



System monitorujący jakość powietrza w Warszawie

**Projekt na przedmiot
Składowanie danych w systemach Big Data**

**Mateusz Krzyżiński
Mikołaj Spytek
Paweł Wojciechowski**

09.01.2023

Cel projektu

1. Stworzenie systemu informatycznego, który stanowi wsparcie w podejmowaniu decyzji i analizach związanych z **problemem zanieczyszczeń powietrza**.
2. Stworzenie rozwiązania na potrzeby miasta stołecznego **Warszawy** z możliwością rozszerzenia dla innych klientów.
3. Umożliwienie stworzenia **dogłębnych analiz** danych dotyczących jakości powietrza **w czasie niemal rzeczywistym**.

Źródła danych

→ AQICN API

- ◆ projekt The World Air Quality Index
- ◆ dane odświeżane co kilka/kilkanaście minut
- ◆ zapytania na podstawie przygotowanej listy identyfikatorów
- ◆ REST API, zwracana odpowiedź w formacie JSON
- ◆ informacje o zanieczyszczeniach:
 - AQI
 - PM2.5
 - PM10

AQI	Air Pollution Level	Health Implications	Cautionary Statement (for PM2.5)
0 - 50	Good	Air quality is considered satisfactory, and air pollution poses little or no risk	None
51 -100	Moderate	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.	Active children and adults, and people with respiratory disease, such as asthma, should limit prolonged outdoor exertion.
101-150	Unhealthy for Sensitive Groups	Members of sensitive groups may experience health effects. The general public is not likely to be affected.	Active children and adults, and people with respiratory disease, such as asthma, should limit prolonged outdoor exertion.
151-200	Unhealthy	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects	Active children and adults, and people with respiratory disease, such as asthma, should avoid prolonged outdoor exertion; everyone else, especially children, should limit prolonged outdoor exertion
201-300	Very Unhealthy	Health warnings of emergency conditions. The entire population is more likely to be affected.	Active children and adults, and people with respiratory disease, such as asthma, should avoid all outdoor exertion; everyone else, especially children, should limit outdoor exertion.
300+	Hazardous	Health alert: everyone may experience more serious health effects	Everyone should avoid all outdoor exertion

