Universidad Rey Juan Carlos

Escuela técnica superior de ingeniería informática

Doble grado en ingeniería informática y matemáticas

Curso Académico 2019/2020

Trabajo fin de grado

DISEÑO Y OPTIMIZACIÓN DE QUERIES DE BIG DATA CON SPARK EN AWS

Autor: *Yihui Xia*

Director: *Juan Manuel Serrano*

26 de abril de 2020

Índice general

Introducción

Durante el desarrollo de la carrera, hemos visto varios lenguajes de programación, como: Pascal, Java, C, Prolog, Scala, Python, etc.
En este trabajo voy a profundizar en Scala, la programación funcional, y sobre todo Spark, librería utilizada para el estudio de Big Data.
El trabajo se puede considerar que tiene dos grandes bloques. En el primer bloque, trabajaré con datos de pequeño tamaño, se hará queries sobre esos datos y sacar conclusiones. Para representar los resultados se hará uso de las librerías gráficas de Scala como Plotly y de la plataforma Databricks. Y en el segundo bloque, utilizaré datos de mayor tamaño, y trabajaré sobre todo con esos datos con los servicios proporcionado por el Amazon Web Services (AWS).
Todos los códigos serán subidos al repositorio de GitHub.

Capítulo 1

Bloque I: Queries con Spark, Plotly y Databricks

1.1 Descripción de los datos

Para el primer bloque voy a trabajar con varias bases de datos. Los datos están relacionados con los desastres naturales: número de mortalidad, tipos de desastres naturales y su respectiva frecuencia. También analizaré qué relación tiene con el cambio climático, sobre todo con el cambio de la temperatura.

Origen: Los bases de datos serán tomadas de estas tres páginas principalmente:

- Desastres naturales: https://ourworldindata.org/natural-disasters actualizado en 2019 - 11 con datos desde 1900 hasta actualidad. Intentaré reproducir resulatados de esta página, y también conclusiones propias.
- Desastres naturales: https://www.kaggle.com/dataenergy/natural-disaster-data actualizado hace 1 año con datos desde 1900 hasta actualidad.
- Cambio climático(temperatura): actualizado hace 3 años con datos desde 1743 hasta actualidad.

Formato: todos en csv.
Los dataset son:

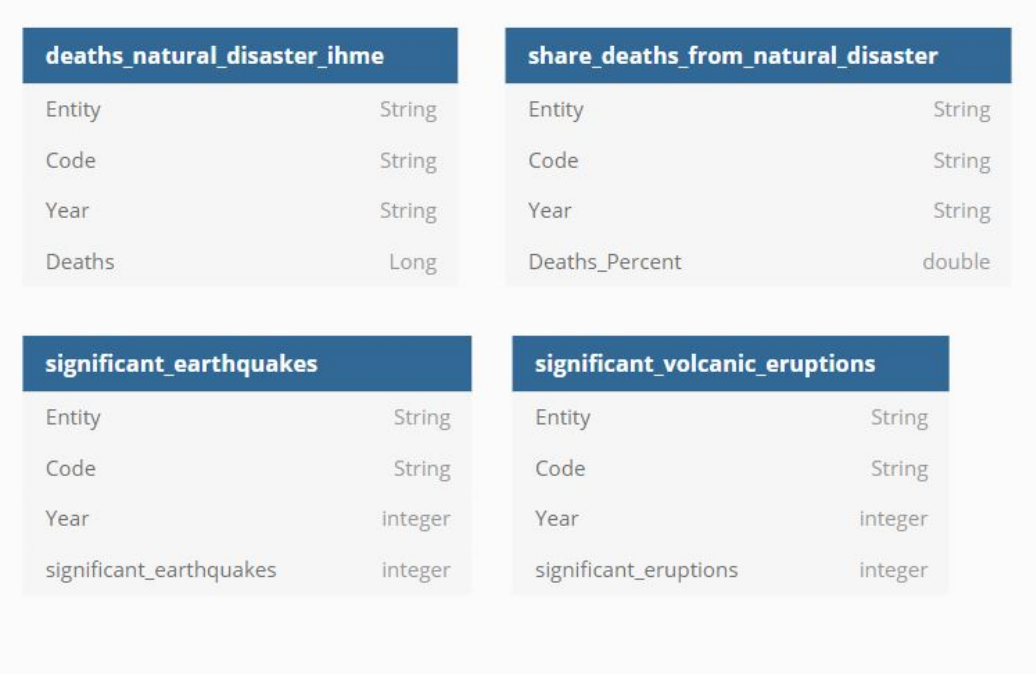
- GlobalLandTemperaturesByCountry.csv (22149 KB): contiene la temperatura media mensual de cada país.
- Number-of-natural-disaster-events.csv (19 KB): número de desastres naturales por año y por tipos.

- economic-damage-from-natural-disasters.csv (18 KB): el daño económico total causado por los desastres naturales por año y por tipo.
- number-of-deaths-from-natural-disasters.csv (20 KB): número de Muertos debido a los desastres naturales por año y por tipo.
- deaths-natural-disasters-ihme.csv (198 KB): número de Muertos debido a los desastres naturales por año y por país.
- share-deaths-from-natural-disasters.csv (206 KB): porcentaje de Muertos debido a los desastres naturales por año y por país.
- significant-earthquakes.csv (2901 KB): número de terremotos significativos por país y año.
- significant-volcanic-eruptions.csv (354 KB): número de erupciones volcanicas significativas por país y año.

Tamaño: suma de los csv utilizados (25851 KB ≅ 25 MB)

1.2 Modelo de datos

Para esta sección voy a utilizar https://dbdiagram.io/home para modelar los datos.



