Group Project Part 3 - ElasticSearch

of Systems and Methods for Big and Unstructured Data Course (SMBUD)

held by

Brambilla Marco

Tocchetti Andrea

Group 14

Banfi Federico 10581441

Carotenuto Alessandro 10803080

Donati Riccardo 10669618

Mornatta Davide 10657647

Zancani Lea 10608972

Academic year 2021/2022

Contents

1 Problem Specification **4**

2 Hypotheses **4**

3 Dataset Schemata **5**
- 3.1 administration-vaccines-latest.csv . . . 6
- 3.2 dpc-covid19-ita-regions.csv . . . 6
- 3.3 SARS-CoV-2-variants-data-in-EU-EEA.csv . . . 7
- 3.4 owid-covid-data.csv . . . 7

4 Queries and Commands **8**
- 4.1 Queries . . . 8
  - 4.1.1 Total number of vaccinations ordered per day [VACCINES INDEX] . . . 9
  - 4.1.2 Total number of vaccinations group by region and provider [VACCINES INDEX] . . . 11
  - 4.1.3 Astrazeneca doses grouped by age range [VACCINES INDEX] . . . 13
  - 4.1.4 Find all the informations about the vaccinations in Lombardia on the 25-11-2021 [VACCINES INDEX] . . . 15
  - 4.1.5 Find all the vaccinations not in the islands in the age range 80+ [VACCINES INDEX] . . . 17
  - 4.1.6 Find the most infected country in the World [GLOBAL_COVID INDEX] . . . 19
  - 4.1.7 Find the number of positive people per region [INFECTION_REGION INDEX] . . . 21
  - 4.1.8 Find the more common variants in a country in a week[VARIANTI INDEX] . . . 23
- 4.2 Commands . . . 26
  - 4.2.1 Update the number of vaccinations in a record (UPDATE) [VACCINES INDEX] . . . 26
  - 4.2.2 Delete a document (DELETE) [VACCINES INDEX] . . . 26

5 DashBoard **27**
- 5.1 Total doses per day in Italy . . . 27
- 5.2 Total doses per region and provider in Italy . . . 28
- 5.3 Total doses per age range and provider in Italy . . . 28
- 5.4 Gender comparison on total vaccinations in Italy . . . 29
- 5.5 Trend of vaccines compared with deaths and infections in Italy 29

5.6 Trend of new cases in the world 31
5.7 Trend of new cases in the italian regions 31
5.8 Trend of intensive care occupation and hospitalization in Italy 32
5.9 Mortality in the italian regions 32
5.10 Percentage spread of variants in a given week 33

6 References & Sources **34**

1 Problem Specification

The use of Big Data is not limited exclusively to the tracking and profiling of vast amounts of data to obtain information useful for practical purposes - as seen in previous deliveries - but are also fundamental in the field of research and of statistical studies. This time, in fact, our project is focused on the use of an Information Retrieval system specifically designed to the collection and the in-depth study of data to obtain more precise and meaningful analyzes.

In particular, the ElasticSearch search engine will be used to monitor the spread of the SARS-CoV-2 virus in recent times from a global point of view and to track the progress of the vaccination campaign from a perspective no longer linked to individuals but to a larger scale. Thanks to the IR-based approach on which it is based, ElasticSearch is the most suitable system for this kind of task as, compared to other SQL and NoSQL solutions, it gives the possibility to compute more accurate searches by exploiting the main features available: the full-text search and the ranking strategy used to evaluate the relevance of the results obtained, which guarantee more wide-ranging query outcomes in order of "importance".

2 Hypotheses

This time, pre-existing datasets released by various government authorities with large amounts of data collected over the last few years were examined, moreover, in order to obtain significant results and realistic information in executing the queries, a time interval was taken at least 3 months. The datasets considered are:

1. administration-vaccines-latest.csv issued by the *"Commissione Straordinaria per la gestione l'emergenza Covid-19"* appointed by the Italian Government,
2. dpc-covid19-ita-regions.csv issued by the *"Dipartimento della Protezione Civile"* in accordance with the rules established by the *"Presidenza del Consiglio dei Ministri"* and the *"Ministero della Salute della Repubblica Italiana"* regarding the Covid-19 emergency,
3. SARS-CoV-2-variants-data-in-EU-EEA.csv issued by the *"European Center for Disease Prevention and Control"*,
4. owid-covid-data.csv released by the Oxford University research group that manages the scientific online publication website *ourworldindata.org*, the data collected comes from various institutions such as the *"Center for Systems Science and Engineering (CSSE) at Johns Hopkins University"*.

3 Dataset Schemata

The datasets are in .csv format and can be found in the *"/datasets"* folder, two of these contain data relating to the administration of vaccines and cases ascertained for the Italian national territory only, while the other two relate to the spread of SARS-CoV-2 variants in Europe and the global pandemic trend around the world.

