Middle East Technical University Department of Computer Engineering

CENG 495

Cloud Computing
Spring 2024–2025
HW - 3

Due date: 2025-05-26 23:59

1 Introduction

For this assignment, you will combine the MapReduce paradigm with the Distributed Task Queue Celery to get insights from the Spotify Songs dataset.

Use the Redis backend for your Celery workers. Your implementation should work for multiple workers. Stream your input data and process it across different workers.

2 Task

Download the dataset. Use the `json` version. With Chains, Groups and Chords implement the following:

- List the total duration (seconds) of each song in the dataset (**total**).
- List the average song duration (seconds) (**average**).
- Identify the 100 artists with the highest number of songs in the dataset (ignore features i.e. Band A, Band B), report their songs' average popularity. (**artist-popularity**)
- Separate the songs according to whether they are *explicit or not* (**yes, no**), then list the average popularity for the songs in those two separate categories (**explicit-popularity**).
- Partition the songs by year; first partition is the songs that came out on or before 2001 (**before-2001**), the second partition is for songs that came out between 2001 and 2012 (**2001-2012**) and final partition is the songs that came out later than 2012 (**after-2012**). Report the average danceability of these 3 partitions (**dancebyyear**).

Create a `results.json` file with the keys given above in **bold**. Use nested `json` objects to report metrics with multiple values:

```
{
  "total": ...,
  "explicit-popularity": {
    "yes": ...,
    "no": ...,
  },
  "artist-popularity": {
    "Band A": ...,
    "Band B": ...,
    ...
  },
  "dancebyyear": {
    "before-2001": ...,
    "2001-2012": ...
  }
}
```

3 Submission

- Use Python programming language using the Celery library.
- Archive your implementation as a `.tar.gz` file named "USERNAME.tar.gz" (e.g. e248205.tar.gz)
- This is an individual assignment. You can discuss your ideas with your peers but using implementation specific code that is not your own is strictly forbidden and constitutes as cheating. The violators will get no grade from this assignment and will be punished according to the department regulations.