Dokumentacja projektu

Wpływ temperatur na jakość powietrza

Autorzy:
Maciej Paczóski
Krzysztof Wolny

Styczeń 2024

Spis treści

1 Cel projektu i potencjalne korzyści z wdrożenia

Projekt ma służyć do analizy zanieczyszczenia powietrza w Europie. Na podstawie corocznych pomiarów jest przedstawiana użytkownikowi ilość szkodliwych substancji w powietrzu na przestrzeni kolejnych lat. Dane reprezentują 34 państwa członkowskie UE, kraje EFTA oraz kraje kandydujące. Dodatkowo można zobaczyć zależność pomiędzy jakością powietrza, a ogrzewaniem oraz klimatyzacją mieszkań. System pozwoli na sprawdzenie, w których krajach zapotrzebowanie na chłodzenie oraz ogrzewanie budynków ma największy wpływ na środowisko. Projekt pozwala na lepsze zrozumienie problemu zanieczyszczenia powietrza w Europie, umożliwienie podejmowania świadomych decyzji dotyczących użytkowania systemów klimatyzacyjnych i ogrzewania.

2 Dane

W projekcie korzystamy z dwóch zbiorów danych, udostępnianych przez Europejski Urząd Statystyczny. Eurostat dopuszcza bezpłatne ponowne wykorzystanie swoich danych. Dostęp do danych można uzyskać przez oficjalną stronę bądź przez niewymagający uwierzytelnienia interfejs API.

2.1 Emisja gazów według sektorów NACE Rev. 2

Dane zostały udostępnione przez Eurostat, a ich kod to env_ac_ainah_r2. Zbiór opisuje emisję gazów cieplarnianych i substancji zanieczyszczających powietrze w podziale na gałęzie przemysłu (sklasyfikowane według NACE Rev. 2) oraz gospodarstwa domowe. Kompletne dane rozpoczynają się od roku referencyjnego 2008. Zbiór ma 4 528 976 rekordów.

Wymiary danych:

- Air pollutant (AIRPOL) - zbierane są dane dotyczące emisji 28 substancji zanieczyszczających
- Geopolitical entity (GEO) - 34 państwa członkowskie UE, kraje EFTA, kraje kandydujące
- Classification of economic activities (NACE_R2) - dane są gromadzone i publikowane w podziale na 84 kategorii według klasyfikacji działalności gospodarczej NACE
- Period of time (TIME) - roczne daty zaczynające się od 1995, a będące kompletne od 2008
- Unit (UNIT) - emisje w tonach i tysiącach ton, a także w gramach na mieszkańca i kilogramach na mieszkańca.

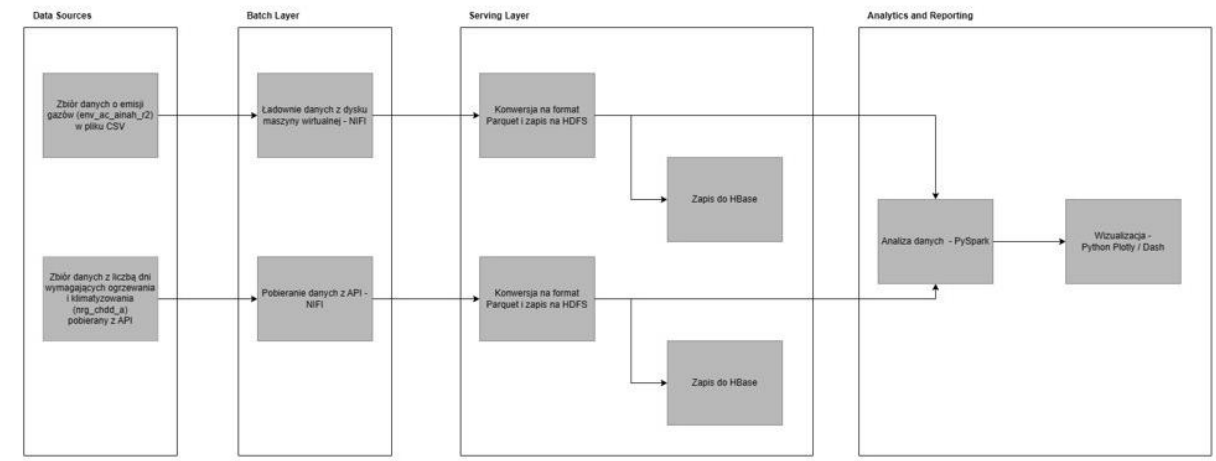
2.2 Wskaźnik stopnia dziennego ogrzewania i klimatyzowania

