Máster Data Science – Universidad Rey Juan Carlos

Bases de Datos NoSQL

El objetivo de este trabajo es utilizar esta información para obtener una serie de medidas que permitirán evaluar a los investigadores en informática.

Quesada Pérez de San Román, Daniel
Nevado López, Desiderio
Martín Olinero, Laura
23-4-2023

Contenido

Introducción

En esta práctica se partirá de una fuente de datos real, y que se caracteriza por un elevado grado de interdependencia interna.

La fuente de datos es la base de datos DBLP Computer Science Bibliography https://dblp.unitrier.de/db/, considerada como la mayor recopilación existente de referencias bibliográficas académicas específicamente centrada en la informática. En particular, almacena los datos relativos a la gran mayoría de las revistas científicas y congresos académicos sobre informática, en muchos casos remontándose hasta publicaciones de los años 60, o incluso anteriores. La recopilación completa de datos se puede descargar como un único fichero XML desde http://dblp.uni-trier.de/xml/. Actualmente, este fichero comprimido tiene un tamaño de 729 MB y expandido ocupa más de 3 GB.

Hay ocho tipos de elementos, aunque para la práctica solo se deben considerar los tres siguientes tipos de publicaciones: artículos de revista, artículos en congresos y artículos en libros.

Transformar el XML a JSON

Para transformar los datos de XML a JSON, eran necesario importar dos tipos de librerías: xmltodict y json.

Una vez importadas esas dos librerías, haríamos la transformación de los datos. Vemos que transformar los datos de XML a JSON tardó 1 hora 27 minutos y 43 segundos, pero la primera vez que se cargaron duró aproximadamente 2 horas y 30 minutos.

```
import xmltodict
import json
%pwd
```

```
'C:\\Users\\DesiG'
```

```
%%time
with open("C:\\Users\\DesiG\\OneDrive\\Escritorio\\dblp.xml","r") as f:
    datos=xmltodict.parse(f.read())
```

```
CPU times: total: 19min 18s
Wall time: 1h 27min 43s
```

```
datos
```


```
{'dblp': {'article': [{'@mdate': '2020-06-25',
    '@key': 'tr/meltdown/s18',
    '@publtype': 'informal',
    'author': ['Paul Kocher',
     'Daniel Genkin',
     'Daniel Gruss',
     'Werner Haas 0004',
     'Mike Hamburg',
     'Moritz Lipp',
     'Stefan Mangard',
     'Thomas Prescher 0002',
     'Michael Schwarz 0001',
     'Yuval Yarom'],
    'title': 'Spectre Attacks: Exploiting Speculative Execution.',
    'journal': 'meltdownattack.com',
    'year': '2018',
    'ee': {'@type': 'oa', '#text': 'https://spectreattack.com/spectre.pdf'}},
   {'@mdate': '2020-06-25',
    '@key': 'tr/meltdown/m18',
```

Limpieza del JSON

En primer lugar, vimos que existían 8 tipos de elementos diferentes de los cuales solo utilizaremos tres ‘article’, ‘inproceedings’ y ‘incollection’. Lo que haremos, será eliminar ‘book’, ‘mastersthesis’, ‘proceedings’, ‘phdthesis’ y ‘data’. Para ellos utilizamos *del:*

```
del datos['dblp']['book']
del datos['dblp']['mastersthesis']
del datos['dblp']['proceedings']
del datos['dblp']['phdthesis']
del datos['dblp']['data']
```

A continuación, vamos a eliminar todas aquellas variables que no serán necesarias en las consultas y crearemos el campo ‘type’, formado por los tres tipos anteriormente nombrados.

Para article:

```
for article in datos['dblp']['article']:
    for key in ['booktitle', '@orcid', 'crossref', 'ee',
                '@key','url', 'volume', 'month', 'note',
                'cdrom', 'publisher', 'pages', 'number']:
        if key in article:
            del article[key]
        else:
            article['type'] = 'article'
```

Para inproceedings:

```
for inproceedings in datos['dblp']['inproceedings']:
    for key in ['booktitle', '@orcid', 'crossref', 'ee',
                '@key','url', 'volume', 'month', 'note',
                'cdrom', 'publisher', 'pages', 'number']:
        if key in inproceedings:
            del inproceedings[key]
        else:
            article['type'] = 'inproceedings'
```

```
for article in datos['dblp']['inproceedings']:
    for key in ['booktitle', '@orcid', 'crossref', 'ee',
                '@key','url', 'volume', 'month', 'note',
                'cdrom', 'publisher', 'pages', 'number']:
        if key in inproceedings:
            del inproceedings[key]
        else:
            article['type'] = 'inproceedings'
```

Para incollection:

```
for incollection in datos['dblp']['incollection']:
    for key in ['booktitle', '@orcid', 'crossref', 'ee',
                '@key','url', 'volume', 'month', 'note',
                'cdrom', 'publisher', 'pages', 'number']:
        if key in incollection:
            del incollection[key]
        else:
            article['type'] = 'incollection'
```

```
for article in datos['dblp']['incollection']:
    for key in ['booktitle', '@orcid', 'crossref', 'ee',
                '@key','url', 'volume', 'month', 'note',
                'cdrom', 'publisher', 'pages', 'number']:
        if key in incollection:
            del incollection[key]
        else:
            article['type'] = 'incollection'
```

Tanto para 'inproceedings' como para 'incollection' utilizamos la primera parte del código para crearlas. La siguiente parte es para meterlas en el campo *type*.