Y su respectiva código en SQL:

```
CREATE TABLE `GlobalLandTemperaturesByCountry` (
  `dt` timestamp,
  `AverageTemperature` double,
  `AverageTemperatureUncertainty` double,
  `Country` String
);

CREATE TABLE `Number_of_natural_disaster_events` (
  `Entity` String,
  `Code` null,
  `Year` integer,
  `Number` integer
);

CREATE TABLE `economic_damage_from_natural_disasters` (
  `Entity` String,
  `Code` null,
  `Year` integer,
  `money` Long
);

CREATE TABLE `Number_of_deaths_from_natural_disasters` (
```

```
    `Entity` String,
    `Code` null,
    `Year` integer,
    `Deaths` integer
);

CREATE TABLE `deaths_natural_disaster_ihme` (
    `Entity` String,
    `Code` String,
    `Year` String,
    `Deaths` Long
);

CREATE TABLE `share_deaths_from_natural_disaster` (
    `Entity` String,
    `Code` String,
    `Year` String,
    `Deaths_Percent` double
);

CREATE TABLE `significant_earthquakes` (
    `Entity` String,
    `Code` String,
    `Year` integer,
    `significant_earthquakes` integer
);

CREATE TABLE `significant_volcanic_eruptions` (
    `Entity` String,
    `Code` String,
    `Year` integer,
    `significant_eruptions` integer
);
```

1.3 Planteamiento inicial

En esta sección voy a listar unas queries que servirán de guía a la hora de implementar las soluciones.

1. ¿Cómo evoluciona el número de desastres naturales según avanza los

años?

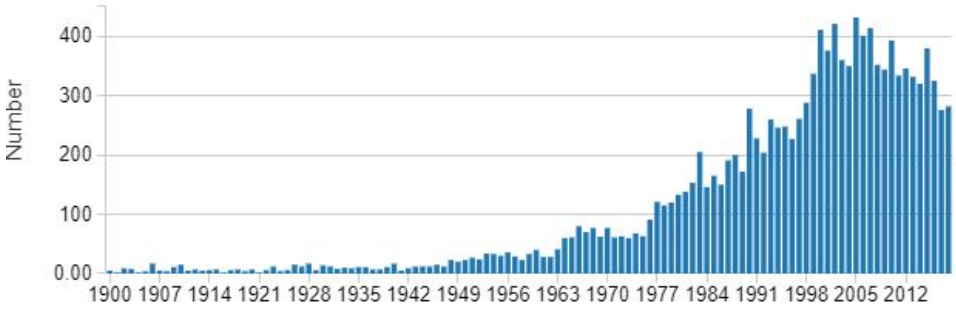
2. ¿Cómo evoluciona el daño económico causado por los desastres naturales?
3. ¿Existe alguna relación entre el número de desastres naturales con el daño económico?
4. ¿Cómo evoluciona la temperatura media anual global?
5. ¿Cómo evoluciona la temperatura media anual de cada país?
6. ¿Existe alguna relación entre la temperatura media anual global con el número de desastres naturales?
7. ¿Cómo evoluciona el número de muertes anual causados por los desastres naturales?
8. ¿Cómo evoluciona el porcentaje de muertes anual causados por los desastres naturales?
9. ¿Cuál es el país con más muertes por los desastres naturales?
10. ¿Existe alguna relación entre el número de desastres naturales con el número total de muertos anuales causados por los desastres naturales?
11. ¿Cómo evolucionan cada tipo de desastres naturales según avanza el tiempo?
12. ¿Qué tipo de desastres naturales provoca mayores muertos?
13. ¿Qué tipo de desastres naturales provoca mayores daños económicos?
14. ¿Cuándo hubo más concentración de terremotos significativos?
15. ¿Cuándo hubo más concentración de erupciones volcánicas?

1.4 Solución con Databricks

Databricks es un plataforma que soporta a Python, R, Scala, SQL. Dispone de un método muy fácil e interactivo para mostrar gráficas en Scala: *display*. La documentación de Databricks: https://docs.databricks.com.

```
1 // 1.   ¿Cómo evoluciona el número de desastres naturales según avanza los años?
2 display(disasterNumber.filter($"Entity" ==="All natural disasters"))
```

▸ (1) Spark Jobs

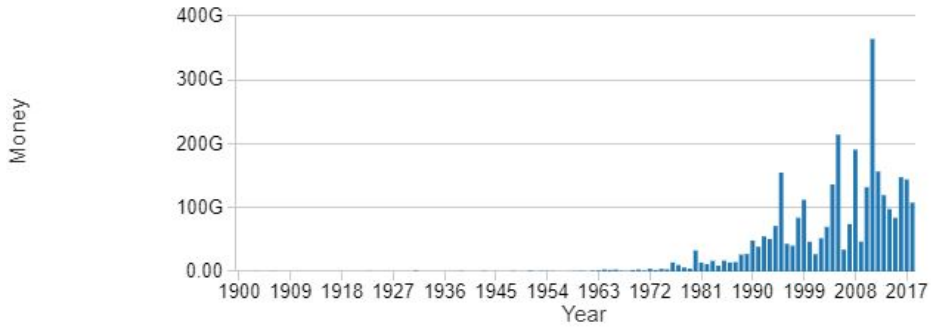


Plot Options...

Figura 1.1: Databricks Solución 1

```
1 // 2.¿Cómo evoluciona el daño económico causado por los desastres naturales?
2 display(disasterEconomic.filter($"Entity"==="All natural disasters"))
```

▸ (1) Spark Jobs



Plot Options...

