Big Data Systems - Assigment 1

Program Title: M Tech Software Engineering
Course No.: SE ZG522
Course Title: Big Data Systems
Course Author: Ashish Narang
Team Members (Group 22):

1. Jayanta H R, 2021MT93120

Introduction

Cricket is one of the most followed sport in the country. With the invent of the IPL T20, a lot of oportunity was opened for data science and analysis. For Big Data System assignment, I have choosen a dataset from the IPL T20 2022 editon and would be performing certain analysis on them to get data which matches the problem statement.

The source code for all the problem statements can be found here

Dataset for this assignment is obtained from https://www.kaggle.com/datasets/vora1011/ipl-2022-match-dataset?resource=download

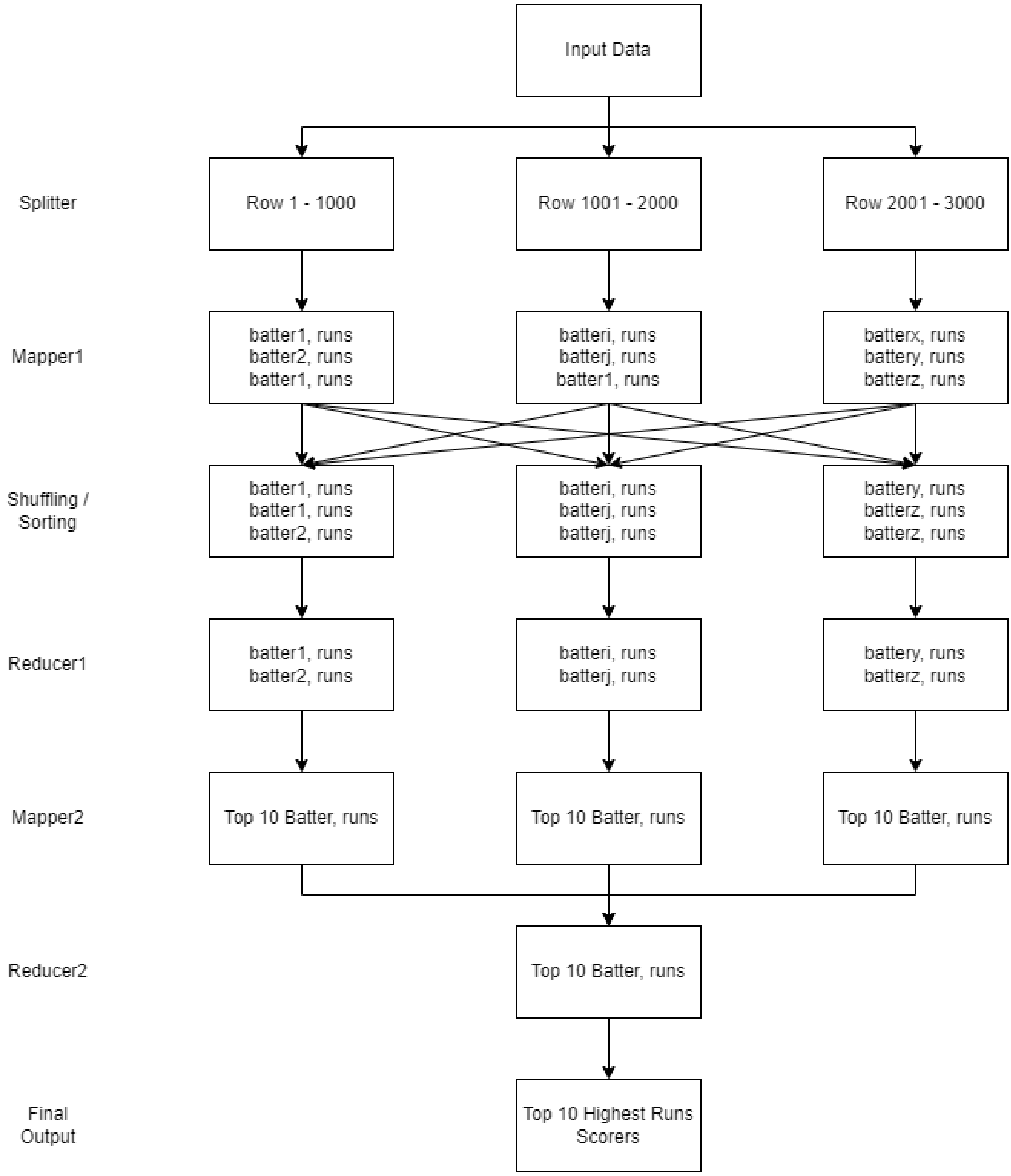
Problem Statements

1. Top 10 run scorers in the tournament
2. Top 10 wicket takers in the tournament
3. Centuries
4. Extras bowled in a match by a bowler
5. Catches taken

1. Top 10 run scorers in the tournament

In the IPLT20 2022, analyze and extract the top 10 run scorers. The output data should contain the batsman names and the total runs scored. Only 10 records should be displayed and should be arranged in the descending order of the runs

Map-Reduce Diagram



Pseudocode

Mapper 1

```
Input: Dataset.csv
For each row
        Extract the Batsman and Runs
Output: Batsman         Runs
```

Reducer 1

```
Input: Output of Mapper1
For each row
        Add the Batsman's score in a match
        return the batsman name and runs
Output: Batsman         runsScored in a match
```

Mapper 2

```
Input: Output of Reducer1
For each row
        Sort the batsman in descending order as per the runs scored
        Extract top 10 run scorers
Output: Batsman         runs scored
```

Reducer 2

```
Input: Output of Mapper2
For each row
        Sort the batsman in descending order as per the runs scored
        Extract top 10 run scorers
Output: Batsman         runs scored
```