During the import of the data present within each dataset, it was preferred to take advantage of the Automatic Mapping feature available on ElasticSearch, subsequently checking that there were no compatibility problems with the types of the various fields. Furthermore, for those fields that expected String data, the keyword type was used instead of full-text to obtain more precise and targeted results by executing the queries. Below are the tables with the diagrams of the various datasets produced following the Mapping, some fields have been omitted for the sake of brevity (to test the queries not all fields have been taken into consideration and reporting them all in the following tables would have been superfluous), however, it is possible to view the complete diagrams of each dataset with all the fields in the *"/documents/report/dataset_schemata"* folder.

In order to correctly display the dashboards relating to each dataset, you must first import the file *"/datasets/Dashboard.ndjson"*, in this the following indexes will be created:

1. varianti
2. infection_region
3. global_covid
4. vaccines

and then, when uploading the datasets, check that the mappings follow the structure specified in *"dataset_schemata.pdf"*. Finally assign them respectively to:

1. SARS-CoV-2-variants-data-in-EU-EEA.csv
2. dpc-covid19-ita-regioni.csv
3. owid-covid-data.csv
4. somministrazioni-vaccini-latest.csv

3.1 administration-vaccines-latest.csv

Field Name	Data Type	Description
fornitore	String	Complete name of the supplier of the vaccine
data_somministrazione	Datetime	Administration date of the vaccines
fascia_anagrafica	String	Age group of the people administered with the vaccines
sesso_maschile	Long	Number of vaccinations administered to males
sesso_femminile	Long	Number of vaccinations administered to females
prima_dose	Long	Number of people administered with the first dose
seconda_dose	Long	Number of people administered with the second dose
pregressa_infezione	Long	Number of people administered with a dose after they have been infected
dose_addizionale_booster	Long	Number of people administered with an additional dose/recall
codice_NUTS1	String	Nomenclature of Territorial Units for Statistics code (Groups of regions)
nome_area	String	Name of the region

3.2 dpc-covid19-ita-regions.csv

Field Name	Data Type	Description
data	Datetime	Date of observation
denominazione_regione	String	Name of the Region
terapia_intensiva	Long	Intensive Care
totale_ospedalizzati	Long	Total hospitalised patients
totale_positivi	Long	Total amount of current positive cases
variazione_totale_positivi	Long	News amount of current positive cases
nuovi_positivi	Long	News amount of current positive cases
deceduti	Long	Death

3.3 SARS-CoV-2-variants-data-in-EU-EEA.csv

Field Name	Data Type	Description
new_cases	Long	Weekly number of new confirmed cases
number_detections_variant	Long	Number of detections reported of the variant
number_sequenced	Long	Weekly number of sequences carried out
number_sequenced _known_variant	Long	Weekly number of known variant sequences carried out
source	String	Data source, either GISAID EpiCoV database or TESSy
variant	String	Each VOC (*"variants of concern"*, i.e. those considered the most lethal), Other or UNK
year_week	String	Year and week of reference

3.4 owid-covid-data.csv

Field Name	Data Type	Description
new_cases	Long	New confirmed cases of COVID-19
new_deaths	Long	New deaths attributed to COVID-19
location	String	Geographical location
date	Datetime	Date of observation
new_cases_per_million	Long	New confirmed cases of COVID-19 per 1,000,000 people

4 Queries and Commands

4.1 Queries

In all the queries performed in an index with a timestamp we are going to filter the documents only considering the 4 months period that we are analysing. For readability we won't represent this part in the next queries.

FILTER

```
  . . .
 "query": {
   "bool": {
     "must": [
       {
         "range": {
           "data_somministrazione": {
             "gte": "2021-08-25T00:00:00Z+01",
             "lte": "2021-12-25T00:00:00Z+01"
           }
         }
       }
     ]
   }
 },
 . . .
```

4.1.1 Total number of vaccinations ordered per day [VACCINES INDEX]

This query allows us to retrieve an ordered list containing the total number of vaccinations (first + second + booster + previous infection) per day.

QUERY

```
GET /vaccines/_search
{
  "size":0,
 "aggs": {
   "vacc_by_date": {
     "terms": {
       "field": "data_somministrazione",
       "order": {"Totale_dosi": "desc"}
     },
     "aggs" : {
        "Totale_dosi" : { "sum" : { "script" : "
         doc['prima_dose'].value +
         doc['seconda_dose'].value +
         doc['dose_addizionale_booster'].value +
         doc['pregressa_infezione'].value" } }
     }
   }
 }
}
```

The output is given by:

- The Date as a key
- The doc count
- The total number of vaccinations

OUTPUT


```
"aggregations" : {
   "vacc_by_date" : {
     "doc_count_error_upper_bound" : -1,
     "sum_other_doc_count" : 48995,
     "buckets" : [
       {
         "key" : 1639612800000,
         "key_as_string" : "2021-12-16T00:00:00.000Z",
         "doc_count" : 413,
         "Totale_dosi" : {
           "value" : 584181.0
         }
       },
       {
         "key" : 1639699200000,
         "key_as_string" : "2021-12-17T00:00:00.000Z",
         "doc_count" : 417,
         "Totale_dosi" : {
           "value" : 578422.0
         }
       },
       . . .
```

4.1.2 Total number of vaccinations group by region and provider [VACCINES INDEX]

This query (performed on the vaccines index) allows us to retrieve the total number of vaccinations, as the previos one, grouped by region and provider of the vaccine.