Źródła danych

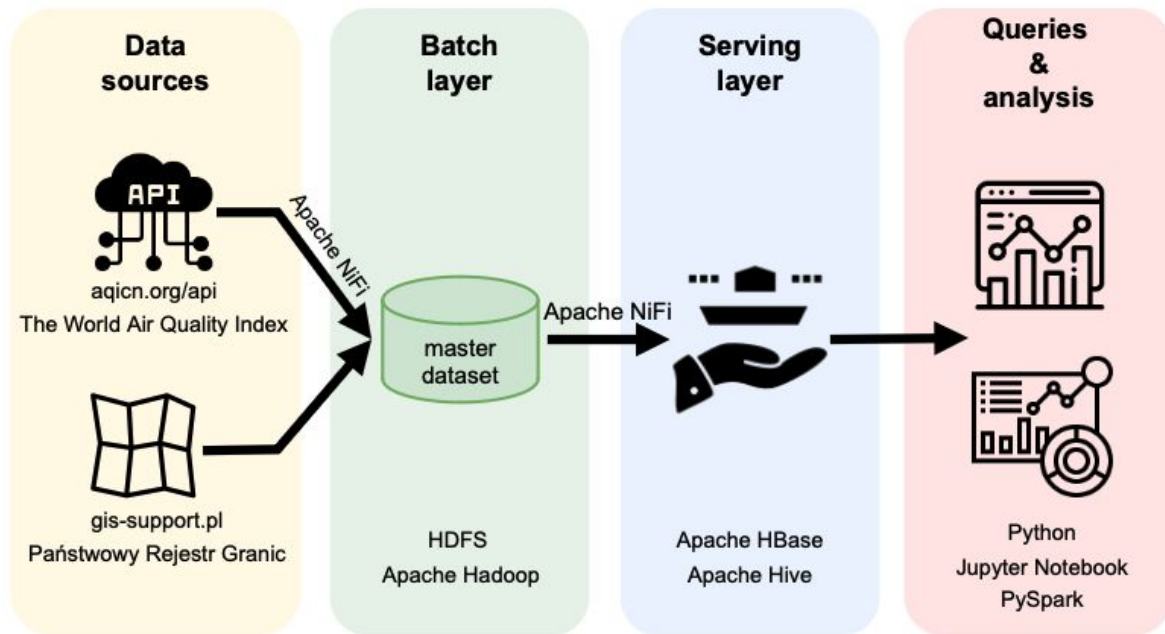
→ gis-support.pl

- ◆ dane geoprzestrzenne (kartograficzne) dla m. st. Warszawy
 - ◆ przygotowane na podstawie danych udostępnionych przez Państwowy Rejestr Granic
- ◆ dane statyczne
- ◆ zawiera informacje o przebiegu granic dzielnic
- ◆ formaty plików:
 - .cpg
 - .dbf
 - .prj
 - .qpj
 - .shp
 - .shx

* możliwe rozszerzenie o inne dane geoprzestrzenne



Architektura rozwiązania



Uproszczony diagram rozwiązania


(przyjęto architekturę lambda z pominięciem speed layer)

Warstwa wsadowa


→ dane o jakości powietrza

API → HDFS


Get files from API to master dataset


	GetStationList GetFile 1.14.0 org.apache.nifi - nifi-standard-nar	
In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name success
Queued 0 (0 bytes)

	SplitStationNames SplitText 1.14.0 org.apache.nifi - nifi-standard-nar	
In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min


Name splits
Queued 0 (0 bytes)

	SendRequestToAPI InvokeHTTP 1.14.0 org.apache.nifi - nifi-standard-nar	
In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min


	ExtractNames ExtractText 1.14.0 org.apache.nifi - nifi-standard-nar	
In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name matched
Queued 0 (0 bytes)


Name Response
Queued 0 (0 bytes)

	MergeOutputsAndConverToAvro MergeRecord 1.14.0 org.apache.nifi - nifi-standard-nar	
In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	5 / 00:00:00.002	5 min

Name merged
Queued 0 (0 bytes)

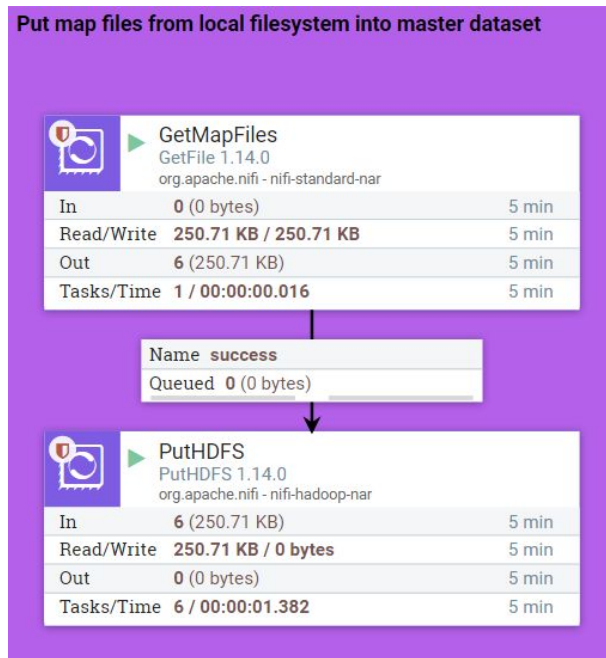
	MakeFileName UpdateAttribute 1.14.0 org.apache.nifi - nifi-update-attribute-nar	
In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name success
Queued 0 (0 bytes)

	PutHDFS PutHDFS 1.14.0 org.apache.nifi - nifi-hadoop-nar	
In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Warstwa wsadowa

