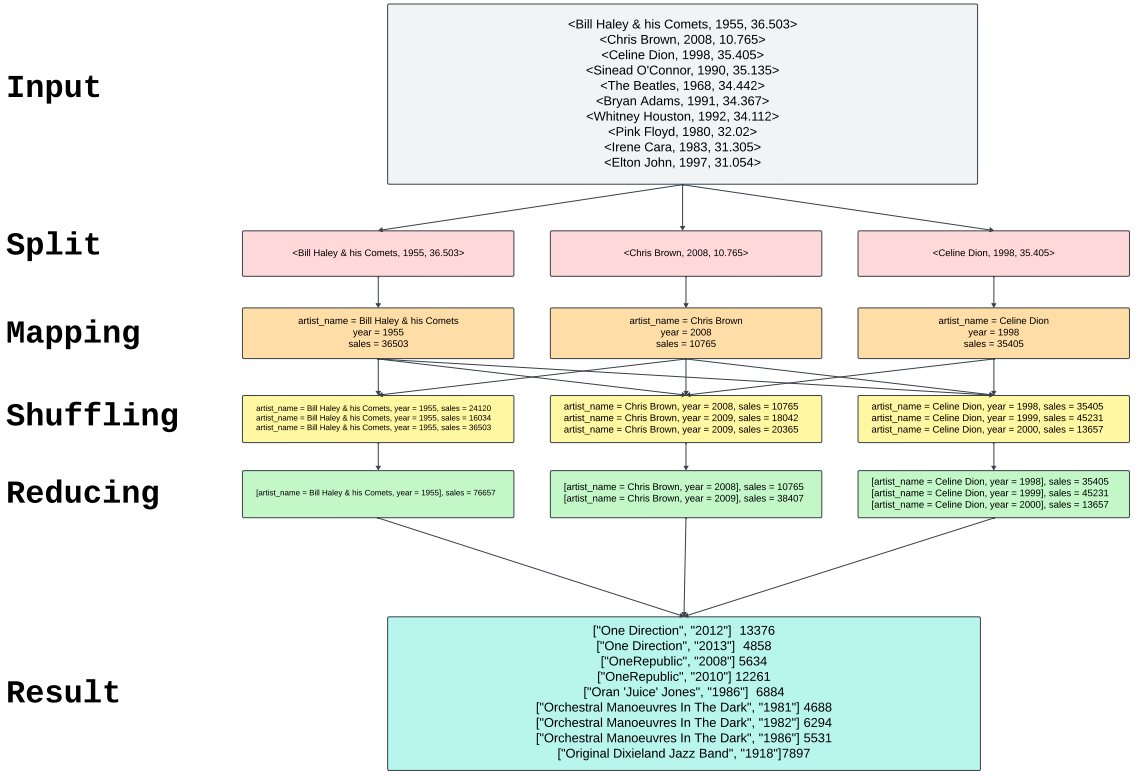
Macquarie University, School of Computing
COMP6210 Big Data

Assignment 1: MapReduce

Name: Do Nam Phong, Phung
ID: 47828013

Flowchart and pseudocode documentation for MapReduce program
All of the variables and shown values are only for presentation, exact values may vary

