# TDT4305 Big Data Project

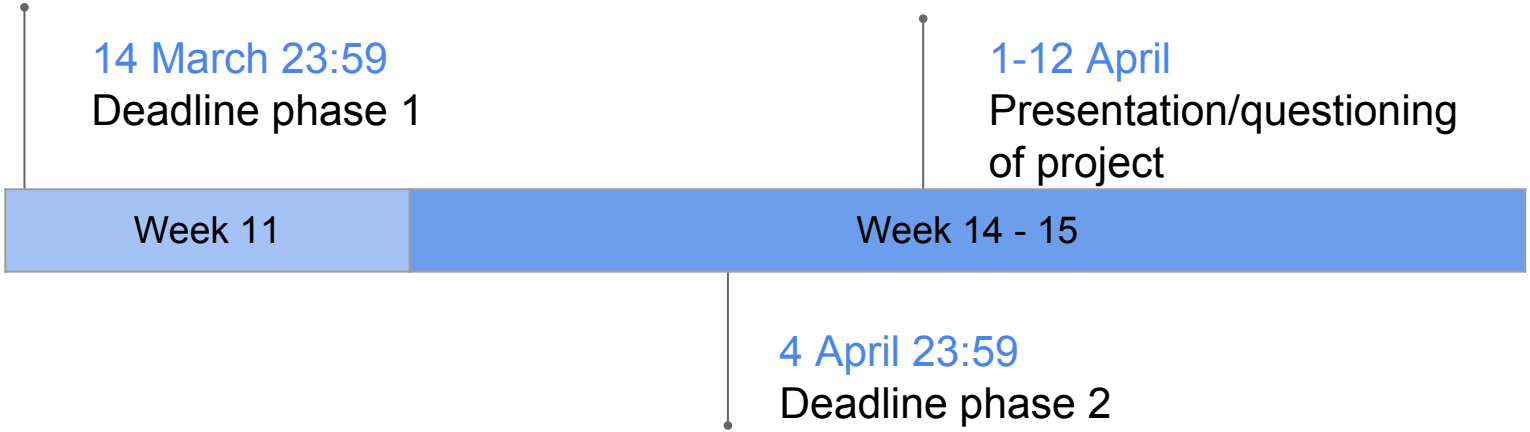
## Phase 1 Data Analysis with Spark

### Due Date: 14 March 23:59

**Project Assistance:**
Stella Maropaki - stella.maropaki@ntnu.no
Shweta Tiwari - shweta.tiwari@ntnu.no
Özer Özdikis - ozer.ozdikis@ntnu.no

# General information

- Consultation hours in room 248 IT-bygget:
  - Wednesday 10-12
  - Friday 10-12



# Objectives




# Music Data

- albums.csv

| id | artist_id | album_title | genre | year_of_pub | num_of_tracks | num_of_sales | rolling_stone_critic | mtv_critic | music_maniac_critic |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1767 | Call me Cat Moneyless That Doggies | Folk | 2006 | 11 | 905193 | 4 | 1.5 | 3 |
| 2 | 23548 | Down Mare | Metal | 2014 | 7 | 969122 | 3 | 4 | 5 |
| 3 | 17822 | Embarrassed Hungry | Latino | 2000 | 11 | 522095 | 2.5 | 1 | 2 |
| 4 | 19565 | Standard Immediate Engineer Slovakia | Pop | 2017 | 4 | 610116 | 1.5 | 2 | 4 |

- artists.csv

| id | real_name | art_name | role | year_of_birth | country | city | email | zip_code |
|---|---|---|---|---|---|---|---|---|
| 1 | Kameko Nelson | | female voice | 1981 | Slovenia | Vedrin | lorem.tristique.aliquet@nonummyFusce.co.uk | 6914 |
| 2 | Sacha Cameron | Bianco Ibureno Chloride | artist | 1955 | Guernsey | Montigny-le-Tilleul | Sed@elementum.com | 0568 |
| 3 | Thane E. Oliver | | rapper | 1993 | Guinea-Bissau | Saint-Sébastien-sur-Loire | Integer.tincidunt.aliquam@libero.com | 813154 |
| 4 | Cole E. Joseph | Persian responsible | bassist | 1994 | Estonia | Newquay | nulla.magna.malesuada@vulputate.com | A4S 2B3 |

# Notes

- Use Spark functions:
  - map(), reduceByKey(), sortByKey(), reduce(), count(), ...
  - no external libraries
- For efficiency:
  - use only data you need
- For faster testing:
  - rdd_sample = rdd.sample(false, 0.1, 5)
- Plagiarism:
  - Use private repositories (Git, BitBucket, ...)
  - Don’t copy

# Notes

Part of the evaluation of the project will be done by running the code files with automatic script.

