ID2221: Data Intensive Computing
Project: COVID-19 vaccination in the EU/EEA

Due on 2021/10/24

Group 1: Yizhan Wu, Yage Hao

Contents

1 Introduction

Nowadays, COVID-19 is continuing to spread around the world, as of Sept 27, 2021, there has been 232,696,764 confirmed cases worldwide [1]. COVID-19 vaccines are the most effective way of avoiding it.

The task is to analyze and visualize the situation of COVID-19 vaccination in the European Union/European Economic Area (EU/EEA). We achieved the goal by building a data pipeline including Kafka, Spark Structured Streaming, Cassandra and visualization using Python Pandas and Pyecharts packages.

2 How to run?

Requirements:

- Python 2.7 and Python 3
- Scala 2.11.12
- Apache Kafka 2.11
- Apache Spark 2.4.3
- Apache Cassandra 3.11.2

1. Start kafka server and zookeeper:

```
zookeeper-server-start.sh $KAFKA_HOME/config/zookeeper.properties
kafka-server-start.sh $KAFKA_HOME/config/server.properties
```

2. Start Cassandra in the foreground:

```
$CASSANDRA_HOME/bin/cassandra -f
```

3. Check list of kafka topics, if topic 'covid' is not in list, create 'covid' topic:

```
kafka-topics.sh --list --zookeeper localhost:2181
kafka-topics.sh --create --zookeeper localhost:2181 --replication-factor 1
    --partitions 1 --topic covid
```

4. In /task1, run built.sbt, start structure streaming processing application for our first task:

```
sbt run
```

5. In /task2, run built.sbt, start structure streaming processing application for our second task:

```
sbt run
```

6. In /producer, run built.sbt, get messages from data.csv and feed them to topic 'covid':

```
sbt run
```

7. In /visualization, run jupyter notebook "Project-Cassandra.ipynb" to visualize the results.

3 Implementation

The data pipeline is shown in Figure 1.

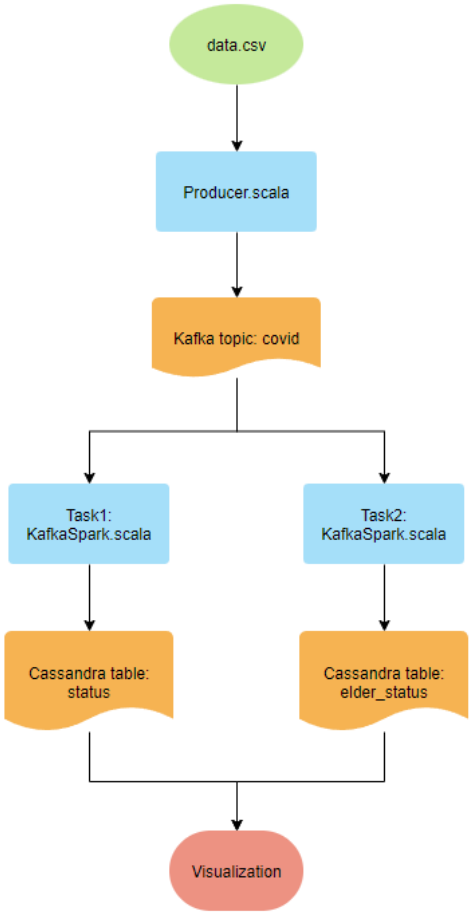

Figure 1: Data pipeline for the project.

3.1 Producer

There are three files in 'Producer' directory:

data.csv: This is our original data file. It contains information on COVID-19 vaccination in the EU/EEA. They are submitted by EU/EEA countries to ECDC through The European Surveillance System (TESSy) twice a week (Tuesdays and Fridays) [2]. Each row contains the corresponding data for a certain week and country, as well as the following information: "YearWeekISO", "ReportingCountry", "Denominator", "NumberDosesReceived", "FirstDose", "FirstDoseRefused", "SecondDose", "UnknownDose", "Region", "TargetGroup", "Vaccine", "Population". For detail explanation of each variable, please check [3].