Figura 1.2: Databricks Solución 2

```scala
//3.¿Existe alguna relación entre el número de desastres naturales con el daño económico?

val diasasterNumberEconomic=disasterNumber.join(disasterEconomic, Seq("Entity","Year"))
display(diasasterNumberEconomic)
```

(1) Spark Jobs

diasasterNumberEconomic: org.apache.spark.sql.DataFrame = [Entity: string, Year: integer ... 2 more fields]

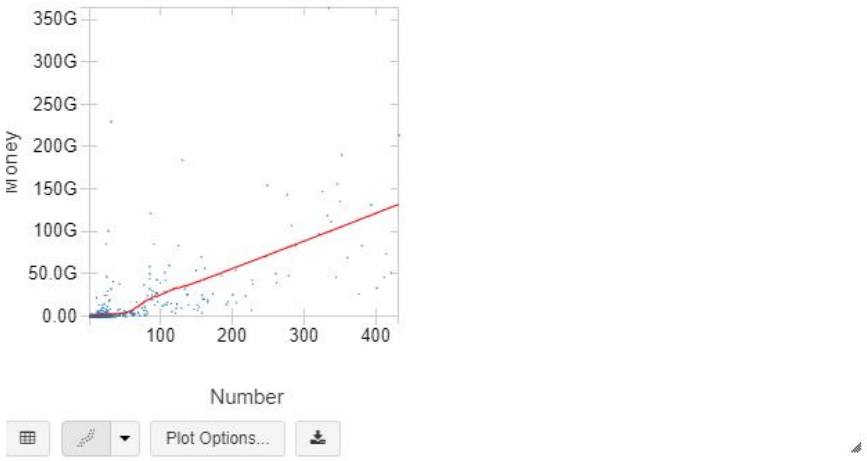


Figura 1.3: Databricks Solución 3

```scala
// Modificar la columna dt de tipo timestamp (yyyy-MM-dd hh-mm-ss) a solo de años
val temperatureByCountryYear:DataFrame = temperatureByCountry.withColumn("dt", date_format(col("dt"), "yyyy").cast("Int"))
//4.¿Cómo evoluciona la temperatura media anual global?
val temperatureAVGTotal = temperatureByCountryYear
                    .groupBy($"dt")
                    .avg("AverageTemperature")
                    .orderBy($"dt")
display(temperatureAVGTotal)
```

(1) Spark Jobs

temperatureByCountryYear: org.apache.spark.sql.DataFrame = [dt: integer, AverageTemperature: double ... 2 more fields]

temperatureAVGTotal: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [dt: integer, avg(AverageTemperature): double]

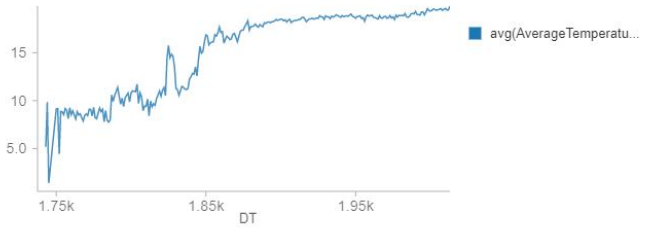


Figura 1.4: Databricks Solución 4

```
// 5.¿Cómo evoluciona la temperatura media anual de cada país?
val temperatureAVG = temperatureByCountryYear
                     .groupBy($"dt",$"Country")
                     .avg("AverageTemperature")
                     .orderBy($"dt", $"Country")
display(temperatureAVG)
```

Show result

md 15

```
display(temperatureAVG.filter($"Country"==="Andorra"))
```

▸ (1) Spark Jobs

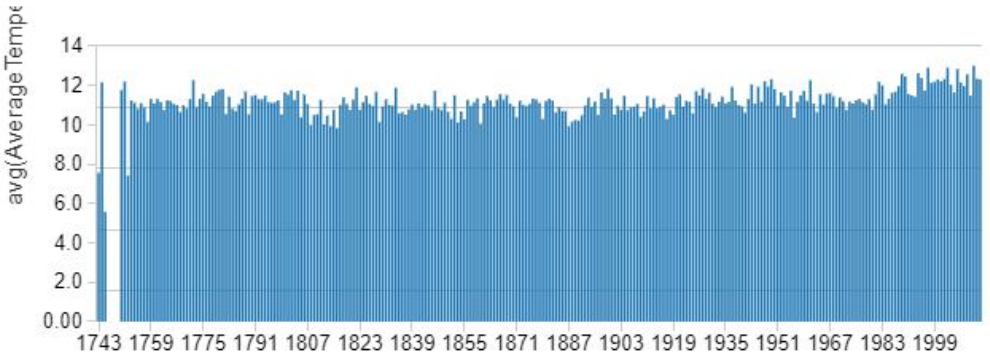


Figura 1.5: Databricks Solución 5

```
//6.¿Existe alguna relación entre la temperatura media anual global con el número de desastres naturales?
val aux = disasterNumber.filter($"Entity" ==="All natural disasters")

display(aux.join(temperatureAVGTotal, aux("Year")===temperatureAVG("dt")))
```

▸ (6) Spark Jobs

▸ aux: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [Entity: string, Year: integer ... 1 more fields]

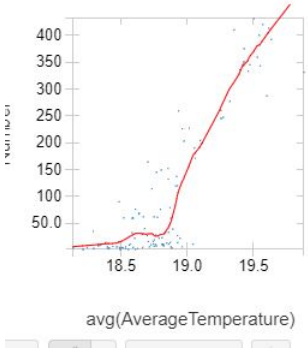


Figura 1.6: Databricks Solución 6

```
//7.¿Cómo evoluciona el número de muertes anual causados por los desastres naturales?
display(deathByCountry
        .groupBy($"Year")
        .sum("Deaths")
        .orderBy($"Year")
)
```