Dane zostały udostępnione przez Eurostat, a ich kod to nrg_chdd_a. Zbiór opisuję wskaźnik stopnia dziennego ogrzewania (HDD), czyli wskaźnik techniczny oparty na pogodzie, mający na celu opisanie zapotrzebowania budynków na energię grzewczą, oraz wskaźnik stopnia chłodu (CDD), czyli wskaźnik techniczny oparty na pogodzie, mający na celu opisanie zapotrzebowania na chłodzenie (klimatyzację) budynków. Dane rozpoczynają się od roku referencyjnego 1979. Zbiór ma 2 552 rekordów.

Wymiary danych:

- Energy indicator (INDIC_NRG) - zbierane są dane dotyczące dwóch wspomnianych wskaźników HDD i CDD
- Geopolitical entity (GEO) - 34 państwa członkowskie UE, kraje EFTA, kraje kandydujące
- Period of time (TIME) - roczne daty zaczynające się od 1979

3 Stos architektoniczny



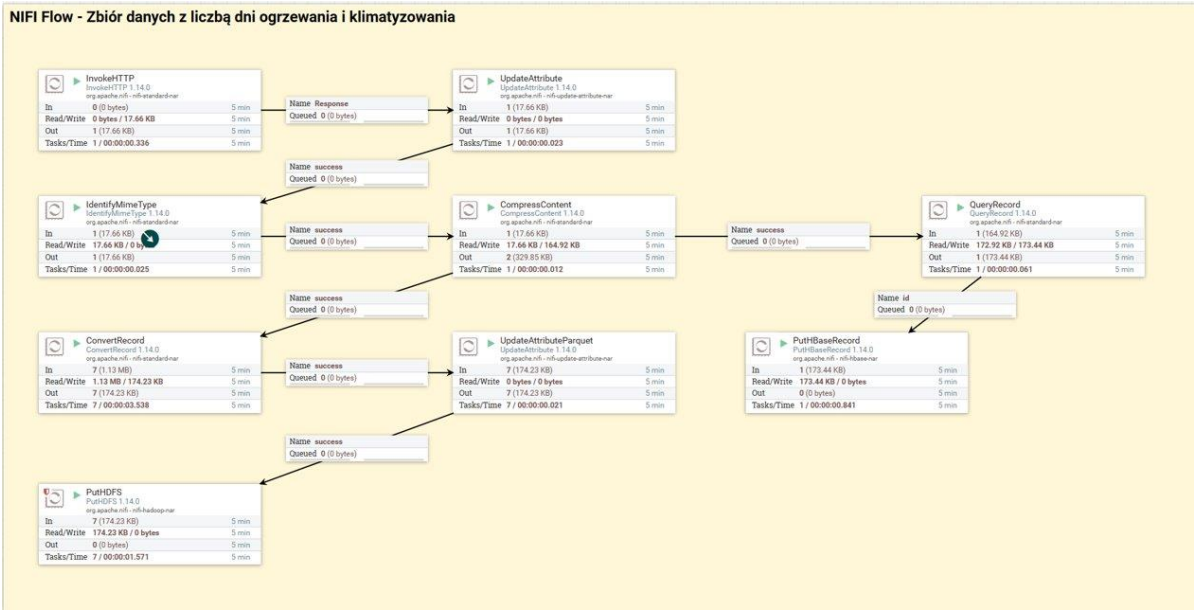
Pierwsze dwie warstwy to źródła danych oraz Batch Layer. Na etapie tych warstw przy pomocy NIFI ładowane są dane. Pierwsze źródło to zbiór o emisji gazów wczytywany z pliku CSV z dysku maszyny wirtualnej. Drugie źródło to zbiór ze wskaźnikami określającym zapotrzebowanie energetyczne ogrzewania i klimatyzowania pobierany z API. Serving Layer jest warstwą w której przeprowadzane jest rozpakowywanie załadowanych plików, ich konwersja na format Parquet i zapis w systemie HDFS. Oprócz tego realizowany jest zapis do wcześniej przygotowanych tabeli HBase. Analiza danych odbywa się w środowisku PySpark oraz Jupyter Notebook, a wizualizacje oraz dashboardy generowane są przy użyciu bibliotek Plotly oraz Dash.