Source Code

mapper1.py

```
import sys

for line in sys.stdin:
    line = line.strip()

matchId,innings,overs,ballnumber,batsman,bowler,nonStriker,extra_type,batsman_runs
```

```
,extras_run,total_run,non_boundary,isWicketDelivery,player_out,kind,fielders_invol
ved,BattingTeam = line.split(',')
    print('{}\t{}'.format(batsman, batsman_runs))
```

reducer1.py

```
import sys

current_batsman = None
current_batsman_runs = 0

for line in sys.stdin:
    line = line.strip()
    batsman, run = line.split('\t')

    try:
        run = int(run)
    except ValueError:
        continue

    if current_batsman == batsman:
        current_batsman_runs += run
    else:
        if current_batsman:
            print('{}\t{}'.format(current_batsman, current_batsman_runs))
        current_batsman = batsman
        current_batsman_runs = run

if current_batsman == batsman:
    print('{}\t{}'.format(current_batsman, current_batsman_runs))
```

mapper2.py

```
import sys

scores=[]
for line in sys.stdin:
    line = line.strip()
    batsman,batsman_runs = line.split('\t')
    try:
        batsman_runs = int(batsman_runs)
    except ValueError:
        continue
    scores.append([batsman_runs, batsman])

top_N=sorted(scores,reverse=True)[0:10]
```

```
for t in top_N:
    print('{}\t{}'.format(t[1], t[0]))
```

reducer2.py

```
import sys

scores=[]
for line in sys.stdin:
    line = line.strip()
    batsman,batsman_runs = line.split('\t')
    try:
        batsman_runs = int(batsman_runs)
        batsman = batsman.strip()
    except ValueError:
        continue
    scores.append([batsman_runs, batsman])

top_N=sorted(scores,reverse=True)[0:10]

for t in top_N:
    print('{}\t{}'.format(t[1], t[0]))
```

Statistics

Part 1

```
2022-10-25 20:14:45,968 INFO mapreduce.Job: Counters: 55
    File System Counters
        FILE: Number of bytes read=261723
        FILE: Number of bytes written=1352900
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=1597171
        HDFS: Number of bytes written=2462
        HDFS: Number of read operations=11
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
    Job Counters
        Killed map tasks=1
        Launched map tasks=2
        Launched reduce tasks=1
        Data-local map tasks=2
        Total time spent by all maps in occupied slots (ms)=5351
        Total time spent by all reduces in occupied slots (ms)=2548
        Total time spent by all map tasks (ms)=5351
        Total time spent by all reduce tasks (ms)=2548
        Total vcore-milliseconds taken by all map tasks=5351
        Total vcore-milliseconds taken by all reduce tasks=2548
        Total megabyte-milliseconds taken by all map tasks=5479424
        Total megabyte-milliseconds taken by all reduce tasks=2609152
    Map-Reduce Framework
        Map input records=17912
        Map output records=17912
        Map output bytes=225893
        Map output materialized bytes=261729
        Input split bytes=188
        Combine input records=0
        Combine output records=0
        Reduce input groups=174
        Reduce shuffle bytes=261729
        Reduce input records=17912
        Reduce output records=174
        Spilled Records=35824
        Shuffled Maps =2
        Failed Shuffles=0
        Merged Map outputs=2
        GC time elapsed (ms)=189
        CPU time spent (ms)=2920
        Physical memory (bytes) snapshot=918667264
        Virtual memory (bytes) snapshot=8397361152
        Total committed heap usage (bytes)=754974720
        Peak Map Physical memory (bytes)=343416832
        Peak Map Virtual memory (bytes)=2796998656
        Peak Reduce Physical memory (bytes)=232030208
        Peak Reduce Virtual memory (bytes)=2803683328
    Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
    File Input Format Counters
        Bytes Read=1596983
    File Output Format Counters
        Bytes Written=2462
2022-10-25 20:14:45,969 INFO streaming.StreamJob: Output directory: /JayLab/1/Part1
```