(1) Spark Jobs

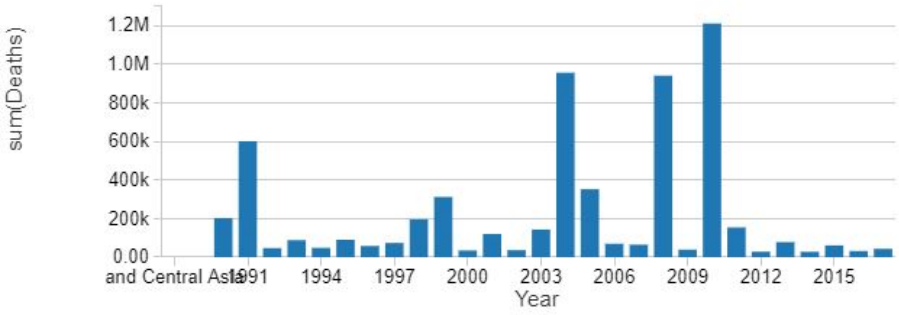


Figura 1.7: Databricks Solución 7

```
//8.¿Cómo evoluciona el porcentaje de muertes anual causados por los desastres naturales?
display(deathPercentByCountry
       .groupBy($"Year")
       .avg("Deaths(Percent) (%)")
       .orderBy($"Year"))
```

(1) Spark Jobs

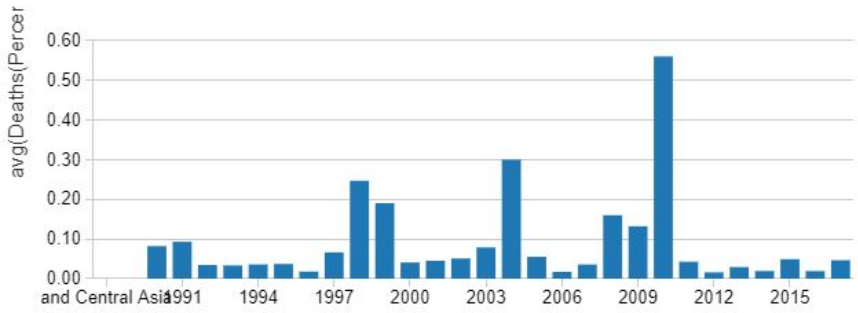


Figura 1.8: Databricks Solución 8

```scala
//9.¿Cuál es el país con más muertes por los desastres naturales?
display(
  deathByCountry
  .groupBy($"Entity")
  .sum("Deaths")
  .orderBy($"sum(Deaths)".desc))
```

▸ (1) Spark Jobs

Entity	sum(Deaths)
World	1438270
Middle SDI	461712
Low SDI	452809
Southeast Asia	407278
South Asia	366492
Low-middle SDI	326298
Latin America and Caribbean	315400
Caribbean	240318
Haiti	227475

Figura 1.9: Databricks Solución 9

```scala
//10.¿Existe alguna relación entre el número de desastres naturales con el número
//total de muertos anuales causados por los desastres naturales?
val disasterNumberDeath =disasterNumber.join(disasterDeath, Seq("Entity","Year"))
display(disasterNumberDeath.filter($"Entity"==="All natural disasters"))
```

▸ (1) Spark Jobs

▸ disasterNumberDeath: org.apache.spark.sql.DataFrame = [Entity: string, Year: integer ... 2 more fields]

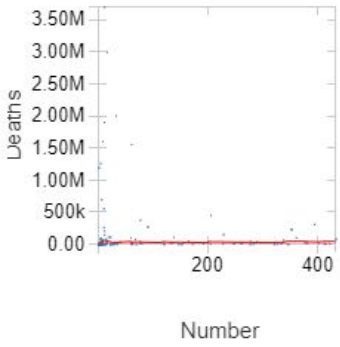


Figura 1.10: Databricks Solución 10

```
// 11.¿Cómo evolucionan cada tipo de desastres naturales según avanza el tiempo?
display(
  disasterNumber.filter($"Entity" =!="All natural disasters")
        .groupBy($"Year", $"Entity")
        .sum("Number")
        .orderBy($"Year".desc)
)
```

▸ (1) Spark Jobs

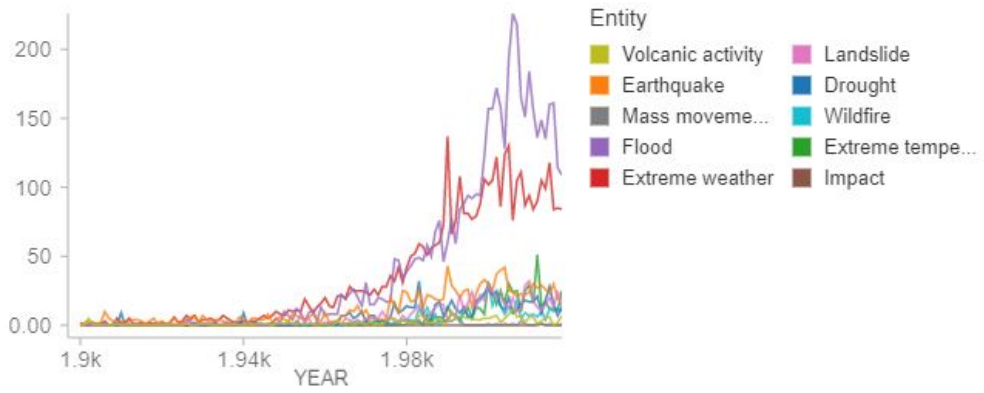


Figura 1.11: Databricks Solución 11

```
//12.¿Qué tipo de desastres naturales provoca mayores muertos?
display(disasterNumberDeath
       .groupBy($"Entity")
       .sum("Deaths")
       .filter($"Entity" =!= "All natural disasters")
       .orderBy($"sum(Deaths)".desc))
```

▸ (1) Spark Jobs

Entity	sum(Deaths)
Drought	11731294
Flood	6960299
Earthquake	2581934
Extreme weather	1398887
Extreme temperature	183143
Volcanic activity	97244
Landslide	65018
Mass movement (dry)	5047
Wildfire	4264

Figura 1.12: Databricks Solución 12

```
1 //13.¿Qué tipo de desastres naturales provoca mayores daños económicos?
2 display(disasterEconomic
3         .groupBy($"Entity")
4         .sum("Money")
5         .filter($"Entity" =!= "All natural disasters")
6         .orderBy($"sum(Money)".desc))
```