Task 1.2 flowchart & pseudocode

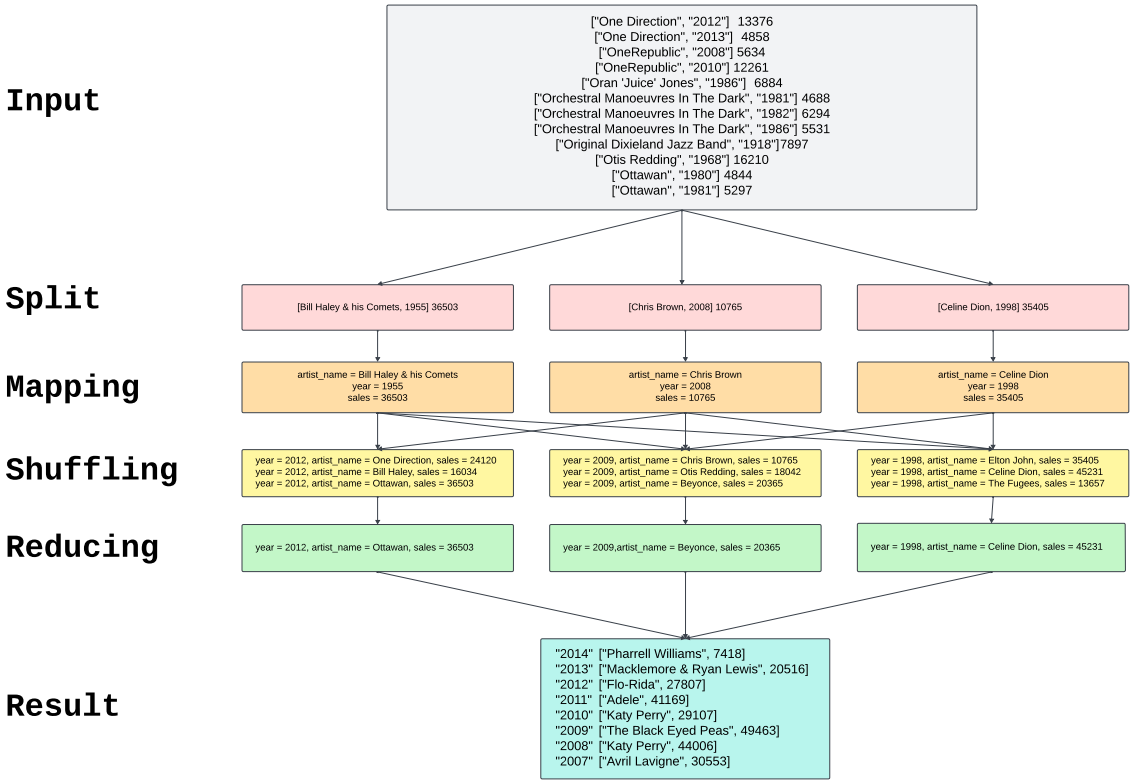


pseudocode

Data: Text file include artist name, year, sales number
Result: Text file include **[artist name, year], sales number** of each year
MAPPER(line)
    IF details(artist_name, year, sales) match **line** re pattern
        YIELD (artist_name, year), sales
    END IF

REDUCER(artist_year, sales)
    Sum sales by artist_name and year
    YIELD (artist_year, sales)
//