3.1 Składowanie

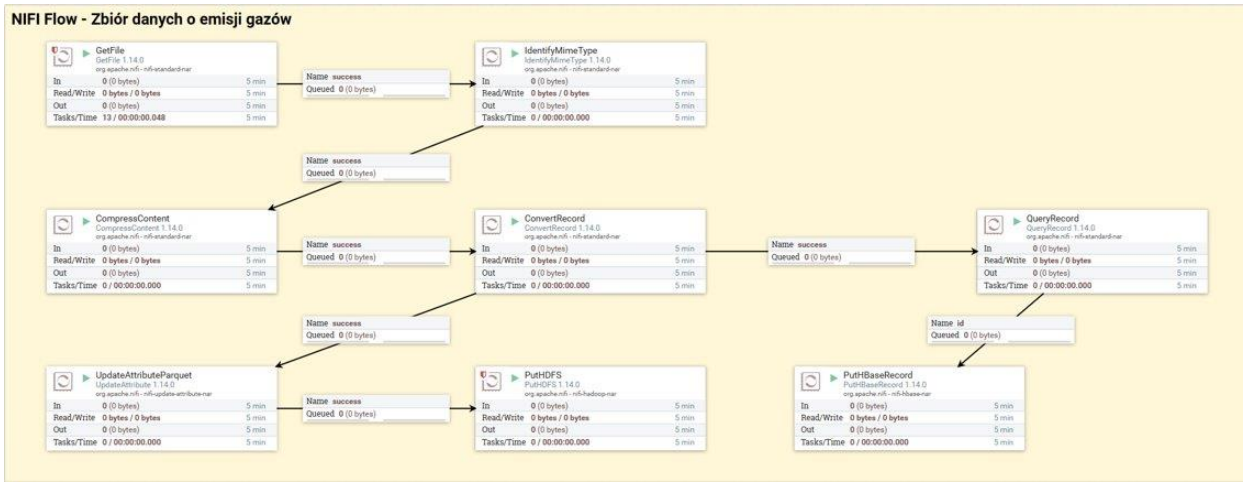
• Apache HDFS- wszystkie niezagregowane dane będą składowane w tabelach HDFS.
• Apache HBase - platforma NoSQL, która będzie składowała przechowywała dane gotowe do przypadków użycia wykorzystujących Random Access.

3.2 Przepływ danych

Przy pomocy Apache NiFi będzie następował zautomatyzowany pre-processing danych.



Przedstawiony przepływ danych odnośnie wskaźnika energetycznego ogrzewania i klimatyzowania zaczyna się od pobrania danych (processor InvokeHTTP) z otwartego API Europejskiego Urzędu Statystycznego (Eurostat). Następnie plik gzip ma identyfikowane rozszerzenie (processor IdentifyMimeType), jest rozpakowywany (processor CompressContent). Następnie zachodzą dwa procesy – pierwszy z nich to dodanie identyfikatora i zapis do tabeli nrg_chdd_a (processor QueryRecord oraz PutHBaseRecord). Drugi to konwersja do formatu Parquet (processor ConvertRecord oraz UpdateAttributeParquet) i zapis pliku w systemie HDFS (processor PutHDFS).



Przepływ danych zaczyna procesor GetFile, który pobiera plik z /home/vagrant/data/project. Następnie następuje podobny proces jak przy przepływie danych z API. Dane sa rozpakowywane oraz zapisywane do HDFS w formacie parquet oraz do Hbase do tabeli env_ac_ainah_r2.