▸ (1) Spark Jobs

Entity	sum(Money)
Extreme weather	1281599825000
Earthquake	814901417000
Flood	788212248000
Drought	166761974000
Wildfire	88234275000
Extreme temperature	62866343000
Landslide	10485334000
Volcanic activity	3933912000
Mass movement (dry)	211600000

Figura 1.13: Databricks Solución 13

```
1 //14.¿Cuándo hubo más concentración de terremotos significativos?
2 display(earthquakes.filter($"Year">1700))
```

▸ (1) Spark Jobs

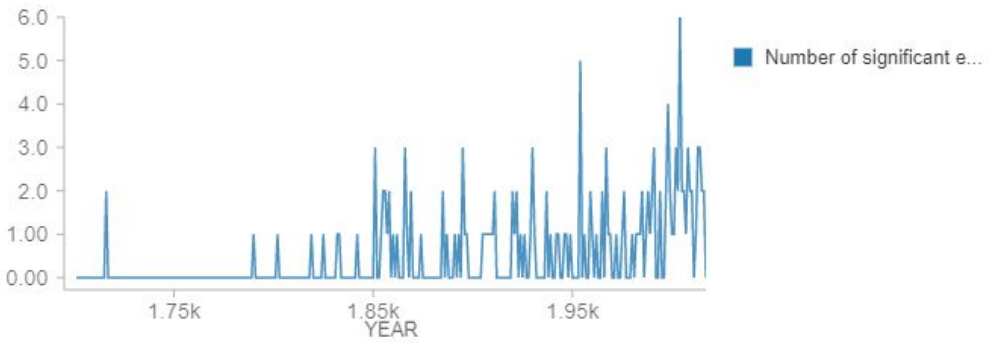


Figura 1.14: Databricks Solución 14

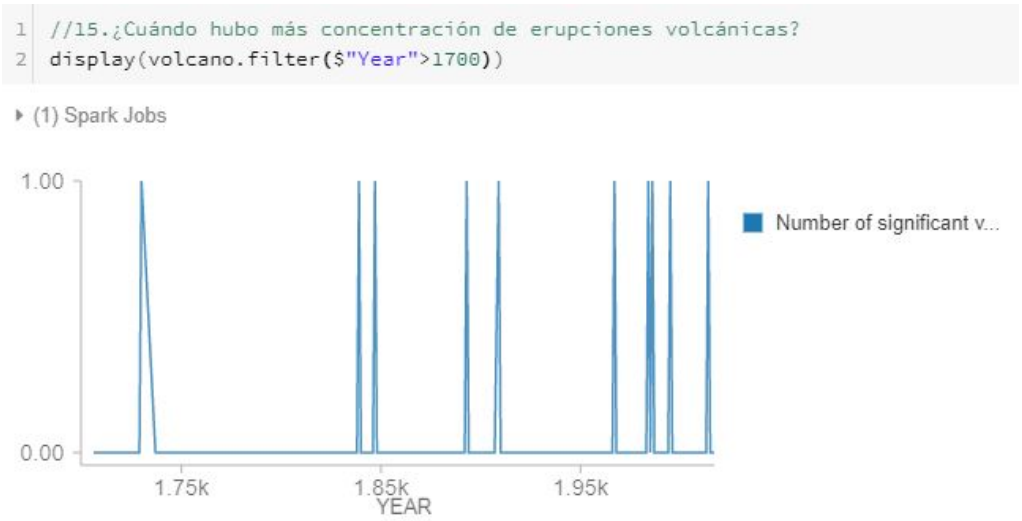


Figura 1.15: Databricks Solución 15

1.5 Solución con Plotly

En esta sección utilizo el Notebook de Jupyter con el Almond 0.9.1. Plotly es una librería que hay para hacer las gráficas. En la introducción he puesto el enlace al repositorio GitHub, aquí pongo otra documentación que hay para Plotly: https://alexarchambault.github.io/plotly-scala/#bar-charts.