→ dane geoprzestrzenne



LFS → HDFS

Master dataset

→ dane geoprzestrzenne

```
vagrant@node1:~/project$ hdfs dfs -ls /user/project/maps
Found 6 items
-rw-r--r--  1 root supergroup          5 2023-01-08 09:46 /user/project/maps/dzielnice_Warszawy.cpg
-rw-r--r--  1 root supergroup    4656 2023-01-08 09:46 /user/project/maps/dzielnice_Warszawy.dbf
-rw-r--r--  1 root supergroup     395 2023-01-08 09:46 /user/project/maps/dzielnice_Warszawy.prj
-rw-r--r--  1 root supergroup     618 2023-01-08 09:46 /user/project/maps/dzielnice_Warszawy.qpj
-rw-r--r--  1 root supergroup  250804 2023-01-08 09:46 /user/project/maps/dzielnice_Warszawy.shp
-rw-r--r--  1 root supergroup     244 2023-01-08 09:46 /user/project/maps/dzielnice_Warszawy.shx
```

→ dane geoprzestrzenne


```
vagrant@node1:~/project$ hdfs dfs -ls /user/project/data
Found 260 items
-rw-r--r--  1 root supergroup    26282 2023-01-03 12:05 /user/project/data/station-list-2023-01-03-12-05-50-875.avro
-rw-r--r--  1 root supergroup   49939 2023-01-03 12:22 /user/project/data/station-list-2023-01-03-12-22-49-776.avro
-rw-r--r--  1 root supergroup   25124 2023-01-03 12:37 /user/project/data/station-list-2023-01-03-12-37-50-852.avro
-rw-r--r--  1 root supergroup   25155 2023-01-03 12:52 /user/project/data/station-list-2023-01-03-12-52-51-078.avro
-rw-r--r--  1 root supergroup   24741 2023-01-03 13:07 /user/project/data/station-list-2023-01-03-13-07-51-166.avro
-rw-r--r--  1 root supergroup   24741 2023-01-03 13:22 /user/project/data/station-list-2023-01-03-13-22-51-319.avro
-rw-r--r--  1 root supergroup   24739 2023-01-03 13:37 /user/project/data/station-list-2023-01-03-13-37-51-384.avro
-rw-r--r--  1 root supergroup   24739 2023-01-03 13:52 /user/project/data/station-list-2023-01-03-13-52-51-495.avro
-rw-r--r--  1 root supergroup   24741 2023-01-03 14:07 /user/project/data/station-list-2023-01-03-14-07-51-583.avro
-rw-r--r--  1 root supergroup   24739 2023-01-03 14:22 /user/project/data/station-list-2023-01-03-14-22-51-682.avro
-rw-r--r--  1 root supergroup   24739 2023-01-03 14:37 /user/project/data/station-list-2023-01-03-14-37-51-841.avro
-rw-r--r--  1 root supergroup   25145 2023-01-04 12:07 /user/project/data/station-list-2023-01-04-12-07-32-629.avro
-rw-r--r--  1 root supergroup   25161 2023-01-04 12:22 /user/project/data/station-list-2023-01-04-12-22-32-662.avro
```


Warstwa dostępu do danych


→ dane o jakości powietrza

HDFS → Hive

Load air quality data from master dataset to serving layer

	GetHDFSfilesFromLast15Mins
GetHDFS 1.14.0 org.apache.nifi - nifi-hadoop-nar	
In	0 (0 bytes)
Read/Write	0 bytes / 26.02 KB
Out	2 (26.02 KB)
Tasks/Time	1 / 00:00:00.070


Name success
Queued 0 (0 bytes)

	ConvertAvroToJson
ConvertAvroToJson 1.14.0 org.apache.nifi - nifi-avro-nar	
In	2 (26.02 KB)
Read/Write	26.02 KB / 38.5 KB
Out	2 (38.5 KB)
Tasks/Time	2 / 00:00:00.052


Name success
Queued 0 (0 bytes)