3.3 Warstwa analityczna

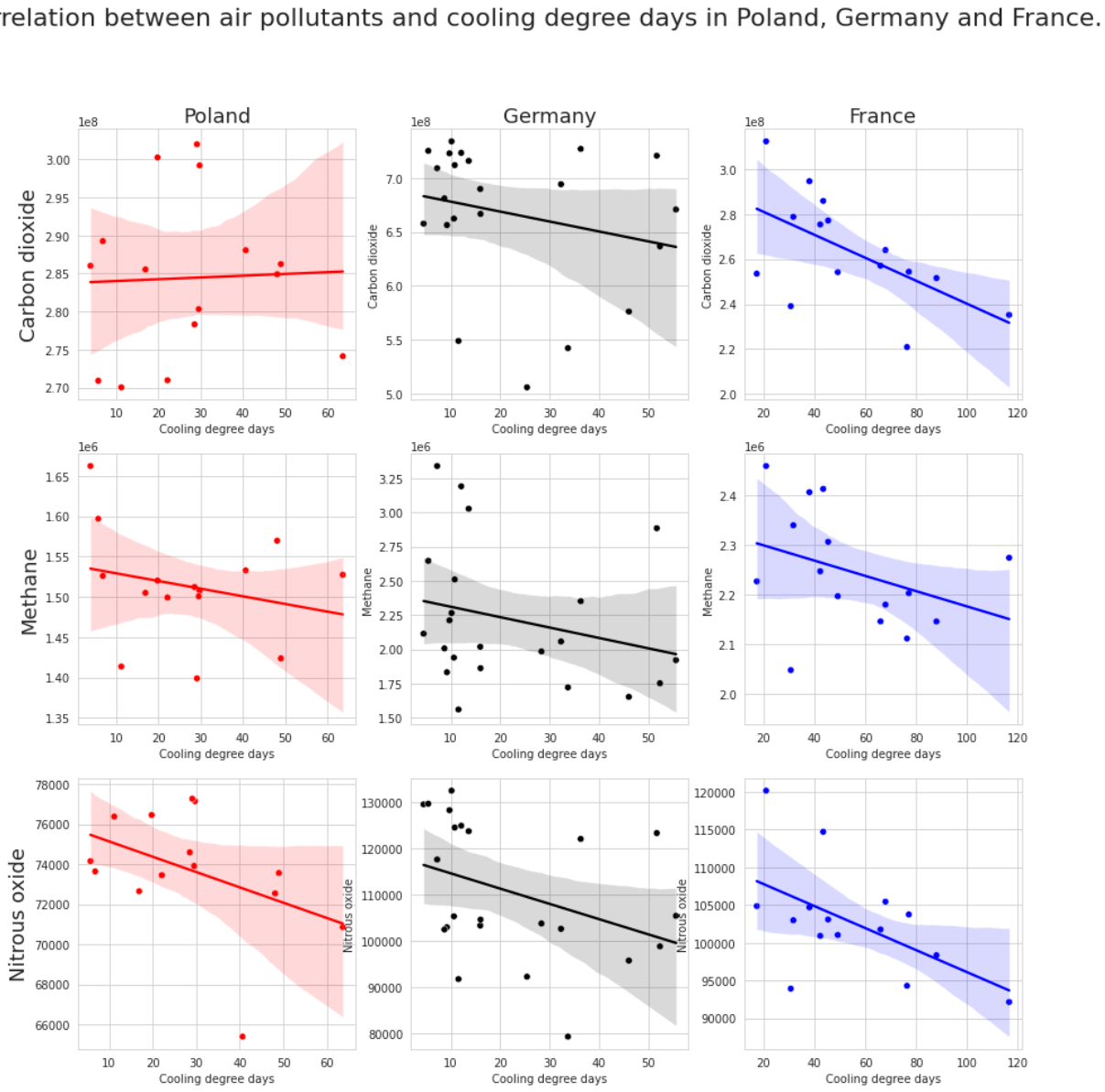
Do analizy korzystamy z narzędzia Jupyter notebook w języku python. Korzystając z Apache Spark wczytujemy dane z platformy HDFS. Łączymy i grupujemy obie bazy danych, aby otrzymać jedną macierz z danymi opisującymi rok, kraj, dane o ogrzewaniu i klimatyzacji oraz dane o szkodliwych substancjach w powietrzu.

```
# load data
ac = spark.read.option("header","true")
                .option("inferschema","true")
                .parquet("hdfs://localhost:8020/users/test/project/env_ac_ainah_r2_linear.parquet")
ac = ac.withColumn("OBS_VALUE", col("OBS_VALUE.member1").cast("double "))
chdd = spark.read.option("header","true")
                    .option("inferschema","true")
                    .parquet("hdfs://localhost:8020/users/test/project/nrg_chdd_a_linear.parquet")
```