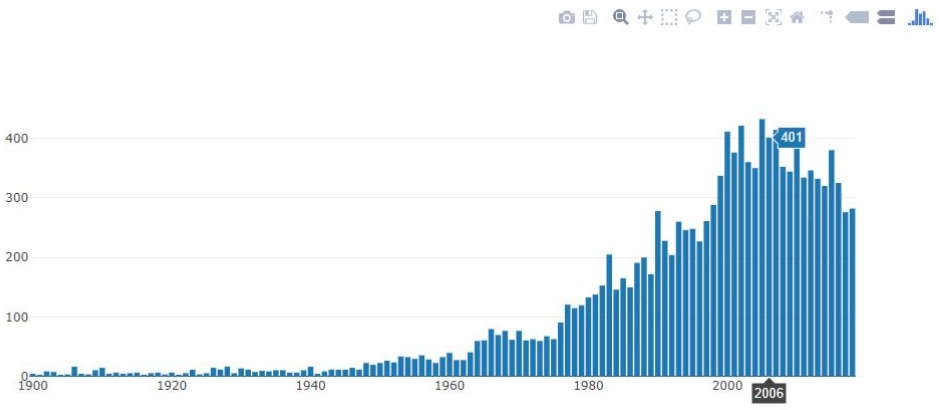


Figura 1.16: Plotly Solución 1

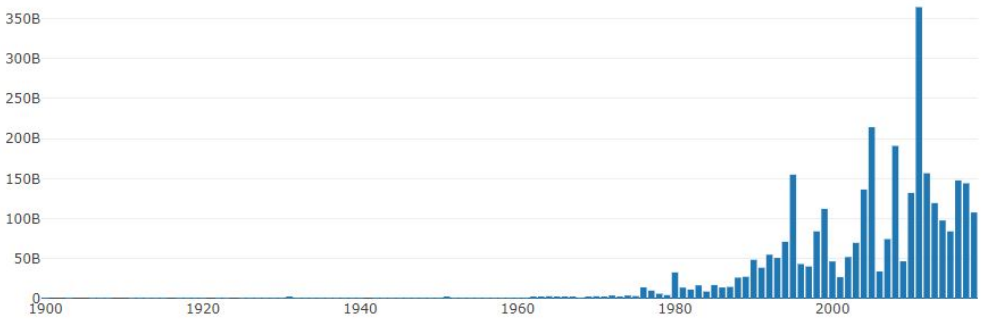


Figura 1.17: Plotly Solución 2

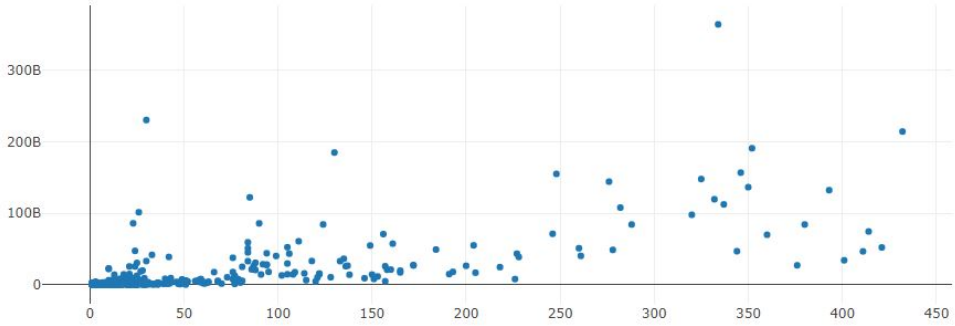


Figura 1.18: Plotly Solución 3

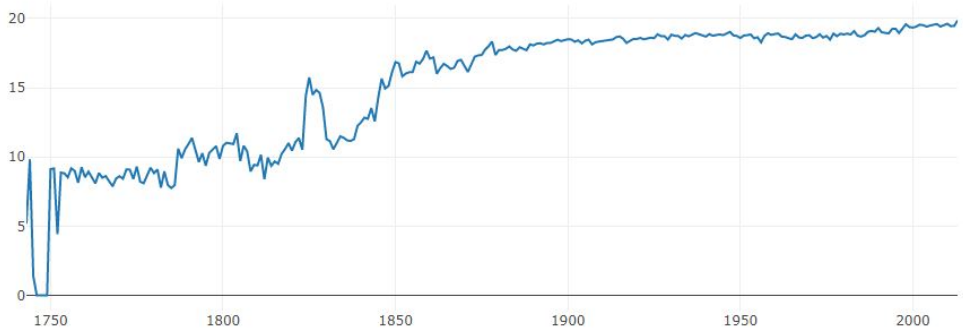


Figura 1.19: Plotly Solución 4

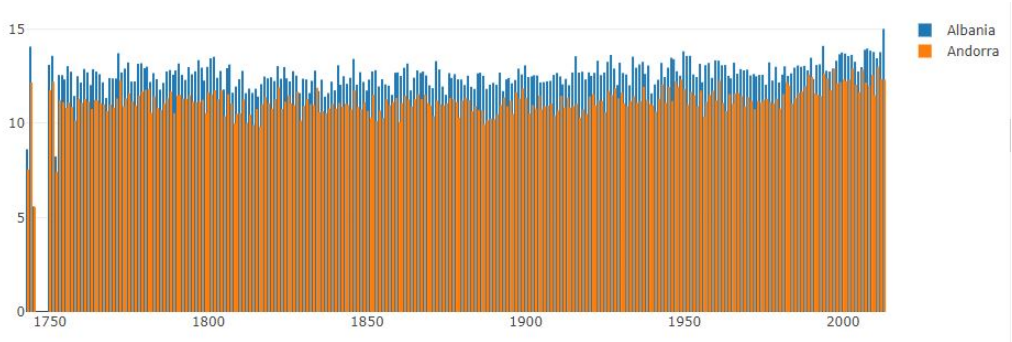


Figura 1.20: Plotly Solución 5

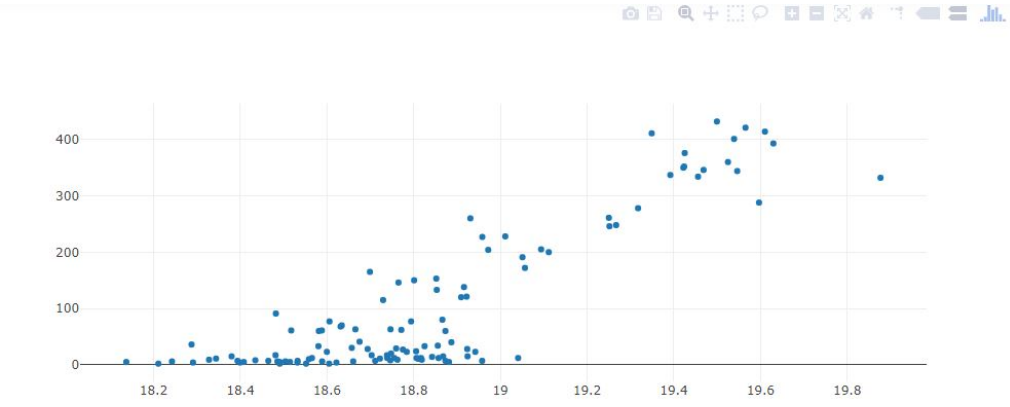


Figura 1.21: Plotly Solución 6

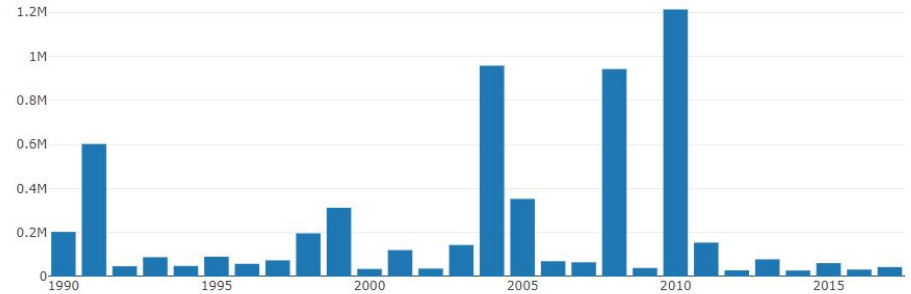


Figura 1.22: Plotly Solución 7

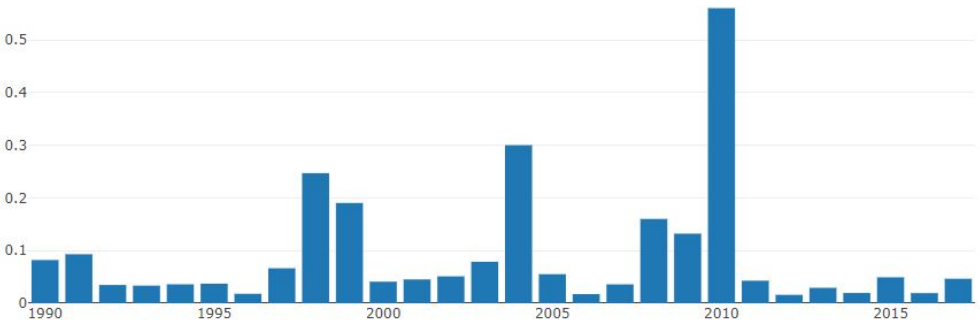


Figura 1.23: Plotly Solución 8

```
//9.¿Cuál es el país con más muertes por los desastres naturales?

  deathByCountry
  .groupBy($"Entity")
  .sum("Deaths")
  .orderBy($"sum(Deaths)".desc)
  .show
```

show at cmd21.sc:4

1 / 1

show at cmd21.sc:4

64 / 64

```
+--------------------+-----------+
|              Entity|sum(Deaths)|
+--------------------+-----------+
|               World|    1438270|
|          Middle SDI|     461712|
|             Low SDI|     452809|
|      Southeast Asia|     407278|