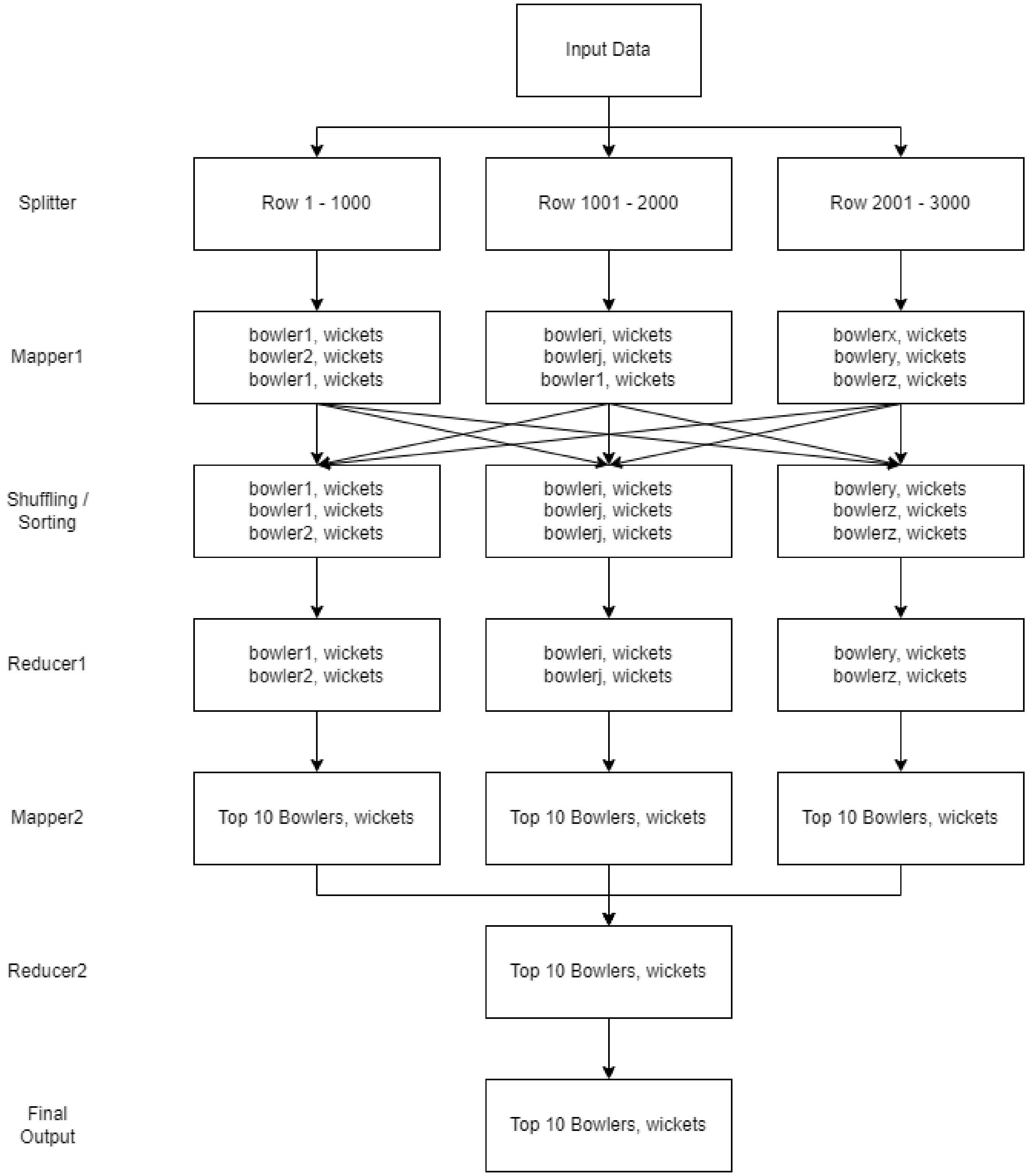
Part 2

```
2022-10-25 20:20:12,352 INFO mapreduce.Job: Counters: 54
    File System Counters
        FILE: Number of bytes read=342
        FILE: Number of bytes written=830159
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=3895
        HDFS: Number of bytes written=144
        HDFS: Number of read operations=11
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
    Job Counters
        Launched map tasks=2
        Launched reduce tasks=1
        Data-local map tasks=2
        Total time spent by all maps in occupied slots (ms)=5188
        Total time spent by all reduces in occupied slots (ms)=2142
        Total time spent by all map tasks (ms)=5188
        Total time spent by all reduce tasks (ms)=2142
        Total vcore-milliseconds taken by all map tasks=5188
        Total vcore-milliseconds taken by all reduce tasks=2142
        Total megabyte-milliseconds taken by all map tasks=5312512
        Total megabyte-milliseconds taken by all reduce tasks=2193408
    Map-Reduce Framework
        Map input records=174
        Map output records=20
        Map output bytes=296
        Map output materialized bytes=348
        Input split bytes=202
        Combine input records=0
        Combine output records=0
        Reduce input groups=20
        Reduce shuffle bytes=348
        Reduce input records=20
        Reduce output records=10
        Spilled Records=40
        Shuffled Maps =2
        Failed Shuffles=0
        Merged Map outputs=2
        GC time elapsed (ms)=199
        CPU time spent (ms)=1790
        Physical memory (bytes) snapshot=862785536
        Virtual memory (bytes) snapshot=8401317888
        Total committed heap usage (bytes)=735051776
        Peak Map Physical memory (bytes)=304414720
        Peak Map Virtual memory (bytes)=2799472640
        Peak Reduce Physical memory (bytes)=258834432
        Peak Reduce Virtual memory (bytes)=2804510720
    Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
    File Input Format Counters
        Bytes Read=3693
    File Output Format Counters
        Bytes Written=144
2022-10-25 20:20:12,352 INFO streaming.StreamJob: Output directory:
/JayLab/1/Part2
```

2. Top 10 wicket takers in the tournament

In the IPLT20 2022, analyze and extract the top 10 wicket takers. The output data should contain the bowler names and the wickets taken. Only 10 records should be displayed and should be arranged in the descending order of the wickets

Map-Reduce Diagram



Pseudocode

Mapper 1

```
Input: Dataset.csv
For each row
        Extract the Bowler and wickets
Output: bowler          wickets
```

Reducer 1

```
Input: Output of Mapper1
For each row
        Add the Bowler's wickets
        return the Bowler name and wickets
Output: Bowler  wickets taken
```

Mapper 2

```
Input: Output of Reducer1
For each row
        Sort as per the wickets taken in ascending Order
        Extract only top 10 wicket takers
Output: Bowler           wickets taken
```

Reducer 2

```
Input: Output of Mapper2
For each row
        Sort as per the wickets taken in ascending Order
        Extract only top 10 wicket takers
Output: Bowler           wickets taken
```

Source Code

mapper1.py

```
import sys

for line in sys.stdin:
    line = line.strip()

matchId,innings,overs,ballnumber,batsman,bowler,nonStriker,extra_type,batsman_runs
```

```
,extras_run,total_run,non_boundary,isWicketDelivery,player_out,kind,fielders_involved,BattingTeam = line.split(',')

    if(isWicketDelivery == '1'):
        print('{}\t{}'.format(bowler, isWicketDelivery))
```

reducer1.py

```
import sys

current_bowler = None
current_bowler_wickets = 0

for line in sys.stdin:
    line = line.strip()
    bowler, wickets = line.split('\t')

    try:
        wickets = int(wickets)
        bowler = bowler.strip()
    except ValueError:
        continue

    if current_bowler == bowler:
        current_bowler_wickets += wickets
    else:
        if current_bowler:
            print('{}\t{}'.format(current_bowler, current_bowler_wickets))
        current_bowler = bowler
        current_bowler_wickets = wickets

if current_bowler == wickets:
    print('{}\t{}'.format(current_bowler, current_bowler_wickets))
```

mapper2.py

```
import sys

scores=[]
for line in sys.stdin:
    line = line.strip()
    batsman,batsman_runs = line.split('\t')
    try:
        batsman_runs = int(batsman_runs)
    except ValueError:
        continue
    scores.append([batsman_runs, batsman])
```

```
top_N=sorted(scores,reverse=True)[0:10]

for t in top_N:
    print('{}\t{}'.format(t[1], t[0]))
```

reducer2.py

```
import sys

scores=[]
for line in sys.stdin:
    line = line.strip()
    batsman,batsman_runs = line.split('\t')
    try:
        batsman_runs = int(batsman_runs)
        batsman = batsman.strip()
    except ValueError:
        continue
    scores.append([batsman_runs, batsman])

top_N=sorted(scores,reverse=True)[0:10]

for t in top_N:
    print('{}\t{}'.format(t[1], t[0]))
```

