POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Systems and Methods for Big and Unstructured Data Project

Author(s): **Stefano Baroni**
Luca Brembilla
Alessia Menozzi
Caterina Giardi
Rocco Agnello

Group Number: **33**

Academic Year: 2022-2023

Contents

Part I

Delivery 3

1 | Import

For the import phase, we created five different json files using the same principle present in the MongoDB import. Through the json library, we opened the raw dataset file, we extracted the relevant information and finally we wrote them on the interested file.

The created files are:

- **publications.json** containing all the publications;
- **venues.json** with all the venues;
- **fos.json** for the fields of study;
- **authors.json** containing all the attributes of the authors;
- **rel_dw.json** which represents the N-M relationship between the fields of study and the publications;

1.1. Libraries used

To work on the raw dataset, in addition to those described in the MongoDB Import section, we added these libraries:

- **hashlib** to generate an id for the fields of study;
- **re** which permits to modify the output files in order to have a well-formatted json file.

1.2. Publications

For the publications, we take the information already present in the raw dataset, like the id or the publisher, and we created the ones which are not.

One of the tricky cases we had, was for the venue id. As a matter or fact, some venues have not the id field, so the script provides one, using a random number between 4.000.000 and 9.999.999. We opted for this range, because plotting all the ids, they are all below the lower threshold we fixed (Figure 1.1). In this way, the probability of having two identical ids created through the *random()* method is very low.

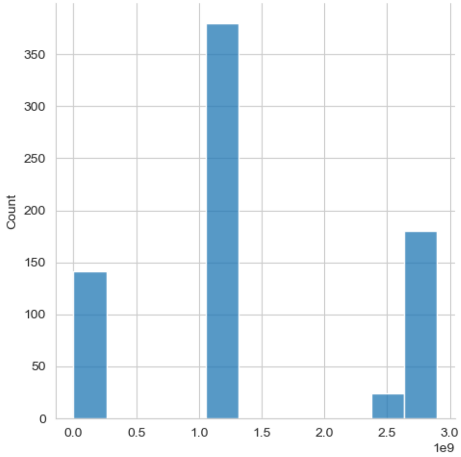


Figure 1.1: Plot of the ids

```
# Publisher part (authors[] filled in author part)
pub['id'] = data[i]['id']
pub['pages'] = random.randint(1, 20)
pub['abstract'] = "Abstract of " + \
    data[i]['title'] + ": " + lorem.words(5)
pub['title'] = data[i]['title']
if data[i]['publisher'] == "":
    data[i]['publisher'] = "PoliPrint, Milano"
pub['publisher'] = data[i]['publisher']
pub['year'] = data[i]['year']
if 'id' in data[i]['venue']:
    pub['venue'] = data[i]['venue']['id']
else:
    data[i]['venue']['id'] = random.randint(4_000_000_000, 9_999_999_999
)
    pub['venue'] = data[i]['venue']['id']

pub['references'] = []
for j in range(random.randint(1, 6)):
    number = vect_id[random.randint(0, n_doc-1)]
    while number in pub['references']:
        number = vect_id[random.randint(0, n_doc-1)]
    pub["references"].append(number)
```

1.3. Venues

The venue part has the same methods viewed before. Notice that before performing the operations, we check if the venue was already present in the one inserted before.

This problem may arise because some publications can derive from the same venue, and if we did not check this, the venue would be inserted many times, creating replicates. This way of solving this issue will be used also for the *authors* and for the *fields of study*.

```
# Venue part
if data[i]['venue']['id'] not in vect_ven_ids:
    vect_ven_ids.append(data[i]['venue']['id'])
    ven['id'] = data[i]['venue']['id']
    ven['raw'] = data[i]['venue']['raw']
    ven['volume'] = random.randint(1, 3)
    ven['number'] = random.randint(1, 20)
    if 'type' in data[i]['venue']:
        ven['type'] = data[i]['venue']['type']
    else:
        data[i]['venue']['type'] = 'J'
        ven['type'] = data[i]['venue']['type']
    ven['date'] = date.fromordinal(random.randint(date(day=1, month=1, year=data[i]['year']).toordinal(
    ), date(day=31, month=12, year=data[i]['year']).toordinal())).isoformat()
```

1.4. Fields of study and relation N-M

In this part of the script we cope with the issue of the *id* field. To avoid that two fields of study have the same id, we used the *md5* hash method, which starting form a seed, which is the name of the fos (that is unique), generates an alphanumeric string.

```
# Relation Deals_With + Field Of Study (FOS) part
for f_o_s in data[i]['fos']:
    obj = {}
    obj['pub_id'] = data[i]['id']
    obj['fos_id'] = hashlib.md5(f_o_s['name'].encode()).hexdigest()
    obj['weight'] = f_o_s['w']
    rel_dw.append(obj)
    if f_o_s['name'] not in vect_fos_ids:
        obj = {}
        vect_fos_ids.append(f_o_s['name'])
        obj['id'] = hashlib.md5(f_o_s['name'].encode()).hexdigest()
        obj['name'] = f_o_s['name']
        fos.append(obj)
```

1.5. Authors

This part has the same structure of the previous part and the same methods presented in the import of MongoDB.

Notice that the last line of the following code presents also the creation of the array of authors of the publication

```
# Author part
pub['authors'] = []
for author in data[i]['authors']:
    if author['id'] not in vect_aut_ids:
        obj = {}
        vect_aut_ids.append(author['id'])
        obj['id'] = author['id']
        obj['name'] = author['name']
        obj['email'] = author['name'].split()[0] + '.' + \
            author['name'].split()[len(
                author['name'].split())-1] + '@mail.com'
        obj['bio'] = 'Bio of ' + author['name'] + ': ' + lorem.words(5)
        if 'org' in author.keys():
            obj['affiliation'] = author['org']
        else:
            obj['affiliation'] = vect_org[random.randint(
                0, len(vect_org) - 1)]
        aut.append(obj)
    pub['authors'].append(author['id'])
```

2 | Queries with Spark

2.1. Data creation/update

Here are the five data creation/update commands.

2.1.1. Query 1

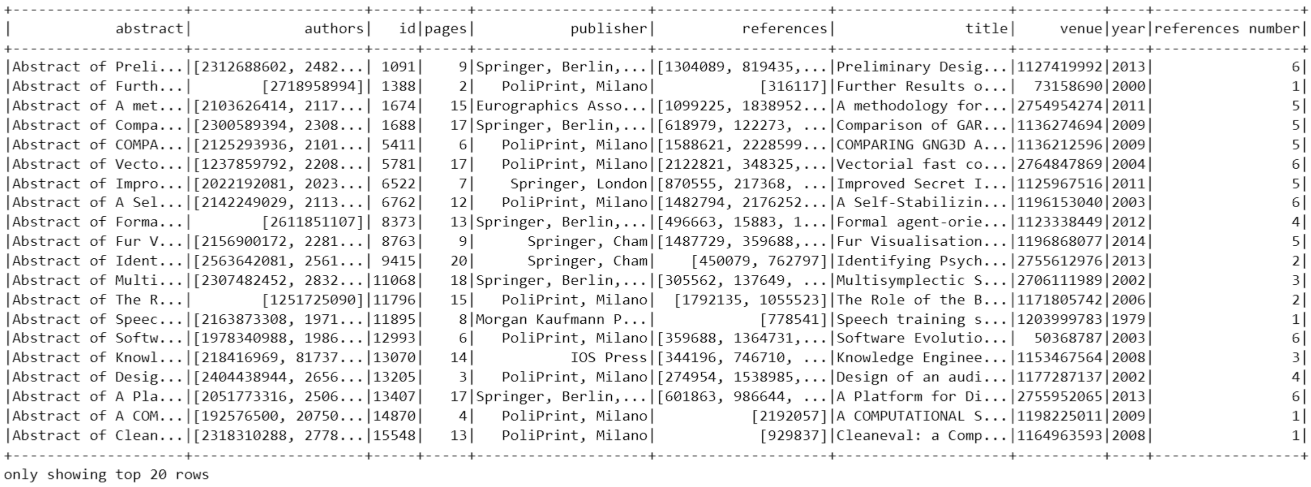
```
columns = ["date", "id", "number", "raw", "type", "volume"]
newRow = spark.createDataFrame([("2022-12-05", 9999999, 99, "New venue
    raw", "C", 3)], columns)
venue_df = venue_df.union(newRow)
venue_df.show(truncate=False)
```