Producer.scala: In this file, we create a Kafka producer. It reads data from 'data.csv', converts each row in original data file into a kafka message and sends it to kafka topic 'covid'.

build.sbt: This file is used to run 'Producer.scala' script with specified dependencies.

3.2 Task1

Through exploratory data analysis using Python, we decide that our first task is to compute the general COVID-19 Vaccination rates, i.e. the percentage of people who have received one dose and two doses of COVID-19 vaccine in different EU/EEA countries. We achieve this by implementing the following two files in 'Task1' directory:

KafkaSpark.scala: In general, the implementation can be divided into three steps. First is to get data from Kafka broker by subscribing to topic 'covid'. Second is to process the streaming data using Spark Structured Streaming. Third is to write result datastream to Cassandra.
In detail, in the second step, we first extract the columns relevant to our computation (including 'ReportingCountry', 'FirstDose', 'SecondDose', 'Population', 'TargetGroup', 'Region') to form a dataframe. Specifically, we set 'ReportingCountry' as key for further aggregation. Then, we defined our own mapping function. It applys stateful operation to accumulate number of first dose and second dose with respective to each country, computes and returns the vaccination rate. Finally, we applied the mapping function to grouped data.
In the last step of writing stream to Cassandra. We need to first create a keyspace 'covid' and a table 'status' in which the task1 results are stored.

build.sbt: We use this file to run 'KafkaSpark.scala' file. We need to carefully specify the dependencies and their versions.

3.3 Task2

In general, Task2 is very similar to Task1. The goal is to compute the COVID-19 Vaccination rates for the elderly, i.e. the percentage of senior people who have received one dose and two doses of COVID-19 vaccine in different EU/EEA countries. The following two files in 'Task2' directory are the implementation:

KafkaSpark.scala: The overall process is the same as Task1. First, read data from Kafka topic 'covid'. Second, process data by structured streaming. Third, output results to Cassandra.
The most significant difference between Task1 and Task2 is in the streaming processing stage. We introduced a filter based on 'TargetGroup' and only kept data from 'age 60+'. The mapping function is the same as Task1's, which computes vaccination rates. And this function is applied to our new dataframe which only contains data from the elderly.
In the output stage, we create a table 'elder_status' which is used to store the results of Task2.

build.sbt: We use this file to run 'KafkaSpark.scala' file. We need to carefully specify the dependencies and their versions.

3.4 Visualization

We visualize our Task1 and Task2 results using Pandas and Pyecharts packages in Python. In detail, we query resulting data from Cassandra and save them to Pandas dataframe. And then use Pyecharts to visualize vaccination rates on world maps. In total, we create four maps which are 'Percentage of people who have received one dose of COVID-19 vaccine in EU/EEA', 'Percentage of people who have received two doses of COVID-19 vaccine in EU/EEA', 'Percentage of senior people(60+) who have received one dose of COVID-19 vaccine in EU/EEA', and 'Percentage of senior people(60+) who have received two doses of COVID-19 vaccine in EU/EEA'.

