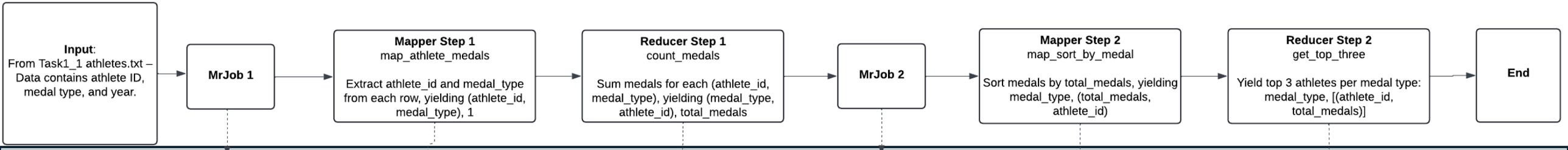
COMP6210 - Big Data Assignment 1
MapReduce
Semester 2, 2024

Task 3: MapReduce Flowcharts

Hassan Mohammad

47892714

Task 2_1: Count Medals and Identify Top Athletes by Medal Type



Goal:
Aggregate the total number of medals for each athlete, categorized by medal type.

- **Input**: Each line of the input data contains records formatted as "athlete_id | country | year | event_name | medal_type".
- **Processing**:
- **Extract Fields**: Split the line by " | " to isolate the athlete_id, medal_type, and other fields.
- **Filter Medals**: Ensure the medal_type is one of Gold, Silver, or Bronze.
- **Yield Key-Value Pair**: Emit a key-value pair where the key is a tuple (athlete_id, medal_type) and the value is 1 to indicate a single medal occurrence.
- **Output**: Key-value pairs where the key is (athlete_id, medal_type) and the value is 1.

- **Input**: The reducer receives keys of the form (athlete_id, medal_type) with associated lists of 1s representing individual medal occurrences.
- **Processing**:
  - **Aggregate Counts**: Sum the values associated with each key to compute the total number of medals for each athlete_id and medal_type.
  - **Yield**: Emit key-value pairs where the key is (medal_type, athlete_id) and the value is the total count of medals.
- **Output**: Key-value pairs where the key is (medal_type, athlete_id) and the value is the total number of medals.

Goal: Identify and rank the top three athletes for each type of medal based on the total number of medals they have won.

- **Input**: The mapper function receives keys of the form (medal_type, athlete_id) with values representing the total number of medals.
- **Processing**:
- **Yield Key-Value Pair**: For each key (medal_type, athlete_id) and value total_medals, emit a key-value pair where the key is medal_type and the value is a tuple (total_medals, athlete_id).
- **Output**: Key-value pairs where the key is medal_type and the value is a tuple (total_medals, athlete_id).

- **Input**: The reducer receives keys of the form medal_type with associated lists of tuples (total_medals, athlete_id).
- **Processing**:
- **Sort Data**: Sort the tuples within each medal_type group by total_medals in descending order.
- **Select Top 3**: Extract the top three tuples based on the total_medals.
- **Yield**: Emit key-value pairs where the key is medal_type and the value is a list containing [athlete_id, total_medals] for the top three athletes.
- **Output**: Key-value pairs where the key is medal_type and the value is a list of the top three athletes with their total medal counts.

Athlete ID	Country	Year	Event	Medal Type
121605	AZE	2012	Heavyweight, Men	Bronze
121606	AZE	2012	Super-Heavyweight, Men	Bronze
65090	BAR	2000	100 metres, Men	Bronze
117135	AZE	2012	Featherweight, Greco-Roman, Men	Silver
126970	AZE	2012	Middleweight, Greco-Roman, Men	Bronze

Athlete ID	Medal Type	Output Key-Value Pair
121605	Bronze	((121605, Bronze), 1)
121606	Bronze	((121606, Bronze), 1)
65090	Bronze	((65090, Bronze), 1)
117135	Silver	((117135, Silver), 1)
126970	Bronze	((126970, Bronze), 1)

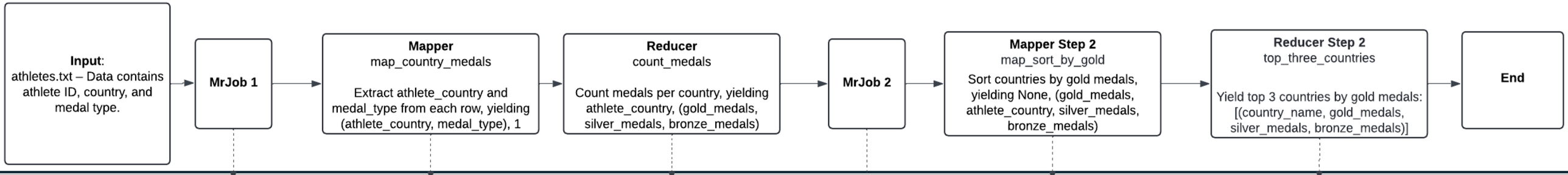
Medal Type	Athlete ID	Total Medals	Output Key-Value Pair
Bronze	121605	1	((Bronze, 121605), 1)
Bronze	121606	1	((Bronze, 121606), 1)
Bronze	65090	1	((Bronze, 65090), 1)
Silver	117135	1	((Silver, 117135), 1)
Bronze	126970	1	((Bronze, 126970), 1)