|          South Asia|     366492|
|      Low-middle SDI|     326298|
|Latin America and...|     315400|
|           Caribbean|     240318|
|               Haiti|     227475|
|           Indonesia|     185201|
```

Figura 1.24: Plotly Solución 9

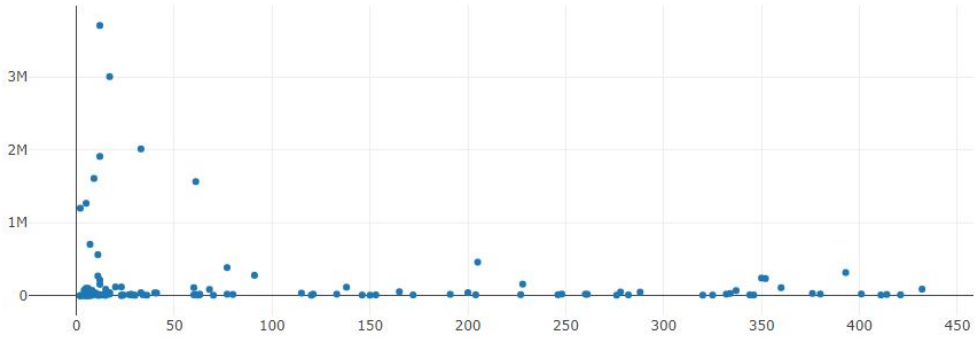


Figura 1.25: Plotly Solución 10

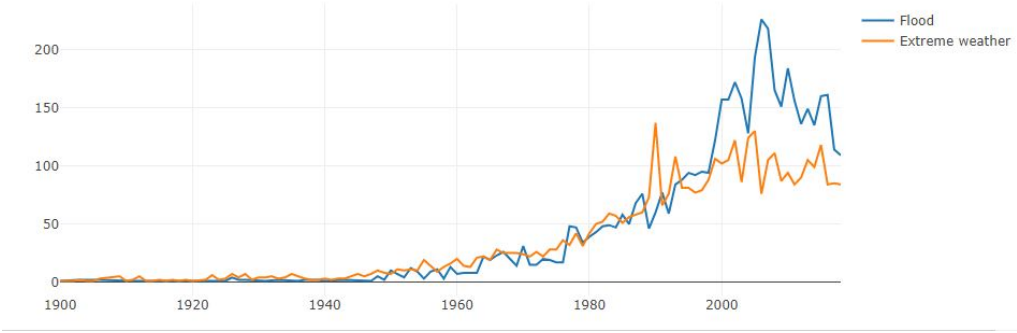


Figura 1.26: Plotly Solución 11

```
: //12.¿Qué tipo de desastres naturales provoca mayores muertos?
disasterNumber.join(disasterDeath, Seq("Entity","Year"))
       .groupBy($"Entity")
       .sum("Deaths")
       .filter($"Entity" =!= "All natural disasters")
       .orderBy($"sum(Deaths)".desc).show
```

run at ThreadPoolExecutor.java:1149

1

show at cmd24.sc:5

1

show at cmd24.sc:5

64

```
+------------------+-----------+
|            Entity|sum(Deaths)|
+------------------+-----------+
|           Drought|   11731294|
|             Flood|    6960299|
|        Earthquake|    2581934|
|   Extreme weather|    1398887|
|Extreme temperature|     183143|
|  Volcanic activity|      97244|
```

