Databases II

Laboratory Exercise 2020/21

Name	**Surname**	**AM**
Nikiforos – George	Papageorgiou	1059633
Nicholas	Stamopoulos	1057764

I certify that I am the author of this work and that I have expressly and specifically cited or referenced in it all sources from which I have used data, ideas, sentences or words, whether they are accurately conveyed (in the original or translated) or paraphrased. I also certify that this assignment was prepared by me personally specifically for this particular course/seminar/curriculum.

I have been informed that according to the internal regulation of the University of Patras, article 50§6, any attempt to copy or in general to tamper with the examination and educational process by any examinee, in addition to nihilism, constitutes a serious disciplinary offense.

Signature	Signature
22 / 09 / 2021	22 / 09 / 2021

Code files attached

We submit the following code files along with this report

file	**Regarding question**	**Description/**
queryX.ipynb	1	**Comment** Contains the code for query X, where X ÿ [1, 10]

Technical characteristics of the operating environment

Technical characteristics of the physical PC used for the work

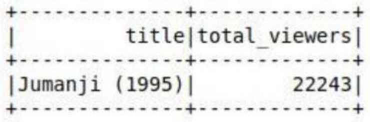
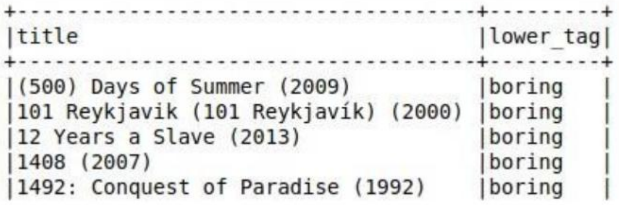
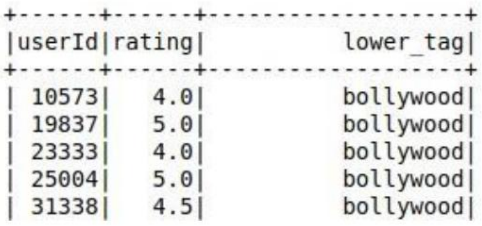
Feature	Price
CPU model	AMD Ryzen 5 2600 Six-Core Processor
CPU clock speed	3.4GHz
Physical CPU cores	6
Logical CPU cores	12
RAM	16Gb
Secondary Storage Type	SSD

Technical characteristics of the virtual machine (VM) used for the job **Feature**

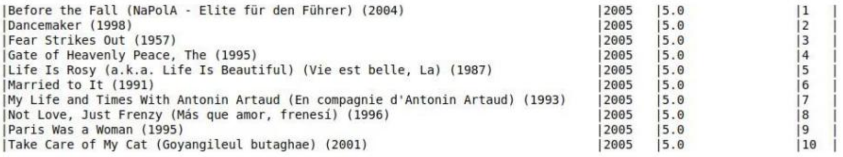
	Price
CPU cores	6
Execution cap	100%
RAM	9Gb
VM OS	Ubuntu 20.04.2 LTS
VM software	VirtualBox Windows 10
Host OS	

Question 1: Question Answers

Note: The requested results are listed in screenshot format, for better illustration.

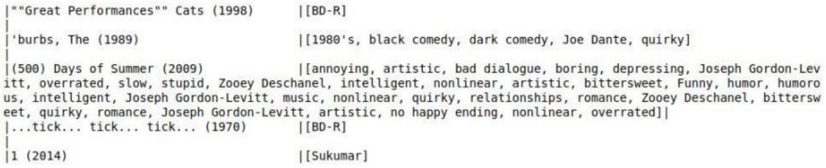
<table>
<tr><th><b>Query</b> Give</th><th>Answer</th></tr>
<tr><td>the number of users who watched the movie “Jumanji”. Give the names of</td><td>
<pre>
+--------------+-------------+
| title|total_viewers|
+--------------+-------------+
|Jumanji (1995)| 22243|
+--------------+-------------+
</pre></td></tr>
<tr><td>movies that users rated as “boring”. Give users who have rated the movie as</td><td>
<pre>
+------------------------------------+---------+
|title |lower_tag|
+------------------------------------+---------+
(500) Days of Summer (2009)	boring
101 Reykjavik (101 Reykjavík) (2000)	boring
12 Years a Slave (2013)	boring
1408 (2007)	boring
1492: Conquest of Paradise (1992)	boring
</pre></td></tr>
<tr><td>"Bollywood" and have rated it with a grade &gt;3.</td><td>
<pre>
+------+------+------------------+
|userId|rating| lower_tag|
+------+------+------------------+
10573	4.0	bollywood
19837	5.0	bollywood
23333	4.0	bollywood
25004	5.0	bollywood
31338	4.5	bollywood
</pre></td></tr>
</table>

Find the top
10 movies
for each year.