Medal Type	Total Medals	Athlete ID	Output Key-Value Pair
Bronze	1	121605	(Bronze, (1, 121605))
Bronze	1	121606	(Bronze, (1, 121606))
Bronze	1	65090	(Bronze, (1, 65090))
Silver	1	117135	(Silver, (1, 117135))
Bronze	1	126970	(Bronze, (1, 126970))

Medal Type	Sorted Values	Output Key-Value Pair
Bronze	[(1, 121605), (1, 121606), (1, 65090), (1, 126970)]	(Bronze, [(121605, 1), (121606, 1), (65090, 1)])
Silver	[(1, 117135)]	(Silver, [(117135, 1)])

Medal Type	Athlete ID	Total Medals
Bronze	121605	1
Bronze	121606	1
Bronze	65090	1
Silver	117135	1

Task 2_2: Top 3 Countries by Gold Medals



Goal: Compute the total number of gold, silver, and bronze medals for each country.

Input: Each line of the input data contains records formatted as "athlete_id | country | other_info | medal_type".
Processing:
Split Data: The line is split by the delimiter " | " to extract the relevant fields.
Extract Fields: Extract the country and the medal type from the split data.
Yield Key-Value Pair: Emit a key-value pair where the key is the country and the value is a tuple containing the medal type and a count of 1.

- **Input**: The reducer receives a key (country) and a list of tuples representing medal types and counts.
- **Processing**:
- **Initialize Counters**: Start counters for gold, silver, and bronze medals.
- **Aggregate Counts**: Iterate over the list of tuples to aggregate the counts of each type of medal.
- **Yield**: Emit a key-value pair where the key is the country and the value is a tuple with the total counts of gold, silver, and bronze medals.

Goal: Identify the top three countries with the highest number of gold medals.

- **Input**: Each record from the first reducer is formatted as ('country', (gold_medals, silver_medals, bronze_medals)).
- **Processing**:
- **Extract Medal Counts**: Unpack the medal counts for gold, silver, and bronze.
- **Yield for Sorting**: Emit a key-value pair where the key is None and the value is a tuple containing gold medals, country, silver medals, and bronze medals.

- **Input**: The reducer receives a list of tuples where each tuple contains gold medals, country, silver medals, and bronze medals.
- **Processing**:
- **Sort Records**: Sort the list of tuples in descending order based on the number of gold medals.
- **Select Top 3**: Extract the top three records from the sorted list.
- **Format Output**: For each of the top three countries, create a dictionary with gold, silver, and bronze medal counts, and emit this information.

Athlete ID	Country	Year	Event	Medal Type
121605	USA	2012	Heavyweight, Men	Gold
121606	USA	2012	Super-Heavyweight, Men	Silver
65090	CHN	2000	100 metres, Men	Bronze
117135	GER	2012	Featherweight, Greco-Roman, Men	Gold
126970	GER	2012	Middleweight, Greco-Roman, Men	Silver

Country	Medal Type	Output Key-Value Pair
USA	Gold	(USA, (Gold, 1))
USA	Silver	(USA, (Silver, 1))
CHN	Bronze	(CHN, (Bronze, 1))
GER	Gold	(GER, (Gold, 1))
GER	Silver	(GER, (Silver, 1))

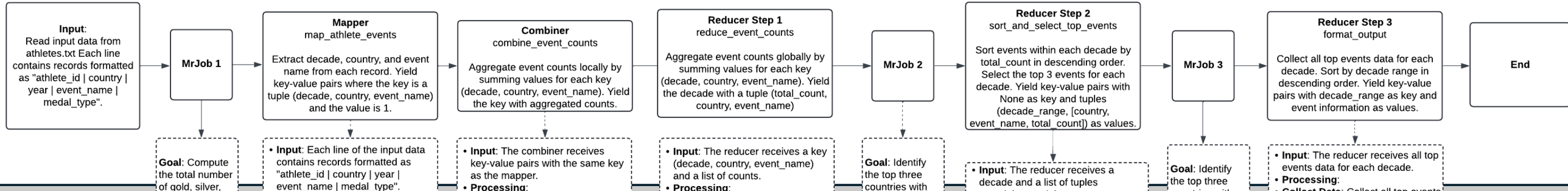
Country	Gold Medals	Silver Medals	Bronze Medals	Output Key-Value Pair
USA	1	1	0	(USA, (1, 1, 0))
CHN	0	0	1	(CHN, (0, 0, 1))
GER	1	1	0	(GER, (1, 1, 0))