	SplitJson
SplitJson 1.14.0 org.apache.nifi - nifi-standard-nar	
In	2 (38.5 KB)
Read/Write	38.5 KB / 35.69 KB
Out	60 (35.69 KB)
Tasks/Time	2 / 00:00:00.019


Name split
Queued 0 (0 bytes)

	RemoveMissingAQI
RouteOnAttribute 1.14.0 org.apache.nifi - nifi-standard-nar	
In	60 (35.69 KB)
Read/Write	0 bytes / 0 bytes
Out	55 (32.68 KB)
Tasks/Time	60 / 00:00:00.167


Name success
Queued 0 (0 bytes)

	FixStationID
UpdateAttribute 1.14.0 org.apache.nifi - nifi-update-attribute-nar	
In	60 (35.69 KB)
Read/Write	0 bytes / 0 bytes
Out	60 (35.69 KB)
Tasks/Time	60 / 00:00:00.139


Name matched
Queued 0 (0 bytes)

	ExtractNecessaryValues
EvaluateJsonPath 1.14.0 org.apache.nifi - nifi-standard-nar	
In	60 (35.69 KB)
Read/Write	35.69 KB / 0 bytes
Out	60 (35.69 KB)
Tasks/Time	60 / 00:00:00.142

Name notNull
Queued 0 (0 bytes)

	CreateValueVectorsAsString
ReplaceText 1.14.0 org.apache.nifi - nifi-standard-nar	
In	55 (32.68 KB)
Read/Write	32.68 KB / 9.82 KB
Out	55 (9.82 KB)
Tasks/Time	55 / 00:00:00.091


Name success
Queued 0 (0 bytes)

	MergeValueVectors
MergeContent 1.14.0 org.apache.nifi - nifi-standard-nar	
In	55 (9.82 KB)
Read/Write	9.82 KB / 9.82 KB
Out	1 (9.82 KB)
Tasks/Time	21 / 00:00:00.071

Name merged
Queued 0 (0 bytes)

	PrepareHQLQueryBeginning
ReplaceText 1.14.0 org.apache.nifi - nifi-standard-nar	
In	1 (9.82 KB)
Read/Write	9.82 KB / 9.95 KB
Out	1 (9.95 KB)
Tasks/Time	1 / 00:00:00.004

Name success
Queued 0 (0 bytes)

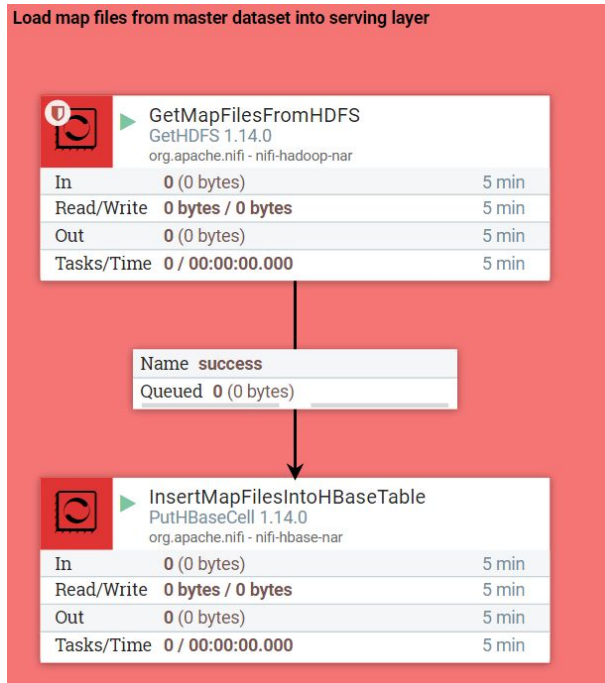
	PutHiveQL
PutHiveQL 1.14.0 org.apache.nifi - nifi-hive-nar	
In	1 (9.95 KB)
Read/Write	9.95 KB / 0 bytes
Out	0 (0 bytes)
Tasks/Time	1 / 00:00:05.960