4 Results

Stream Processing Results in Cassandra: Figure 2 and Figure 3.

```
cqlsh:covid> select * from status;

 key | per1     | per2
-----+----------+----------
  HR | 0.452801 | 0.400827
  AT | 0.626111 | 0.560687
  FR | 0.697618 | 0.609797
  EL | 0.587633 | 0.528289
  BE | 0.694061 | 0.650148
  HU | 0.569978 | 0.534157
  IT | 0.729086 | 0.638284
  SK | 0.436379 | 0.408683
  IS |  0.71472 | 0.559011
  DK | 0.713928 |  0.69733
  LU | 0.617526 | 0.541721
  DE | 0.685317 | 0.613656
  FI | 0.703985 | 0.637874
  LV | 0.452122 | 0.353181
  IE | 0.705176 | 0.648196
  SE | 0.672402 |  0.63751
  PL |  0.51405 | 0.442508
  LT | 0.612114 | 0.505689
  MT | 0.771171 | 0.705982
  SI | 0.525824 | 0.423524
  RO | 0.294661 | 0.235566
  LI | 0.613518 | 0.542648
  NO | 0.725345 | 0.683525
  BG | 0.205935 | 0.166605
  EE | 0.545001 | 0.476311
  ES | 0.747226 | 0.642731
  PT |  0.81826 | 0.645666
  NL | 0.702985 | 0.590952
  CZ | 0.548217 | 0.516926
  CY | 0.654072 | 0.612343
```

Figure 2: Vaccination rate of one dose and two doses in different countries.

```
cqlsh:covid> select * from elder_status;

 key | per1     | per2
-----+----------+----------
  HR | 0.703286 | 0.661299
  AT | 0.892389 |  0.84461
  FR | 0.917661 | 0.835064
  EL | 0.796569 | 0.766452
  BE |  0.93875 | 0.908917
  HU | 0.811822 | 0.783969
  IT |  0.93314 | 0.848929
  SK | 0.670328 | 0.650265
  IS |  1.03926 |  1.01712
  DK | 0.991448 | 0.992736
  LU | 0.867765 | 0.827194
  FI | 0.953876 | 0.917201
  LV | 0.553748 | 0.459611
  IE |  1.04068 |  1.02258
  SE | 0.943071 | 0.923042
  PL | 0.776156 | 0.715291
  LT | 0.761618 | 0.689241
  MT |  1.00657 | 0.981737
  SI | 0.797653 |  0.72422
  RO | 0.364155 | 0.321712
  NO | 0.986849 | 0.970864
  BG | 0.306573 | 0.275721
  EE | 0.732817 | 0.685788
  ES | 0.997053 | 0.937481
  PT |  1.03993 | 0.949431
  CZ | 0.821092 | 0.781834
  CY | 0.936745 | 0.909448
```

Figure 3: Vaccination rate of one dose and two doses in senior group in different countries.

Visualization Results: Figure 4, 5, 6, 7.

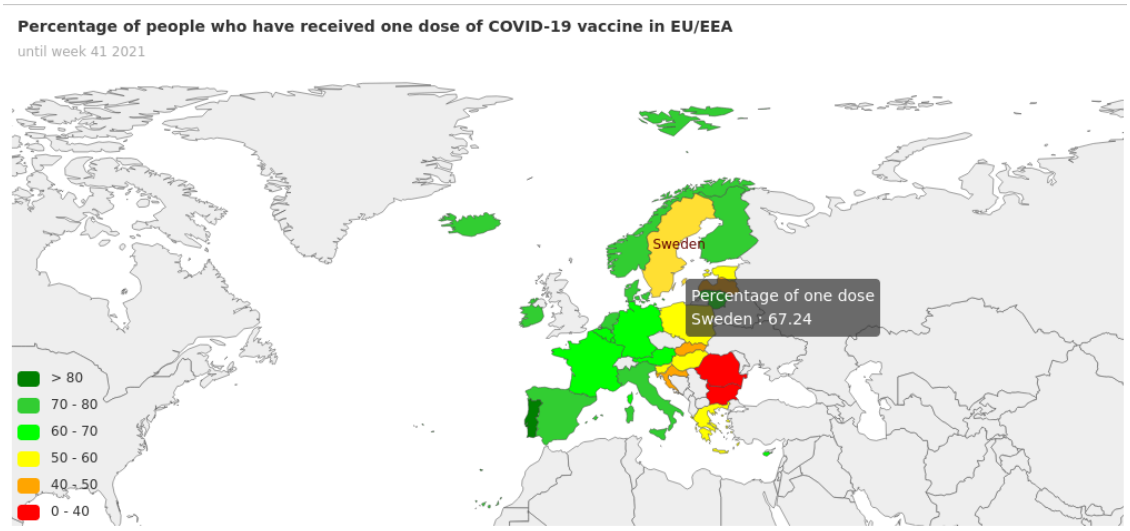

Figure 4: Percentage of people who have received one dose of COVID-19 vaccine in EU/EEA.