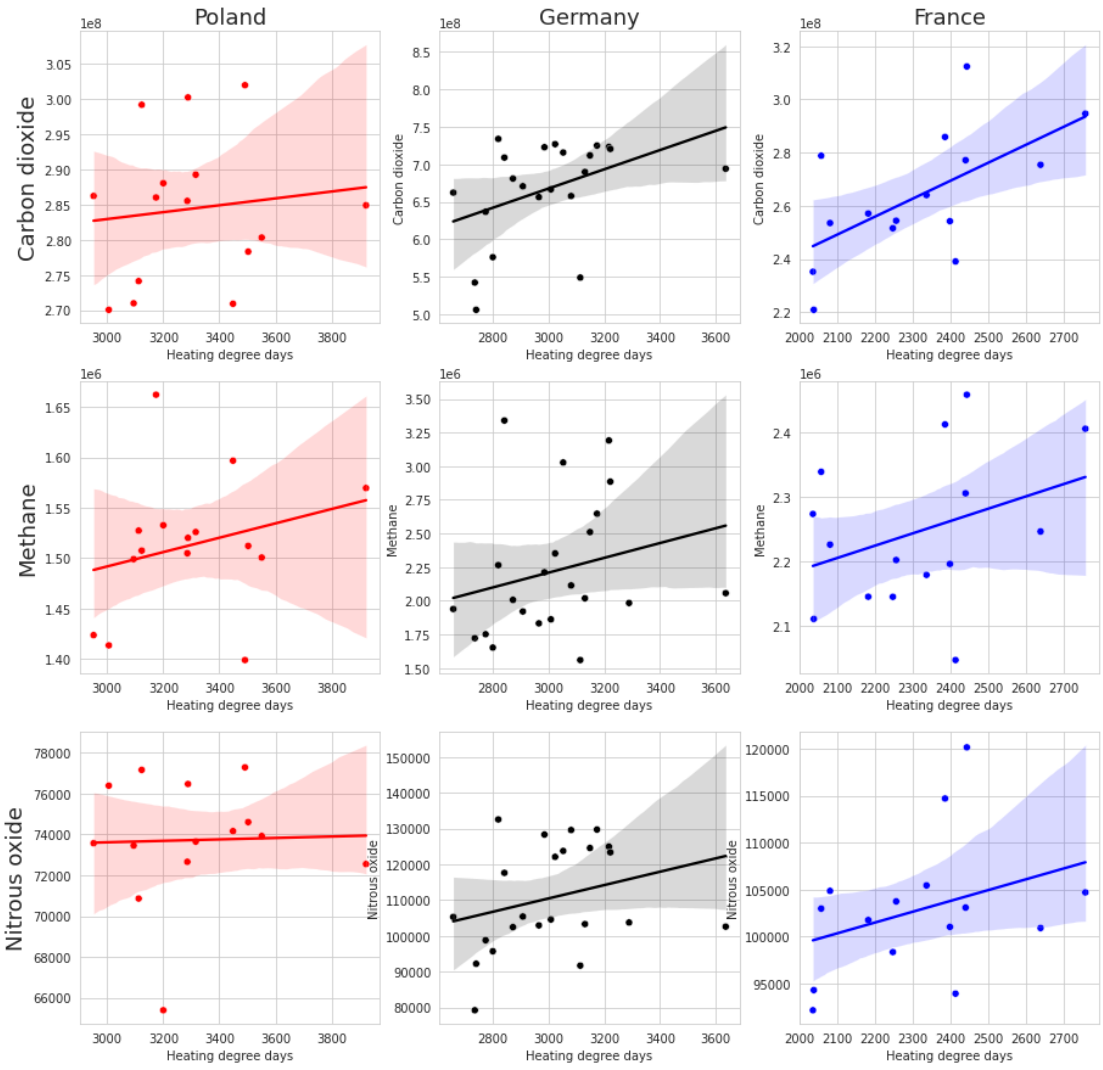
```
# Calculate corelation matrix
ac_martix = ac.select(col('airpol'),col('geo'),col('TIME_PERIOD'),col('OBS_VALUE'))
              .filter((col("nace_r2") == "TOTAL") & (col("unit") == "T")).withColumnRenamed('airpol','var')
chdd_matrix = chdd.select(col('indic_nrg'),col('geo'),col('TIME_PERIOD'),col('OBS_VALUE'))
              .filter(col("unit") == "NR").withColumnRenamed('indic_nrg','var')
matrix = ac_martix.union(chdd_matrix)
matrix = matrix.groupBy("TIME_PERIOD","geo").pivot("var").sum("OBS_VALUE")
matrix = matrix.toPandas()
matrix = map_code(matrix)
# Map column names
matrix.columns = map_code(map_code(pd.DataFrame(matrix.columns,columns=['airpol']))
                          .rename({'airpol': 'indic_nrg'}, axis=1)).iloc[:,0].to_list()
```

Następnie tworzymy wykresy korzystając z bibliotek matplotlib, seaborn, plotly oraz dash w języku python.

