

Assignment 1: MapReduce

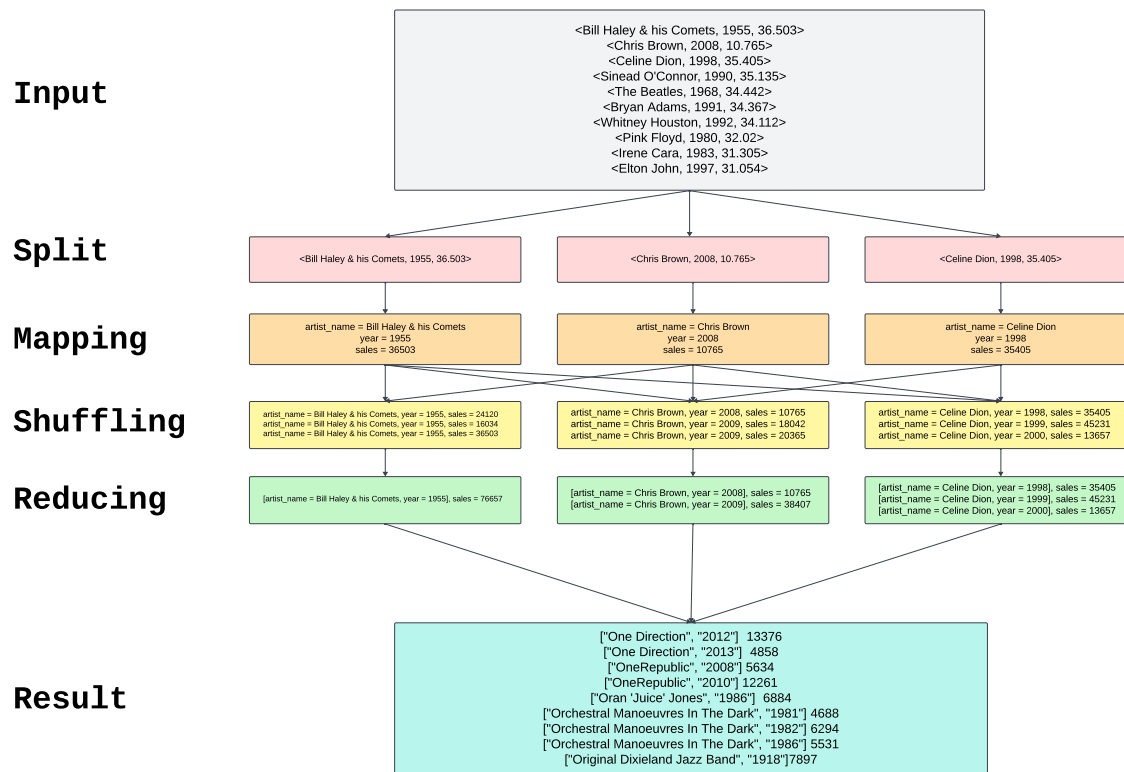
Name: Do Nam Phong, Phung

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)

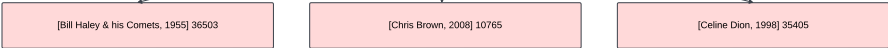
//

Task 2.1 flow chart

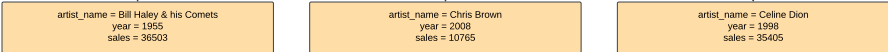
Input



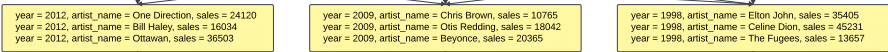
Split



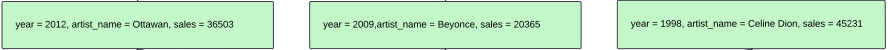
Mapping



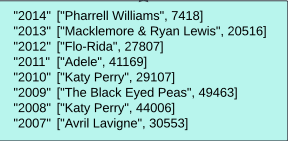
Shuffling



Reducing



Result



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
```

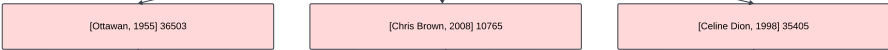
//

Task 2.2 flow chart

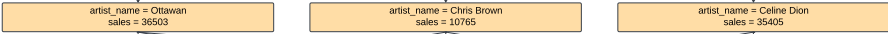
Input



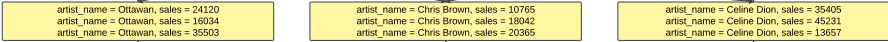
Split



Mapping



Shuffling



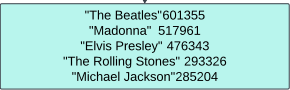
Combining



Reducing



Result



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

Input