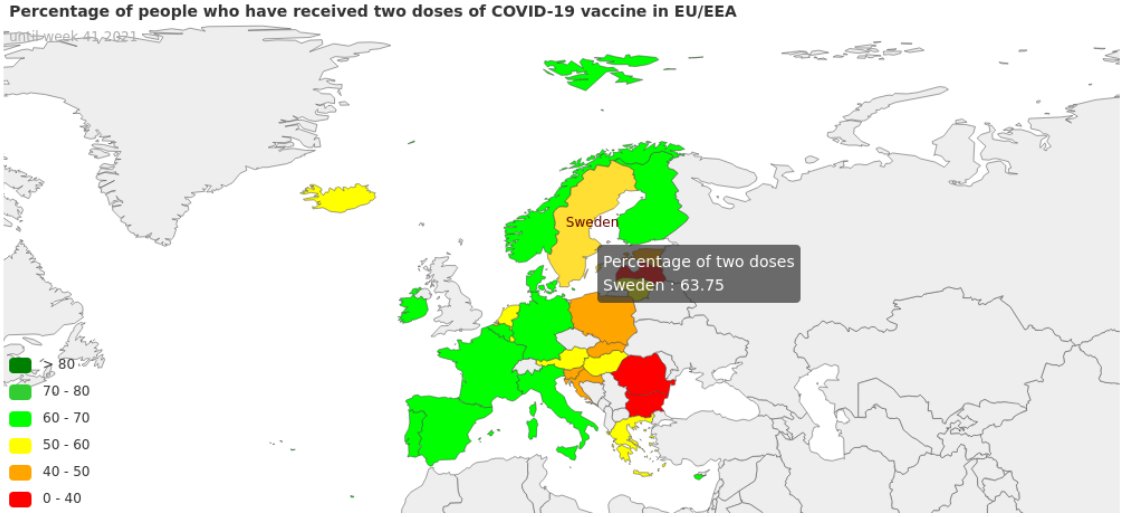

Figure 5: Percentage of people who have received two doses of COVID-19 vaccine in EU/EEA.

