This project provides a Python implementation of the MapReduce programming model to analyze a large dataset of songs. Specifically, it computes the maximum song duration for each artist from a dataset assumed to be a CSV format. The script can be run using multiple processes for mapping and reducing to handle large amounts of data efficiently.

Dataset

The dataset should be in CSV format with at least the following columns:

1. Song Title
2. Artist's Name
3. Duration (in seconds)

Methodology

Split Data

The data is either read directly from a file or piped from standard input, then split into smaller chunks. Each chunk is processed by a separate map process.

Map Function

Each map process reads a chunk of the data, parsing each row to extract the artist's name and song duration. It outputs key-value pairs where the key is the artist's name and the value is the duration of a song.

Shuffle Function

The shuffle function collects all key-value pairs from the map processes and distributes them to reduce processes. The distribution ensures that all values associated with the same key (artist's name) are sent to the same reducer.

Reduce Function

Each reduce process receives key-value pairs for certain artists. It computes the maximum song duration for each artist and outputs the result.

Output

The final output is printed to the console, showing the artist's name and their maximum song duration.