638435600  column=weather_data:main.feels_like, timestamp=1638019987294, value=273.01
1638435600  column=weather_data:main.grnd_level, timestamp=1638019987294, value=980
1638435600  column=weather_data:main.humidity, timestamp=1638019987294, value=81
1638435600  column=weather_data:main.pressure, timestamp=1638019987294, value=994
1638435600  column=weather_data:main.sea_level, timestamp=1638019987294, value=994
1638435600  column=weather_data:main.temp, timestamp=1638019987294, value=277.7
1638435600  column=weather_data:main.temp_kf, timestamp=1638019987294, value=0
1638435600  column=weather_data:main.temp_max, timestamp=1638019987294, value=277.7
1638435600  column=weather_data:main.temp_min, timestamp=1638019987294, value=277.7
1638435600  column=weather_data:pop, timestamp=1638019987294, value=0.36
1638435600  column=weather_data:rain.3h, timestamp=1638019987294, value=0.42
1638435600  column=weather_data:sys.pod, timestamp=1638019987294, value=d
1638435600  column=weather_data:visibility, timestamp=1638019987294, value=10000
1638435600  column=weather_data:weather[0].description, timestamp=1638019987294, value=light rain
1638435600  column=weather_data:weather[0].icon, timestamp=1638019987294, value=10d
1638435600  column=weather_data:weather[0].id, timestamp=1638019987294, value=500
1638435600  column=weather_data:weather[0].main, timestamp=1638019987294, value=Rain
1638435600  column=weather_data:wind.deg, timestamp=1638019987294, value=266
1638435600  column=weather_data:wind.gust, timestamp=1638019987294, value=14.55
1638435600  column=weather_data:wind.speed, timestamp=1638019987294, value=7.21
```

Analysis - modelling

- Goal: predicting delays in the nearest future
- Model will be updated constantly with incoming data
- Model will utilize all data sources
- We plan to use following variables:

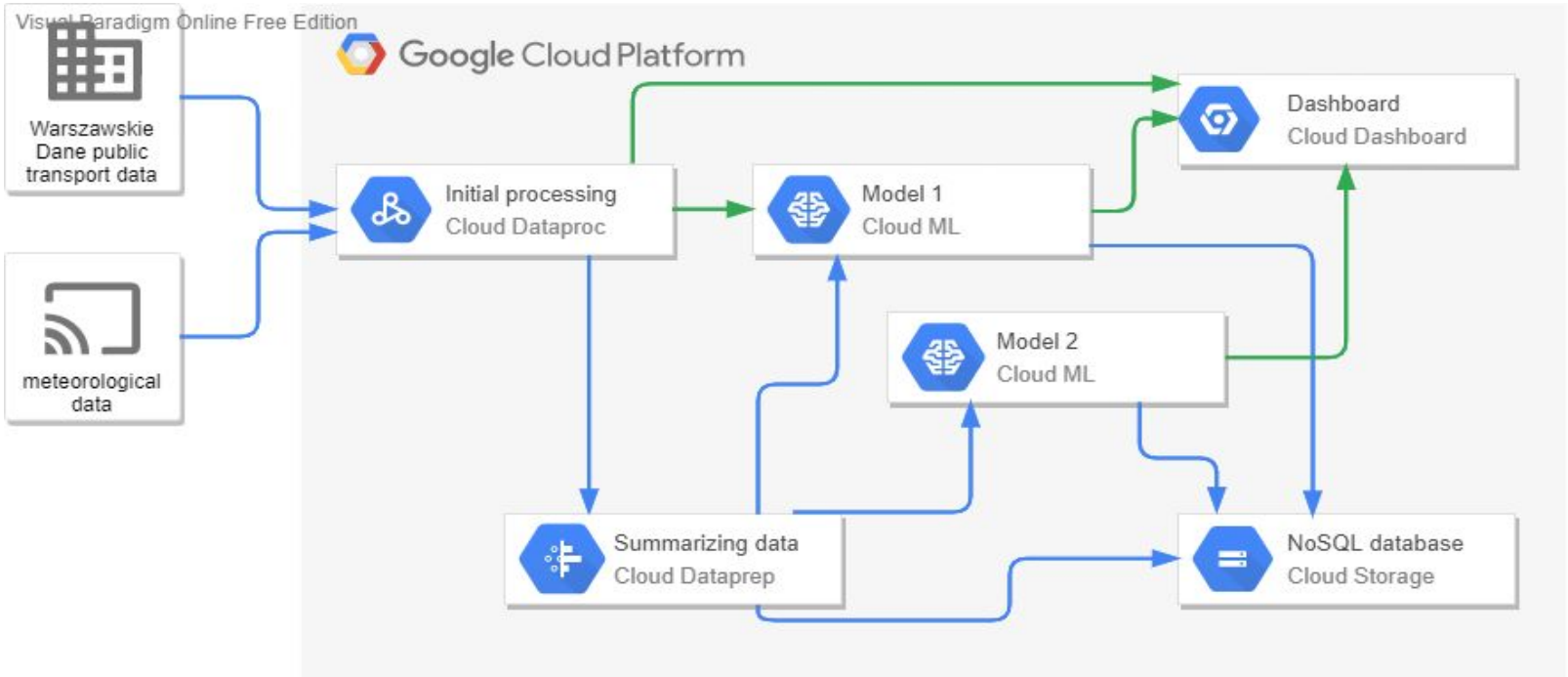
➔ time of day
➔ line number
➔ current vehicle position
➔ last delay
➔ next stop
➔ direction of the line
➔ temperature

➔ pressure
➔ humidity
➔ cloudiness
➔ wind
➔ visibility
➔ precipitation

Dashboard

- Offline:
  - average delays against time of day / week day / month
  - average delays against the particular line
  - heatmap of Warsaw with average delays
- Online:
  - current / future average delay
  - current / future delay for each line
  - current / future heatmap of Warsaw with information about delays

Architecture



Thank you for your attention

Laura Bąkała
Wojciech Kretowicz
Karol Pysiak
Mateusz Szysz