Practice - Big Data Analytics and Reasoning

1 Song dataset

Given songs dataset reported in table 1, implements the following queries:

- Find the most popular song
- Find the total number of sold albums for *author1*
- For each album computes the mean number of streams, the number of tracks and the standard deviation of streams' count

2 University dataset

Given the following csv files representing the information abount students, exams and courses build a spark application that compute the following analysis:

- For each year, computes the count of exam attempts made by bachelor and master students
- Transform the `score` column into a column `result` that have two possible values `passed` or `failed`
- For each year, computes both the count of successful and failed exam attempts made by bachelor and master students

title	popularity	#streams	author	album	year	#sold_copies
song1	8	1000	author1	album1	2020	20
song2	7	2000	author1	album1	2020	20
song3	5	1500	author1	album1	2020	20
song4	6	5000	author1	album2	2021	40
song5	9	4000	author1	album2	2021	40
song6	10	3000	author1	album2	2021	40

Table 1: Song Dataset

3 MapReduce - WordCount

Read the content of txt files and count the occurrences of each word except for "*a*", "*an*" and "*the*".
Find the words that appears more than once and the occurrences of the word *spark*