Task 2.1 flow chart



pseudocode

Task 2_1

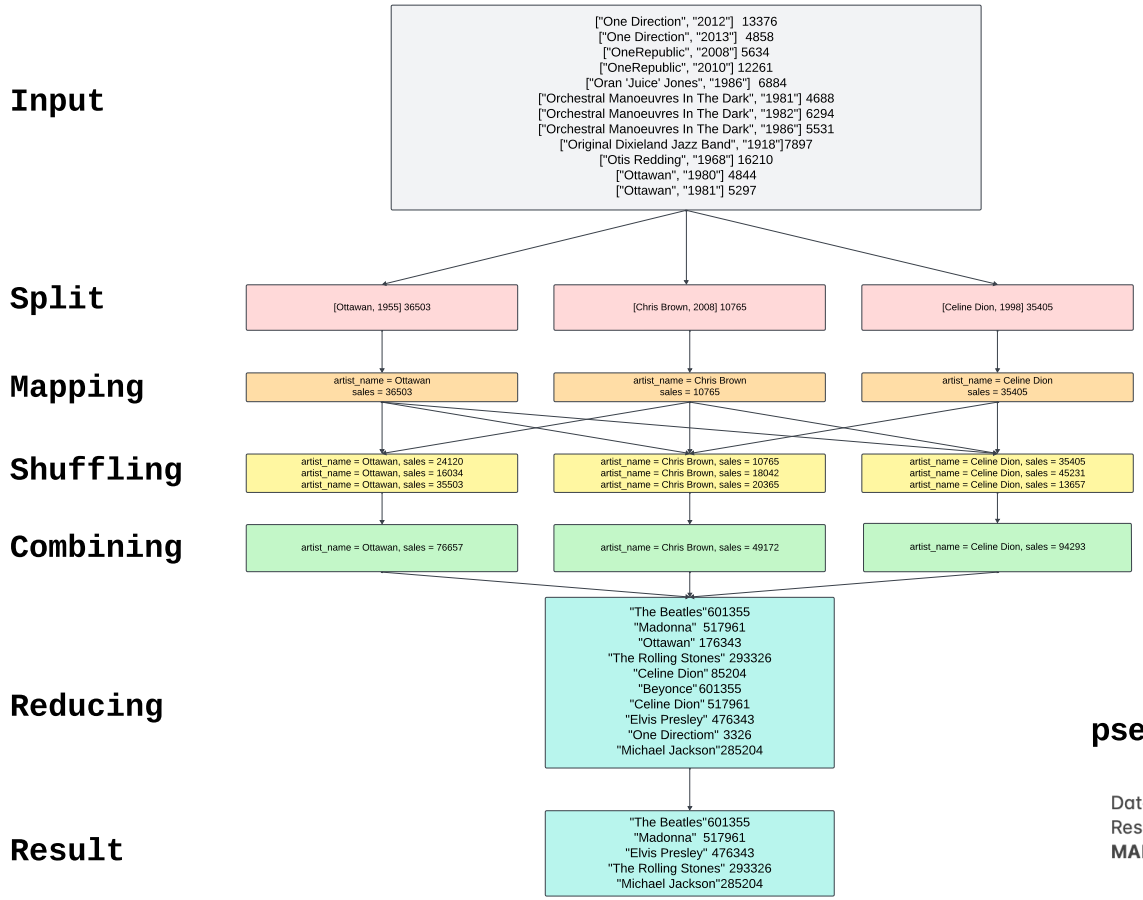
Data: Text file include artist name, year, sales number
Result: Text file include **year, [artist name, sales number]** of artist with highest sales each year

```
MAPPER(line)
    IF details(artist_name, year, sales) match line re pattern
        YIELD year, (artist_name, sales)
    END IF

REDUCER(year, artist_sales)
    OBTAIN max sales each year
    YIELD  None, (year, max_sales)

REDUCER(year_max_sales)
    DISPLAY year_max_sales (year, max_sales) by year in descending order as `sorted_year`
    FOR year, max_sales in each `sorted_year`
        YIELD year, max_sales
    END FOR
//
```