Statistics

Part 1

```
2022-10-25 20:25:50,370 INFO mapreduce.Job: Counters: 54
    File System Counters
        FILE: Number of bytes read=13988
        FILE: Number of bytes written=857427
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=1597171
        HDFS: Number of bytes written=1403
        HDFS: Number of read operations=11
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
    Job Counters
        Launched map tasks=2
        Launched reduce tasks=1
        Data-local map tasks=2
        Total time spent by all maps in occupied slots (ms)=5220
        Total time spent by all reduces in occupied slots (ms)=2110
        Total time spent by all map tasks (ms)=5220
        Total time spent by all reduce tasks (ms)=2110
        Total vcore-milliseconds taken by all map tasks=5220
        Total vcore-milliseconds taken by all reduce tasks=2110
        Total megabyte-milliseconds taken by all map tasks=5345280
        Total megabyte-milliseconds taken by all reduce tasks=2160640
    Map-Reduce Framework
        Map input records=17912
        Map output records=912
        Map output bytes=12158
        Map output materialized bytes=13994
        Input split bytes=188
        Combine input records=0
        Combine output records=0
        Reduce input groups=105
        Reduce shuffle bytes=13994
        Reduce input records=912
        Reduce output records=104
        Spilled Records=1824
        Shuffled Maps =2
        Failed Shuffles=0
        Merged Map outputs=2
        GC time elapsed (ms)=237
        CPU time spent (ms)=2170
        Physical memory (bytes) snapshot=937807872
        Virtual memory (bytes) snapshot=8398974976
        Total committed heap usage (bytes)=744488960
        Peak Map Physical memory (bytes)=361345024
        Peak Map Virtual memory (bytes)=2797686784
        Peak Reduce Physical memory (bytes)=218529792
        Peak Reduce Virtual memory (bytes)=2804658176
    Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
    File Input Format Counters
        Bytes Read=1596983
    File Output Format Counters
        Bytes Written=1403
2022-10-25 20:25:50,370 INFO streaming.StreamJob: Output directory: /JayLab/2/Part1
```

Part 2

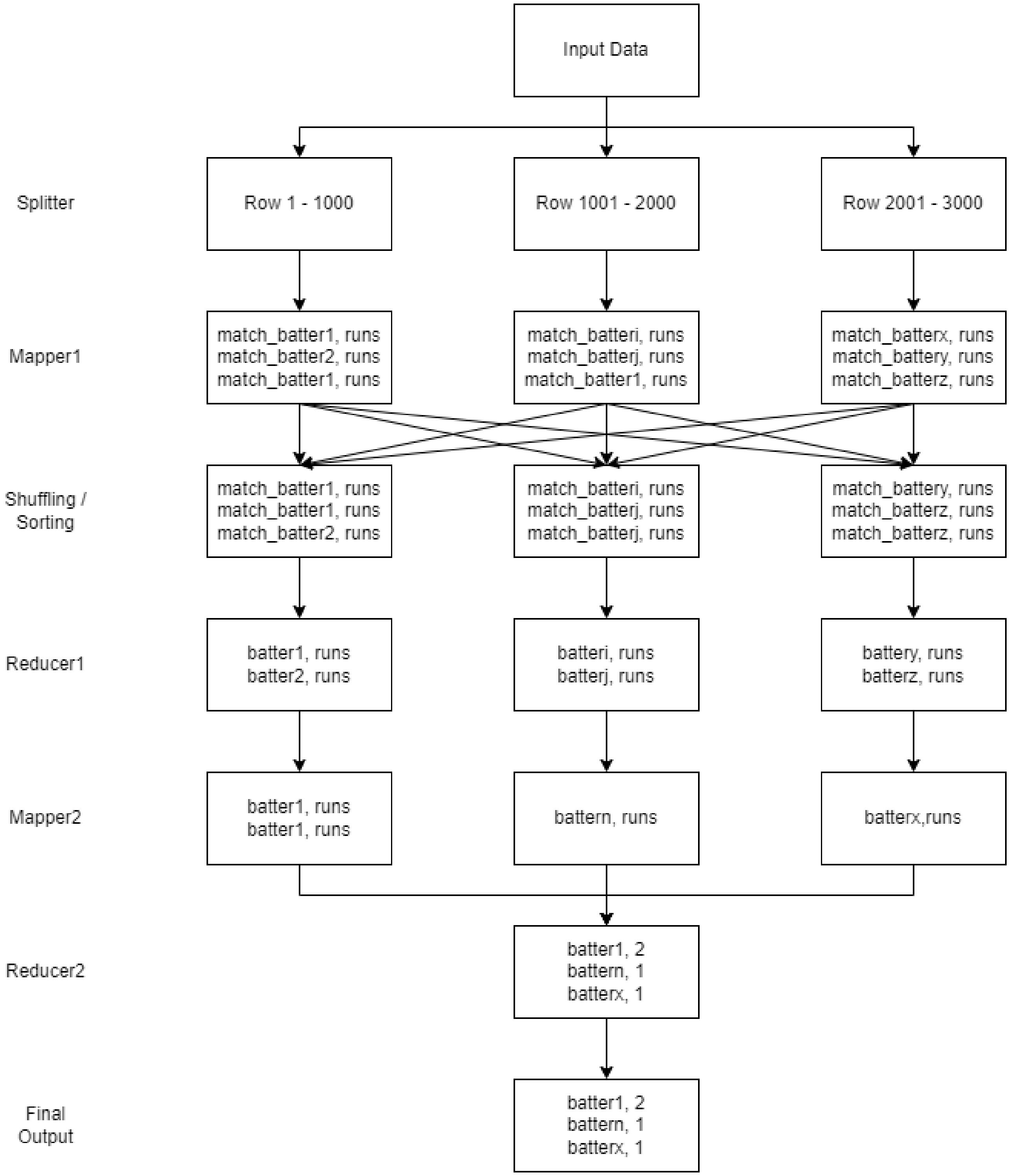
```

2022-10-25 20:30:11,717 INFO mapreduce.Job: Counters: 54
    File System Counters
        FILE: Number of bytes read=334
        FILE: Number of bytes written=830143
```

```
            FILE: Number of read operations=0
            FILE: Number of large read operations=0
            FILE: Number of write operations=0
            HDFS: Number of bytes read=2307
            HDFS: Number of bytes written=154
            HDFS: Number of read operations=11
            HDFS: Number of large read operations=0
            HDFS: Number of write operations=2
            HDFS: Number of bytes read erasure-coded=0
        Job Counters
            Launched map tasks=2
            Launched reduce tasks=1
            Data-local map tasks=2
            Total time spent by all maps in occupied slots (ms)=5029
            Total time spent by all reduces in occupied slots (ms)=2241
            Total time spent by all map tasks (ms)=5029
            Total time spent by all reduce tasks (ms)=2241
            Total vcore-milliseconds taken by all map tasks=5029
            Total vcore-milliseconds taken by all reduce tasks=2241
            Total megabyte-milliseconds taken by all map tasks=5149696
            Total megabyte-milliseconds taken by all reduce tasks=2294784
        Map-Reduce Framework
            Map input records=104
            Map output records=20
            Map output bytes=288
            Map output materialized bytes=340
            Input split bytes=202
            Combine input records=0
            Combine output records=0
            Reduce input groups=20
            Reduce shuffle bytes=340
            Reduce input records=20
            Reduce output records=10
            Spilled Records=40
            Shuffled Maps =2
            Failed Shuffles=0
            Merged Map outputs=2
            GC time elapsed (ms)=203
            CPU time spent (ms)=1880
            Physical memory (bytes) snapshot=864178176
            Virtual memory (bytes) snapshot=8397783040
            Total committed heap usage (bytes)=745013248
            Peak Map Physical memory (bytes)=339836928
            Peak Map Virtual memory (bytes)=2797289472
            Peak Reduce Physical memory (bytes)=234184704
            Peak Reduce Virtual memory (bytes)=2804183040
        Shuffle Errors
            BAD_ID=0
            CONNECTION=0
            IO_ERROR=0
            WRONG_LENGTH=0
            WRONG_MAP=0
            WRONG_REDUCE=0
        File Input Format Counters
            Bytes Read=2105
        File Output Format Counters
            Bytes Written=154
    2022-10-25 20:30:11,717 INFO streaming.StreamJob: Output directory:
    /JayLab/2/Part2
```