QUERY

```
GET /vaccines/_search
{
  "size":0,
 "aggs": {
   "vacc_by_date": {
     "terms": {
       "field": "nome_area"
     },
     "aggs": {
       "s": {
         "terms": {
           "field": "fornitore",
            "order": {"Totale_dosi": "desc"}
         },
         "aggs" : {
            "Totale_dosi" : { "sum" : { "script" : "
            doc['prima_dose'].value +
            doc['seconda_dose'].value +
            doc['dose_addizionale_booster'].value +
            doc['pregressa_infezione'].value" } }
        }
       }
     }
   }
 }
}
```

The output is given by:

- The name of the region
- The names of the providers with the total related vaccinations

OUTPUT


```
"buckets" : [
        {
          "key" : "Lazio",
          "doc_count" : 9852,
          "s" : {
            "doc_count_error_upper_bound" : 0,
            "sum_other_doc_count" : 0,
            "buckets" : [
              {
                "key" : "Pfizer/BioNTech",
                "doc_count" : 3238,
                "Totale_dosi" : {
                  "value" : 7622311.0
                }
              },
              {
                "key" : "Moderna",
                "doc_count" : 2984,
                "Totale_dosi" : {
                  "value" : 1418839.0
                }
              },
              . . .
```

4.1.3 Astrazeneca doses grouped by age range [VACCINES INDEX]

This query allows us the monitor the vaccinations of the provider Astrazeneca in the different age ranges.

QUERY

```
GET /vaccines/_search
{
  "size":0,
  "aggs": {
    "aggr1": {
      "terms": {
        "field": "fascia_anagrafica"
      },
      "aggs" : {
            "Totale_dosi" : { "sum" : { "script" : "
            doc['prima_dose'].value +
            doc['seconda_dose'].value +
            doc['dose_addizionale_booster'].value +
            doc['pregressa_infezione'].value" } }
        }
    }
  }
}
```

The output is given by:

- The age range
- The total number of Astrazeneca vaccinations

OUTPUT

```
"buckets" : [
        {
          "key" : "60-69",
          "doc_count" : 481,
          "Totale_dosi" : {
            "value" : 7382.0
          }
        },
        {
          "key" : "70-79",
          "doc_count" : 417,
          "Totale_dosi" : {
            "value" : 3314.0
          }
        },
        . . .
```

4.1.4 Find all the informations about the vaccinations in Lombardia on the 25-11-2021 [VACCINES INDEX]

This query allows us the retrieve some specific information about the vaccinations in a certain region, a certain day.

QUERY

```
GET /vaccines/_search
{
  "query": {
    "bool": {
      "must": [
        {
          "term": {
            "data_somministrazione": {
              "value": "2021-11-25T00:00:00"
            }
          }
        },
        {
          "term": {
            "nome_area": {
              "value": "Lombardia"
            }
          }
        }
      ]
    }
  }
}
```

The output is given by:

- The number of documents that match the query
- The fields of the documents that match the query

OUTPUT


```
"hits" : {
  "total" : {
    "value" : 20,
    "relation" : "eq"
  },
  "max_score" : 3.8716726,
  "hits" : [
    {
      "_index" : "vaccines",
      "_type" : "_doc",
      "_id" : "mVXh7H0BdN5GzFyWWMIo",
      "_score" : 3.8716726,
      "_source" : {
        "area" : "LOM",
        "codice_regione_ISTAT" : 3,
        "nome_area" : "Lombardia",
        "data_somministrazione" : "2021-11-25",
        . . .
```

4.1.5 Find all the vaccinations not in the islands in the age range 80+ [VACCINES INDEX]

This query also allows to perform a search of specific documents in the database. i.e. the vaccinations that didn't take place in the islands (NUTS code != ITG) for the age range of 80+ years.