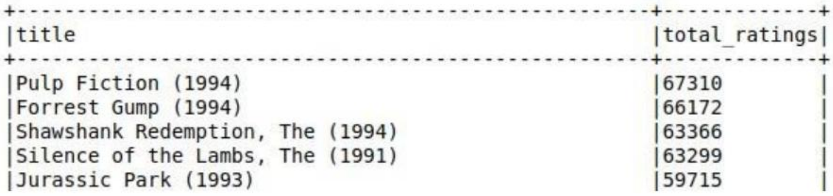
```
|Before the Fall (NaPolA - Elite für den Führer) (2004)                      |2005   |5.0              |1   |
|Dancemaker (1998)                                                           |2005   |5.0              |2   |
|Fear Strikes Out (1957)                                                     |2005   |5.0              |3   |
|Gate of Heavenly Peace, The (1995)                                          |2005   |5.0              |4   |
|Life Is Rosy (a.k.a. Life Is Beautiful) (Vie est belle, La) (1987)          |2005   |5.0              |5   |
|Married to It (1991)                                                        |2005   |5.0              |6   |
|My Life and Times With Antonin Artaud (En compagnie d'Antonin Artaud) (1993) |2005   |5.0              |7   |
|Not Love, Just Frenzy (Más que amor, frenesí) (1996)                        |2005   |5.0              |8   |
|Paris Was a Woman (1995)                                                    |2005   |5.0              |9   |
|Take Care of My Cat (Goyangileul butaghae) (2001)                           |2005   |5.0              |10  |
```

Give the
tags for each
movie and the
name of the
movie for the
year 2015.



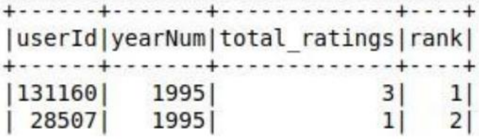
```
|""Great Performances"" Cats (1998)      |[BD-R]
|
|'burbs, The (1989)                      |[1980's, black comedy, dark comedy, Joe Dante, quirky]
|
|(500) Days of Summer (2009)             |[annoying, artistic, bad dialogue, boring, depressing, Joseph Gordon-Lev
itt, overrated, slow, stupid, Zooey Deschanel, intelligent, nonlinear, artistic, bittersweet, Funny, humor, humoro
us, intelligent, Joseph Gordon-Levitt, music, nonlinear, quirky, relationships, romance, Zooey Deschanel, bittersw
eet, quirky, romance, Joseph Gordon-Levitt, artistic, no happy ending, nonlinear, overrated]|
|...tick... tick... tick... (1970)       |[BD-R]
|
|1 (2014)                                |[Sukumar]
```

Give the number
of ratings for each
movie.



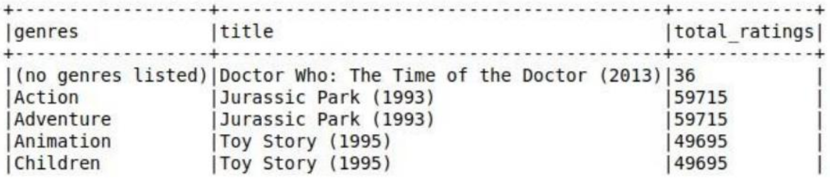
```
+------------------------------------------------------+-------------+
|title                                                 |total_ratings|
+------------------------------------------------------+-------------+
|Pulp Fiction (1994)                                   |67310        |
|Forrest Gump (1994)                                   |66172        |
|Shawshank Redemption, The (1994)                      |63366        |
|Silence of the Lambs, The (1991)                      |63299        |
|Jurassic Park (1993)                                  |59715        |
```

Find the first
10 users
with
the most
ratings for
each year. Find



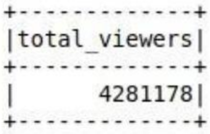
```
+------+-------+-------------+----+
|userId|yearNum|total_ratings|rank|
+------+-------+-------------+----+
|131160|   1995|            3|   1|
| 28507|   1995|            1|   2|
```

the movies with
the most ratings
for each movie
category.



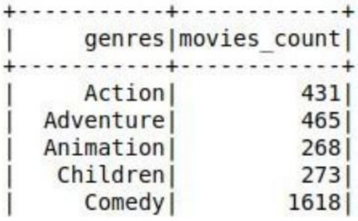
```
+------------------+----------------------------------------+-------------+
|genres            |title                                   |total_ratings|
+------------------+----------------------------------------+-------------+
|(no genres listed)|Doctor Who: The Time of the Doctor (2013)|36           |
|Action            |Jurassic Park (1993)                    |59715        |
|Adventure         |Jurassic Park (1993)                    |59715        |
|Animation         |Toy Story (1995)                        |49695        |
|Children          |Toy Story (1995)                        |49695        |
```