```
[ "One Direction", "2012" ] 13376
[ "One Direction", "2013" ] 4858
[ "OneRepublic", "2008" ] 5634
[ "OneRepublic", "2010" ] 12261
[ "Oran 'Juice' Jones", "1986" ] 6884
[ "Orchestral Manoeuvres In The Dark", "1981" ] 4688
[ "Orchestral Manoeuvres In The Dark", "1982" ] 6294
[ "Orchestral Manoeuvres In The Dark", "1986" ] 5531
[ "Original Dixieland Jazz Band", "1918" ] 7897
[ "Otis Redding", "1968" ] 16210
[ "Ottawan", "1980" ] 4844
[ "Ottawan", "1981" ] 5297
```

Split

```
[Ottawan, 1955] 36503
[Chris Brown, 2008] 10765
[Celine Dion, 1998] 35405
```

Mapping

```
artist_name = Ottawan
year = 1955
sales = 36503
decade = 1950-1959

artist_name = Chris Brown
year = 2008
sales = 10765
decade = 2000-2009

artist_name = Celine Dion
year = 1998
sales = 35405
decade = 1990-1999
```

Shuffling

```
decade = 1950-1959, artist_name = One Direction, sales = 24120
decade = 1950-1959, artist_name = Elvis Presley, sales = 16034
decade = 1950-1959, artist_name = Ottawan, sales = 36503
decade = 1950-1959, artist_name = OneRepublic, sales = 392
decade = 1950-1959, artist_name = Ronaldo, sales = 5092

decade = 2000-2009, artist_name = Chris Brown, sales = 10765
decade = 2000-2009, artist_name = Otis Redding, sales = 18042
decade = 2000-2009, artist_name = Beyonce, sales = 20365
decade = 2000-2009, artist_name = Elvis Presley, sales = 6034
decade = 2000-2009, artist_name = Ottawan, sales = 6503

decade = 1990-1999, artist_name = Elton John, sales = 35405
decade = 1990-1999, artist_name = Celine Dion, sales = 45231
decade = 1990-1999, artist_name = The Fugees, sales = 13657
decade = 1990-1999, artist_name = Elvis Presley, sales = 6034
decade = 1990-1999, artist_name = Ottawan, sales = 3503
```

Reducing

```
decade = 1950-1959, artist_name = Ottawan, sales = 36503
decade = 1950-1959, artist_name = One Direction, sales = 24120
decade = 1950-1959, artist_name = Elvis Presley, sales = 16034

decade = 2000-2009, artist_name = Beyonce, sales = 20365
decade = 2000-2009, artist_name = Otis Redding, sales = 18042
decade = 2000-2009, artist_name = Chris Brown, sales = 10765

decade = 1990-1999, artist_name = Celine Dion, sales = 45231
decade = 1990-1999, artist_name = Elton John, sales = 35405
decade = 1990-1999, artist_name = The Fugees, sales = 13657
```

Result

```
[ "2010-2019", "Bruno Mars" ] 59539
[ "2010-2019", "Rihanna" ] 56447
[ "2010-2019", "Katy Perry" ] 54766
[ "2000-2009", "The Black Eyed Peas" ] 151896
[ "2000-2009", "Eminem" ] 142194
[ "2000-2009", "Britney Spears" ] 129805
[ "1990-1999", "Madonna" ] 201710
[ "1990-1999", "Mariah Carey" ] 192422
[ "1990-1999", "Whitney Houston" ] 112255
```

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

//