QUERY

```
GET /vaccines/_search
{
   "query": {
     "bool": {
       "filter": [
         {
           "bool": {
             "must_not":[
               {
                 "term":{
                   "codice_NUTS1":"ITG"
                 }
               }
              ],
             "should":[
                  {
                    "term":{
                      "fascia_anagrafica":"80-89"
                    }
                  },
                  {
                     "term":{
                      "fascia_anagrafica":"90+"
                    }
                  }
               ]
           }
         }
       ]
     }
   }
}
```

The output is given by:

- The number of documents that match the query
- The fields of the documents that match the query

OUTPUT


```
"hits" : {
    "total" : {
      "value" : 10000,
      "relation" : "gte"
    },
    "max_score" : 0.0,
    "hits" : [
      {
        "_index" : "vaccines",
        "_type" : "_doc",
        "_id" : "zVXh7H0BdN5GzFyWUlDT",
        "_score" : 0.0,
        "_source" : {
          "area" : "CAM",
          "codice_regione_ISTAT" : 15,
          "nome_area" : "Campania",
          "data_somministrazione" : "2021-09-17",
          "dose_addizionale_booster" : 2,
          "codice_NUTS1" : "ITF",
          . . .
```

4.1.6 Find the most infected country in the World [GLOBAL_COVID INDEX]

This query allows us to find the country with the biggest number of new infected people per million in the world on the 20-12-2021.

QUERY

```
GET /global_covid/_search
{
  "query": {
    "bool": {
      "must": [
        {
          "term": {
            "date": {
              "value": "2021-12-20T00:00:00"
            }
          }
        }
      ]
    }
  },
  "sort": [
  {
    "new_cases_per_million": {
    "order": "desc"
  }
  }
],
"size": 1
}
```

The output is given by:

- The 20-12-2021 document of the country (ANDORRA)

OUTPUT


```
"continent" : "Europe",
          "date" : "2021-12-20",
          "new_cases_per_million" : 6631.848,
          "location" : "Andorra",
          "total_tests" : "259560.0",
          "iso_code" : "AND",
          "tests_per_case" : "7.3",
          "total_tests_per_thousand" : "3355.483"
          . . .
```

4.1.7 Find the number of positive people per region [INFECTION_REGION INDEX]

This query allows us to find the total number of positive people exploiting the variation of the positives (new cases - healed cases), until the last record in the database, grouped by region and ordered in decreasing order.

QUERY

```
GET /infection_region/_search
{
  "size":0,
  "aggs": {
    "agg1": {
      "terms": {
        "field": "denominazione_regione",
        "order": {"attuali_positivi": "desc"}
      },
      "aggs" : {
            "attuali_positivi" : { "sum" : { "script" :
            "doc['variazione_totale_positivi'].value" } }
        }
    }
  }
}
```

The output is given by:

- The name of the region
- The number of actual positives

OUTPUT


```
"buckets" : [
        {
          "key" : "Lombardia",
          "doc_count" : 669,
          "attuali_positivi" : {
            "value" : 94367.0
          }
        },
        {
          "key" : "Veneto",
          "doc_count" : 669,
          "attuali_positivi" : {
            "value" : 65479.0
          }
        },
        . . .
```

4.1.8 Find the more common variants in a country in a week[VARIANTI INDEX]

This query allows us to find the more commons variants of COVID-19 in Italy the 32th week of 2021 calculating the percentage for each variants on the total of the sequenced people and sorting in descending order.

QUERY

```
GET /varianti/_search
{
  "query": {
    "bool": {
      "must": [
        {
          "term": {
            "country": {
              "value": "Italy"
            }
          }
        },
        {
          "term": {
            "year_week": {
              "value": "2021-32"
            }
          }
        }
      ]
    }
  },
  "sort": {
    "_script": {
        "script": "Math.round(
        (double)doc['number_detections_variant'].value /
        (double)doc['number_sequenced'].value
         *100*100)/100.0",
        "type": "number",
        "order": "desc"
    }
}
}
```

The output is given by:

- The document of the week in the country
- The diffusion's percentage of the variant

OUTPUT


```
{
        "_index" : "varianti",
        "_type" : "_doc",
        "_id" : "gE1M_H0BPJPV_obP69wN",
        "_score" : null,
        "_source" : {
          "country" : "Italy",
          "new_cases" : 43135,
          "percent_variant" : 98.8,
          "source" : "GISAID",
          "number_sequenced_known_variant" : 1688,
          "year_week" : "2021-32",
          "country_code" : "IT",
          "number_detections_variant" : 1668,
          "number_sequenced" : 1704,
          "percent_cases_sequenced" : 4.0,
          "variant" : "B.1.617.2",
          "valid_denominator" : "Yes"
        },
        "sort" : [
          97.89
        ]
      },
      . . .
```

4.2 Commands

4.2.1 Update the number of vaccinations in a record (UPDATE) [VACCINES INDEX]

In this command we simulate the need for changing the number of vaccinations in a specific document.