Name success
Queued 0 (0 bytes)

	PrepareHQLQueryEnd
ReplaceText 1.14.0 org.apache.nifi - nifi-standard-nar	
In	1 (9.95 KB)
Read/Write	9.95 KB / 9.95 KB
Out	1 (9.95 KB)
Tasks/Time	1 / 00:00:00.000

Warstwa dostępu do danych

→ dane geoprzestrzenne



HDFS → HBase

Warstwa raportowa

→ odczytane dane o jakości powietrza

	station_id	station_name	latitude	longitude	location	datetime	aqi	pm10	pm25
0	365221	Rudzka	52.278000	20.972000	Rudzka, Marymont-Kaskada, Bielany, Warsaw, Mas...	2023-01-04 16:30:48	40	17	40
1	93862	Jana Piekalkiewiczza	52.193804	21.051250	Jana Piekalkiewiczza, Marcelin, Mokotów, Warsaw...	2023-01-04 16:00:45	57	19	57
2	76228	Graniczna	52.238812	21.002685	Graniczna, IV, Śródmieście, Warsaw, Masovian V...	2023-01-04 15:00:35	0	0	0
3	81664	Motycka	52.275087	21.044866	Motycka 23, 23, Motycka, Targówek Mieszkaniowy...	2023-01-04 15:16:46	57	29	57
4	358753	Koszycka	52.242000	20.944000	Koszycka, Koło-Górczewska, Koło, Wola, Warsaw,...	2023-01-04 16:17:59	26	9	26