Listing 2.1: Creation of a new row

2.1.2. Query 2

```
    publication_df.withColumn("references number", size(col("references
    "))).show(truncate=False)
```

Listing 2.2: Creation of a new column that counts the number of references



```
+--------------------+--------------------+-----+-----+--------------------+--------------------+--------------------+----------+----+-----------------+
|            abstract|             authors|   id|pages|           publisher|          references|               title|     venue|year|references number|
+--------------------+--------------------+-----+-----+--------------------+--------------------+--------------------+----------+----+-----------------+
|Abstract of Preli...|[2312688602, 2482...| 1091|    9|Springer, Berlin,...|[1304089, 819435,...|Preliminary Desig...|1127419992|2013|                6|
|Abstract of Furth...|        [2718958994]| 1388|    2|   PoliPrint, Milano|            [316117]|Further Results o...|  73158690|2000|                1|
|Abstract of A met...|[2103626414, 2117...| 1674|   15|Eurographics Asso...|[1099225, 1838952...|A methodology for...|2754954274|2011|                5|
|Abstract of Compa...|[2300589394, 2308...| 1688|   17|Springer, Berlin,...|[618979, 122273, ...|Comparison of GAR...|1136274694|2009|                5|
|Abstract of COMPA...|[2125293936, 2101...| 5411|    6|   PoliPrint, Milano|[1588621, 2228599...|COMPARING GNG3D A...|1136212596|2009|                5|
|Abstract of Vecto...|[1237859792, 2208...| 5781|   17|   PoliPrint, Milano|[2122821, 348325,...|Vectorial fast co...|2764847869|2004|                6|
|Abstract of Impro...|[2022192081, 2023...| 6522|    7|    Springer, London|[870555, 217368, ...|Improved Secret I...|1125967516|2011|                5|
|Abstract of A Sel...|[2142249029, 2113...| 6762|   12|   PoliPrint, Milano|[1482794, 2176252...|A Self-Stabilizin...|1196153040|2003|                6|
|Abstract of Forma...|        [2611851107]| 8373|   13|Springer, Berlin,...|[496663, 15883, 1...|Formal agent-orie...|1123338449|2012|                4|
|Abstract of Fur V...|[2156900172, 2281...| 8763|    9|      Springer, Cham|[1487729, 359688,...|Fur Visualisation...|1196868077|2014|                5|
|Abstract of Ident...|[2563642081, 2561...| 9415|   20|      Springer, Cham|    [450079, 762797]|Identifying Psych...|2755612976|2013|                2|
|Abstract of Multi...|[2307482452, 2832...|11068|   18|Springer, Berlin,...|[305562, 137649, ...|Multisymplectic S...|2706111989|2002|                3|
|Abstract of The R...|        [1251725090]|11796|   15|   PoliPrint, Milano|  [1792135, 1055523]|The Role of the B...|1171805742|2006|                2|
|Abstract of Speec...|[2163873308, 1971...|11895|    8|Morgan Kaufmann P...|            [778541]|Speech training s...|1203999783|1979|                1|
|Abstract of Softw...|[1978340988, 1986...|12993|    6|   PoliPrint, Milano|[359688, 1364731,...|Software Evolutio...|  50368787|2003|                6|
|Abstract of Knowl...|[218416969, 81737...|13070|   14|           IOS Press|[344196, 746710, ...|Knowledge Enginee...|1153467564|2008|                3|
|Abstract of Desig...|[2404438944, 2656...|13205|    3|   PoliPrint, Milano|[274954, 1538985,...|Design of an audi...|1177287137|2002|                4|
|Abstract of A Pla...|[2051773316, 2506...|13407|   17|Springer, Berlin,...|[601863, 986644, ...|A Platform for Di...|2755952065|2013|                6|
|Abstract of A COM...|[192576500, 20750...|14870|    4|   PoliPrint, Milano|           [2192057]|A COMPUTATIONAL S...|1198225011|2009|                1|
|Abstract of Clean...|[2318310288, 2778...|15548|   13|   PoliPrint, Milano|            [929837]|Cleaneval: a Comp...|1164963593|2008|                1|
+--------------------+--------------------+-----+-----+--------------------+--------------------+--------------------+----------+----+-----------------+
only showing top 20 rows
```

Figure 2.1: Result of query 9.1.2

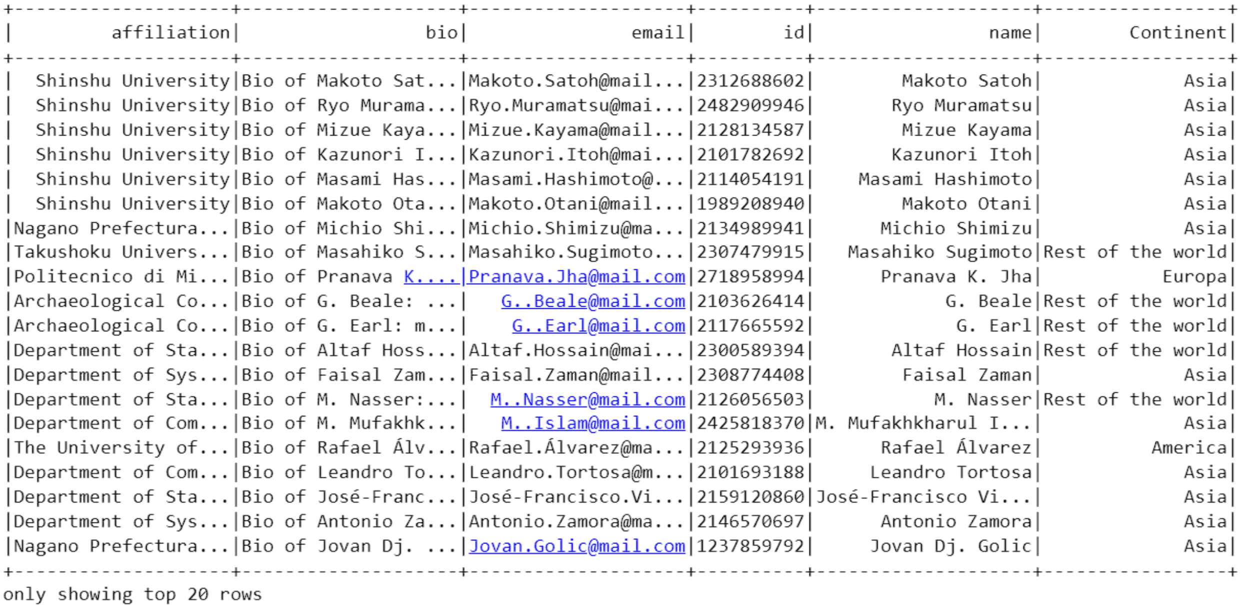
2.1.3. Query 3

Creation of a new column that represents the region of the university based on the vectors vectASIA, vectEU, vectAMERICA that we manually initialized in the python file

```
author_df = author_df.withColumn('Continent',
            when(author_df.affiliation.isin(vectASIA), lit("Asia"))\
            .when(author_df.affiliation.isin(vectAMERICA), lit("America"))\

            .when(author_df.affiliation.isin(vectEU), lit("Europa"))\

            .otherwise(lit("Rest of the world")))\
            .show(truncate=False)
```