Figura 1.27: Plotly Solución 12

```
//13.¿Qué tipo de desastres naturales provoca mayores daños económicos?
disasterEconomic
       .groupBy($"Entity")
       .sum("Money")
       .filter($"Entity" =!= "All natural disasters")
       .orderBy($"sum(Money)".desc)
       .show
```

show at cmd25.sc:5

1 / 1

show at cmd25.sc:5

64 / 64

```
+-------------------+-------------+
|             Entity|   sum(Money)|
+-------------------+-------------+
|    Extreme weather|1281599825000|
|         Earthquake| 814901417000|
|              Flood| 788212248000|
|            Drought| 166761974000|
|           Wildfire|  88234275000|
|Extreme temperature|  62866343000|
|          Landslide|  10485334000|
|  Volcanic activity|   3933912000|
```

Figura 1.28: Plotly Solución 13

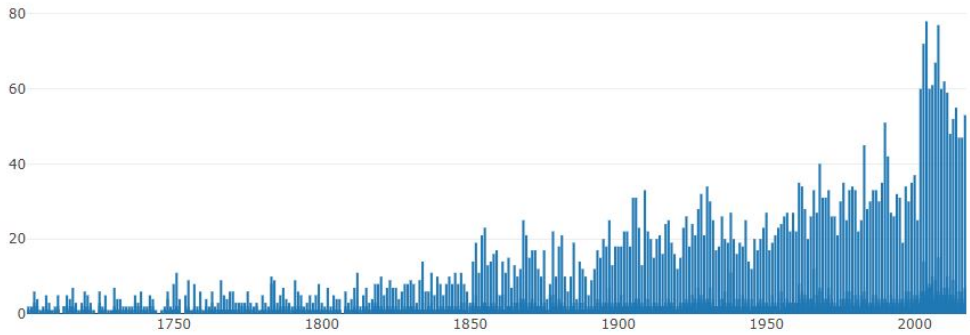


Figura 1.29: Plotly Solución 14

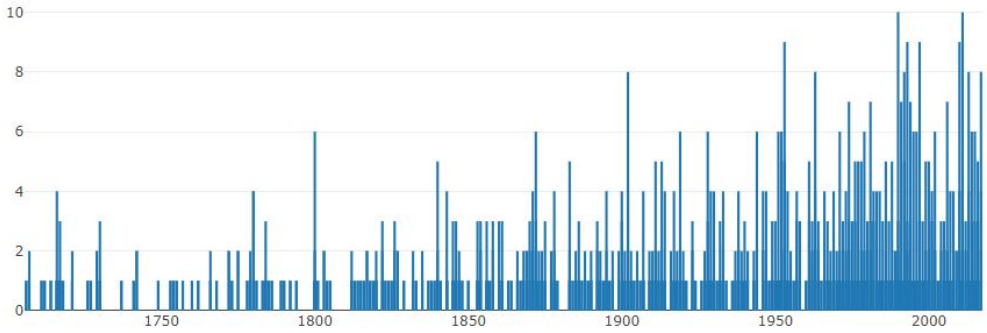


Figura 1.30: Plotly Solución 15

1.6 Conclusión

1. ¿Cómo evoluciona el número de desastres naturales según avanza los años?
   - Podemos ver que, aunque hay fluctuaciones, la tendencia en general es que el número de desastres naturales es cada vez mayor.

2. ¿Cómo evoluciona el daño económico causado por los desastres naturales?
   - Aunque hay picos, pero a grandes rascos podemos afirmar que el coste es creciente a lo largo de los años.

3. ¿Existe alguna relación entre el número de desastres naturales con el daño económico?
   - Podemos observar que cuando el número de desastres naturales aumenta también aumenta el daño económico causado.

4. ¿Cómo evoluciona la temperatura media anual global?
   - Podemos ver que la temperatura media global va aumentando.

5. ¿Cómo evoluciona la temperatura media anual de cada país?
   - La temperatura media de los países fluctúa entre 15 y 25 grados centígrados, pero a partir de los años 1800 vemos que algunos países superan los 25 grados centígrados. Debido que son muchos países y no se puedo ver bien, he cogido una gráfica de ejemplo con Andorra y se puede ver que hay aumentos en los últimos años.

6. ¿Existe alguna relación entre la temperatura media anual global con el número de desastres naturales?
   - Podemos ver que hay una relación exponencial entre la temperatura media con el número de desastres naturales.

7. ¿Cómo evoluciona el número de muertes anual causados por los desastres naturales?
   - El número de las muertes no es cada vez más, eso puede ser debido a los avances que hubo en los últimos años , lo que hace que el número de muertos sea menor.

8. ¿Cómo evoluciona el porcentaje de muertes anual causados por los desastres naturales?
   - En general no ha sobrepasado 0.6 % de la población.

9. ¿Cuál es el país con más muertes por los desastres naturales?
   - Según la tabla es Haití con un total de 227475.

10. ¿Existe alguna relación entre el número de desastres naturales con el número total de muertos anuales causados por los desastres naturales?
    - Al parecer no hay una relación entre ellas. En el apartado 1 y 7, hemos visto que el número de desastres naturales es cada vez más pero el de las muertes no ocurre lo mismo.

11. ¿Cómo evolucionan cada tipo de desastres naturales según avanza el tiempo?
    - Como en el apartado 1 hemos visto el número de desastres naturales aumenta cada vez más, podemos observar que todos los tipos de desastres han aumentado, sobre todo las inundaciones y las meteorologías extremas.

12. ¿Qué tipo de desastres naturales provoca mayores muertos?
    - Según los datos, son las inundaciones, con 11731294 muertos.

13. ¿Qué tipo de desastres naturales provoca mayores daños económicos?
    - Según los datos es la meteorología extrema, con un coste total de 1281599825000 $.

14. ¿Cuándo hubo más concentración de terremotos significativos?
    - Debido a que los datos lejanos no son completos, voy a analizar a partir de 1700. Se puede observar hay una mayor frecuencia en los últimos años.

15. ¿Cuándo hubo más concentración de erupciones volcánicas?
    - Igual que en el caso de terremotos.