Exportar datos a MongoDB

Para exportar los datos a MongoDB lo que hicimos fue lo siguiente:

1º Importamos las librerías y lo conectamos a nuestro localhost:

```
import pymongo
from pymongo import MongoClient
conex = pymongo.MongoClient('localhost',27017)
```

2º Lo conectamos a MongoDB y le ponemos de nombre BDPracticaFinal

```
db = conex.BDPracticaFinal
conexion = pymongo.MongoClient()
db= conexion.BDPracticaFinal
```

3º Creamos la colección, e insertamos los datos de article, inproceedings e incollection:

```
coleccion=db.publi
```

```
coleccion.insert_many(datos['dblp']['article'])
```

```
<pymongo.results.InsertManyResult at 0x1b32121bb20>
```

```
coleccion.insert_many(datos['dblp']['inproceedings'])
```

```
<pymongo.results.InsertManyResult at 0x1b558db48b0>
```

```
coleccion.insert_many(datos['dblp']['incollection'])
```

```
<pymongo.results.InsertManyResult at 0x1b4eb0df1c0>
```

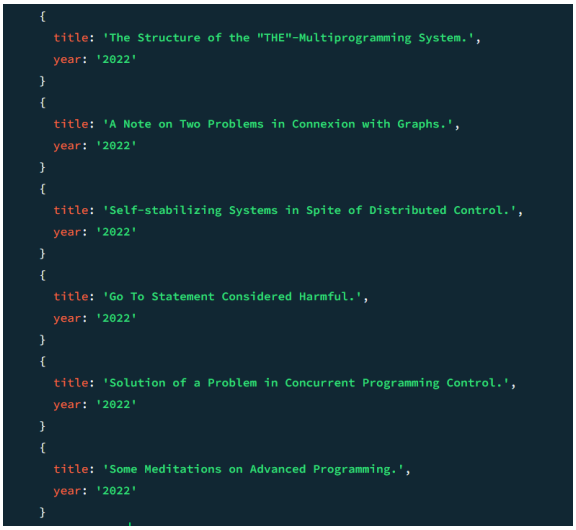
Consultas

Consulta 1. Listado de todas las publicaciones del autor “Edsger W. Dijkstra” ordenado por fecha de publicación.

Las primeras publicaciones del autor Edsger W.Dijkstra son:

```
> db.publi.aggregate([
    {'$match': {'author': 'Edsger W. Dijkstra'}},
    {'$sort': {'year': 1}},
    {'$project': {'_id': 0, 'title': 1, 'year': 1}}]
  )
<  {
     title: 'A note on two problems in connexion with graphs.',
     year: '1959'
   }
   {
     title: 'ALGOL Sub-Committee Report - Extensions.',
     year: '1959'
   }
   {
     title: 'Letter to the editor: defense of ALGOL 60.',
     year: '1961'
   }
   {
     title: 'Operating Experience with ALGOL 60.',
     year: '1962'
   }
   {
     title: 'Some Meditations on Advanced Programming.',
     year: '1962'
   }
```

Las últimas publicaciones del autor Edsger W. Dijkstra son:



```
   {
     title: 'The Structure of the "THE"-Multiprogramming System.',
     year: '2022'
   }
   {
     title: 'A Note on Two Problems in Connexion with Graphs.',
     year: '2022'
   }
   {
     title: 'Self-stabilizing Systems in Spite of Distributed Control.',
     year: '2022'
   }
   {
     title: 'Go To Statement Considered Harmful.',
     year: '2022'
   }
   {
     title: 'Solution of a Problem in Concurrent Programming Control.',
     year: '2022'
   }
   {
     title: 'Some Meditations on Advanced Programming.',
     year: '2022'
   }
```

Todas las publicaciones del autor Edsger W. Dijkstra son:

1. Listado de todas las publicaciones del autor "Edsger W. Dijkstra" ordenado por fecha de publicación.

```
%%time
Ejercicio1 = [{'$match': {'author': 'Edsger W. Dijkstra'}},
              {'$sort': {'year':1}},
              {'$project': {'_id': 0, 'title': 1, 'year': 1}}]

Resultado1 = db.publi.aggregate(Ejercicio1, allowDiskUse = True)
print(list(Resultado1))
```


```
[{'title': 'A note on two problems in connexion with graphs.', 'year': '1959'}, {'title': 'ALGOL Sub-Committee Report - Extensi
ons.', 'year': '1959'}, {'title': 'Letter to the editor: defense of ALGOL 60.', 'year': '1961'}, {'title': 'Some Meditations on
Advanced Programming.', 'year': '1962'}, {'title': 'Operating Experience with ALGOL 60.', 'year': '1962'}, {'title': 'Some comm
ents on the aims of MIRFAC.', 'year': '1964'}, {'title': 'Solution of a problem in concurrent programming control.', 'year': '1
965'}, {'title': 'The structure of the "THE"-multiprogramming system.', 'year': '1967'}, {'title': 'The Structure of "THE"-Mult
iprogramming System.', 'year': '1968'}, {'title': 'Letters to the editor: go to statement considered harmful.', 'year': '196
8'}, {'title': 'Letters to the editor: The go to statement reconsidered.', 'year': '1968'}, {'title': 'Hierarchical Ordering of
Sequential Processes.', 'year': '1971'}, {'title': 'A class of allocation strategies inducing bounded delays only.', 'year': '1
972'}, {'title': 'Information Streams Sharing a Finite Buffer.', 'year': '1972'}, {'title': 'The Humble Programmer.', 'year':
'1972'}, {'title': 'Self-stabilizing Systems in Spite of Distributed Control.', 'year': '1974'}, {'title': 'A time-wise hierarc
hy imposed upon the use of a two-level store.', 'year': '1975'}, {'title': 'On-the-fly garbage collection: an exercise in coope
ration.', 'year': '1975'}, {'title': 'Guarded commands, non-determinancy and a calculus for the derivation of programs.', 'yea
r': '1975'}, {'title': 'On the teaching of programming, i. e. on the teaching of thinking.', 'year': '1975'}, {'title': 'Crafts
man or Scientist.', 'year': '1975'}, {'title': 'Correctness concerns and, among other things, why they are resented.', 'year':
'1975'}, {'title': 'Guarded commands, non-determinacy and a calculus for the derivation of programs.', 'year': '1975'}, {'titl
e': 'Guarded Commands, Nondeterminacy and Formal Derivation of Programs.', 'year': '1975'}, {'title': 'The Effective Arrangemen
t of Logical Systems.', 'year': '1976'}, {'title': 'Formal Techniques and Sizeable Programs.', 'year': '1976'}, {'title': 'On a
Gauntlet Thrown by David Gries.', 'year': '1976'}, {'title': 'Programming: From Craft to Scientific Discipline.', 'year': '197
7'}, {'title': 'A position paper on software reliability.', 'year': '1977'}, {'title': 'Finding the Correctness Proof of a Conc
urrent Program.', 'year': '1978'}, {'title': 'Stationary Behaviour of Some Ternary Networks.', 'year': '1978'}, {'title': 'A Th
eorem about Odd Powers of Odd Integers.', 'year': '1978'}, {'title': 'In Honour of Fibonacci.', 'year': '1978'}, {'title': 'Fin
ding the Correctness Proof of a Concurrent Program.', 'year': '1978'}, {'title': 'On the Foolishness of "Natural Language Progr
amming".', 'year': '1978'}, {'title': 'Program Inversion.', 'year': '1978'}, {'title': 'A More Formal Treatment of a Less Simpl
e Example.', 'year': '1978'}, {'title': 'On the Interplay between Mathematics and Programming.', 'year': '1978'}, {'title': 'Do
D-I: the summing up.', 'year': '1978'}, {'title': 'On a political pamphlet from the middle ages.', 'year': '1978'}, {'title':
'On-the-Fly Garbage Collection: An Exercise in Cooperation.', 'year': '1978'}, {'title': 'Software Engineering : As It Should B
e.', 'year': '1979'}, {'title': 'Termination Detection for Diffusing Computations.', 'year': '1980'}, {'title': 'Some Beautiful
Arguments Using Mathematical Induction.', 'year': '1980'}, {'title': "American programming's plight.", 'year': '1981'}, {'titl
e': 'A Word of Welcome.', 'year': '1981'}, {'title': 'How do we tell truths that might hurt?', 'year': '1982'}, {'title': 'An I
ntroduction to Three Algorithms for Sorting in Situ.', 'year': '1982'}, {'title': 'Smoothsort, an Alternative for Sorting In Si
tu.', 'year': '1982'}, {'title': 'Derivation of a Termination Detection Algorithm for Distributed Computations.', 'year': '198
3'}, {'title': 'The Structure of "THE"-Multiprogramming System (Reprint).', 'year': '1983'}, {'title': 'Solutions of a Problem
in Concurrent Programming Control (Reprint).', 'year': '1983'}, {'title': 'The fruits of misunderstanding.', 'year': '1983'},
{'title': 'A Belated Proof of Self-Stabilization.', 'year': '1986'}, {'title': 'A Simple Fixpoint Argument Without the Restrict
ion to Continuity.', 'year': '1986'}, {'title': "A Heuristic Explanation of Batcher's Baffler.", 'year': '1987'}, {'title': 'On
Binary Operators and Their Derived Relations.', 'year': '1988'}, {'title': 'Position paper on "fairness".', 'year': '1988'},
{'title': 'The Linear Search Revisited.', 'year': '1989'}, {'title': 'Making a Fair Roulette From a Possibly Biased Coin.', 'ye
ar': '1990'}, {'title': 'On the Economy of doing Mathematics.', 'year': '1992'}, {'title': 'The Unification of Three Calculi.',
'year': '1992'}, {'title': 'Heuristics for a Calculational Proof.', 'year': '1995'}, {'title': 'On two equations that have the
same extreme solution.', 'year': '1996'}, {'title': "Bulterman's theorem on shortest trees.", 'year': '1996'}, {'title': 'The b
alance and the coins.', 'year': '1996'}, {'title': 'Fibonacci and the greatest common divisor.', 'year': '1996'}, {'title': 'Th
e argument about the arithmetic mean and the geometric mean, heuristics included.', 'year': '1996'}, {'title': "A bagatelle on
Euclid's algorithm.", 'year': '1996'}, {'title': 'An alternative of the ETAC to EWD1163.', 'year': '1996'}, {'title': 'A prime
is in at most 1 way the sum of 2 squares.', 'year': '1996'}, {'title': 'On the transitive closure of a wellfounded relation.',
'year': '2000'}, {'title': "Under the spell of Leibniz's dream.", 'year': '2001'}, {'title': "Designing a Calculational Proof o
f Cantor's Theorem.", 'year': '2001'}, {'title': 'The end of computing science?', 'year': '2001'}, {'title': 'EWD 1308: What Le
d to "Notes on Structured Programming".', 'year': '2002'}, {'title': 'Go to Statement Considered Harmful (Reprint).', 'year':
'2002'}, {'title': 'Solution of a Problem in Concurrent Programming Control (Reprint).', 'year': '2002'}, {'title': 'EWD1300: T
he Notational Conventions I Adopted, and Why.', 'year': '2002'}, {'title': 'My recollections of operating system design.', 'yea
r': '2005'}, {'title': 'The Humble Programmer.', 'year': '2022'}, {'title': 'On-the-Fly Garbage Collection: An Exercise in Coop
eration.', 'year': '2022'}, {'title': 'On the Reliability of Programs.', 'year': '2022'}, {'title': 'Recursive Programming.',
'year': '2022'}, {'title': 'The Structure of the "THE"-Multiprogramming System.', 'year': '2022'}, {'title': 'A Note on Two Pro
blems in Connexion with Graphs.', 'year': '2022'}, {'title': 'Self-stabilizing Systems in Spite of Distributed Control.', 'yea
r': '2022'}, {'title': 'Go To Statement Considered Harmful.', 'year': '2022'}, {'title': 'Solution of a Problem in Concurrent P
rogramming Control.', 'year': '2022'}, {'title': 'Some Meditations on Advanced Programming.', 'year': '2022'}]
CPU times: total: 0 ns
Wall time: 27.4 s
```


Esta consulta tardó en ejecutarse en Python un total de 27,4 segundos.

Consulta 2. Número de publicaciones de la autora "Elena Ferrari".

```
> db.publi.aggregate(
    [{'$match': {'author': 'Elena Ferrari'}},
                 {'$count': 'Total'}])
<   {
      Total: 124
    }
```

La autora Elena Ferrari tiene un total de 124 publicaciones en total.

Esta consulta tardó en ejecutarse en Python un total de 6,35 segundos.