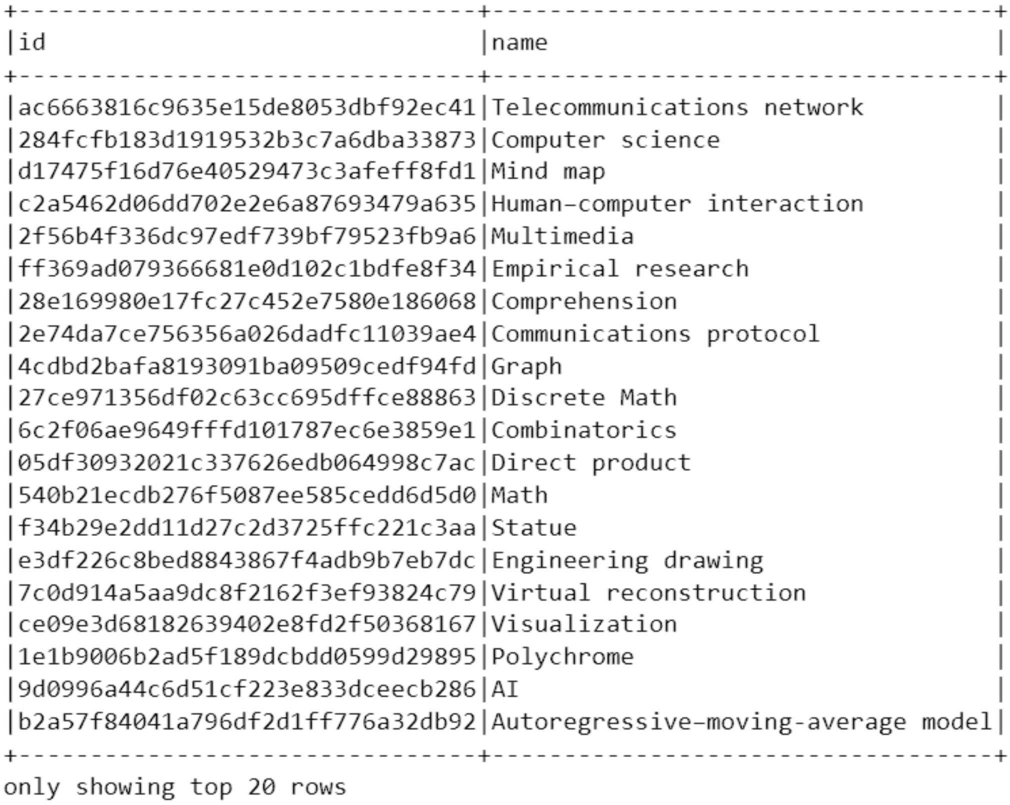
```
+--------------------+--------------------+--------------------+----------+--------------------+-----------------+
|         affiliation|                 bio|               email|        id|                name|        Continent|
+--------------------+--------------------+--------------------+----------+--------------------+-----------------+
|  Shinshu University|Bio of Makoto Sat...|Makoto.Satoh@mail...|2312688602|        Makoto Satoh|             Asia|
|  Shinshu University|Bio of Ryo Murama...|Ryo.Muramatsu@mai...|2482909946|       Ryo Muramatsu|             Asia|
|  Shinshu University|Bio of Mizue Kaya...|Mizue.Kayama@mail...|2128134587|        Mizue Kayama|             Asia|
|  Shinshu University|Bio of Kazunori I...|Kazunori.Itoh@mai...|2101782692|       Kazunori Itoh|             Asia|
|  Shinshu University|Bio of Masami Has...|Masami.Hashimoto@...|2114054191|    Masami Hashimoto|             Asia|
|  Shinshu University|Bio of Makoto Ota...|Makoto.Otani@mail...|1989208940|        Makoto Otani|             Asia|
|Nagano Prefectura...|Bio of Michio Shi...|Michio.Shimizu@ma...|2134989941|      Michio Shimizu|             Asia|
|Takushoku Univers...|Bio of Masahiko S...|Masahiko.Sugimoto...|2307479915|   Masahiko Sugimoto|Rest of the world|
|Politecnico di Mi...|Bio of Pranava K....|Pranava.Jha@mail.com|2718958994|      Pranava K. Jha|           Europa|
|Archaeological Co...|Bio of G. Beale: ...|   G..Beale@mail.com|2103626414|            G. Beale|Rest of the world|
|Archaeological Co...|Bio of G. Earl: m...|    G..Earl@mail.com|2117665592|             G. Earl|Rest of the world|
|Department of Sta...|Bio of Altaf Hoss...|Altaf.Hossain@mai...|2300589394|       Altaf Hossain|Rest of the world|
|Department of Sys...|Bio of Faisal Zam...|Faisal.Zaman@mail...|2308774408|        Faisal Zaman|             Asia|
|Department of Sta...|Bio of M. Nasser:...|  M..Nasser@mail.com|2126056503|           M. Nasser|Rest of the world|
|Department of Com...|Bio of M. Mufakhk...|   M..Islam@mail.com|2425818370|M. Mufakhkharul I...|             Asia|
|The University of...|Bio of Rafael Álv...|Rafael.Álvarez@ma...|2125293936|      Rafael Álvarez|          America|
|Department of Com...|Bio of Leandro To...|Leandro.Tortosa@m...|2101693188|     Leandro Tortosa|             Asia|
|Department of Sta...|Bio of José-Franc...|José-Francisco.Vi...|2159120860|José-Francisco Vi...|             Asia|
|Department of Sys...|Bio of Antonio Za...|Antonio.Zamora@ma...|2146570697|      Antonio Zamora|             Asia|
|Nagano Prefectura...|Bio of Jovan Dj. ...|Jovan.Golic@mail.com|1237859792|     Jovan Dj. Golic|             Asia|
+--------------------+--------------------+--------------------+----------+--------------------+-----------------+
only showing top 20 rows
```

Figure 2.2: Result of query 9.1.3

2.1.4. Query 4

```
fos_df = fos_df.withColumn('name',
                           when(fos_df.name.contains("mathematics"),
                                regexp_replace(fos_df.name, 'Discrete mathematics', 'Discrete Math'))
                           .when(fos_df.name.contains("Mathematics"),
                                 regexp_replace(fos_df.name, ' Mathematics', 'Math'))
                           .when(fos_df.name.contains("Artificial intelligence"),
                                 regexp_replace(fos_df.name, 'Artificial intelligence', 'AI'))
                           .otherwise(fos_df.name)).show()
```