Pierwsze dwa wykresy są to wykresy rozproszone. Na każdym z tych wykresów analizujemy trzy wybrane polutanty: 'Carbon dioxide', 'Methane', 'Nitrous oxide' w trzech różnych państwach Unii Europejskiej: Polski, Niemiec i Francji. Na pierwszym z tych wykresów analizujemy dane wraz ze wskaźnikiem klimatyzacji. Wykres pokazuje dane z wszystkich dostepnych lat oraz regresje liniową dla posiadanych danych. Drugi wykres przedstawia to samo, ale dla wskaźnika ogrzewania.

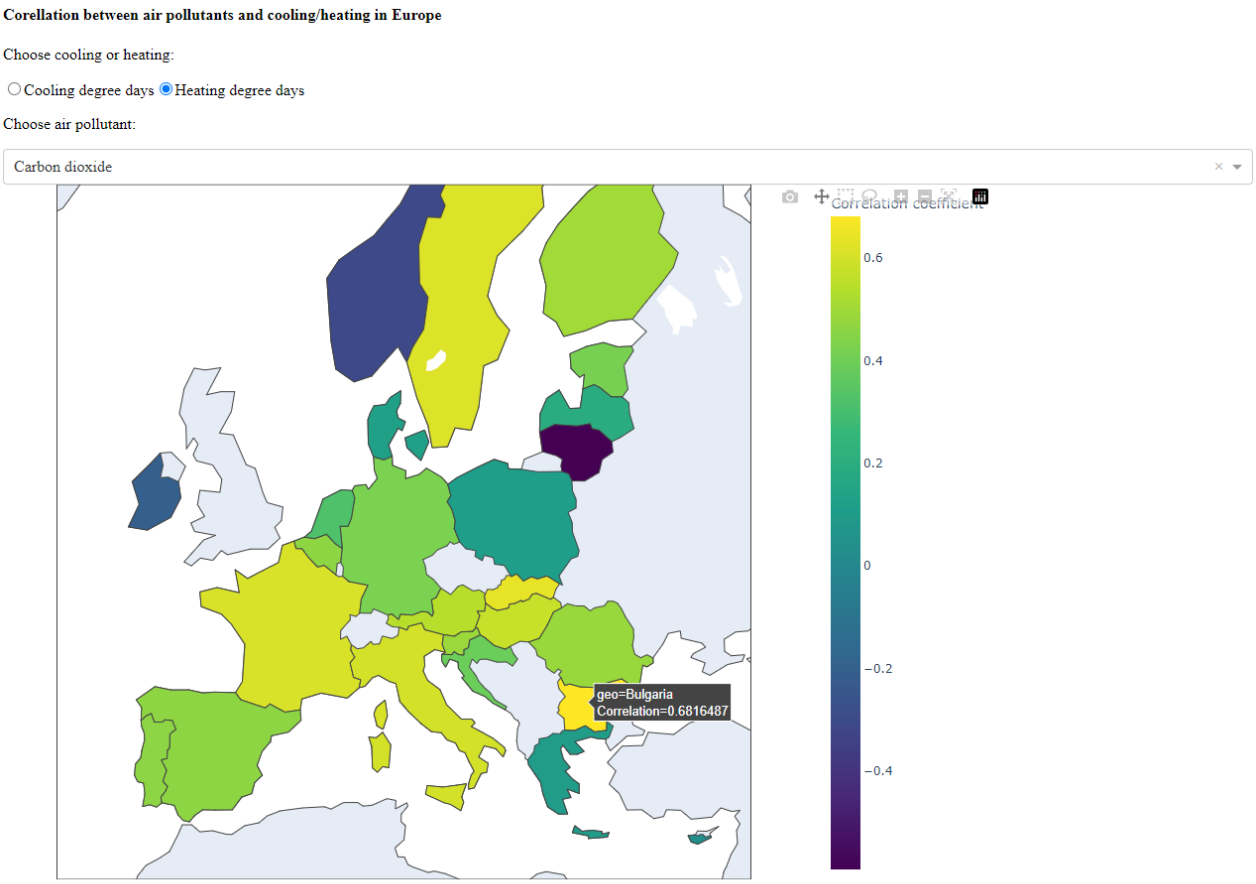


Correlation between air pollutants and heating degree days in Poland, Germany and France.