Consulta 3. Número de publicaciones de la autora "Elisa Bertino" agrupadas por tipo de publicación (article, inproceedings, incollections).

```
> db.publi.aggregate(
    [{'$match': {'author': 'Elisa Bertino'}},
               {'$group': {'_id': '$type', 'Total': {'$sum': 1}}},
               {'$sort': {'Total': -1}}])
< {
    _id: 'inproceedings',
    Total: 651
  }
  {
    _id: 'article',
    Total: 388
  }
  {
    _id: 'incollection',
    Total: 13
  }
```

La autora Elisa Bertino tiene un total de 1052 publicaciones en total, divididas de la siguiente manera:

- Inproceedings con un total de 651 publicaciones.
- Article con un total de 388 publicaciones.
- Incollection con un total de 13 publicaciones.

Esta consulta tardó en ejecutarse en Python un total de 5,01 segundos.

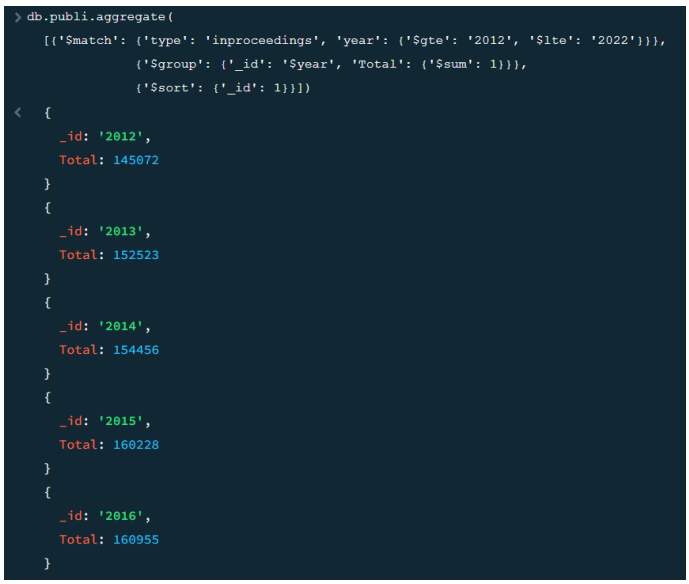
Consulta 4. Número de artículos en revista (article) para el año 2010.

```
> db.publi.aggregate(
    [{'$match': {'type': 'article', 'year': '2010'}},
               {'$count': 'Total'}])
< {
    Total: 91327
  }
```

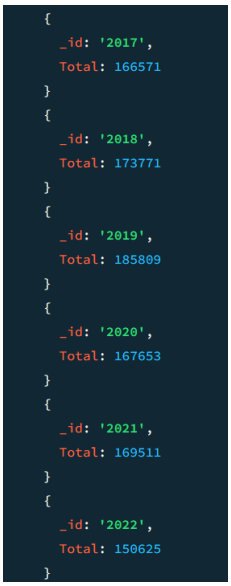
El número de artículos en revista totales son 91.327 artículos.

Esta consulta tardó en ejecutarse en Python un total de 5,36 segundos.

Consulta 5. Número de publicaciones en congresos (inproceedings) por año desde el 2012 hasta el 2022.



```
> db.publi.aggregate(
    [{'$match': {'type': 'inproceedings', 'year': {'$gte': '2012', '$lte': '2022'}}},
               {'$group': {'_id': '$year', 'Total': {'$sum': 1}}},
               {'$sort': {'_id': 1}}])
< {
    _id: '2012',
    Total: 145072
  }
  {
    _id: '2013',
    Total: 152523
  }
  {
    _id: '2014',
    Total: 154456
  }
  {
    _id: '2015',
    Total: 160228
  }
  {
    _id: '2016',
    Total: 160955
  }
```



```
  {
    _id: '2017',
    Total: 166571
  }
  {
    _id: '2018',
    Total: 173771
  }
  {
    _id: '2019',
    Total: 185809
  }
  {
    _id: '2020',
    Total: 167653
  }
  {
    _id: '2021',
    Total: 169511
  }
  {
    _id: '2022',
    Total: 150625
  }
```

El número de publicaciones en congresos por años son los siguientes:

- **2012** = 145.072 publicaciones
- **2013** = 152.523 publicaciones.
- **2014** = 154.456 publicaciones.
- **2015** = 160.228 publicaciones.
- **2016** = 160.955 publicaciones.
- **2017** = 166.571 publicaciones.
- **2018** = 173.771 publicaciones.
- **2019** = 185.809 publicaciones.
- **2020** = 167.653 publicaciones.
- **2021** = 169.511 publicaciones.
- **2022** =150.625 publicaciones.

Esta consulta tardó en ejecutarse en Python un total de 5,76 segundos.

Consulta 6. Número de artículos (article) por revista y año desde el año 2015.

```
> db.publi.aggregate(
  [{'$match': {'type': 'article', 'year': {'$gte': '2015'}}},
              {'$group': {'_id': {'revista': '$journal', 'year': '$year'}, 'Total': {'$sum': 1}}},
              {'$sort': {'_id.year': 1, 'Total': -1}}])
<  {
    _id: {
      revista: 'CoRR',
      year: '2015'
    },
    Total: 18432
  }
  {
    _id: {
      revista: 'Sensors',
      year: '2015'
    },
    Total: 1661
  }
  {
    _id: {
      revista: 'Appl. Math. Comput.',
      year: '2015'
    },
    Total: 1385
```

