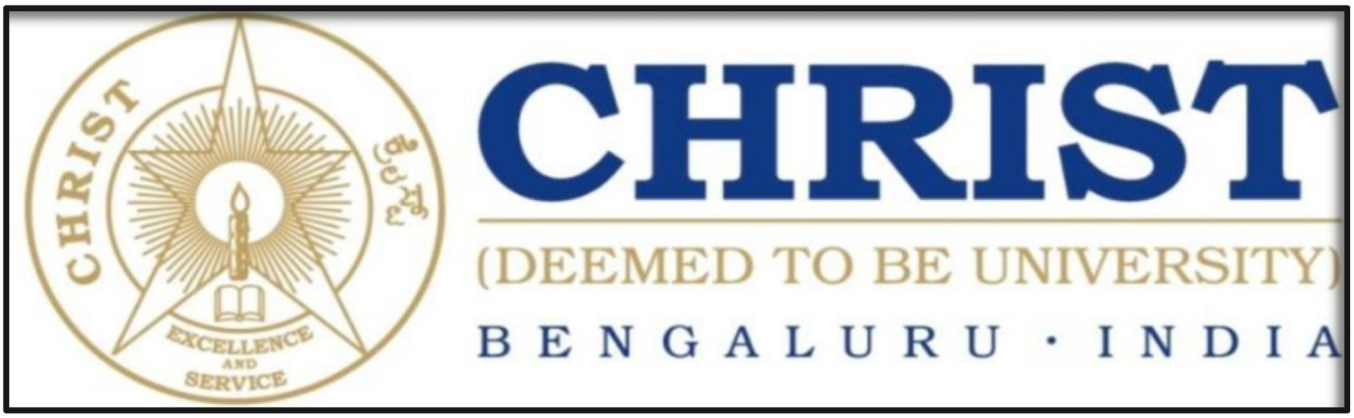


USER ACTIVITY ANALYSIS AND SEGMENTATION USING HADOOP MAPREDUCE

(DEMONSTRATION JOINING OF TWO BIG DATA SETS)

SANJAY R

2348055

5th November 2024

MDS571

Big Data Analytics

Department of Statistics and Data Science

CHRIST (Deemed to be University)

Department of Statistics and Data Science

Course:MDS571-Big Data Analytics **Lab :4**

R.SANJAY **2348055**

User Activity Analysis and Segmentation Using Hadoop MapReduce

Demonstration joining of two big data sets

✧ INTRODUCTION:

In the field of data analytics, joining large datasets is a fundamental operation that enables the integration of information from multiple sources. This becomes particularly important when analyzing sports data, where player statistics and personal information are often stored separately. By combining these datasets, analysts can generate more comprehensive insights, allowing for better decision-making and performance evaluation.

The Cricket Dataset Join program is a Hadoop-based MapReduce application designed to join two large cricket-related datasets. The first dataset contains player details such as names, countries, and roles, while the second holds performance metrics like runs scored, wickets taken, and match statistics. Joining these datasets helps provide a complete picture of each player's contribution, enabling more informed analyses.MapReduce is a suitable tool for this task due to its ability to handle large datasets distributed across many nodes. The parallel processing nature of MapReduce ensures that the join operation is both efficient and scalable, even for extensive cricket data. This approach is essential for processing and analyzing big data in real-time or at scale, making it a key component of modern data analytics.

✧ PROBLEM DESCRIPTION:

In cricket data analysis, valuable information about players, such as their personal details and performance statistics, is often stored in separate datasets. For example, one dataset may contain player names, countries, and roles, while another may include performance metrics like runs scored, wickets taken, and matches played. To fully understand a player's career, these datasets need to be combined.

The main challenge is how to efficiently join these two datasets based on a common key, which in this case is the player_id. Without combining these datasets, it is difficult to gain insights that provide a complete profile of each player. The objective of this program is to merge these datasets into a unified output that shows both personal and performance details for every player.

The solution uses the MapReduce framework, which is designed to handle large amounts of data by distributing the workload across multiple machines. This approach ensures that the joining process is both scalable and efficient, making it ideal for working with big data in cricket analytics.

✧ DATASET DESCRIPTION:

The program utilizes two key datasets: player_info and performance_info, which together provide a comprehensive view of a cricketer's profile.

The **player_info** dataset contains essential player details, including the following fields:

- ✓ **player_id:** A unique identifier for each player.
- ✓ **name:** The name of the player.
- ✓ **country:** The country the player represents.
- ✓ **role:** The player's role in the team (e.g., Batsman, Bowler, All-Rounder, Wicketkeeper).

The **performance_info** dataset, on the other hand, contains crucial performance metrics for each player, such as:

- ✓ **matches:** The number of matches the player has participated in.
- ✓ **runs:** The total runs scored by the player.
- ✓ **wickets:** The total wickets taken by the player.
- ✓ **average:** The player's batting or bowling average.

Both datasets are structured in a tab-separated format, with each record corresponding to a specific player. The player_id field acts as the common key between the two datasets, which is used to merge the data. For effective integration, the data in these datasets are preprocessed to ensure consistency in formatting and are aligned for easy use during the MapReduce process.

These datasets reflect real-world cricket statistics, offering a solid basis for analyzing player performance. The design and structure of the datasets ensure they are both relevant and applicable for joining operations, enabling insightful analysis of player performance when combined.