Listing 2.3: Update of fos names with replacing

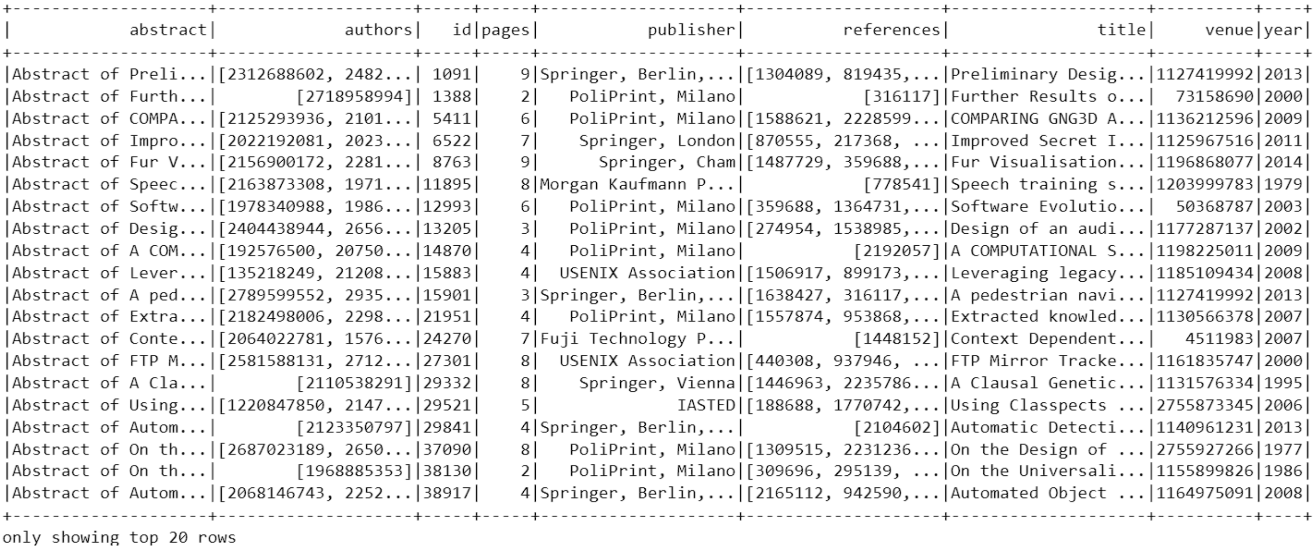


```
+--------------------------------+-----------------------------------+
|id                              |name                               |
+--------------------------------+-----------------------------------+
|ac6663816c9635e15de8053dbf92ec41|Telecommunications network         |
|284fcfb183d1919532b3c7a6dba33873|Computer science                   |
|d17475f16d76e40529473c3afeff8fd1|Mind map                           |
|c2a5462d06dd702e2e6a87693479a635|Human-computer interaction         |
|2f56b4f336dc97edf739bf79523fb9a6|Multimedia                         |
|ff369ad079366681e0d102c1bdfe8f34|Empirical research                 |
|28e169980e17fc27c452e7580e186068|Comprehension                      |
|2e74da7ce756356a026dadfc11039ae4|Communications protocol            |
|4cdbd2bafa8193091ba09509cedf94fd|Graph                              |
|27ce971356df02c63cc695dffce88863|Discrete Math                      |
|6c2f06ae9649fffd101787ec6e3859e1|Combinatorics                      |
|05df30932021c337626edb064998c7ac|Direct product                     |
|540b21ecdb276f5087ee585cedd6d5d0|Math                               |
|f34b29e2dd11d27c2d3725ffc221c3aa|Statue                             |
|e3df226c8bed8843867f4adb9b7eb7dc|Engineering drawing                |
|7c0d914a5aa9dc8f2162f3ef93824c79|Virtual reconstruction             |
|ce09e3d68182639402e8fd2f50368167|Visualization                      |
|1e1b9006b2ad5f189dcbdd0599d29895|Polychrome                         |
|9d0996a44c6d51cf223e833dceecb286|AI                                 |
|b2a57f84041a796df2d1ff776a32db92|Autoregressive-moving-average model|
+--------------------------------+-----------------------------------+
only showing top 20 rows
```

Figure 2.3: Result of query 9.1.4

2.1.5. Query 5

```
publication_df= publication_df.where(publication_df.pages < 10).show()
```



```
+--------------------+--------------------+-----+-----+--------------------+--------------------+--------------------+----------+----+
|            abstract|             authors|   id|pages|           publisher|          references|               title|     venue|year|
+--------------------+--------------------+-----+-----+--------------------+--------------------+--------------------+----------+----+
|Abstract of Preli...|[2312688602, 2482...| 1091|    9|Springer, Berlin,...|[1304089, 819435,...|Preliminary Desig...|1127419992|2013|
|Abstract of Furth...|        [2718958994]| 1388|    2|   PoliPrint, Milano|            [316117]|Further Results o...|  73158690|2000|
|Abstract of COMPA...|[2125293936, 2101...| 5411|    6|   PoliPrint, Milano|[1588621, 2228599...|COMPARING GNG3D A...|1136212596|2009|
|Abstract of Impro...|[2022192081, 2023...| 6522|    7|    Springer, London|[870555, 217368, ...|Improved Secret I...|1125967516|2011|
|Abstract of Fur V...|[2156900172, 2281...| 8763|    9|      Springer, Cham|[1487729, 359688,...|Fur Visualisation...|1196868077|2014|
|Abstract of Speec...|[2163873308, 1971...|11895|    8|Morgan Kaufmann P...|            [778541]|Speech training s...|1203999783|1979|
|Abstract of Softw...|[1978340988, 1986...|12993|    6|   PoliPrint, Milano|[359688, 1364731,...|Software Evolutio...|  50368787|2003|
|Abstract of Desig...|[2404438944, 2656...|13205|    3|   PoliPrint, Milano|[274954, 1538985,...|Design of an audi...|1177287137|2002|
|Abstract of A COM...|[192576500, 20750...|14870|    4|   PoliPrint, Milano|           [2192057]|A COMPUTATIONAL S...|1198225011|2009|
|Abstract of Lever...|[135218249, 21208...|15883|    4|  USENIX Association|[1506917, 899173,...|Leveraging legacy...|1185109434|2008|
|Abstract of A ped...|[2789599552, 2935...|15901|    3|Springer, Berlin,...|[1638427, 316117,...|A pedestrian navi...|1127419992|2013|
|Abstract of Extra...|[2182498006, 2298...|21951|    4|   PoliPrint, Milano|[1557874, 953868,...|Extracted knowled...|1130566378|2007|
|Abstract of Conte...|[2064022781, 1576...|24270|    7|Fuji Technology P...|           [1448152]|Context Dependent...|   4511983|2007|
|Abstract of FTP M...|[2581588131, 2712...|27301|    8|  USENIX Association|[440308, 937946, ...|FTP Mirror Tracke...|1161835747|2000|
|Abstract of A Cla...|        [2110538291]|29332|    8|    Springer, Vienna|[1446963, 2235786...|A Clausal Genetic...|1131576334|1995|
|Abstract of Using...|[1220847850, 2147...|29521|    5|              IASTED|[188688, 1770742,...|Using Classpects ...|2755873345|2006|
|Abstract of Autom...|        [2123350797]|29841|    4|Springer, Berlin,...|           [2104602]|Automatic Detecti...|1140961231|2013|
|Abstract of On th...|[2687023189, 2650...|37090|    8|   PoliPrint, Milano|[1309515, 2231236...|On the Design of ...|2755927266|1977|
|Abstract of On th...|        [1968885353]|38130|    2|   PoliPrint, Milano|[309696, 295139, ...|On the Universali...|1155899826|1986|
|Abstract of Autom...|[2068146743, 2252...|38917|    4|Springer, Berlin,...|[2165112, 942590,...|Automated Object ...|1164975091|2008|
+--------------------+--------------------+-----+-----+--------------------+--------------------+--------------------+----------+----+
only showing top 20 rows
```