6. Número de artículos (article) por revista y año desde el año 2015.

```
%%time
Ejercicio6 = [{'$match': {'type': 'article', 'year': {'$gte': '2015'}}},
              {'$group': {'_id': {'revista': '$journal', 'year': '$year'}, 'Total': {'$sum': 1}}},
              {'$sort': {'_id.year': 1, 'Total': -1}}]

Resultado6 = db.publi.aggregate(Ejercicio6, allowDiskUse=True)
print(list(Resultado6))
```

```
a': 'Int. J. Ambient Comput. Intell.', 'year': '2023'}, 'Total': 1}, {'_id': {'revista': 'J. Data Sci. Stat. Vis.', 'year':
'2023'}, 'Total': 1}, {'_id': {'revista': 'Found. Trends Comput. Graph. Vis.', 'year': '2023'}, 'Total': 1}, {'_id': {'revis
ta': 'Compositionality', 'year': '2023'}, 'Total': 1}, {'_id': {'revista': 'Informing Sci. Int. J. an Emerg. Transdiscipl.',
'year': '2023'}, 'Total': 1}, {'_id': {'revista': 'J. Cases Inf. Technol.', 'year': '2023'}, 'Total': 1}, {'_id': {'revist
a': 'Int. J. Knowl. Based Organ.', 'year': '2023'}, 'Total': 1}, {'_id': {'revista': 'Trans. Int. Soc. Music. Inf. Retr.',
'year': '2023'}, 'Total': 1}, {'_id': {'revista': 'Int. J. Virtual Pers. Learn. Environ.', 'year': '2023'}, 'Total': 1}, {'_
id': {'revista': 'EURASIP J. Image Video Process.', 'year': '2023'}, 'Total': 1}, {'_id': {'revista': 'Found. Trends Electro
n. Des. Autom.', 'year': '2023'}, 'Total': 1}, {'_id': {'revista': 'J. Satisf. Boolean Model. Comput.', 'year': '2023'}, 'To
tal': 1}, {'_id': {'revista': 'Informatica (Slovenia)', 'year': '2023'}, 'Total': 1}, {'_id': {'revista': 'Int. J. Comput. S
ci. Sport', 'year': '2023'}, 'Total': 1}, {'_id': {'revista': 'Int. J. Big Data Intell. Appl.', 'year': '2023'}, 'Total':
1}, {'_id': {'revista': 'Int. J. Syst. Softw. Secur. Prot.', 'year': '2023'}, 'Total': 1}, {'_id': {'revista': 'Int. J. Web
Serv. Res.', 'year': '2023'}, 'Total': 1}, {'_id': {'revista': 'J. Braz. Comput. Soc.', 'year': '2023'}, 'Total': 1}, {'_i
d': {'revista': 'Algorithms Mol. Biol.', 'year': '2023'}, 'Total': 1}, {'_id': {'revista': 'Netw. Commun. Technol.', 'year':
'2023'}, 'Total': 1}, {'_id': {'revista': 'Artif. Intell. Eng. Des. Anal. Manuf.', 'year': '2023'}, 'Total': 1}, {'_id': {'r
evista': 'Found. Trends Commun. Inf. Theory', 'year': '2023'}, 'Total': 1}, {'_id': {'revista': 'Found. Trends Hum. Comput.
Interact.', 'year': '2023'}, 'Total': 1}, {'_id': {'revista': 'Int. J. Fog Comput.', 'year': '2023'}, 'Total':
1}]
```

El resultado era demasiado amplio por eso no se visualiza toda la solución. Se ve algunas revistas y su número de artículos de 2015 y de 2023.

Esta consulta tardó en ejecutarse en Python un total de 6,65 segundos.

Consulta 7. Nombre de la revista con más publicaciones (es decir, más artículos publicados).

```
> db.publi.aggregate(
    [{'$match': {'type': 'article'}},
                {'$group': {'_id': '$journal', 'Total': {'$sum': 1}}},
                {'$sort': {'Total': -1}}, {'$limit': 1}])
<   {
      _id: 'CoRR',
      Total: 494105
    }
```

La revista con más publicaciones es CoRR con un total de 494.105 publicaciones.

Esta consulta tardó en ejecutarse en Python un total de 6,18 segundos.

Consulta 8. Nombre de la revista en la que hayan publicado más autores diferentes.

```
> db.publi.aggregate(
    [{'$unwind':"$author"},
    {'$match': {'type': 'article'}},
    {'$group': {'_id': '$journal', 'Autores': {'$addToSet': '$author'}}},
    {'$project': {'_id': 1, 'AutoresTotales': {'$size': '$Autores'}}},
    {'$sort': {'AutoresTotales': -1}},
    {'$limit': 1}])
<   {
      _id: 'CoRR',
      AutoresTotales: 482103
    }
```

El nombre de la revista que tiene más autores en total es CoRR con un total de 482.103 autores

Esta consulta tardó en ejecutarse en Python un total de 43,8 segundos

Consulta 9. Porcentaje de publicaciones en revistas con respecto al total de publicaciones.

```
> db.publi.aggregate(
    [{'$group': {'_id': 'None', 'Total': {'$sum': 1}, 'Revistas': {'$sum': {'$cond': [{'$eq': ['$type', 'article']}, 1, 0]}}}},
    {'$project': {'Porcentaje': {'$multiply': [{'$divide': ['$Revistas', '$Total']}, 100]}}},
   {'$project': {'_id': 0}}])
<   {
      Porcentaje: 48.68838378299506
    }
```

El porcentaje de publicaciones en revistas es 48.68839940547105 aproximadamente un 48,69%.

Esta consulta tardó en ejecutarse en Python un total de 26,3 segundos

Consulta 10. Número de autores ocasionales, es decir, que tengan menos de 5 publicaciones en total.

```
> db.publi.aggregate(
  [{'$unwind': "$author"},
  {'$group': {'_id': "$author", 'Publicaciones': {'$sum': 1}}},
  {'$match': {'Publicaciones': {'$lt': 5}}},
  {'$group': {'_id': 'null', 'AutoresOcasionales': {'$sum': 1}}}])
<   {
      _id: 'null',
      AutoresOcasionales: 3013611
    }
```

La cantidad de autores ocasionales (que tengan menos de 5 publicaciones) son: 3.013.611