3. Centuries

In the IPLT20 2022, analyze and extract the batsman who have scored atleast a century. The output data should contain the batsman names and the number of centuries. The output should be arranged in the descending order of centuries scored

Map-Reduce Diagram



Pseudocode

Mapper 1

```
Input: Dataset.csv
For each row
        Extract the matchID_Batsman and Runs
Output: matchID_Batsman         Runs
```

Reducer 1

```
Input: Output of Mapper1
For each row
        Add the Batsman's score in a match
        return the batsman name and runs
Output: Batsman         runsScored in a match
```

Mapper 2

```
Input: Output of Reducer1
For each row
        Filter only the runs which is equal to or more than 100
        Return the batsman name and runs
Output: Batsman         Scores 100 and above
```

Reducer 2

```
Input: Output of Mapper2
For each row
        return the batsman name and count of centuries
Output: Batsman         count of centuries
```

Source Code

mapper1.py

```
import sys

for line in sys.stdin:
    line = line.strip()

matchId,innings,overs,ballnumber,batsman,bowler,nonStriker,extra_type,batsman_runs
,extras_run,total_run,non_boundary,isWicketDelivery,player_out,kind,fielders_invol
```

```
ved,BattingTeam = line.split(',')
    print('{}\t{}'.format(matchId + '_' + batsman, batsman_runs))
```

reducer1.py

```
import sys

current_batsman = None
current_batsman_runs = 0

for line in sys.stdin:
    line = line.strip()
    batsman, run = line.split('\t')

    try:
        run = int(run)
    except ValueError:
        continue

    if current_batsman == batsman:
        current_batsman_runs += run
    else:
        if current_batsman:
            print('{}\t{}'.format(current_batsman, current_batsman_runs))
        current_batsman = batsman
        current_batsman_runs = run

if current_batsman == batsman:
    print('{}\t{}'.format(current_batsman, current_batsman_runs))
```

mapper2.py

```
import sys

for line in sys.stdin:
    line = line.strip()
    batsman,batsman_runs = line.split('\t')

    try:
        batsman_runs = int(batsman_runs)
        batsman = batsman.strip()
        matchId, batsman = batsman.split('_')
    except ValueError:
        continue

    if(batsman_runs >= 100):
        print('{}\t{}'.format(batsman, batsman_runs))
```

reducer2.py

```
import sys

current_batsman = None
count = 0

for line in sys.stdin:
    line = line.strip()
    batsman, run = line.split('\t')

    if current_batsman == batsman:
        count = count+1
    else:
        if current_batsman:
            print (current_batsman, '\t' ,count)
        current_batsman = batsman
        count = 1

if current_batsman == batsman:
    print('{}\t{}'.format(current_batsman, count))
```

Statistics

Part 1

```
2022-10-25 20:14:45,968 INFO mapreduce.Job: Counters: 55
    File System Counters
        FILE: Number of bytes read=261723
        FILE: Number of bytes written=1352900
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=1597171
        HDFS: Number of bytes written=2462
        HDFS: Number of read operations=11
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
    Job Counters
        Killed map tasks=1
        Launched map tasks=2
        Launched reduce tasks=1
        Data-local map tasks=2
        Total time spent by all maps in occupied slots (ms)=5351
        Total time spent by all reduces in occupied slots (ms)=2548
        Total time spent by all map tasks (ms)=5351
        Total time spent by all reduce tasks (ms)=2548
        Total vcore-milliseconds taken by all map tasks=5351
        Total vcore-milliseconds taken by all reduce tasks=2548
        Total megabyte-milliseconds taken by all map tasks=5479424
        Total megabyte-milliseconds taken by all reduce tasks=2609152
    Map-Reduce Framework
        Map input records=17912
        Map output records=17912
        Map output bytes=225893
        Map output materialized bytes=261729
        Input split bytes=188
        Combine input records=0
        Combine output records=0
        Reduce input groups=174
        Reduce shuffle bytes=261729
        Reduce input records=17912
        Reduce output records=174
        Spilled Records=35824
        Shuffled Maps =2
        Failed Shuffles=0
        Merged Map outputs=2
        GC time elapsed (ms)=189
        CPU time spent (ms)=2920
        Physical memory (bytes) snapshot=918667264
        Virtual memory (bytes) snapshot=8397361152
        Total committed heap usage (bytes)=754974720
        Peak Map Physical memory (bytes)=343416832
        Peak Map Virtual memory (bytes)=2796998656
        Peak Reduce Physical memory (bytes)=232030208
        Peak Reduce Virtual memory (bytes)=2803683328
    Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
    File Input Format Counters
        Bytes Read=1596983
    File Output Format Counters
        Bytes Written=2462
2022-10-25 20:14:45,969 INFO streaming.StreamJob: Output directory:
/JayLab/1/Part1
```