Figure 2.4: Result of query 9.1.5

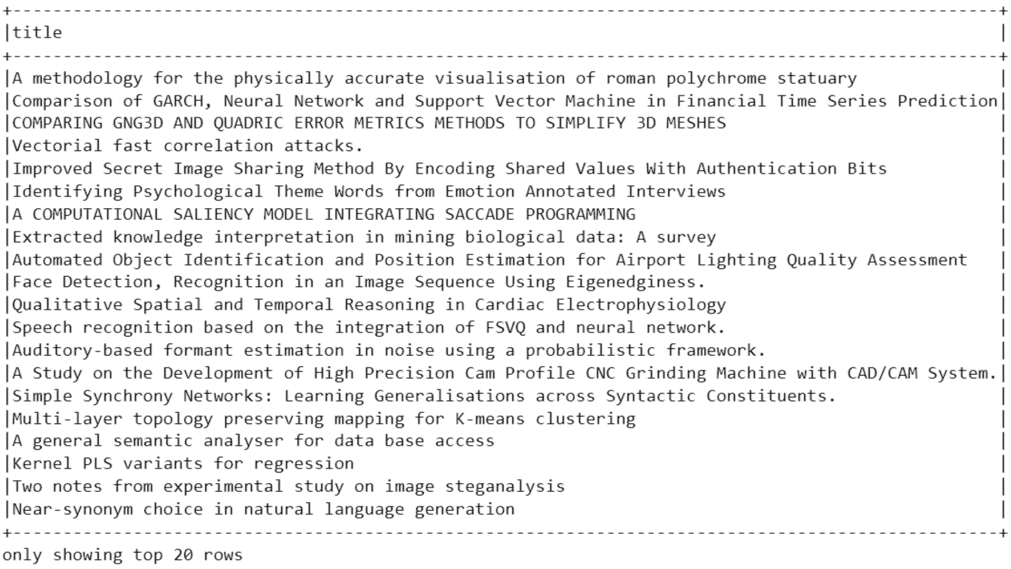
2.2. Queries

2.2.1. WHERE, JOIN

Return the papers that have as a field of study "Artificial Intelligence"

```
fosDF
    .filter(col("name") == "Artificial intelligence")
    .join(rel_dwDF, fosDF.id == rel_dwDF.fos_id, "inner")
    .join(publicationsDF, rel_dwDF.pub_id == publicationsDF.id, "inner")
    .select("title")
    .show()
```

Listing 2.4: Query 9.2.1



```
+--------------------------------------------------------------------------------------------------+
|title                                                                                             |
+--------------------------------------------------------------------------------------------------+
|A methodology for the physically accurate visualisation of roman polychrome statuary              |
|Comparison of GARCH, Neural Network and Support Vector Machine in Financial Time Series Prediction|
|COMPARING GNG3D AND QUADRIC ERROR METRICS METHODS TO SIMPLIFY 3D MESHES                           |
|Vectorial fast correlation attacks.                                                               |
|Improved Secret Image Sharing Method By Encoding Shared Values With Authentication Bits           |
|Identifying Psychological Theme Words from Emotion Annotated Interviews                           |
|A COMPUTATIONAL SALIENCY MODEL INTEGRATING SACCADE PROGRAMMING                                    |
|Extracted knowledge interpretation in mining biological data: A survey                            |
|Automated Object Identification and Position Estimation for Airport Lighting Quality Assessment   |
|Face Detection, Recognition in an Image Sequence Using Eigenedginess.                             |
|Qualitative Spatial and Temporal Reasoning in Cardiac Electrophysiology                           |
|Speech recognition based on the integration of FSVQ and neural network.                           |
|Auditory-based formant estimation in noise using a probabilistic framework.                       |
|A Study on the Development of High Precision Cam Profile CNC Grinding Machine with CAD/CAM System.|
|Simple Synchrony Networks: Learning Generalisations across Syntactic Constituents.                |
|Multi-layer topology preserving mapping for K-means clustering                                    |
|A general semantic analyser for data base access                                                  |
|Kernel PLS variants for regression                                                                |
|Two notes from experimental study on image steganalysis                                           |
|Near-synonym choice in natural language generation                                                |
+--------------------------------------------------------------------------------------------------+
only showing top 20 rows
```

Figure 2.5: Result of query 9.2.1

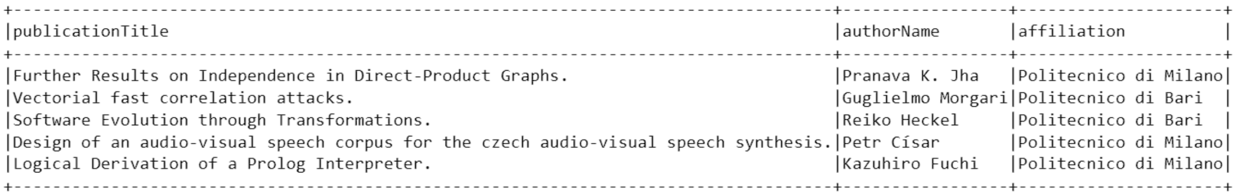
2.2.2. WHERE, LIMIT, LIKE

Return the first 5 authors name of our database that have affiliation with a "Politecnico" and return also the titles of their publications.

```
# import expr
from pyspark.sql.functions import expr

authorsDF
    # Rename the column of the author id for the join
    .withColumnRenamed("id", "authorId")
    # Affiliation with Politecnico in it  .filter(col("affiliation").like("%Politecnico%"))
    # Limit the df got at 5 authors
    .limit(5)
    # Join the publicationsDF to get these 5 authors publications
    .join(publicationsDF, expr("array_contains(authors, authorId)"))
    # Select what we want
    .select(col("title").alias("publicationTitle"), col("name").alias("authorName"), "affiliation")
    .show(truncate=False)
```

Listing 2.5: Query 9.2.2



```
+------------------------------------------------------------------------------------+-----------------+---------------------+
|publicationTitle                                                                    |authorName       |affiliation          |
+------------------------------------------------------------------------------------+-----------------+---------------------+
|Further Results on Independence in Direct-Product Graphs.                           |Pranava K. Jha   |Politecnico di Milano|
|Vectorial fast correlation attacks.                                                 |Guglielmo Morgari|Politecnico di Bari  |
|Software Evolution through Transformations.                                         |Reiko Heckel     |Politecnico di Bari  |
|Design of an audio-visual speech corpus for the czech audio-visual speech synthesis.|Petr Císar       |Politecnico di Milano|
|Logical Derivation of a Prolog Interpreter.                                         |Kazuhiro Fuchi   |Politecnico di Milano|
+------------------------------------------------------------------------------------+-----------------+---------------------+
```

Figure 2.6: Result of query 9.2.2

Notice that the result contains only 5 elements because the authors to which we've limited the query published only one paper each (or, at least, we have one paper for each of them in our database).