Task 2.2 flow chart



pseudocode

—

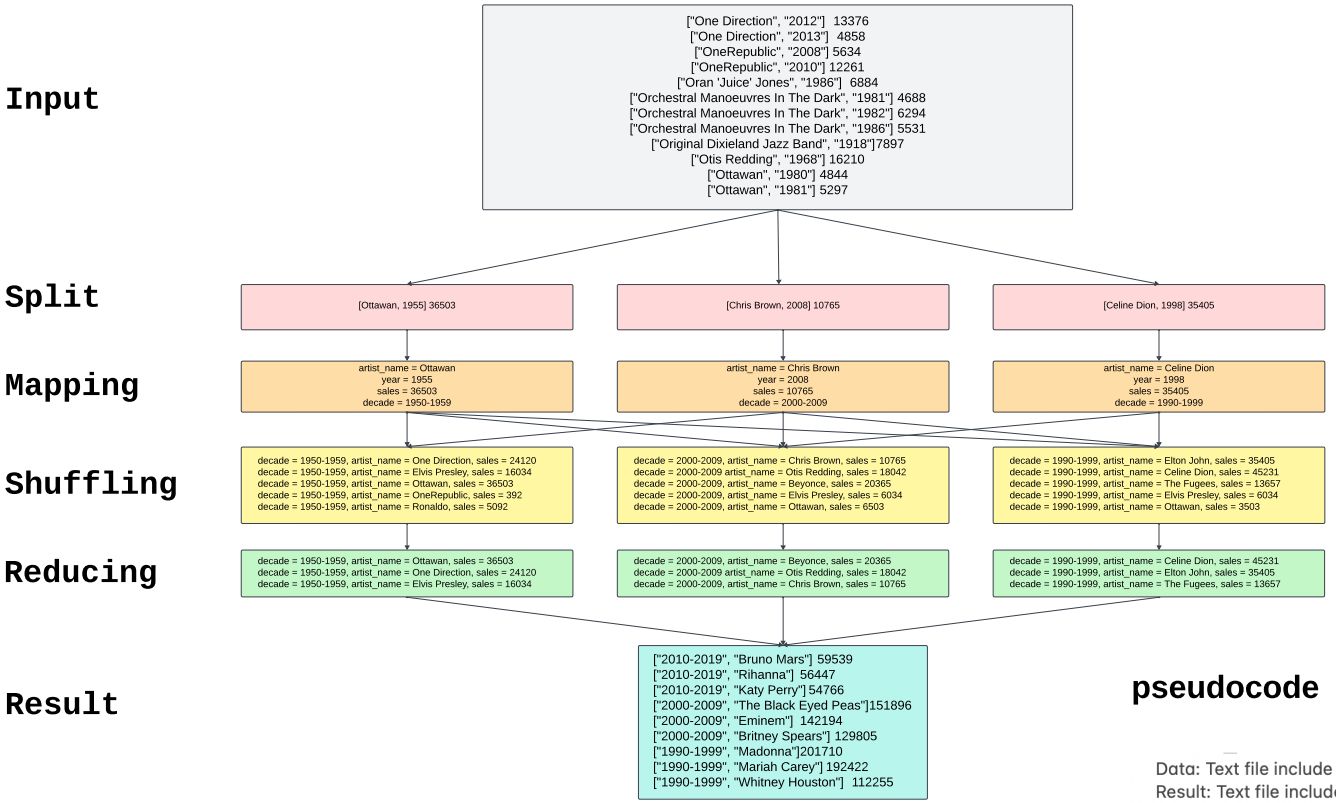
Data: Text file include artist name, year, sales number
Result: Text file include **artist name, sales number** of top 5 highest sales of all time
MAPPER(line)
    IF details(artist_name, sales) match line re pattern
        YIELD (artist_name, sales)
    END IF

COMBINER(artist_name, sales)
    COMPUTE sum of sales by artist_name in each local group
    YIELD (artist_name, sum(sales))

REDUCER(artist_name, sales)
    COMPUTE sum of sales by artist_name globally
    YIELD None, (artist_name, sum(sales))

REDUCER(artist_all_sales)
    DISPLAY artist_all_sales (artist_name, sales) by 5 highest sales in descending order as `top_5_sales`
    FOR artist_name, sales in each `top_5_sales`
        YIELD artist_name, sales
    END FOR
//

Task 2.3 flow chart



pseudocode

Data: Text file include artist name, year, sales number
Result: Text file include **[decade, artist name], sales number** of top 3 highest sales of each decade

MAPPER(line)
    IF details(artist_name, sales) match line re pattern
        GET decade from year
        YIELD (decade, artist_name), sales
    END IF

COMBINER(decade_artist, sales)
    COMPUTE sum of sales by decade_artist(decade, artist_name) in each local group
    YIELD (decade_artist, sum(sales))

REDUCER(decade_artist, sales)
    COMPUTE sum of sales by decade_artist globally
    SET decade_artist as decade, artist_name
    YIELD decade, (artist_name, sum(sales))

REDUCER(decade, artist_total_sales)
    DISPLAY artist_total_sales(artist_name, sales) by 3 highest sales in descending order as `top_three_sales`
    FOR artist_name, sales in each `top_three_sales`
        YIELD (decade,artist_name), sales
    END FOR
//