Part 2

```
2022-10-25 20:20:12,352 INFO mapreduce.Job: Counters: 54
    File System Counters
        FILE: Number of bytes read=342
        FILE: Number of bytes written=830159
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
```

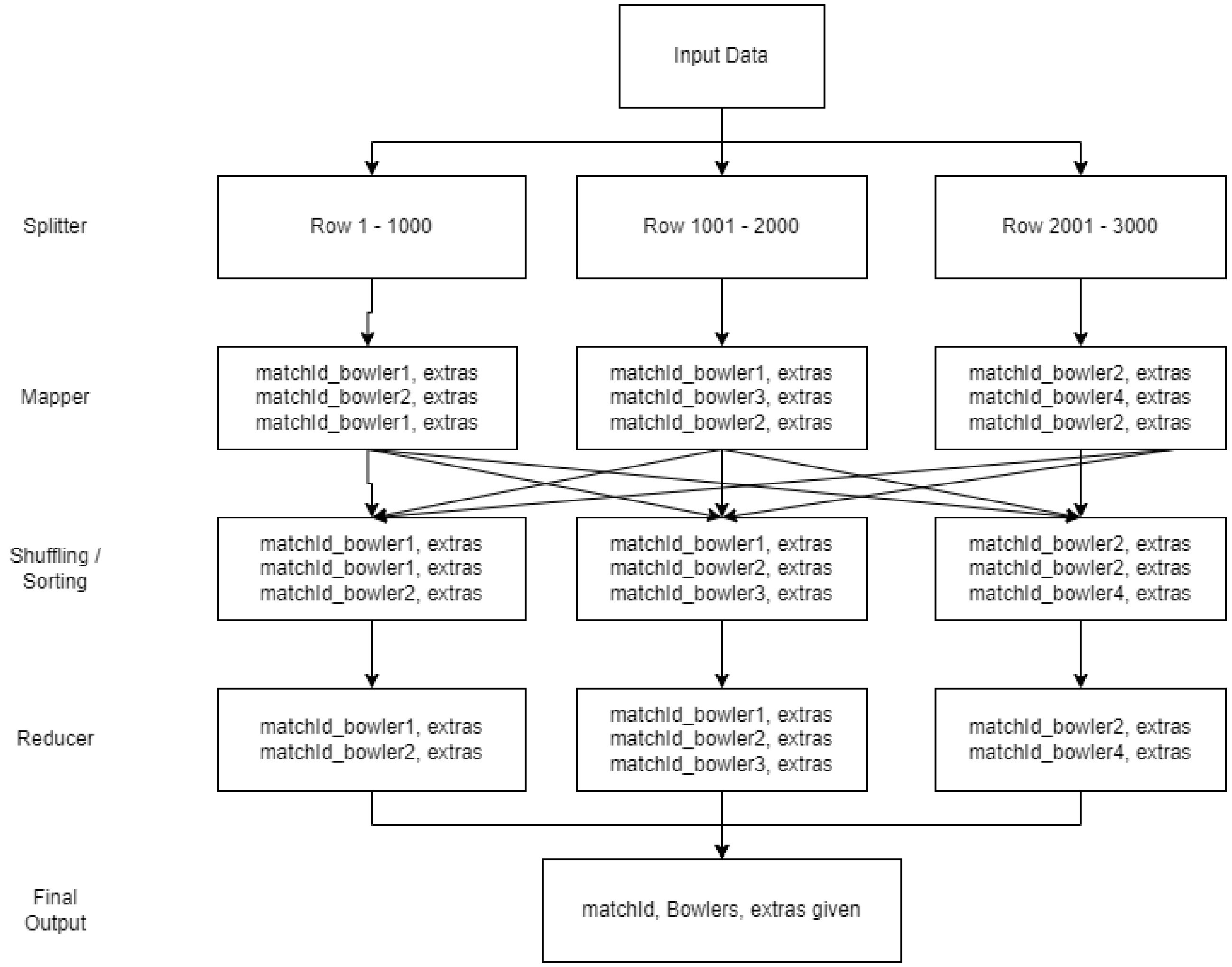
```
          FILE: Number of write operations=0
          HDFS: Number of bytes read=3895
          HDFS: Number of bytes written=144
          HDFS: Number of read operations=11
          HDFS: Number of large read operations=0
          HDFS: Number of write operations=2
          HDFS: Number of bytes read erasure-coded=0
      Job Counters
          Launched map tasks=2
          Launched reduce tasks=1
          Data-local map tasks=2
          Total time spent by all maps in occupied slots (ms)=5188
          Total time spent by all reduces in occupied slots (ms)=2142
          Total time spent by all map tasks (ms)=5188
          Total time spent by all reduce tasks (ms)=2142
          Total vcore-milliseconds taken by all map tasks=5188
          Total vcore-milliseconds taken by all reduce tasks=2142
          Total megabyte-milliseconds taken by all map tasks=5312512
          Total megabyte-milliseconds taken by all reduce tasks=2193408
      Map-Reduce Framework
          Map input records=174
          Map output records=20
          Map output bytes=296
          Map output materialized bytes=348
          Input split bytes=202
          Combine input records=0
          Combine output records=0
          Reduce input groups=20
          Reduce shuffle bytes=348
          Reduce input records=20
          Reduce output records=10
          Spilled Records=40
          Shuffled Maps =2
          Failed Shuffles=0
          Merged Map outputs=2
          GC time elapsed (ms)=199
          CPU time spent (ms)=1790
          Physical memory (bytes) snapshot=862785536
          Virtual memory (bytes) snapshot=8401317888
          Total committed heap usage (bytes)=735051776
          Peak Map Physical memory (bytes)=304414720
          Peak Map Virtual memory (bytes)=2799472640
          Peak Reduce Physical memory (bytes)=258834432
          Peak Reduce Virtual memory (bytes)=2804510720
      Shuffle Errors
          BAD_ID=0
          CONNECTION=0
          IO_ERROR=0
          WRONG_LENGTH=0
          WRONG_MAP=0
          WRONG_REDUCE=0
      File Input Format Counters
          Bytes Read=3693
      File Output Format Counters
```

```
        Bytes Written=144
2022-10-25 20:20:12,352 INFO streaming.StreamJob: Output directory:
/JayLab/1/Part2
```