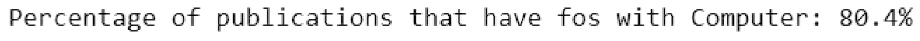
2.2.3. WHERE, IN, Nested Query

This query return the percentage of papers that have "Computer" in their fos, e.g. paper with "Computer Science" or "Computer Vision".

```
    computerScience_fos = fos_df.filter(col("name").contains("Computer")).select(col("id")).collect()
    computerScience_fos = [csf[0] for csf in computerScience_fos]

    count_cs_publications = rel_df.filter(col("fos_id").isin(computerScience_fos)).select(col("pub_id")).distinct().count()
    print("percentage of publications that have fos about Computer:" + str(count_cs_publications/2500*100) + "%")
```

Listing 2.6: Query 9.2.3



```
Percentage of publications that have fos with Computer: 80.4%
```

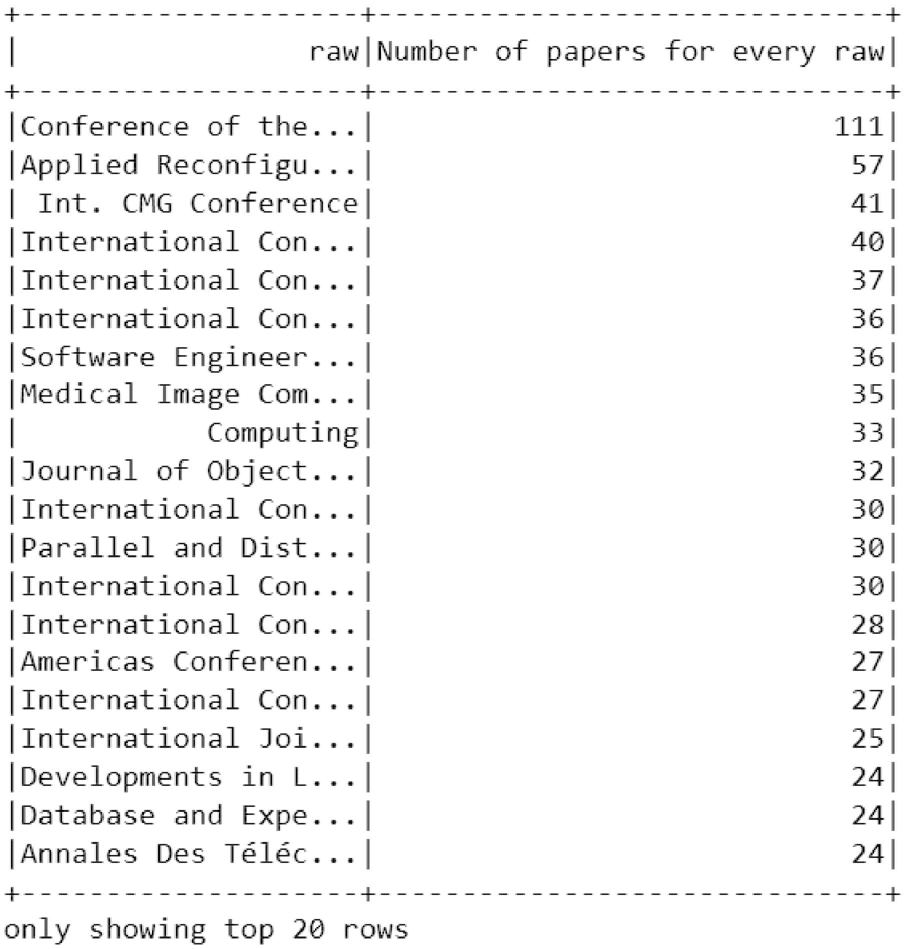
Figure 2.7: Result of query 9.2.3

2.2.4. GROUP BY, 1 JOIN, AS

The query counts for each name of the venue (that has different ID since it can have different editions or volumes) the number of papers that were presented there

```
 venue_df.join(publication_df, venue_df.id == publication_df.venue, "inner")\
     .groupby("raw")\
         .count()\
             .select(venue_df.raw, col("count").alias("Number of papers for every raw")).orderBy(col("Number of papers for every raw").desc()).show()
```

Listing 2.7: Query 9.2.4



```
+--------------------+------------------------------+
|                 raw|Number of papers for every raw|
+--------------------+------------------------------+
|Conference of the...|                           111|
|Applied Reconfigu...|                            57|
| Int. CMG Conference|                            41|
|International Con...|                            40|
|International Con...|                            37|
|International Con...|                            36|
|Software Engineer...|                            36|
|Medical Image Com...|                            35|
|           Computing|                            33|
|Journal of Object...|                            32|
|International Con...|                            30|
|Parallel and Dist...|                            30|
|International Con...|                            30|
|International Con...|                            28|
|Americas Conferen...|                            27|
|International Con...|                            27|
|International Joi...|                            25|
|Developments in L...|                            24|
|Database and Expe...|                            24|
|Annales Des Téléc...|                            24|
+--------------------+------------------------------+
only showing top 20 rows
```

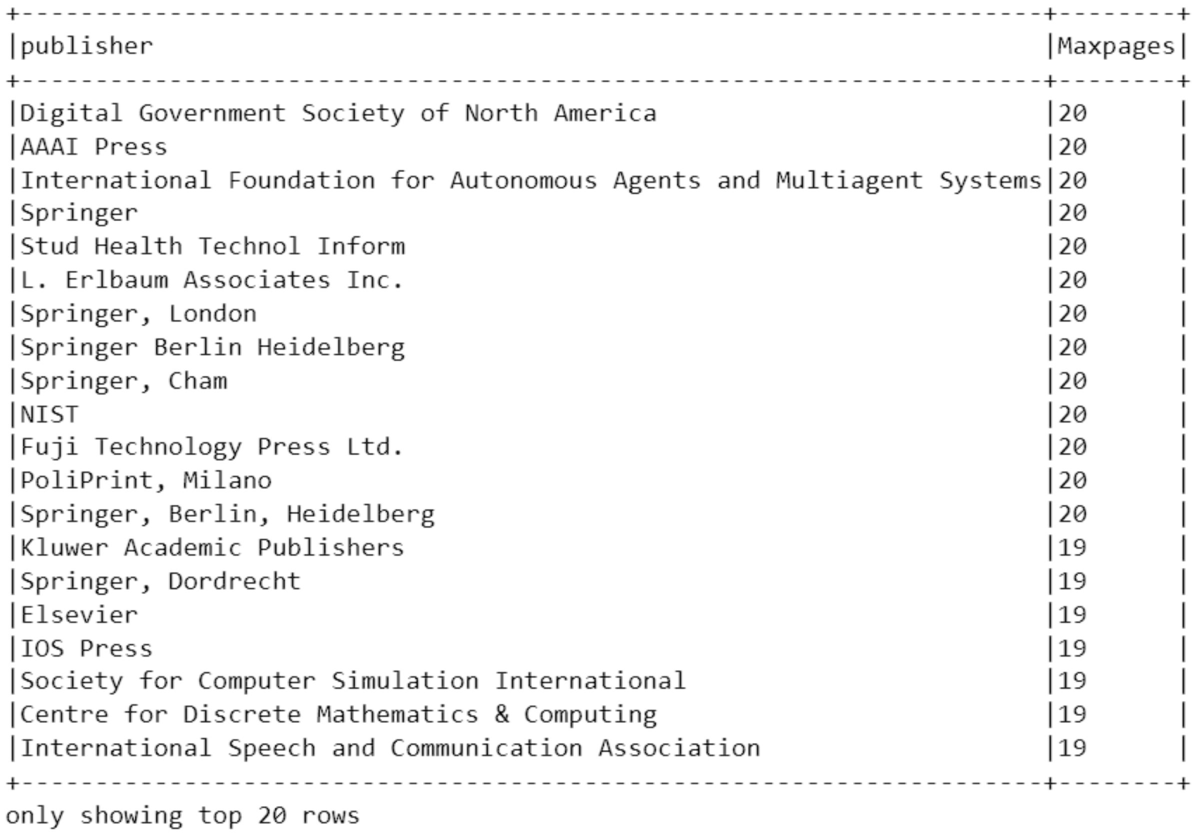
Figure 2.8: Result of query 9.2.4

2.2.5. WHERE, GROUP BY

Filter the papers that have at least 3 authors, then it shows for every publisher the max number of pages between the papers he published, shown in descending order

```
    publication_df.filter(size('authors') >= 3)\
        .groupBy('publisher').max('pages')\
            .select(publication_df.publisher, col("max(pages)").alias("Maxpages"))\
                .orderBy(col("Maxpages").desc())\
                .show(truncate=false)
```