COMMAND

```
PUT vaccines/_doc/B1Xh7H0BdN5GzFyWW-4i
{
  "area" : "BAS",
  "codice_regione_ISTAT" : 17,
  "nome_area" : "Basilicata",
  "data_somministrazione" : "2021-12-23",
  "dose_addizionale_booster" : 0,
  "codice_NUTS1" : "ITF",
  "fascia_anagrafica" : "40-49",
  "prima_dose" : 1,
  "pregressa_infezione" : 0,
  "fornitore" : "Janssen",
  "@timestamp" : "2021-12-23T00:00:00.000+01:00",
  "seconda_dose" : 0,
  "sesso_maschile" : 0,
  "codice_NUTS2" : "ITF5",
  "sesso_femminile" : 2
}
```

4.2.2 Delete a document (DELETE) [VACCINES INDEX]

In this command we delete a document from the database via document _id that we previously retrieved with a GET.

COMMAND

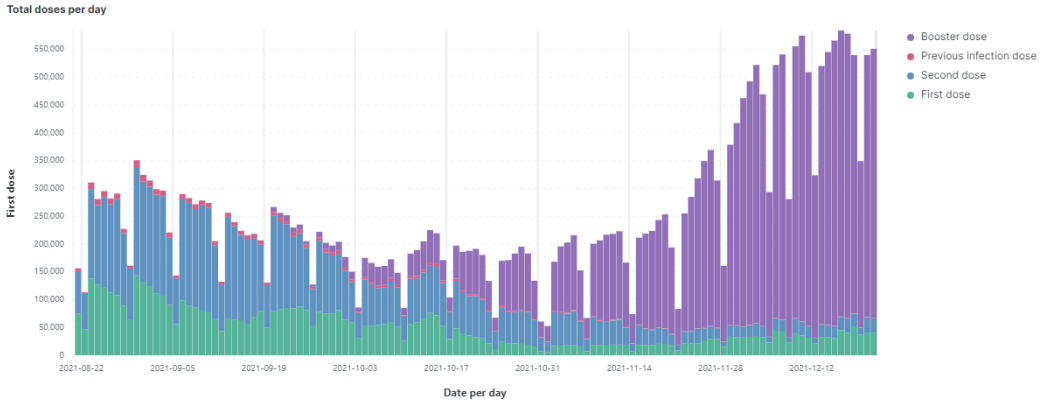
```
DELETE vaccines/_doc/B1Xh7H0BdN5GzFyWW-4i
```

5 DashBoard

In this section we are going to describe the dashboard that we realized in Kibana. It is an informative dashboard about the latest data about the COVID-19 (in the dashboard the time horizon is fixed between 21-08-2021 and 21-12-2021), comprehending the vaccinations, tests, variants in the regional level of Italy but also in a generic level in the World.
All the datasets have been used in this dashboard.

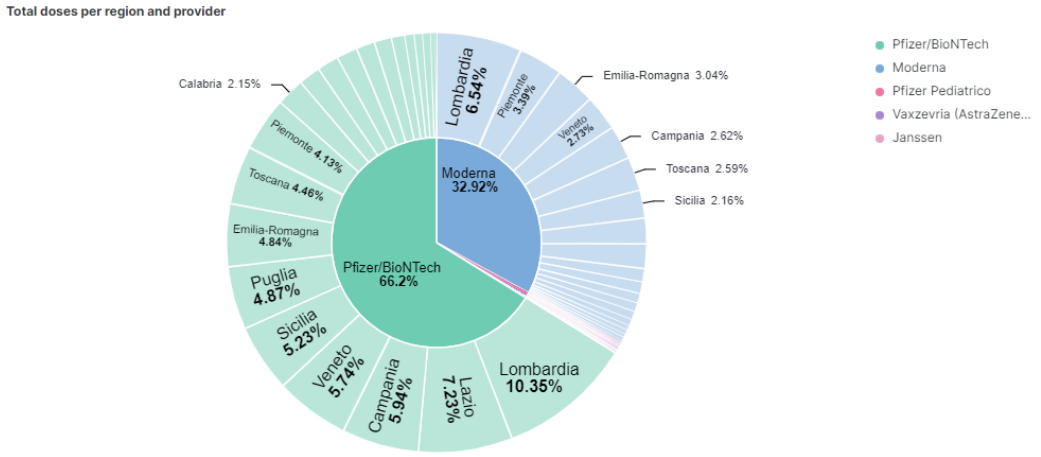
5.1 Total doses per day in Italy

This representation is derived from the first query and it shows the distribution of the vaccinations in Italy. We can modify the time horizon as we please and focus on just a limited types of vaccinations.
We can see a large increase in booster doses over the past few months as expected.