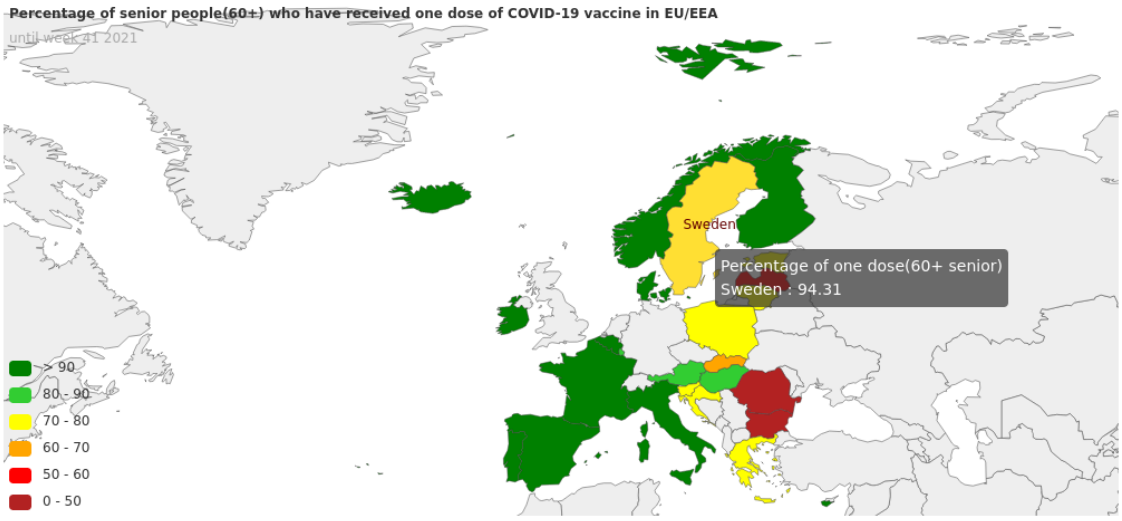

Figure 6: Percentage of senior people(60+) who have received one dose of COVID-19 vaccine in EU/EEA.

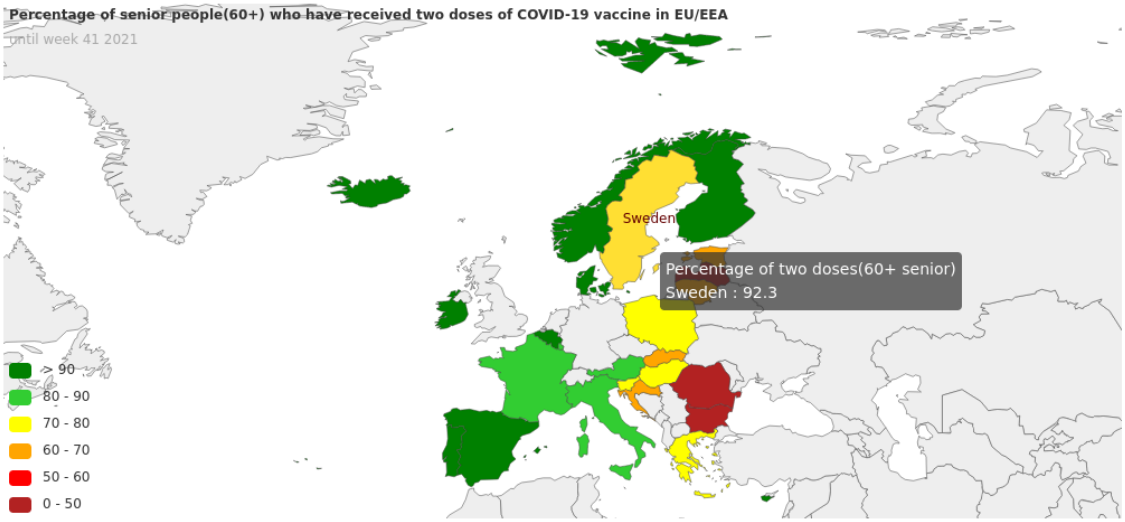

Figure 7: Percentage of senior people(60+) who have received two doses of COVID-19 vaccine in EU/EEA.

References

[1] Worldometer, *Worldometers.info. 2021. COVID Live Update: 232,696,764 Cases and 4,764,064 Deaths from the Coronavirus*, Available: https://www.worldometers.info/coronavirus/?utm_campaign=homeAdvegas1?

[2] European Centre for Disease Prevention and Control, *2021. Data on COVID-19 vaccination in the EU/EEA*, Available: https://www.ecdc.europa.eu/en/publications-data/data-covid-19-vaccination-eu-eea

[3] European Centre for Disease Prevention and Control, *Data dictionary for the COVID-19 vaccination data in the EU/EEA*, Available: https://www.ecdc.europa.eu/sites/default/files/documents/Variable_Dictionary_VaccineTracker-20-08-2021.pdf

[4] Amir H. Payberah, *Lab 2 - Stream and Graph Processing with Spark, Kafka, and Cassandra*, Available: https://github.com/fstaccone/DataIntensiveComputing/blob/master/lab2Assignment/lab2.pdf

[5] Pyecharts.org., *pyecharts - A Python Echarts Plotting Library built with love.*., Available: https://pyecharts.org/#/