# Introduction to Spark

Database Administration
Lab Guide 5

2022/2023

Consider the schema of the IMDB dataset, which is stored as multiple *tab-separated values* (TSV) files:

**title.basics:** tconst, titleType, primaryTitle, originalTitle, isAdult, startYear, endYear, runtimeMinutes, genres

**title.principals:** tconst, ordering, nconst, category, job, characters

**title.ratings:** tconst, averageRating, numVotes

**name.basics:** nconst, primaryName, birthYear, deathYear, primaryProfession, knownForTitles

Explore processing and exporting data in Spark.

**Steps**

1. Deploy Spark and execute the `main.py` job, by following the instructions in the appendix. Keep using the `main.py` file for the remaining steps.
2. Compute the number of titles per `titleType`.
3. Using the DataFrame function `write.parquet(`*`out_folder`*`)`[1], export the `titles` DataFrame to Parquet.[2]
4. Export the DataFrame again, using `gzip` compression, by providing the `compression='gzip'` parameter to `write.parquet`.
5. Export the DataFrame again, partitioned by the `startYear` column (`partitionBy='startYear'`).
6. Using the previous export and the function `spark.read.parquet(`*`in_folder`*`)`, obtain the number of titles released in 2022.

**Questions**

1. What is the size difference between TSV and Parquet? And Parquet using `snappy` (default) and `gzip` compressions?
2. What are the tradeoffs between `snappy` and `gzip` compressions?
3. What is the benefit of data partitioning?

**Learning Outcomes** Get familiarized with the Spark framework and with data storage formats.

[1] `https://spark.apache.org/docs/3.3.2/sql-data-sources-parquet.html`
[2] Make sure to export it to somewhere in `/app` so it becomes accessible in the local file system.

# Spark HowTo

**Download and extract the supplementary files.**

**Deploy the cluster:**

```
$ docker-compose -p spark up -d
```

**Execute the main.py job:**

```
$ docker exec spark_spark_1 python3 main.py
```

**To stop the cluster:**

```
$ docker-compose -p spark stop
```

**To delete the cluster:**

```
$ docker-compose -p spark down
```