Listing 2.8: Query 9.2.5



```
+---------------------------------------------------------------------+--------+
|publisher                                                            |Maxpages|
+---------------------------------------------------------------------+--------+
|Digital Government Society of North America                          |20      |
|AAAI Press                                                           |20      |
|International Foundation for Autonomous Agents and Multiagent Systems|20      |
|Springer                                                             |20      |
|Stud Health Technol Inform                                           |20      |
|L. Erlbaum Associates Inc.                                           |20      |
|Springer, London                                                     |20      |
|Springer Berlin Heidelberg                                           |20      |
|Springer, Cham                                                       |20      |
|NIST                                                                 |20      |
|Fuji Technology Press Ltd.                                           |20      |
|PoliPrint, Milano                                                    |20      |
|Springer, Berlin, Heidelberg                                         |20      |
|Kluwer Academic Publishers                                           |19      |
|Springer, Dordrecht                                                  |19      |
|Elsevier                                                             |19      |
|IOS Press                                                            |19      |
|Society for Computer Simulation International                        |19      |
|Centre for Discrete Mathematics & Computing                          |19      |
|International Speech and Communication Association                   |19      |
+---------------------------------------------------------------------+--------+
only showing top 20 rows
```

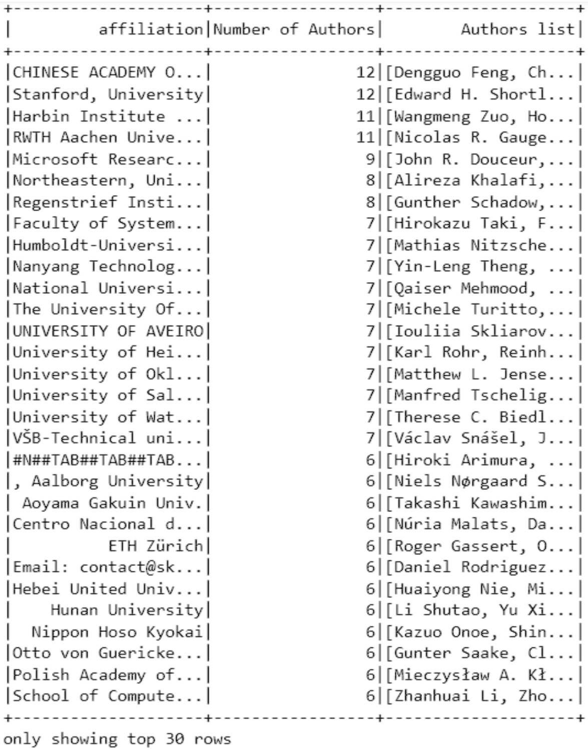
Figure 2.9: Result of query 9.2.5

2.2.6. GROUP BY, HAVING, AS

The query groups the venues by affiliation, for each of them it counts the number of authors by id and collects the names of the authors in a list, then it filters (HAVING) the affiliations by the number of authors in the list that should be between 5 and 15. We order the table by descending order for the number of authors, and in case of tie they show the affiliation name in alphabetic order

```
 author_df.groupBy("affiliation")\
     .agg(
         countDistinct("id").alias("Number of Authors"),
         collect_list(author_df.name).alias("Authors list")
         )\
             .filter((col("Number of Authors") < 15) & (col("Number of Authors") > 5))\
                 .orderBy(col("Number of Authors").desc(), col("affiliation").asc())\
                     .show(truncate=False)
```

Listing 2.9: Query 9.2.6



```
+--------------------+-----------------+--------------------+
|         affiliation|Number of Authors|        Authors list|
+--------------------+-----------------+--------------------+
|CHINESE ACADEMY O...|               12|[Dengguo Feng, Ch...|
|Stanford, University|               12|[Edward H. Shortl...|
|Harbin Institute ...|               11|[Wangmeng Zuo, Ho...|
|RWTH Aachen Unive...|               11|[Nicolas R. Gauge...|
|Microsoft Researc...|                9|[John R. Douceur,...|
|Northeastern, Uni...|                8|[Alireza Khalafi,...|
|Regenstrief Insti...|                8|[Gunther Schadow,...|
|Faculty of System...|                7|[Hirokazu Taki, F...|
|Humboldt-Universi...|                7|[Mathias Nitzsche...|
|Nanyang Technolog...|                7|[Yin-Leng Theng, ...|
|National Universi...|                7|[Qaiser Mehmood, ...|
|The University Of...|                7|[Michele Turitto,...|
|UNIVERSITY OF AVEIRO|                7|[Iouliia Skliarov...|
|University of Hei...|                7|[Karl Rohr, Reinh...|
|University of Okl...|                7|[Matthew L. Jense...|
|University of Sal...|                7|[Manfred Tschelig...|
|University of Wat...|                7|[Therese C. Biedl...|
|VŠB-Technical uni...|                7|[Václav Snášel, J...|
|#N##TAB##TAB##TAB...|                6|[Hiroki Arimura, ...|
|, Aalborg University|                6|[Niels Nørgaard S...|
| Aoyama Gakuin Univ.|                6|[Takashi Kawashim...|
|Centro Nacional d...|                6|[Núria Malats, Da...|
|         ETH Zürich|                6|[Roger Gassert, O...|
|Email: contact@sk...|                6|[Daniel Rodriguez...|
|Hebei United Univ...|                6|[Huaiyong Nie, Mi...|
|    Hunan University|                6|[Li Shutao, Yu Xi...|
|  Nippon Hoso Kyokai|                6|[Kazuo Onoe, Shin...|
|Otto von Guericke...|                6|[Gunter Saake, Cl...|
|Polish Academy of...|                6|[Mieczysław A. Kł...|
|School of Compute...|                6|[Zhanhuai Li, Zho...|
+--------------------+-----------------+--------------------+
only showing top 30 rows
```

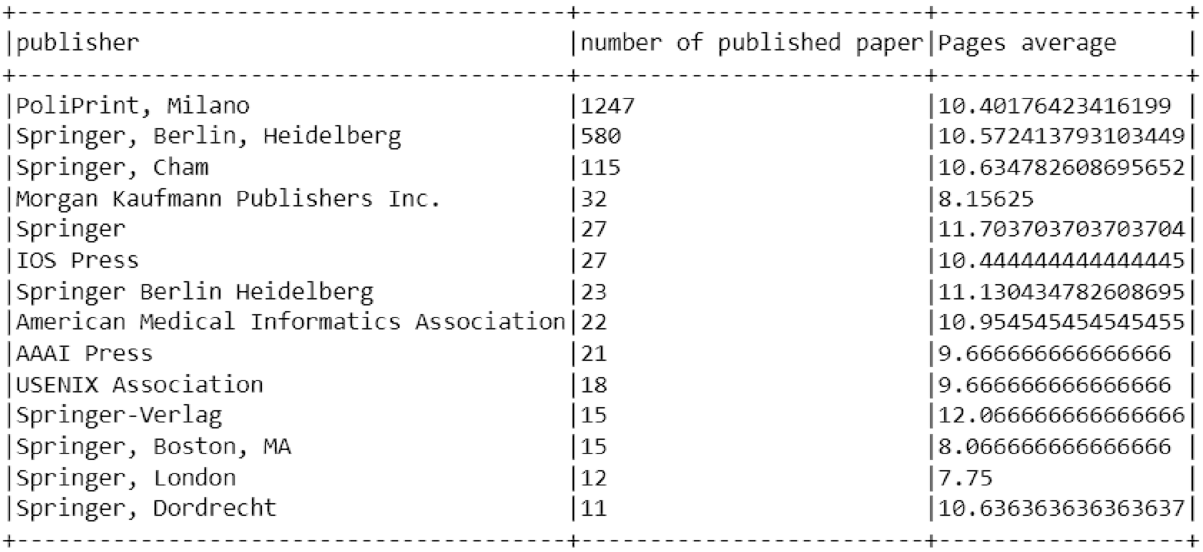
Figure 2.10: Result of query 9.2.6

2.2.7. WHERE, GROUP BY, HAVING, AS

This query lists the Publisher that have published at least 2 papers presented in a Conference, and shows the number of published paper, as well as the average number of pages of a paper published by them.

```
    publication_df.filter(publication_df.type == "Conference").groupBy("publisher").agg(
    count("id").alias("number of published paper"),
    avg("pages").alias("Pages average"),
    ).filter(
        col("number of published paper") > 1
      ).orderBy(
          col("number of published paper").desc(),
          col("pages average").desc()
          ).show(truncate = False)
```