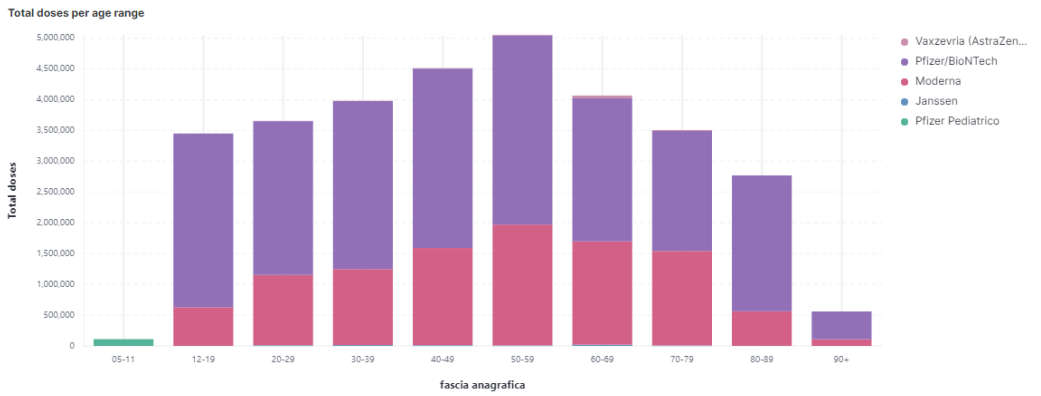
5.2 Total doses per region and provider in Italy

This representation is also derived from the second query and it shows for each vaccine provider the percentage of vaccinations for each region. We can focus on a specific region or a specific provider.
As we can see Astrazeneca in the last months has almost vanished, but if we extend the time horizon to the previous months the slice gets bigger



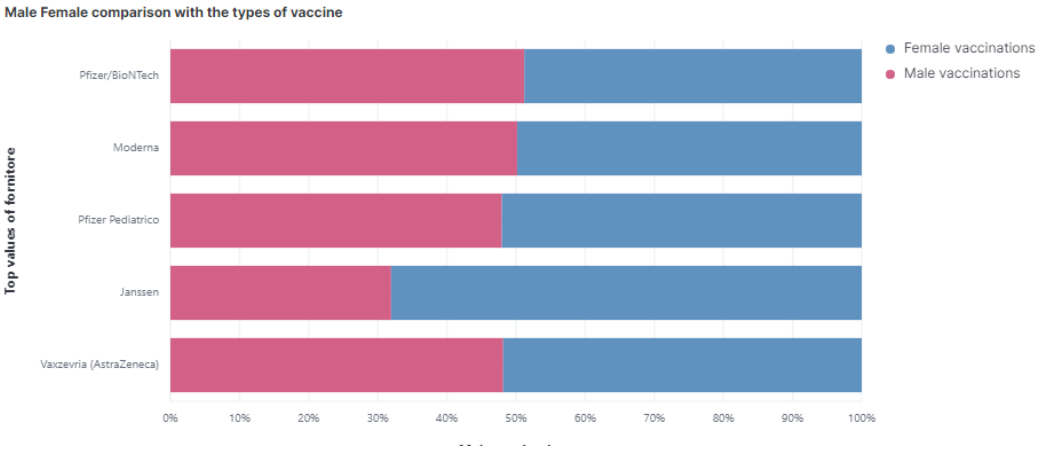
5.3 Total doses per age range and provider in Italy

This representation is derived from a generalization of the third query, it shows the total doses divided per age range and breaked down by vaccine type. We can modify the time horizon as we please and focus on just a limited types of vaccinations or age ranges.
We can make the same conclusion about the Astrazeneca vaccine.



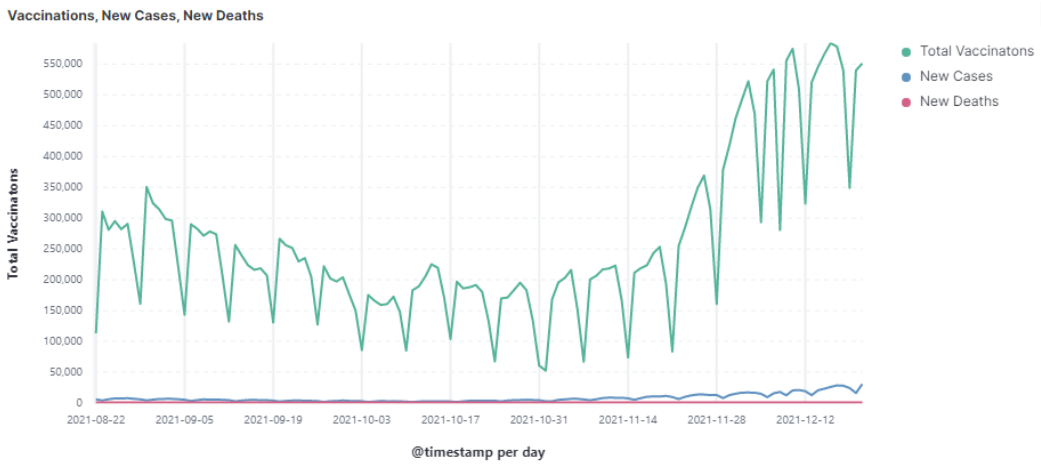
5.4 Gender comparison on total vaccinations in Italy

This representation shows the total vaccinations breaked down by vaccine type in percentage, between male and female.
We can focus on just one type of vaccine or gender.