Use the requested:

- Code file names
- Output formats
- Output file names

Make sure your code is running in terminal.

# Deliver on Blackboard

- Code files:
  - 9 tasks using RDD API
  - 1 task using Dataset API
- Report (pdf):
  - short description of code for each task
  - results for needed tasks
- Output files (.tsv):
  - for the tasks needed
  - in the described format
- Do not upload the datasets!!

**The following expected results are not the correct ones!**

**They only have the expected format!**

# Task 1

How many genres are there?

| id | artist_id | album_title | genre | year_of_pub | num_of_tracks | num_of_sales | rolling_stone_critic | mtv_critic | music_maniac_critic |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1767 | Call me Cat Moneyless That Doggies | Folk | 2006 | 11 | 905193 | 4 | 1.5 | 3 |
| 2 | 23548 | Down Mare | Metal | 2014 | 7 | 969122 | 3 | 4 | 5 |
| 3 | 17822 | Embarrassed Hungry | Latino | 2000 | 11 | 522095 | 2.5 | 1 | 2 |
| 4 | 19565 | Standard Immediate Engineer Slovakia | Pop | 2017 | 4 | 610116 | 1.5 | 2 | 4 |

Expected format of result: 25

# Task 2

What is the year of birth of the oldest artist?

| id | real_name | art_name | role | year_of_birth | country | city | email | zip_code |
|---|---|---|---|---|---|---|---|---|
| 1 | Kameko Nelson | | female voice | 1981 | Slovenia | Vedrin | lorem.tristique.aliquet@nonummyFusce.co.uk | 6914 |
| 2 | Sacha Cameron | Bianco Ibureno Chloride | artist | 1955 | Guernsey | Montigny-le-Tilleul | Sed@elementum.com | 0568 |
| 3 | Thane E. Oliver | | rapper | 1993 | Guinea-Bissau | Saint-Sébastien-sur-Loire | Integer.tincidunt.aliquam@libero.com | 813154 |
| 4 | Cole E. Joseph | Persian responsible | bassist | 1994 | Estonia | Newquay | nulla.magna.malesuada@vulputate.com | A4S 2B3 |

Expected format of results: 1955

# Task 3

Find the total number of artists coming from each country and sort them in descending order of artists counts.

| id | real_name | art_name | role | year_of_birth | country | city | email | zip_code |
|---|---|---|---|---|---|---|---|---|
| 1 | Kameko Nelson | | female voice | 1981 | Slovenia | Vedrin | lorem.tristique.aliquet@nonummyFusce.co.uk | 6914 |
| 2 | Sacha Cameron | Bianco Ibureno Chloride | artist | 1955 | Guernsey | Montigny-le-Tilleul | Sed@elementum.com | 0568 |
| 3 | Thane E. Oliver | | rapper | 1993 | Guinea-Bissau | Saint-Sébastien-sur-Loire | Integer.tincidunt.aliquam@libero.com | 813154 |
| 4 | Cole E. Joseph | Persian responsible | bassist | 1994 | Estonia | Newquay | nulla.magna.malesuada@vulputate.com | A4S 2B3 |

Expected format of results:

Botswana 156

China 156

Armenia 130

# Task 4

Find the total number of albums each artist has and sort them in descending order of album counts.

| id | artist_id | album_title | genre | year_of_pub | num_of_tracks | num_of_sales | rolling_stone_critic | mtv_critic | music_maniac_critic |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1767 | Call me Cat Moneyless That Doggies | Folk | 2006 | 11 | 905193 | 4 | 1.5 | 3 |
| 2 | 23548 | Down Mare | Metal | 2014 | 7 | 969122 | 3 | 4 | 5 |
| 3 | 17822 | Embarrassed Hungry | Latino | 2000 | 11 | 522095 | 2.5 | 1 | 2 |
| 4 | 19565 | Standard Immediate Engineer Slovakia | Pop | 2017 | 4 | 610116 | 1.5 | 2 | 4 |

Expected format of results:

10708 11

1608 9

47714 9

# Task 5

Find the total number of sales per genre and sort them in descending order of album sales.