Listing 2.10: Query 9.2.7



```
+----------------------------------+-------------------------+------------------+
|publisher                         |number of published paper|Pages average     |
+----------------------------------+-------------------------+------------------+
|PoliPrint, Milano                 |1247                     |10.40176423416199 |
|Springer, Berlin, Heidelberg      |580                      |10.572413793103449|
|Springer, Cham                    |115                      |10.634782608695652|
|Morgan Kaufmann Publishers Inc.   |32                       |8.15625           |
|Springer                          |27                       |11.703703703703704|
|IOS Press                         |27                       |10.444444444444445|
|Springer Berlin Heidelberg        |23                       |11.130434782608695|
|American Medical Informatics Association|22                 |10.954545454545455|
|AAAI Press                        |21                       |9.666666666666666 |
|USENIX Association                |18                       |9.666666666666666 |
|Springer-Verlag                   |15                       |12.066666666666666|
|Springer, Boston, MA              |15                       |8.066666666666666 |
|Springer, London                  |12                       |7.75              |
|Springer, Dordrecht               |11                       |10.636363636363637|
+----------------------------------+-------------------------+------------------+
```

Figure 2.11: Result of query 9.2.7

2.2.8. WHERE, Nested Query (i.e., 2-step Queries), GROUP BY

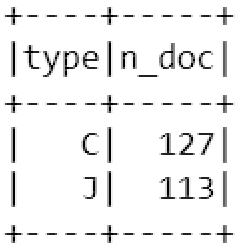
Count the publications grouping by the type of the venue they are taken from ('C' for Conference, 'J' for Journal). The publications counted are only the ones with at least one author from the 'Politecnico di Bari'.

```
authors_from_bari = author_df.filter(col('affiliation') == 'Politecnico di Bari').select('id').collect()
authors_from_bari = [barese[0] for barese in authors_from_bari]

exploded_pub = publication_df.select(publication_df.id, publication_df.venue, explode(publication_df.authors))
exploded_pub = exploded_pub.withColumnRenamed("col", "author")
exploded_pub = exploded_pub.withColumnRenamed("id", "pub_id")

exploded_pub.filter(col('author').isin(authors_from_bari))\
    .join(venue_df, exploded_pub.venue == venue_df.id, "inner")\
    .groupBy('type').agg(countDistinct('pub_id').alias('n_doc')).show()
```

Listing 2.11: Query 9.2.8



```
+----+-----+
|type|n_doc|
+----+-----+
|   C|  127|
|   J|  113|
+----+-----+
```

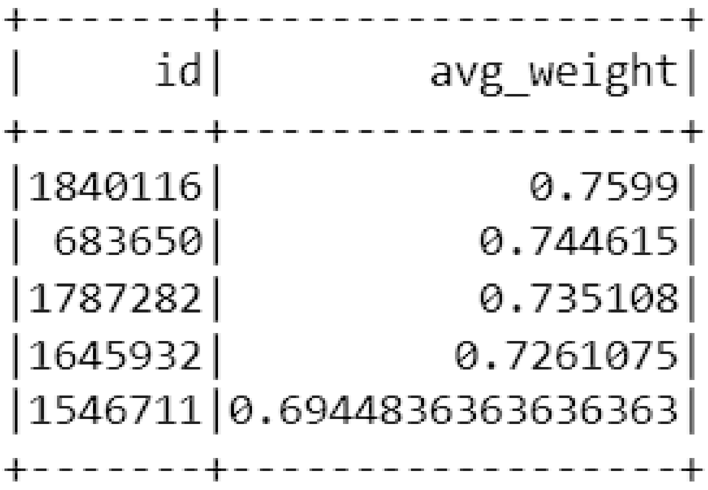
Figure 2.12: Result of Query 9.2.8

2.2.9. WHERE, GROUP BY, HAVING, 1 JOIN

From all the publications wrote before the 2000s, compute the average weight of the weights of its Field of Studies and display the top five of them, having the average above 0.5.

```
publication_df.filter(col('year')<2000)\
    .join(rel_df, publication_df.id == rel_df.pub_id, "inner")\
    .groupBy('id').agg(avg('weight').alias('avg_weight')).filter(col('avg_weight')>0.5)\
    .orderBy(col('avg_weight').desc())\
    .limit(5).show()
```

Listing 2.12: Query 9.2.9



```
+-------+------------------+
|     id|        avg_weight|
+-------+------------------+
|1840116|            0.7599|
| 683650|          0.744615|
|1787282|          0.735108|
|1645932|         0.7261075|
|1546711|0.6944836363636363|
+-------+------------------+
```

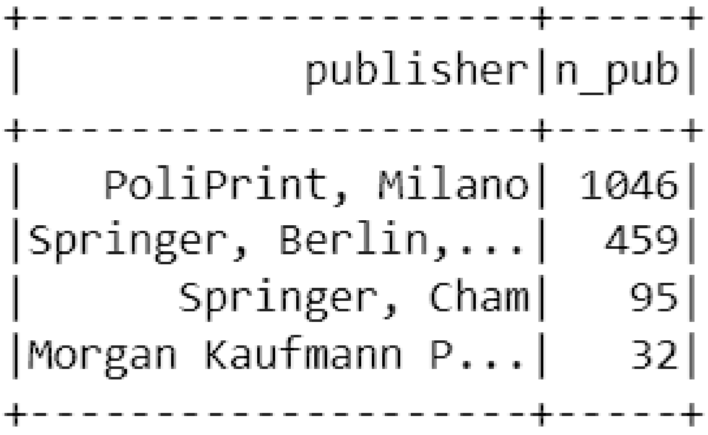
Figure 2.13: Result of Query 9.2.9

2.2.10. WHERE, GROUP BY, HAVING, 2 JOINs

Count the number of publications, grouping them by publisher, having the Field of Study dealing with Computers (its name must contain the word ’Computer’).

```
fos_df.filter(col(’name’).contains(’Computer’))\
    .join(rel_df, fos_df.id == rel_df.fos_id, "inner")\
    .join(publication_df, rel_df.pub_id == publication_df.id, "inner")\
    .groupBy(’publisher’).agg(countDistinct(’pub_id’).alias(’n_pub’)).
   filter(col(’n_pub’)>25)\
    .orderBy(col(’n_pub’).desc())\
    .limit(5).show()
```

Listing 2.13: Query 9.2.10



```
+--------------------+-----+
|           publisher|n_pub|
+--------------------+-----+
|   PoliPrint, Milano| 1046|
|Springer, Berlin,...|  459|
|      Springer, Cham|   95|
|Morgan Kaufmann P...|   32|
+--------------------+-----+
```

Figure 2.14: Result of Query 9.2.10