Esta consulta tardó en ejecutarse en Python un total de 2 minutos 29 segundos

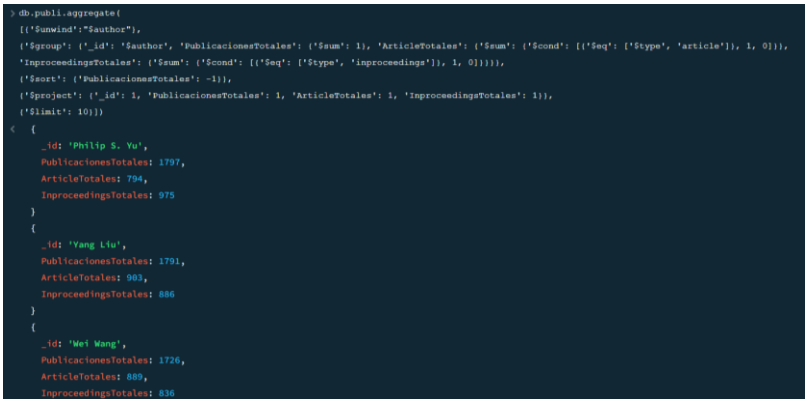
Consulta 11. Autor más prolífico y número total de publicaciones que tiene.

```
> db.publi.aggregate(
  [{'$unwind': "$author"},
  {'$group': {'_id': '$author', 'PublicacionesTotales': {'$sum': 1}}},
  {'$sort': {'PublicacionesTotales': -1}},
  {'$limit': 1}])
<   {
      _id: 'Philip S. Yu',
      PublicacionesTotales: 1797
    }
```

El autor más prolífico es Philip S. Yu con un total de 1797 publicaciones.

Esta consulta tardó en ejecutarse en Python un total de 2 minutos 29 segundos.

Consulta 12. Número de artículos de revista (article) y número de artículos en congresos (inproceedings) de los diez autores con más publicaciones totales. Para cada uno de estos diez autores debe aparecer su nombre, el número total de publicaciones, el número de artículos en revista y el número de artículos en congresos.



Los 10 autores con más publicaciones totales son:

Autor	PublicacionesTotales	ArticleTotales	InproceedingsTotales
Philip S. Yu	1797	794	975
Yang Liu	1791	903	886
Wei Wang	1726	889	836
Mohamed-Slim Alouini	1700	1118	591
Yu Zhang	1589	816	773
Wei Zhang	1551	772	779
H. Vincent Poor	1451	957	493
Lei Wang	1443	759	684
Dacheng Tao	1437	936	495
Xin Wang	1390	676	712

Esta consulta tardó en ejecutarse en Python un total de 4 minutos 12 segundos.

Consulta 13. Número medio de autores de todas las publicaciones que tenga en su conjunto de datos.

```
> db.publi.aggregate(
  [{'$unwind':'$author'},
  {'$group':{'_id':"$_id", 'Numero_Medio_Autores_Publicacion': {'$sum':1}}},
  {'$group':{'_id':"None", 'Numero_Medio_Autores_Publicacion': {'$avg':'$Numero_Medio_Autores_Publicacion' }}},
  {'$project': {'_id': 0, 'Numero_Medio_Autores_Publicacion': 1}}])
<   {
      Numero_Medio_Autores_Publicacion: 3.2634572517267886
    }
```

El número medio de autores en todas las publicaciones es de 3.2634572517267886 aproximadamente 3 autores por publicación.

Esta consulta tardó en ejecutarse en Python un total de 1 minutos 47 segundos.

Consulta 14. Número medio de autores de las publicaciones en congresos y revistas cada año.

```
> db.publi.aggregate([
    {'$unwind': '$author'},
    {'$match': {'type': {'$in': ["article", "inproceedings"]}}},
    {"$group": {"_id": {"id": "$_id", "year": "$year"}, 'PublicacionesTotales': {'$sum': 1}, 'Autores_Año': {'$sum': 1}}},
    {'$group': {'_id': "$_id.year", 'Media_Autores': {'$avg': '$Autores_Año'}, 'PublicacionesTotales': {'$sum': '$PublicacionesTotales'}}},
    {'$sort':{"_id":1 }},
    {'$project': {'_id': 1,'Media_Autores': 1, 'PublicacionesTotales': 1}}])
<   {
      _id: '1936',
      Media_Autores: 1,
      PublicacionesTotales: 12
    }
    {
      _id: '1937',
      Media_Autores: 1,
      PublicacionesTotales: 16
    }
    {
      _id: '1938',
      Media_Autores: 1.1,
      PublicacionesTotales: 11
    }
    {
      _id: '1939',
      Media_Autores: 1,
      PublicacionesTotales: 18
```

```
  {
    _id: '2019',
    Media_Autores: 3.7576991220910556,
    PublicacionesTotales: 1510084
  }
  {
    _id: '2020',
    Media_Autores: 3.8454841336361185,
    PublicacionesTotales: 1610408
  }
  {
    _id: '2021',
    Media_Autores: 3.9352868107407257,
    PublicacionesTotales: 1736099
  }
  {
    _id: '2022',
    Media_Autores: 4.088344919767515,
    PublicacionesTotales: 1764869
  }
  {
    _id: '2023',
    Media_Autores: 4.1109766661624985,
    PublicacionesTotales: 298979
  }
```

```
%%time
Ejercicio14 = [
  {'$unwind': '$author'},
  {'$match': {'type': {'$in': ["article", "inproceedings"]}}},
  {"$group": {"_id": {"id": "$_id", "year": "$year"}, 'PublicacionesTotales': {'$sum': 1}, 'Autores_Año': {'$sum': 1}}},
  {'$group': {'_id': "$_id.year", 'Media_Autores': {'$avg': '$Autores_Año'}, 'PublicacionesTotales': {'$sum': '$PublicacionesTot
  {'$sort':{"_id":1 }},
  {'$project': {'_id': 1,'Media_Autores': 1, 'PublicacionesTotales': 1}}]

Resultado14 = db.publi.aggregate(Ejercicio14, allowDiskUse=True)
print(list(Resultado14))
```