4. Extras bowled in a match by a bowler

In the IPLT20 2022, analyze and extract the extras bowled by a bowler in a match. The output data should contain the matchID, bowler and extras given. Output to be arranged as per the match ID in ascending order

Map-Reduce Diagram



Pseudocode

Mapper

```
Input: Dataset.csv
For each row
    Extract the extra run given deliveries
    return the bowler name and the extra runs conceded
Output: matchID_bowler          extra runs conceded
```

Reducer

```
Input: Output of Mapper1
For each row
    Add the runs conceded by the bowler
Output: matchID         bowler      total runs conceded
```

Source Code

mapper.py

```
import sys

for line in sys.stdin:
    line = line.strip()

matchId,innings,overs,ballnumber,batsman,bowler,nonStriker,extra_type,batsman_runs
,extras_run,total_run,non_boundary,isWicketDelivery,player_out,kind,fielders_invol
ved,BattingTeam = line.split(',')

    if(extra_type != 'NA'):
        print('{}\t{}'.format(matchId + '_' + bowler, extras_run))
```

reducer.py

```
import sys

current_matchId_bowler = None
current_extras = 0

for line in sys.stdin:
    line = line.strip()
    matchId_bowler, extras = line.split('\t')
    try:
        extras = int(extras)
    except ValueError:
        continue

    if current_matchId_bowler == matchId_bowler:
        current_extras += extras
    else:
        if current_matchId_bowler:
            matchId, bowler = matchId_bowler.split('_')
            print('{}\t{}\t{}'.format(matchId, bowler, current_extras))
        current_matchId_bowler = matchId_bowler
        current_extras = extras

if current_matchId_bowler == matchId_bowler:
    matchId, bowler = matchId_bowler.split('_')
    print('{}\t{}\t{}'.format(matchId, bowler, current_extras))
```

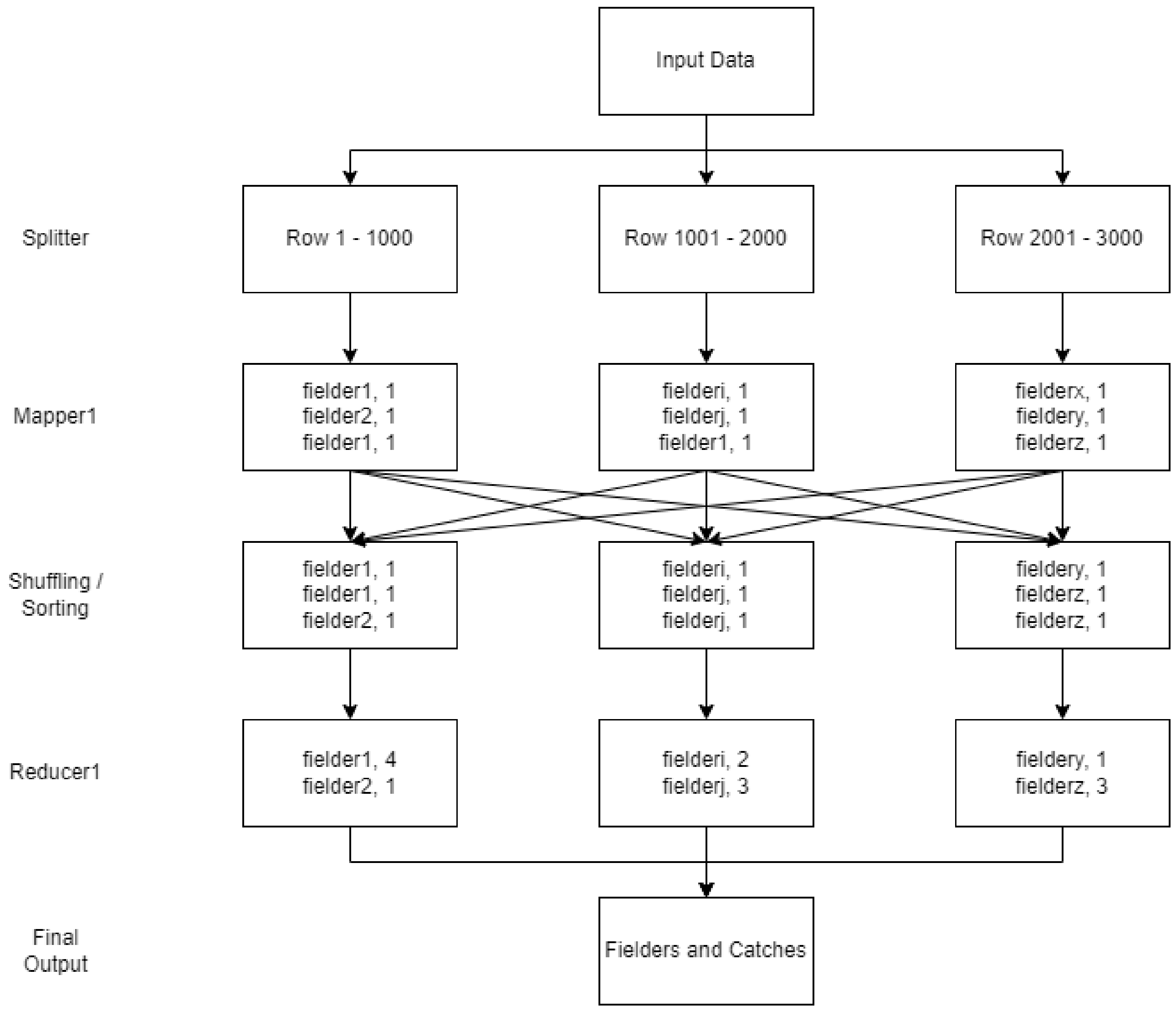
Statistics

```
2022-10-25 20:42:15,615 INFO mapreduce.Job: Counters: 54
    File System Counters
        FILE: Number of bytes read=9474
        FILE: Number of bytes written=848294
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=1597171
        HDFS: Number of bytes written=3917
        HDFS: Number of read operations=11
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
    Job Counters
        Launched map tasks=2
        Launched reduce tasks=1
        Data-local map tasks=2
        Total time spent by all maps in occupied slots (ms)=5221
        Total time spent by all reduces in occupied slots (ms)=2210
        Total time spent by all map tasks (ms)=5221
        Total time spent by all reduce tasks (ms)=2210
        Total vcore-milliseconds taken by all map tasks=5221
        Total vcore-milliseconds taken by all reduce tasks=2210
        Total megabyte-milliseconds taken by all map tasks=5346304
        Total megabyte-milliseconds taken by all reduce tasks=2263040
    Map-Reduce Framework
        Map input records=17912
        Map output records=650
        Map output bytes=8168
        Map output materialized bytes=9480
        Input split bytes=188
        Combine input records=0
        Combine output records=0
        Reduce input groups=157
        Reduce shuffle bytes=9480
        Reduce input records=650
        Reduce output records=157
        Spilled Records=1300
        Shuffled Maps =2
        Failed Shuffles=0
        Merged Map outputs=2
        GC time elapsed (ms)=219
        CPU time spent (ms)=2090
        Physical memory (bytes) snapshot=903471104
        Virtual memory (bytes) snapshot=8400039936
        Total committed heap usage (bytes)=741867520
        Peak Map Physical memory (bytes)=344375296
        Peak Map Virtual memory (bytes)=2797801472
        Peak Reduce Physical memory (bytes)=217579520
        Peak Reduce Virtual memory (bytes)=2804768768
    Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
    File Input Format Counters
        Bytes Read=1596983
    File Output Format Counters
        Bytes Written=3917
2022-10-25 20:42:15,615 INFO streaming.StreamJob: Output directory: /JayLab/4/Part1
```