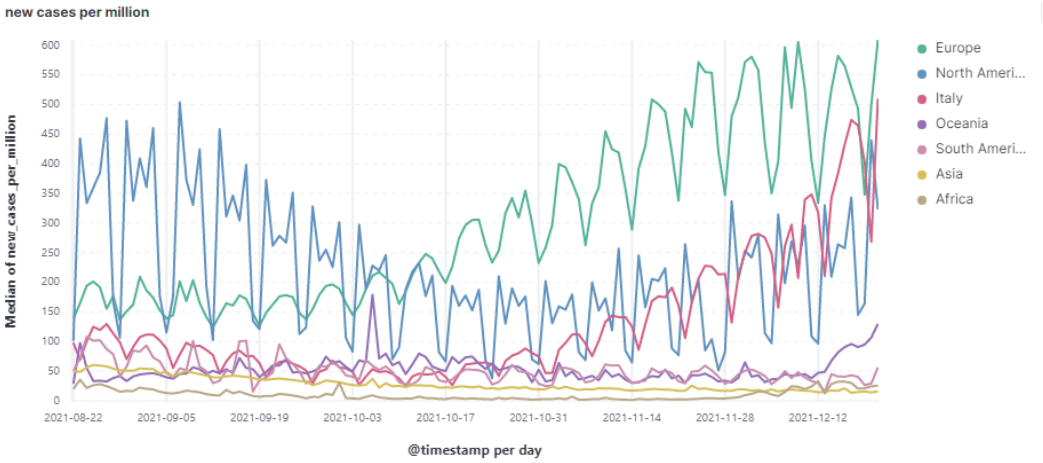
5.5 Trend of vaccines compared with deaths and infections in Italy

In this representation we mixed two different indexes (vaccinations and global infections/deaths) in order to get a comparison of this data in Italy in the same graph.
We can navigate the graph exploiting the informations of just one index for example hiding the information about the vaccinations to get a clearer graph about infections and deaths.


Vaccinations, New Cases, New Deaths
Total Vaccinatons
New Cases
New Deaths
550,000
500,000
450,000
400,000
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
Total Vaccinatons
2021-08-22
2021-09-05
2021-09-19
2021-10-03
2021-10-17
2021-10-31
2021-11-14
2021-11-28
2021-12-12
@timestamp per day

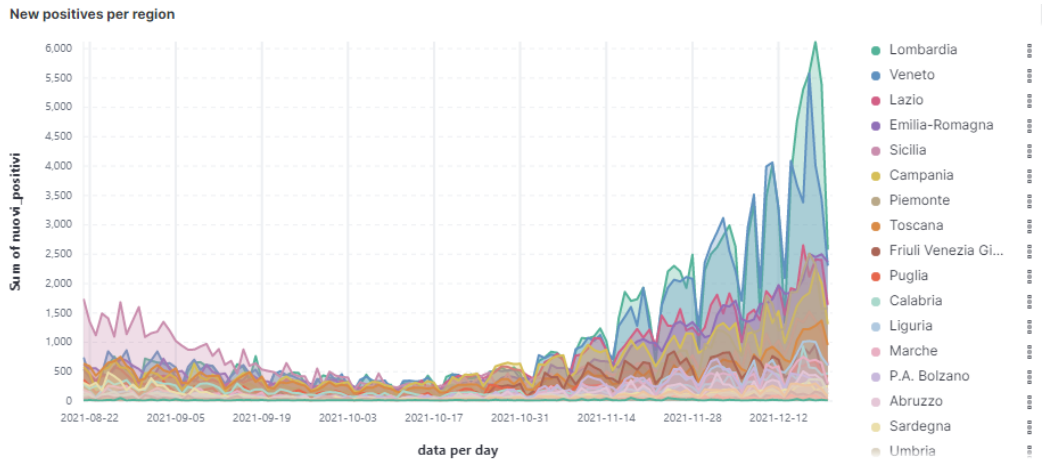
5.6 Trend of new cases in the world

In this representation we exploits the global index to get a trend of the continents new cases per million and we compare it with the new cases in Italy (red line).