| id | artist_id | album_title | genre | year_of_pub | num_of_tracks | num_of_sales | rolling_stone_critic | mtv_critic | music_maniac_critic |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1767 | Call me Cat Moneyless That Doggies | Folk | 2006 | 11 | 905193 | 4 | 1.5 | 3 |
| 2 | 23548 | Down Mare | Metal | 2014 | 7 | 969122 | 3 | 4 | 5 |
| 3 | 17822 | Embarrassed Hungry | Latino | 2000 | 11 | 522095 | 2.5 | 1 | 2 |
| 4 | 19565 | Standard Immediate Engineer Slovakia | Pop | 2017 | 4 | 610116 | 1.5 | 2 | 4 |

Expected format of results:

Indie 23848059

Pop 17166379

Rap 8889141

# Task 6

Find the top 10 albums with the best average critic.

$$\text{Average critic} = \frac{\text{rolling\_stone\_critic} + \text{mtv\_ctitic} + \text{music\_maniac\_critic}}{3}$$

| id | artist_id | album_title | genre | year_of_pub | num_of_tracks | num_of_sales | rolling_stone_critic | mtv_critic | music_maniac_critic |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1767 | Call me Cat Moneyless That Doggies | Folk | 2006 | 11 | 905193 | 4 | 1.5 | 3 |
| 2 | 23548 | Down Mare | Metal | 2014 | 7 | 969122 | 3 | 4 | 5 |
| 3 | 17822 | Embarrassed Hungry | Latino | 2000 | 11 | 522095 | 2.5 | 1 | 2 |
| 4 | 19565 | Standard Immediate Engineer Slovakia | Pop | 2017 | 4 | 610116 | 1.5 | 2 | 4 |

Expected format of results:

| | |
|---|---|
| 156 | 5.0 |
| 189 | 5.0 |
| 267 | 4.9 |

# Task 7

For the 10 albums of task 6, find the countries of their artists.

| id | artist_id | album_title | genre | year_of_pub | num_of_tracks | num_of_sales | rolling_stone_critic | mtv_critic | music_maniac_critic |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1767 | Call me Cat Moneyless That Doggies | Folk | 2006 | 11 | 905193 | 4 | 1.5 | 3 |
| 2 | 23548 | Down Mare | Metal | 2014 | 7 | 969122 | 3 | 4 | 5 |
| 3 | 17822 | Embarrassed Hungry | Latino | 2000 | 11 | 522095 | 2.5 | 1 | 2 |
| 4 | 19565 | Standard Immediate Engineer Slovakia | Pop | 2017 | 4 | 610116 | 1.5 | 2 | 4 |

| id | real_name | art_name | role | year_of_birth | country | city | email | zip_code |
|---|---|---|---|---|---|---|---|---|
| 1 | Kameko Nelson | | female voice | 1981 | Slovenia | Vedrin | lorem.tristique.aliquet@nonummyFusce.co.uk | 6914 |
| 2 | Sacha Cameron | Bianco Ibureno Chloride | artist | 1955 | Guernsey | Montigny-le-Tilleul | Sed@elementum.com | 0568 |
| 3 | Thane E. Oliver | | rapper | 1993 | Guinea-Bissau | Saint-Sébastien-sur-Loire | Integer.tincidunt.aliquam@libero.com | 813154 |
| 4 | Cole E. Joseph | Persian responsible | bassist | 1994 | Estonia | Newquay | nulla.magna.malesuada@vulputate.com | A4S 2B3 |

Expected format of results:

156 5.0 Slovenia

189 5.0 Guernsey

267 4.9 Estonia

# Task 8

Find the artists that have an album with the highest (5.0) MTV critic, sorted alphabetically.

| id | artist_id | album_title | genre | year_of_pub | num_of_tracks | num_of_sales | rolling_stone_critic | mtv_critic | music_maniac_critic |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1767 | Call me Cat Moneyless That Doggies | Folk | 2006 | 11 | 905193 | 4 | 1.5 | 3 |
| 2 | 23548 | Down Mare | Metal | 2014 | 7 | 969122 | 3 | 4 | 5 |
| 3 | 17822 | Embarrassed Hungry | Latino | 2000 | 11 | 522095 | 2.5 | 1 | 2 |
| 4 | 19565 | Standard Immediate Engineer Slovakia | Pop | 2017 | 4 | 610116 | 1.5 | 2 | 4 |