5. Catches taken

In the IPLT20 2022, analyze and extract the total catches taken by a fielder. The output data should contain the fieldername and the count of catches taken. Output to be arranged in the descending order of the catches taken

Map-Reduce Diagram



Pseudocode

Mapper

```
Input: Dataset.csv
For each row
        Extract the catches and the caught and bowled deliveries
Output: fielder         1
```

Reducer

```
Input: Output of Mapper1
For each row
        Add the catches caught by a fielder
Output: fielder         catches
```

Source Code

mapper.py

```python
import sys

for line in sys.stdin:
    line = line.strip()

matchId,innings,overs,ballnumber,batsman,bowler,nonStriker,extra_type,batsman_runs,extras_run,total_run,non_boundary,isWicketDelivery,player_out,kind,fielders_involved,BattingTeam = line.split(',')

    if isWicketDelivery == '1':
        if kind == 'caught':
            print('{}\t{}'.format(fielders_involved, 1))
        elif kind == 'caught and bowled':
            print('{}\t{}'.format(bowler, 1))
```

reducer.py

```python
import sys

current_fielder = None
current_fielder_catches = 0

for line in sys.stdin:
    line = line.strip()
    fielder, catches = line.split('\t')

    try:
        catches = int(catches)
    except ValueError:
        continue

    if current_fielder == fielder:
        current_fielder_catches += catches
    else:
        if current_fielder:
            print('{}\t{}'.format(current_fielder, current_fielder_catches))
        current_fielder = fielder
        current_fielder_catches = catches

if current_fielder == fielder:
    print('{}\t{}'.format(current_fielder, current_fielder_catches))
```

Statistics

```
2022-10-25 20:48:15,645 INFO mapreduce.Job: Counters: 54
    File System Counters
        FILE: Number of bytes read=9474
        FILE: Number of bytes written=848294
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=1597171
        HDFS: Number of bytes written=3917
        HDFS: Number of read operations=11
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=2
        HDFS: Number of bytes read erasure-coded=0
    Job Counters
        Launched map tasks=2
        Launched reduce tasks=1
        Data-local map tasks=2
        Total time spent by all maps in occupied slots (ms)=5221
        Total time spent by all reduces in occupied slots (ms)=2210
        Total time spent by all map tasks (ms)=5221
        Total time spent by all reduce tasks (ms)=2210
        Total vcore-milliseconds taken by all map tasks=5221
        Total vcore-milliseconds taken by all reduce tasks=2210
        Total megabyte-milliseconds taken by all map tasks=5346304
        Total megabyte-milliseconds taken by all reduce tasks=2263040
    Map-Reduce Framework
        Map input records=17912
        Map output records=650
        Map output bytes=8168
        Map output materialized bytes=9480
        Input split bytes=188
        Combine input records=0
        Combine output records=0
        Reduce input groups=157
        Reduce shuffle bytes=9480
        Reduce input records=650
        Reduce output records=157
        Spilled Records=1300
        Shuffled Maps =2
        Failed Shuffles=0
        Merged Map outputs=2
        GC time elapsed (ms)=219
        CPU time spent (ms)=2090
        Physical memory (bytes) snapshot=903471104
        Virtual memory (bytes) snapshot=8400039936
        Total committed heap usage (bytes)=741867520
        Peak Map Physical memory (bytes)=344375296
        Peak Map Virtual memory (bytes)=2797801472
        Peak Reduce Physical memory (bytes)=217579520
        Peak Reduce Virtual memory (bytes)=2804768768
    Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
    File Input Format Counters
        Bytes Read=1596983
    File Output Format Counters
        Bytes Written=3917
2022-10-25 20:48:15,645 INFO streaming.StreamJob: Output directory:
/JayLab/5/Part1
```

References

- https://towardsdatascience.com/chaining-multiple-mapreduce-jobs-with-hadoop-java-832a326cbfa7
- https://www.edureka.co/blog/hadoop-streaming-mapreduce-program/