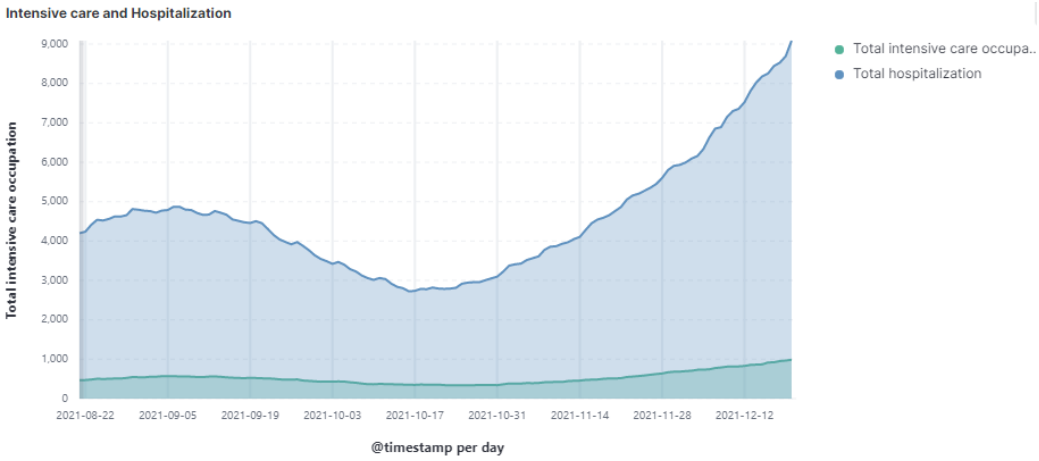
5.7 Trend of new cases in the italian regions

In this representation we exploits the regional infection index to get a trend of the italian regional new cases.
We can navigate the graph and select each region, or set of regions, one by one to get a clearer view of the trend.



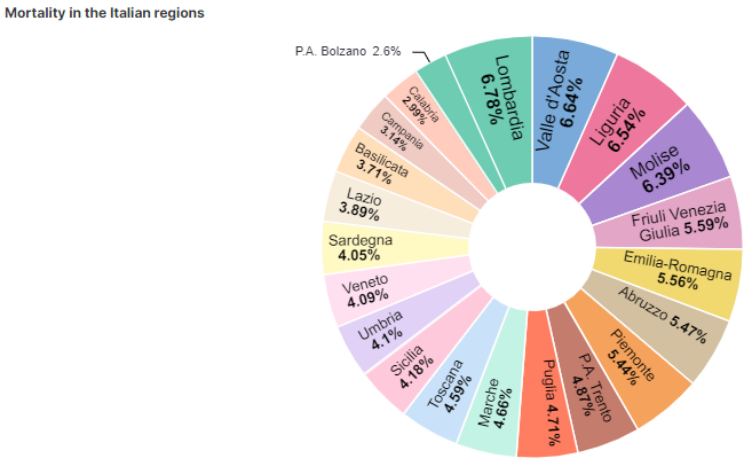
5.8 Trend of intensive care occupation and hospitalization in Italy

In this representation we exploits the regional infection index to get a trend of the hospitalization compared with the intensive care occupations.
It is paramount important to keep checked this information in order to avoid health crisis.



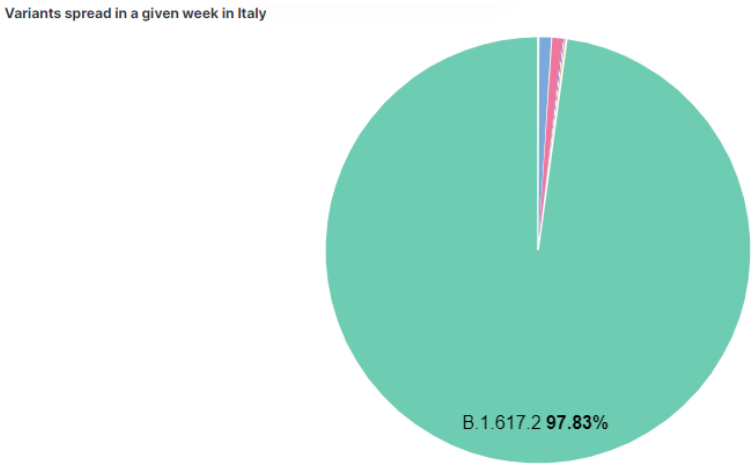
5.9 Mortality in the italian regions

In this representation we can compare the percentage mortality region by region. It is interesting trying to understand the meaning behind this graph.



5.10 Percentage spread of variants in a given week

In this last representation we exploit the variants index in order to visualize the query number 8, that is, the spread on a pie chart of the variants in Italy the 32th week of the 2021.



6 References & Sources

[1] Course Slides

[2] https://github.com/italia/covid19-opendata-vaccini/

[3] https://www.ecdc.europa.eu/en/publications-data/data-virus-variants-covid-19-eueea

[4] https://github.com/pcm-dpc/COVID-19

[5] https://github.com/owid/covid-19-data

[6] https://www.elastic.co/guide/en/elasticsearch/reference/current/index.html