[{'_id': '1936', 'Media_Autores': 1.0, 'PublicacionesTotales': 12}, {'_id': '1937', 'Media_Autores': 1.0, 'PublicacionesTotale
s': 16}, {'_id': '1938', 'Media_Autores': 1.1, 'PublicacionesTotales': 11}, {'_id': '1939', 'Media_Autores': 1.0, 'Publicacione
sTotales': 18}, {'_id': '1940', 'Media_Autores': 1.2, 'PublicacionesTotales': 12}, {'_id': '1941', 'Media_Autores': 1.0, 'Publi
cacionesTotales': 13}, {'_id': '1942', 'Media_Autores': 1.0769230769230769, 'PublicacionesTotales': 14}, {'_id': '1943', 'Media
_Autores': 1.0, 'PublicacionesTotales': 8}, {'_id': '1944', 'Media_Autores': 1.0, 'PublicacionesTotales': 5}, {'_id': '1945',
'Media_Autores': 1.2222222222222223, 'PublicacionesTotales': 11}, {'_id': '1946', 'Media_Autores': 1.1612903225806452, 'Publica
cionesTotales': 36}, {'_id': '1947', 'Media_Autores': 1.2, 'PublicacionesTotales': 12}, {'_id': '1948', 'Media_Autores': 1.15,
'PublicacionesTotales': 46}, {'_id': '1949', 'Media_Autores': 1.34, 'PublicacionesTotales': 67}, {'_id': '1950', 'Media_Autore
s': 1.08, 'PublicacionesTotales': 27}, {'_id': '1951', 'Media_Autores': 1.236842105263158, 'PublicacionesTotales': 47}, {'_id':
'1952', 'Media_Autores': 1.287037037037037, 'PublicacionesTotales': 139}, {'_id': '1953', 'Media_Autores': 1.187878787878788,
'PublicacionesTotales': 196}, {'_id': '1954', 'Media_Autores': 1.3217821782178218, 'PublicacionesTotales': 267}, {'_id': '195
5', 'Media_Autores': 1.2694300518134716, 'PublicacionesTotales': 245}, {'_id': '1956', 'Media_Autores': 1.3240740740740742, 'Pu
blicacionesTotales': 429}, {'_id': '1957', 'Media_Autores': 1.4481707317073171, 'PublicacionesTotales': 475}, {'_id': '1958',
'Media_Autores': 1.3501144164759726, 'PublicacionesTotales': 590}, {'_id': '1959', 'Media_Autores': 1.4619799139167862, 'Public
acionesTotales': 1019}, {'_id': '1960', 'Media_Autores': 1.3623931623931624, 'PublicacionesTotales': 797}, {'_id': '1961', 'Med
ia_Autores': 1.347926267281106, 'PublicacionesTotales': 1170}, {'_id': '1962', 'Media_Autores': 1.355379188712522, 'Publicacion
esTotales': 1537}, {'_id': '1963', 'Media_Autores': 1.4149043303121853, 'PublicacionesTotales': 1405}, {'_id': '1964', 'Media_A
utores': 1.4249049429657794, 'PublicacionesTotales': 1499}, {'_id': '1965', 'Media_Autores': 1.3799837266069976, 'Publicaciones
Totales': 1696}, {'_id': '1966', 'Media_Autores': 1.4152005629838142, 'PublicacionesTotales': 2011}, {'_id': '1967', 'Media_Aut
ores': 1.4054696789536267, 'PublicacionesTotales': 2364}, {'_id': '1968', 'Media_Autores': 1.4272076372315037, 'PublicacionesTo
tales': 2990}, {'_id': '1969', 'Media_Autores': 1.3990963855421688, 'PublicacionesTotales': 2787}, {'_id': '1970', 'Media_Autor
es': 1.4270531400966184, 'PublicacionesTotales': 2954}, {'_id': '1971', 'Media_Autores': 1.5177647464464298, 'PublicacionesTotal
es': 4400}, {'_id': '1972', 'Media_Autores': 1.4693819424665338, 'PublicacionesTotales': 5159}, {'_id': '1973', 'Media_Autore
s': 1.4868995631187774, 'PublicacionesTotales': 6129}, {'_id': '1974', 'Media_Autores': 1.5498521335023236, 'PublicacionesTotal
es': 7337}, {'_id': '1975', 'Media_Autores': 1.5203812614074224, 'PublicacionesTotales': 7497}, {'_id': '1976', 'Media_Autore
s': 1.6000747523827321, 'PublicacionesTotales': 8562}, {'_id': '1977', 'Media_Autores': 1.5760425909494233, 'PublicacionesTotal
es': 8881}, {'_id': '1978', 'Media_Autores': 1.592326870867124, 'PublicacionesTotales': 10210}, {'_id': '1979', 'Media_Autore
s': 1.6384474327628362, 'PublicacionesTotales': 10722}, {'_id': '1980', 'Media_Autores': 1.6294360385144429, 'PublicacionesTota
les': 11846}, {'_id': '1981', 'Media_Autores': 1.6829640947288007, 'PublicacionesTotales': 13218}, {'_id': '1982', 'Media_Autor
es': 1.7195420186749666, 'PublicacionesTotales': 15469}, {'_id': '1983', 'Media_Autores': 1.7401090028259991, 'PublicacionesTot
ales': 17241}, {'_id': '1984', 'Media_Autores': 1.7607120493782986, 'PublicacionesTotales': 19683}, {'_id': '1985', 'Media_Auto
res': 1.7765714735268665, 'PublicacionesTotales': 22582}, {'_id': '1986', 'Media_Autores': 1.838432495576878, 'PublicacionesTota
les': 28242}, {'_id': '1987', 'Media_Autores': 1.854858624521546, 'PublicacionesTotales': 30845}, {'_id': '1988', 'Media_Autore
s': 1.9466174416284174, 'PublicacionesTotales': 39018}, {'_id': '1989', 'Media_Autores': 1.9450380296993843, 'PublicacionesTota
les': 42962}, {'_id': '1990', 'Media_Autores': 2.005922165820643, 'PublicacionesTotales': 52162}, {'_id': '1991', 'Media_Autore
s': 2.0349952897665817, 'PublicacionesTotales': 58325}, {'_id': '1992', 'Media_Autores': 2.085984468937876, 'PublicacionesTotal
es': 66618}, {'_id': '1993', 'Media_Autores': 2.1451054537707771, 'PublicacionesTotales': 80553}, {'_id': '1994', 'Media_Autore
s': 2.204876083969837, 'PublicacionesTotales': 93059}, {'_id': '1995', 'Media_Autores': 2.2402076983858223, 'PublicacionesTota
les': 99230}, {'_id': '1996', 'Media_Autores': 2.2713286430535334, 'PublicacionesTotales': 112349}, {'_id': '1997', 'Media_Autor
es': 2.340939370625434, 'PublicacionesTotales': 124751}, {'_id': '1998', 'Media_Autores': 2.413001224987916, 'PublicacionesTot
ales': 147736}, {'_id': '1999', 'Media_Autores': 2.439606971924436, 'PublicacionesTotales': 163622}, {'_id': '2000', 'Media_Aut
ores': 2.472051388961276, 'PublicacionesTotales': 189725}, {'_id': '2001', 'Media_Autores': 2.5441924812030074, 'PublicacionesT
otales': 211486}, {'_id': '2002', 'Media_Autores': 2.6122769862662047, 'PublicacionesTotales': 244227}, {'_id': '2003', 'Media_
Autores': 2.7066787248473885, 'PublicacionesTotales': 299291}, {'_id': '2004', 'Media_Autores': 2.754931445580592, 'Publicacion
esTotales': 357254}, {'_id': '2005', 'Media_Autores': 2.814599422469948, 'PublicacionesTotales': 419122}, {'_id': '2006', 'Medi
a_Autores': 2.880714901715883, 'PublicacionesTotales': 484349}, {'_id': '2007', 'Media_Autores': 2.9102178626400605, 'Publicaci
onesTotales': 527509}, {'_id': '2008', 'Media_Autores': 2.960775400370798, 'PublicacionesTotales': 566648}, {'_id': '2009', 'Med
ia_Autores': 3.0127122350402846, 'PublicacionesTotales': 619974}, {'_id': '2010', 'Media_Autores': 3.059677188709447, 'Publicac
ionesTotales': 663286}, {'_id': '2011', 'Media_Autores': 3.1187687197888554, 'PublicacionesTotales': 736176}, {'_id': '2012',
'Media_Autores': 3.191500610165719, 'PublicacionesTotales': 802889}, {'_id': '2013', 'Media_Autores': 3.261624708563644 5, 'Publ
icacionesTotales': 871542}, {'_id': '2014', 'Media_Autores': 3.3382557916420783, 'PublicacionesTotales': 922083}, {'_id': '201
5', 'Media_Autores': 3.3985029597891256, 'PublicacionesTotales': 981158}, {'_id': '2016', 'Media_Autores': 3.4664483635577867,
'PublicacionesTotales': 1042725}, {'_id': '2017', 'Media_Autores': 3.5410972237242577, 'PublicacionesTotales': 1146023}, {'_id':
'2018', 'Media_Autores': 3.6491085310336344, 'PublicacionesTotales': 1304965}, {'_id': '2019', 'Media_Autores': 3.757699122091 0
556, 'PublicacionesTotales': 1510084}, {'_id': '2020', 'Media_Autores': 3.8454841336361185, 'PublicacionesTotales': 1610408},
{'_id': '2021', 'Media_Autores': 3.9352868107407257, 'PublicacionesTotales': 1736099}, {'_id': '2022', 'Media_Autores': 4.08834
4919767515, 'PublicacionesTotales': 1764869}, {'_id': '2023', 'Media_Autores': 4.1109766661624985, 'PublicacionesTotales': 2989
79}]