Hive → PySpark → Jupyter

Warstwa raportowa

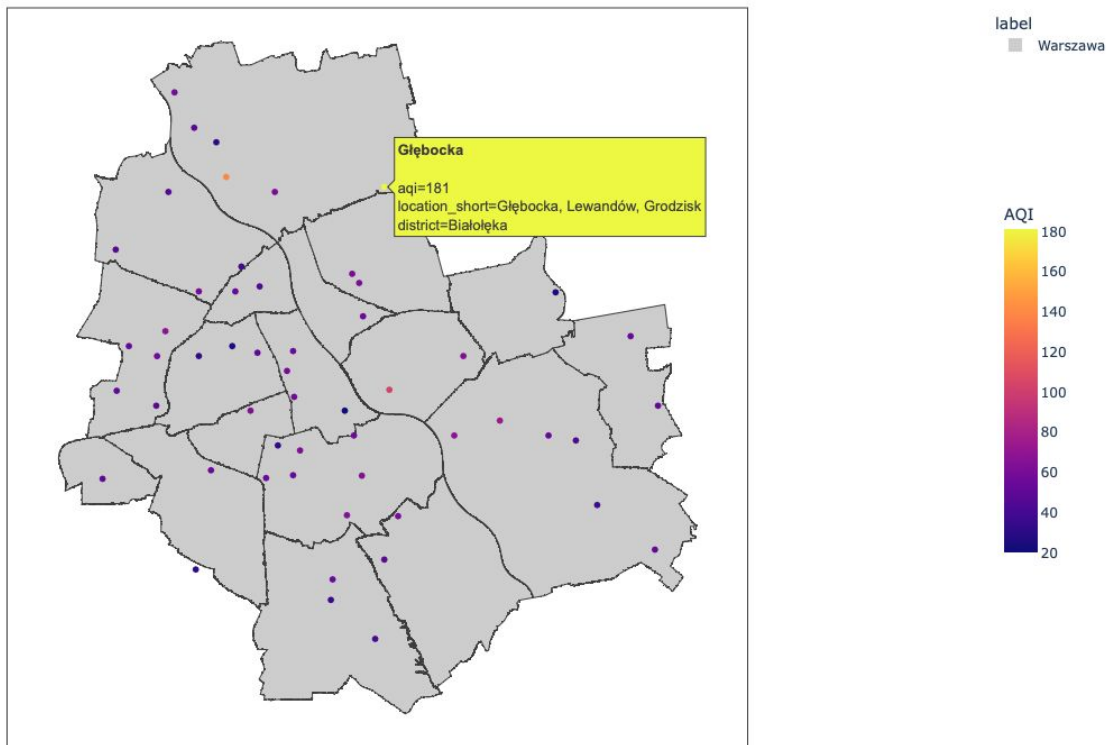
→ odczytane dane geoprzestrzenne

geometry	
nazwa_dzie	
Żoliborz	POLYGON ((20.95755 52.26693, 20.95760 52.26712...
Praga-Południe	POLYGON ((21.10347 52.22220, 21.10326 52.22206...
Mokotów	POLYGON ((21.02160 52.21305, 21.02172 52.21307...
Wola	POLYGON ((20.98121 52.25855, 20.98124 52.25838...
Wilanów	POLYGON ((21.10500 52.19428, 21.10509 52.19404...
Wesoła	POLYGON ((21.19234 52.23658, 21.18920 52.23786...
Wawer	POLYGON ((21.10801 52.18913, 21.10798 52.18918...
Włochy	POLYGON ((20.92000 52.21420, 20.92019 52.21408...
Ursynów	POLYGON ((20.98653 52.16231, 20.98702 52.16252...
Śródmieście	POLYGON ((21.06187 52.21720, 21.06172 52.21703...
Praga-Północ	POLYGON ((21.00546 52.26812, 21.00542 52.26826...
Ursus	POLYGON ((20.87858 52.20955, 20.87822 52.20932...
Targówek	POLYGON ((21.06202 52.30838, 21.06205 52.30830...
Rembertów	POLYGON ((21.14226 52.27855, 21.14906 52.27935...
Ochota	POLYGON ((21.00181 52.22771, 21.00197 52.22747...
Bielany	POLYGON ((20.92511 52.32894, 20.92511 52.32846...
Białołęka	POLYGON ((21.01270 52.29430, 21.01267 52.29426...
Bemowo	POLYGON ((20.88915 52.28046, 20.88917 52.28044...

HBase → HappyBase → Jupyter

Warstwa raportowa

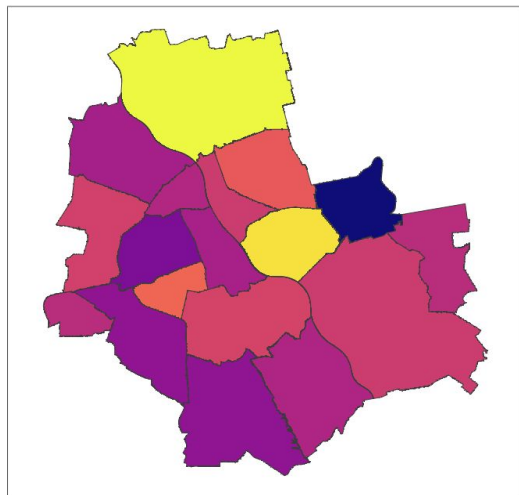
→ Analiza #1 - pomiary dla poszczególnych stacji