Give the total
number of users
watching the
same movie, on
the same day and
time. Give the



```
+-------------+
|total_viewers|
+-------------+
|      4281178|
+-------------+
```

number of movies,
for each category,
that users rated
as "funny" and
with a rating > 3.5.



```
+-----------+------------+
|     genres|movies_count|
+-----------+------------+
|     Action|         431|
|  Adventure|         465|
|  Animation|         268|
|   Children|         273|
|     Comedy|        1618|
```

Question 2: Performance comparison on single node/virtual cluster/Livy

Virtual cluster settings

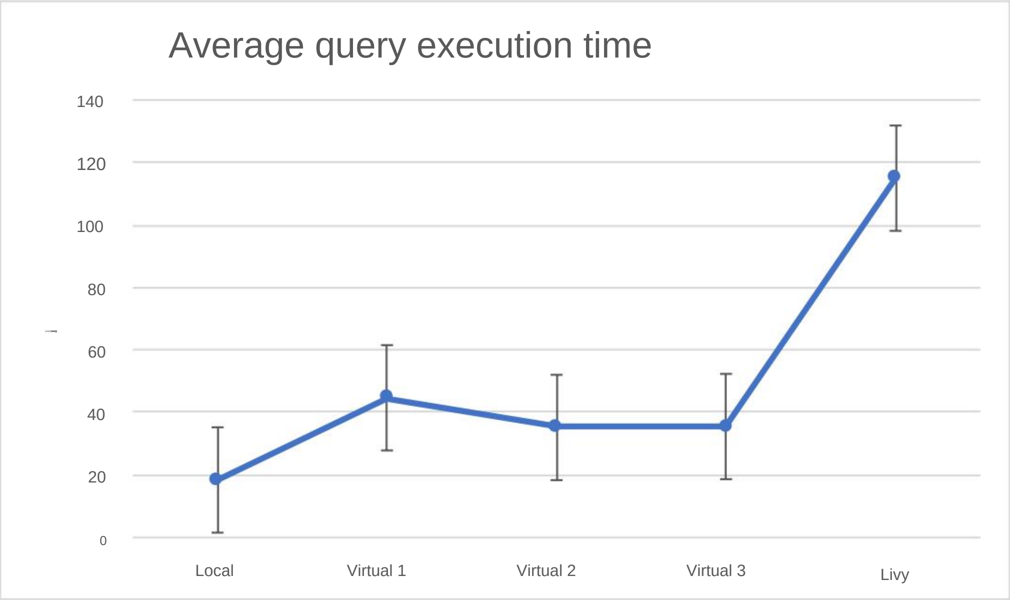
A/A 1	Executor cores Executor mem Driver cores			Driver mem
	1	1G	1	1G
2	2	2G	1	1G
3	2	2G	2	2G

Execution Times

Note: The execution times below were measured in seconds. Timing was done using the sparkMeasure library.

Question 1	Local	Virtual 1	Virtual 2	Virtual 3	Livy
	14	40	28	29	120
2	5	19	15	15	15
3	20	46	40	38	123
4	26	55	44	52	180
5	4	18	15	13	16
6	15	40	36	34	108
7	27	53	41	41	168
8	25	48	41	38	119
9	32	90	62	63	181
10	16	38	30	32	120

Analysis of results



After measuring the time and analyzing the results, we make the following observations:

- On a single node machine (local) we achieve fast query execution, as all the computing power we have assigned to the VM is used.
- Between Virtual 1, Virtual 2 and Virtual 3, we notice that the first consumes more time to execute queries than the other two and this is logically due to the assignment of only one core for each worker. We also notice that between Virtual 2 and 3 the differences are negligible, therefore the increase in cores and memory that the driver binds in Virtual 3 did not bring better results than 2.
- The execution of queries on the Livy server consumes the maximum time.
- Regarding the queries, we notice that specifically queries 2 and 5 always run in much less time compared to the rest. We suspect that this is due to the fact that these two queries do not make use of the rating.csv file, which contains the largest number of records of all and is therefore more "expensive" in terms of operations.

Bibliography

1. *PySpark 3.1.2 Documentation.* http://spark.apache.org/docs/latest/api/python/
2. A. Komnenos. *Tutorial 6 – Introduction to Apache Spark.* https://eclass.upatras.gr/modules/document/index.php?course=CEID1176&openDir =/5e6f65ear83d/60756472Ab51
3. Nishant Bahri. *Movie Lens Data Analysis Using PySpark [for beginners].* https://medium.com/analytics-vidhya/movie-lens-data-analysis-using-pyspark-for beginners-9c0f5f21eaf5
4. Mauro Krikorian. *Movie Data Statistics with Apache Spark.* https://medium.com/southworks/movie-data-statistics-with-apache-spark 58c2ef8fe452