| id | real_name | art_name | role | year_of_birth | country | city | email | zip_code |
|---|---|---|---|---|---|---|---|---|
| 1 | Kameko Nelson | | female voice | 1981 | Slovenia | Vedrin | lorem.tristique.aliquet@nonummyFusce.co.uk | 6914 |
| 2 | Sacha Cameron | Bianco Ibureno Chloride | artist | 1955 | Guernsey | Montigny-le-Tilleul | Sed@elementum.com | 0568 |
| 3 | Thane E. Oliver | | rapper | 1993 | Guinea-Bissau | Saint-Sébastien-sur-Loire | Integer.tincidunt.aliquam@libero.com | 813154 |
| 4 | Cole E. Joseph | Persian responsible | bassist | 1994 | Estonia | Newquay | nulla.magna.malesuada@vulputate.com | A4S 2B3 |

Expected format of results:

Sacha Cameron

Thane E. Oliver

Cole E. Joseph

# Task 9

Find the average MTV critic of all albums for each artist from Norway (country='Norway').

| id | artist_id | album_title | genre | year_of_pub | num_of_tracks | num_of_sales | rolling_stone_critic | mtv_critic | music_maniac_critic |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1767 | Call me Cat Moneyless That Doggies | Folk | 2006 | 11 | 905193 | 4 | 1.5 | 3 |
| 2 | 23548 | Down Mare | Metal | 2014 | 7 | 969122 | 3 | 4 | 5 |
| 3 | 17822 | Embarrassed Hungry | Latino | 2000 | 11 | 522095 | 2.5 | 1 | 2 |
| 4 | 19565 | Standard Immediate Engineer Slovakia | Pop | 2017 | 4 | 610116 | 1.5 | 2 | 4 |

| id | real_name | art_name | role | year_of_birth | country | city | email | zip_code |
|---|---|---|---|---|---|---|---|---|
| 1 | Kameko Nelson | | female voice | 1981 | Slovenia | Vedrin | lorem.tristique.aliquet@nonummyFusce.co.uk | 6914 |
| 2 | Sacha Cameron | Bianco Ibureno Chloride | artist | 1955 | Guernsey | Montigny-le-Tilleul | Sed@elementum.com | 0568 |
| 3 | Thane E. Oliver | | rapper | 1993 | Guinea-Bissau | Saint-Sébastien-sur-Loire | Integer.tincidunt.aliquam@libero.com | 813154 |
| 4 | Cole E. Joseph | Persian responsible | bassist | 1994 | Estonia | Newquay | nulla.magna.malesuada@vulputate.com | A4S 2B3 |

Expected format of results:

Erin F. Evans Norway 5.0

Hayley Russell Norway 4.8

Ahmed Rosa Norway 3.2

# Task 10

Explore using Spark SQL and Dataset API.

| id | artist_id | album_title | genre | year_of_pub | num_of_tracks | num_of_sales | rolling_stone_critic | mtv_critic | music_maniac_critic |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1767 | Call me Cat Moneyless That Doggies | Folk | 2006 | 11 | 905193 | 4 | 1.5 | 3 |
| 2 | 23548 | Down Mare | Metal | 2014 | 7 | 969122 | 3 | 4 | 5 |
| 3 | 17822 | Embarrassed Hungry | Latino | 2000 | 11 | 522095 | 2.5 | 1 | 2 |
| 4 | 19565 | Standard Immediate Engineer Slovakia | Pop | 2017 | 4 | 610116 | 1.5 | 2 | 4 |

| id | real_name | art_name | role | year_of_birth | country | city | email | zip_code |
|---|---|---|---|---|---|---|---|---|
| 1 | Kameko Nelson | | female voice | 1981 | Slovenia | Vedrin | lorem.tristique.aliquet@nonummyFusce.co.uk | 6914 |
| 2 | Sacha Cameron | Bianco Ibureno Chloride | artist | 1955 | Guernsey | Montigny-le-Tilleul | Sed@elementum.com | 0568 |
| 3 | Thane E. Oliver | | rapper | 1993 | Guinea-Bissau | Saint-Sébastien-sur-Loire | Integer.tincidunt.aliquam@libero.com | 813154 |
| 4 | Cole E. Joseph | Persian responsible | bassist | 1994 | Estonia | Newquay | nulla.magna.malesuada@vulputate.com | A4S 2B3 |

Expected format of results:

450

1589

# QUESTIONS?