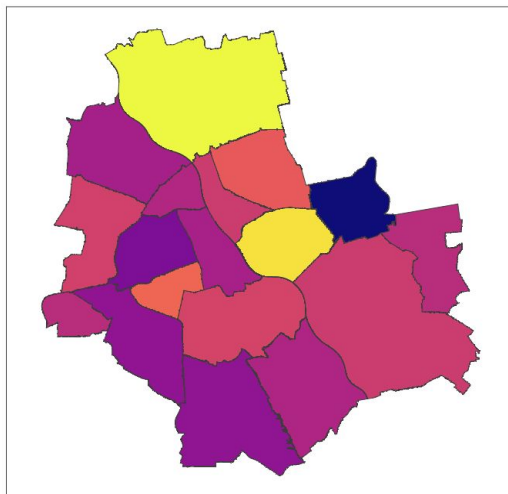
Warstwa raportowa

→ Analiza #2 - średnie dla dzielnic

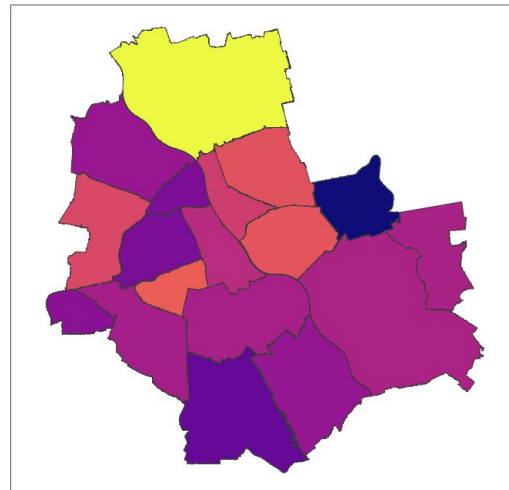
AQI



PM2.5

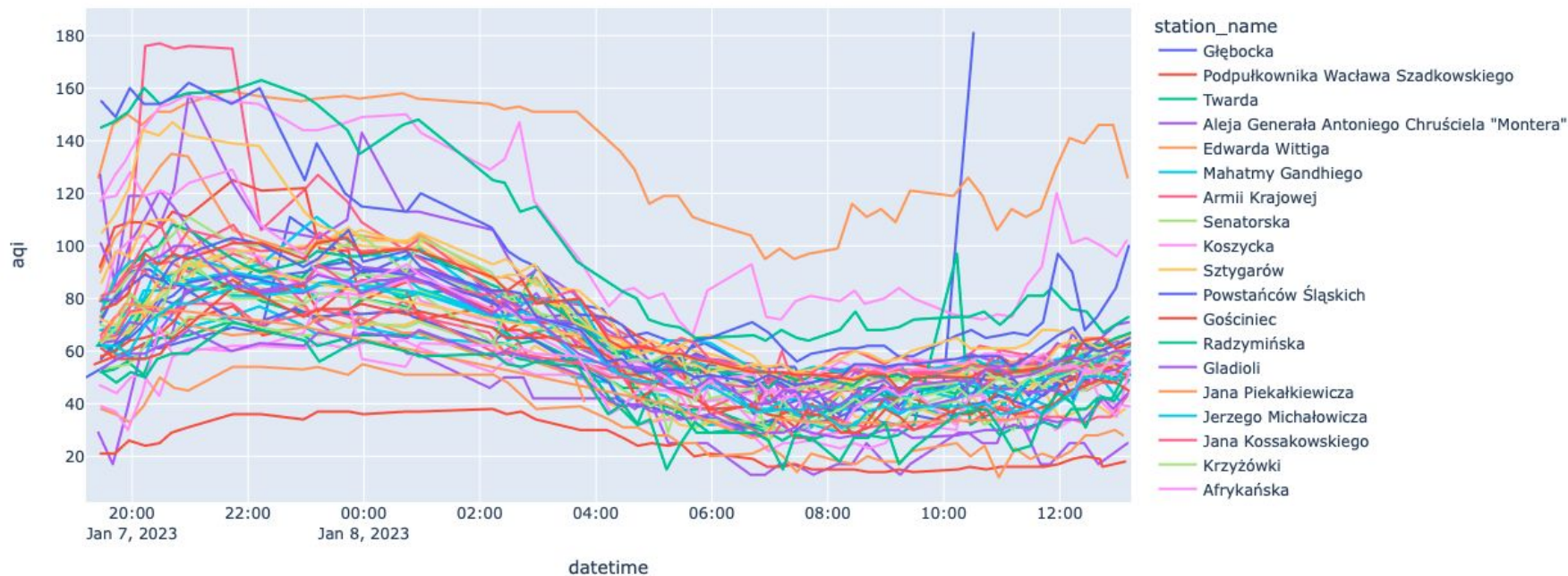


PM10



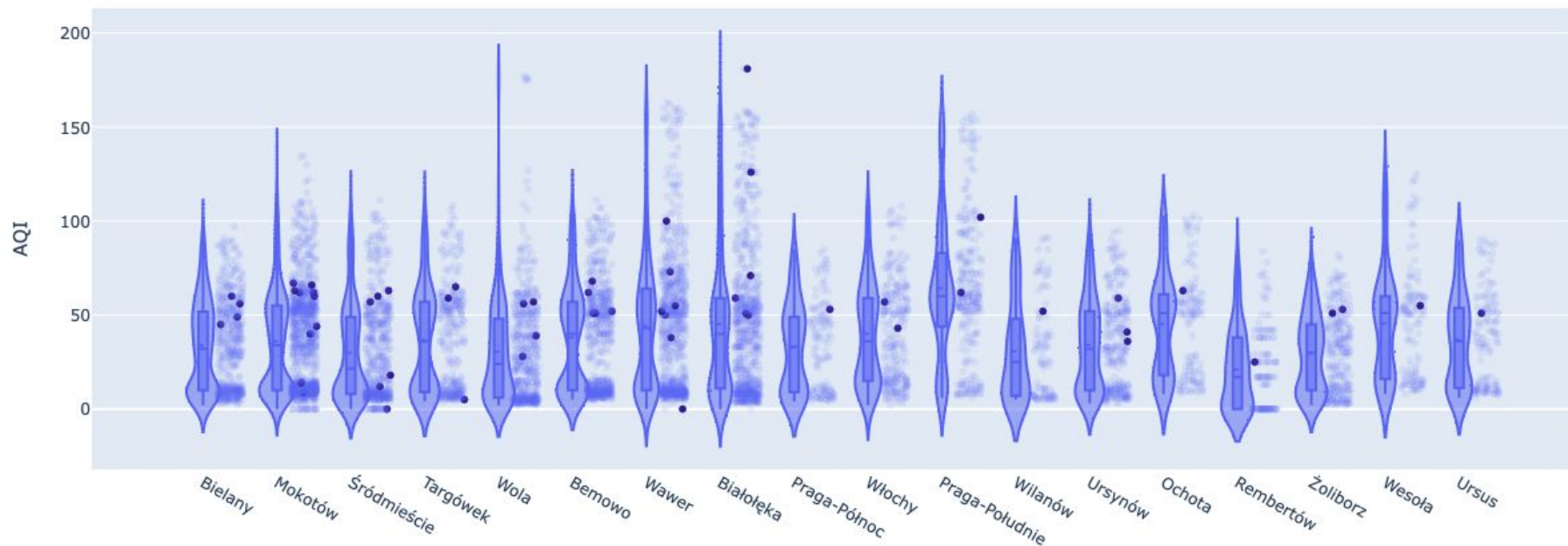
Warstwa raportowa

→ Analiza #3 - szereg czasowy pomiarów (24h)



Warstwa raportowa

→ Analiza #4 - rozkład pomiarów (historycznie)



Problemy i ograniczenia

1. API zwraca czasem złe wartości AQI bez komunikatu o błędach
 - potencjalnie rozwiązane
2. Trudności z odczytywaniem plików binarnych z danymi geoprzestrzennymi
 - rozwiązane przez korzystanie z tymczasowych zapisów

Podsumowanie

- Udało się zrealizować cel projektu - został stworzony **system monitorujący jakość powietrza w Warszawie**.
- Z punktu widzenia biznesu najważniejsza jest warstwa raportowa, umożliwiająca **wykorzystanie analiz w procesach decyzyjnych**.
- Opracowane rozwiązanie jest **wydajne** i ma **duży potencjał biznesowy**.



System monitorujący jakość powietrza w Warszawie

Projekt na przedmiot
Składowanie danych w systemach Big Data

Pytania?

Mateusz Krzyżiński
Mikołaj Spytek
Paweł Wojciechowski

09.01.2023