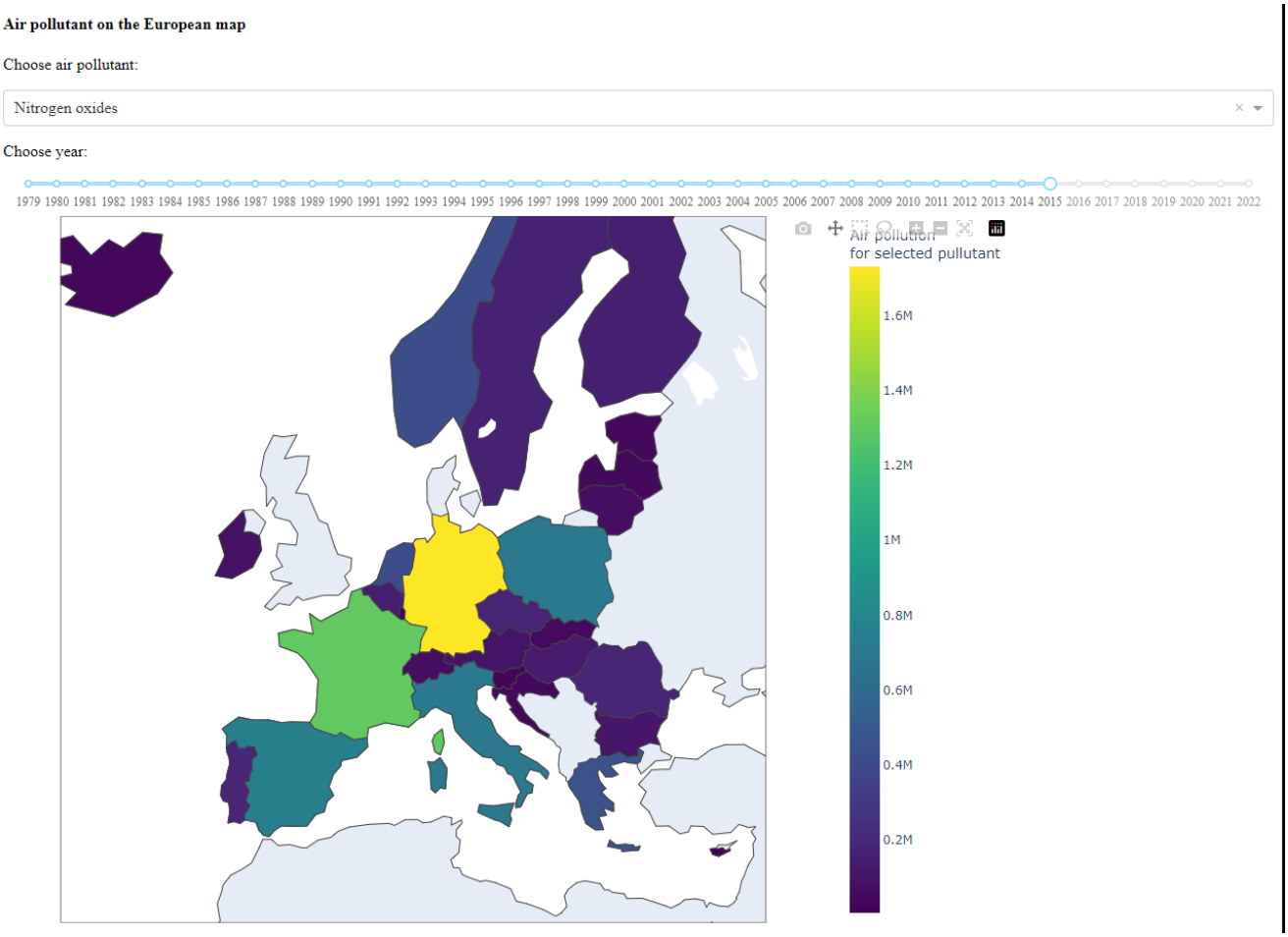


Następnie tworzone są trzy dashboardy na których jest przedstawiona mapa europy:

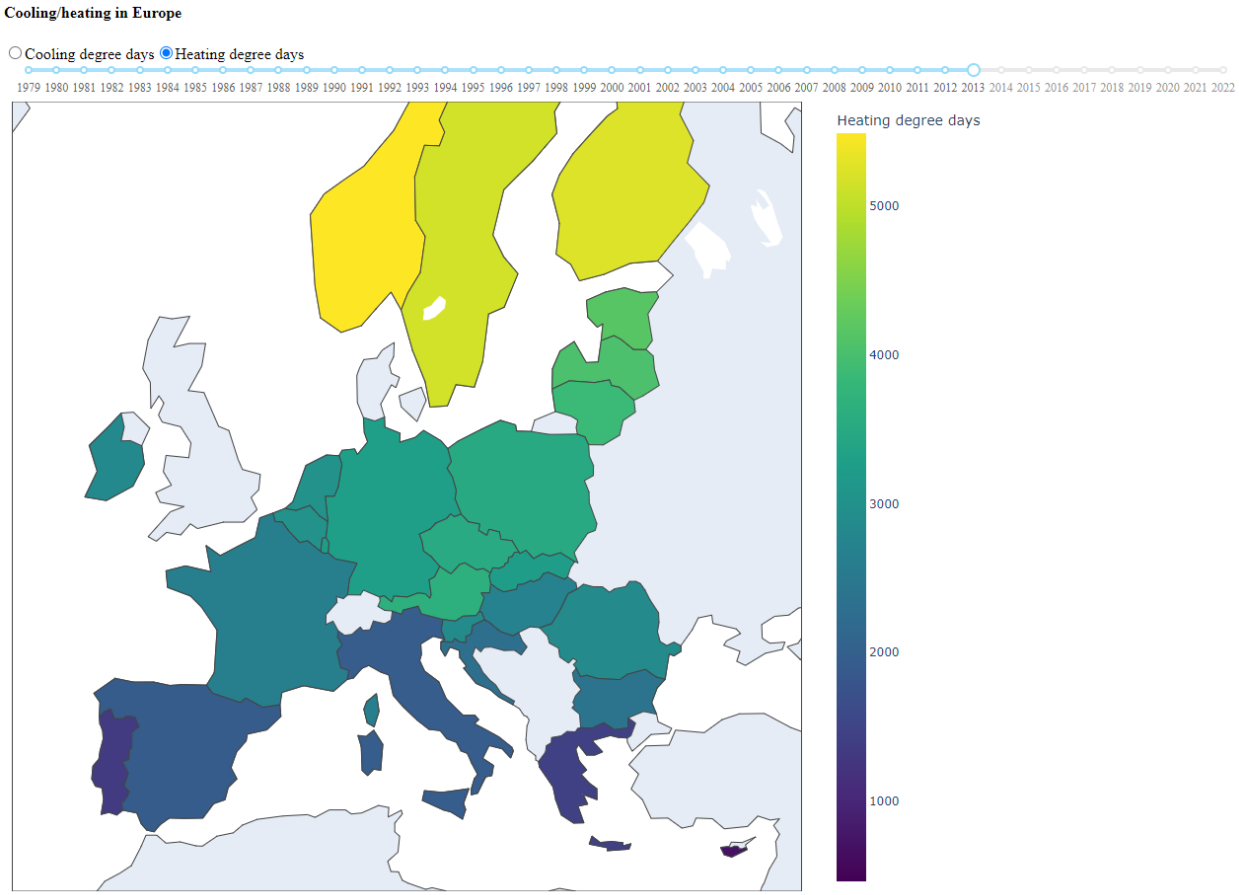
1. Możliwość wyboru polutanta oraz wskaźnika ogrzewania/klimatyzacji. Dashboard jest kolorowany w zależności od współczynnika korelacji pomiedzy dwoma wybranymi opcjami.



2. Możliwośc wyboru polutanta oraz roku. Dashboard jest kolorowany w zależności od poziomu wybranego polutanta w danym roku.



3. Możliwośc wyboru wskaźnika ogrzewania/klimatyzacji oraz roku. Dashboard jest kolorowany w zależności od poziomu wskaźnika w danym roku.



4 Podział pracy

Opis zadania	Maciej Paczóski	Krzysztof Wolny
Zaprojektowanie projektu	+	+
Utrzymanie projektu na GitHub	+	+
Dokumentacja projektu	+	+
Pozyskanie danych	+	
Automatyzacja przepływu danych		+
Składowanie danych HDFS	+	
Składowanie agregacji Hbase		+
Wsadowa analiza danych	+	+