De una de las cosas más importantes que podemos darnos cuenta en esta consulta, es que cuanto más antiguo es en año, menos media de autores por publicaciones.

Esta consulta tardó en ejecutarse en Python un total de 2 minutos 36 segundos.

Consulta 15. Los diez autores que tengan más publicaciones en solitario.

```
> db.publi.aggregate([
   {'$unwind':"$author"},
   { '$match': { '$or': [{ 'type': "article" }, { 'type': "inproceeding" }, { 'type': "incollection" }] } },
   { '$match': { 'type': "article" } },
   { '$group': { '_id': "$author", 'PublicacionesTotales': { '$sum': 1 } } },
   { '$sort': { 'PublicacionesTotales': -1 } },
   { '$limit': 10 }])
<  {
     _id: {
       '@orcid': '0000-0003-4827-1793',
       '#text': 'Mohamed-Slim Alouini'
     },
     PublicacionesTotales: 1118
   }
   {
     _id: {
       '@orcid': '0000-0002-2636-5214',
       '#text': 'Lajos Hanzo'
     },
     PublicacionesTotales: 1035
   }
```

Los diez autores con más publicaciones en solitario son:

- Mohamed-Slim Alouini = 1118 publicaciones.
- Lajos Hanzo = 1035 publicaciones.
- Witold Pedrycz = 965 publicaciones.
- H. Vincent Poor = 957 publicaciones.
- Dacheng Tao = 936 publicaciones.
- Yang Liu = 903 publicaciones.
- H. Vincent Poor = 895 publicaciones.
- Wei Wang = 889 publicaciones.
- Yu Zhang = 817 publicaciones.
- Philip S. Yu = 794 publicaciones.
- Chin-Chen Chang = 772 publicaciones.

Chin-Chen Chang ha sido añadido ya que H. Vincent Poor aparece en dos ocasiones.

Esta consulta tardó en ejecutarse en Python un total de 1 minutos 56 segundos.

Evaluación del tiempo empleado

En la fase de desarrollo de las consultas, hemos empleado bastante tiempo en hacer las pruebas con el código para Python con el que se pueden visualizar mejor las consultas y un código para visualizarlo en Mongo. No obstante, consideramos que ha sido tiempo útil de aprendizaje ya que hemos podido informarnos mucho sobre la asignatura.

Entorno al 70% de nuestro tiempo ha sido creando el código nuevo, investigando sobre las consultas y ejecutando, … y de ese 70%, aproximadamente la mitad de tiempo hemos estado probando las consultas en MongoDB.

También hemos estado esperando durante mucho tiempo la transformación de los datos de XML a JSON, la limpieza y pasar los datos a Mongo.

Un 20% de nuestro tiempo ha sido requerido para hacer la memoria y un 10% poner todo ordenado, renombrado, …

Algunas mejoras que se puede hacer al trabajo podrían ser:

- Emplear más tiempo en la realización de la memoria.

- Dedicar un poco más de tiempo en explicar todo mejor.

- Que los resultados de las consultas se visualicen mejor.