Gold Medals	Country	Silver Medals	Bronze Medals	Output Key-Value Pair
1	USA	1	0	(None, (1, USA, 1, 0))
0	CHN	0	1	(None, (0, CHN, 0, 1))
1	GER	1	0	(None, (1, GER, 1, 0))

Sorted Values	Output Key-Value Pair
[(1, USA, 1, 0), (1, GER, 1, 0), (0, CHN, 0, 1)]	(USA, {"Gold": 1, "Silver": 1, "Bronze": 0})
	(GER, {"Gold": 1, "Silver": 1, "Bronze": 0})
	(CHN, {"Gold": 0, "Silver": 0, "Bronze": 1})

Country	Gold Medals	Silver Medals	Bronze Medals
USA	1	1	0
GER	1	1	0
CHN	0	0	1

Task 2_3: Top 3 Countries by Gold Medals



Athlete ID	Country	Year	Event Name	Medal Type
121605	USA	2020	Baseball, Men	Gold
121606	JPN	2020	Baseball, Men	Gold
65090	DOM	2020	Baseball, Men	Gold
117135	BRA	2010	Football, Men	Gold
126970	CAN	2010	Football, Women	Silver

Decade	Country	Event Name	Output Key-Value Pair
2020	USA	Baseball, Men	((2020, USA, Baseball, Men), 1)
2020	JPN	Baseball, Men	((2020, JPN, Baseball, Men), 1)
2020	DOM	Baseball, Men	((2020, DOM, Baseball, Men), 1)
2010	BRA	Football, Men	((2010, BRA, Football, Men), 1)
2010	CAN	Football, Women	((2010, CAN, Football, Women), 1)

Decade	Country	Event Name	Aggregated Count	Output Key-Value Pair
2020	USA	Baseball, Men	1	((2020, USA, Baseball, Men), 1)
2020	JPN	Baseball, Men	1	((2020, JPN, Baseball, Men), 1)
2020	DOM	Baseball, Men	1	((2020, DOM, Baseball, Men), 1)
2010	BRA	Football, Men	1	((2010, BRA, Football, Men), 1)
2010	CAN	Football, Women	1	((2010, CAN, Football, Women), 1)

Decade	Total Count	Country	Event Name	Output Key-Value Pair
2020	1	USA	Baseball, Men	(2020, (1, USA, Baseball, Men))
2020	1	JPN	Baseball, Men	(2020, (1, JPN, Baseball, Men))
2020	1	DOM	Baseball, Men	(2020, (1, DOM, Baseball, Men))
2010	1	BRA	Football, Men	(2010, (1, BRA, Football, Men))
2010	1	CAN	Football, Women	(2010, (1, CAN, Football, Women))

Decade Range	Sorted Values	Output Key-Value Pair
2020-2029	[(1, USA, Baseball, Men), (1, JPN, Baseball, Men), (1, DOM, Baseball, Men)]	(None, (2020-2029, [USA, Baseball, Men, 1]))
		(None, (2020-2029, [JPN, Baseball, Men, 1]))
		(None, (2020-2029, [DOM, Baseball, Men, 1]))
2010-2019	[(1, BRA, Football, Men), (1, CAN, Football, Women)]	(None, (2010-2019, [BRA, Football, Men, 1]))
		(None, (2010-2019, [CAN, Football, Women, 1]))

Decade Range	Event Information	Output Key-Value Pair
2020-2029	[USA, Baseball, Men, 1]	(2020-2029, [USA, Baseball, Men, 1])
	[JPN, Baseball, Men, 1]	(2020-2029, [JPN, Baseball, Men, 1])
	[DOM, Baseball, Men, 1]	(2020-2029, [DOM, Baseball, Men, 1])
2010-2019	[BRA, Football, Men, 1]	(2010-2019, [BRA, Football, Men, 1])
	[CAN, Football, Women, 1]	(2010-2019, [CAN, Football, Women, 1])

Decade Range	Country	Event Name	Medal Count
2020-2029	USA	Baseball, Men	1
2020-2029	JPN	Baseball, Men	1
2020-2029	DOM	Baseball, Men	1
2010-2019	BRA	Football, Men	1
2010-2019	CAN	Football, Women	1