✧ CODE:

```
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.NullWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

import java.io.IOException;
import java.util.ArrayList;
import java.util.List;

public class CricketJoin {

    /**
     * Mapper Class
     * - Reads input data from both datasets (player_info and performance_info).
     * - Extracts the ID (first column) as the key.
```

```
     * - Emits the ID as the key and the full record as the value.
     */
    public static class CricketJoinMapper extends Mapper<Object, Text, IntWritable,
Text> {
        private IntWritable joinKey = new IntWritable();
        private Text valueText = new Text();

        @Override
        public void map(Object key, Text value, Context context) throws
IOException, InterruptedException {
            String line = value.toString();                    // Convert the input line
to a string
            String[] parts = line.split("\t");                 // Split the line into parts
by tabs

            if (parts.length > 1) {                            // Ensure the line
contains at least two columns
                joinKey.set(Integer.parseInt(parts[0])); // Set the join key (player
ID)
                valueText.set(value.toString());               // Set the entire record
as value
                context.write(joinKey, valueText);             // Emit the key-value
pair
            }
        }
    }
    /**
     * Reducer Class
     * - Groups records by ID (key).
     * - Separates player_info and performance_info data based on the number of
columns in each record.
     * - Joins matching records from both datasets.
     * - Formats the combined record and writes it to the output.
     */
```

```
public static class CricketJoinReducer extends Reducer<IntWritable, Text,
NullWritable, Text> {
        private Text result = new Text();
        @Override
        public void reduce(IntWritable key, Iterable<Text> values, Context context)
throws IOException, InterruptedException {
            List<String> playerInfo = new ArrayList<>();              // List to store
player_info records
            List<String> performanceInfo = new ArrayList<>();    // List to store
performance_info records

            for (Text value : values) {
                String[] parts = value.toString().split("\t");

                if (parts.length == 4) {              // Identify records from
player_info (4 columns)
                    playerInfo.add(value.toString());
                } else if (parts.length == 5) { // Identify records from
performance_info (5 columns)
                    performanceInfo.add(value.toString());
                }
            }

            // Perform the join operation
            for (String player : playerInfo) {
                for (String performance : performanceInfo) {
                    String[] playerParts = player.split("\t");
                    String[] performanceParts = performance.split("\t");

                    // Build the joined record in the desired format
                    StringBuilder joinedRecord = new StringBuilder();
                    joinedRecord.append(playerParts[0]).append("\t"); // ID
                    joinedRecord.append(playerParts[1]).append("\t"); // Name
```

```
                    joinedRecord.append(playerParts[2]).append("\t"); //
Country
                    joinedRecord.append(playerParts[3]).append("\t"); // Role
                    joinedRecord.append(performanceParts[1]).append("\t"); //
Matches
                    joinedRecord.append(performanceParts[2]).append("\t"); //
Runs
                    joinedRecord.append(performanceParts[3]).append("\t"); //
Wickets
                    joinedRecord.append(performanceParts[4]);
// Average

                    // Write the formatted result
                    result.set(joinedRecord.toString());
                    context.write(NullWritable.get(), result);
                }
            }
        }
    }

    /**
     * Driver Class
     * - Sets up and configures the MapReduce job.
     * - Specifies the Mapper, Reducer, input/output data formats, and paths.
     */
    public static void main(String[] args) throws Exception {
        // Verify that three arguments are passed (player_info path,
performance_info path, and output path)
        if (args.length != 3) {
            System.err.println("Usage: CricketJoin <player_info path>
<performance_info path> <output path>");
            System.exit(-1);
        }
```

```
        Configuration conf = new Configuration();                    // Create a new
Hadoop configuration
        Job job = Job.getInstance(conf, "Cricket Join");          // Initialize the job
with a name

        job.setJarByClass(CricketJoin.class);                        // Specify the
main class
        job.setMapperClass(CricketJoinMapper.class);                 // Set the
Mapper class
        job.setReducerClass(CricketJoinReducer.class);             // Set the Reducer
class

        job.setMapOutputKeyClass(IntWritable.class);                 // Set the output
key type for the Mapper
        job.setMapOutputValueClass(Text.class);                      // Set the output
value type for the Mapper

        job.setOutputKeyClass(NullWritable.class);                   // Set the output
key type for the Reducer
        job.setOutputValueClass(Text.class);                         // Set the output
value type for the Reducer

        // Add input paths for player_info and performance_info datasets
        FileInputFormat.addInputPath(job, new Path(args[0])); // Input path 1:
player_info
        FileInputFormat.addInputPath(job, new Path(args[1])); // Input path 2:
performance_info
        FileOutputFormat.setOutputPath(job, new Path(args[2])); // Output path

        // Exit with job completion status
        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }
}
```

✧ PROGRAM DESCRIPTION:

The Cricket Dataset Join Program utilizes Hadoop's MapReduce framework to efficiently integrate two large datasets containing cricket-related information. This program is specifically designed to handle the challenges of merging datasets in a distributed environment, where the volume of data can be substantial. By leveraging the MapReduce paradigm, the program ensures scalability, fault tolerance, and high performance during the data processing stage. The task involves joining two datasets—player_info and performance_info—on a common key, player_id, to produce a unified output that consolidates player demographics and performance statistics.

The program's architecture is divided into three primary components—Mapper, Reducer, and Driver. Each of these components plays a distinct role in transforming and integrating the data, from reading and categorizing it to joining and outputting the final results. Below is a detailed explanation of how these components collaborate to achieve the program's objectives.

✓ MAPPER CLASS:

The Mapper class serves as the entry point for processing the input datasets. It reads the player_info and performance_info files line by line, where each record is tab-separated. The key task of the Mapper is to emit player_id as the key and the entire record as the value. This enables the MapReduce framework to group all records associated with a particular player_id during the shuffle and sort phase. Unlike traditional approaches where data may be pre-categorized, the Mapper is designed to treat input records uniformly, deferring dataset differentiation to the Reducer. This approach simplifies the Mapper logic and ensures flexibility when dealing with datasets of varying structures. By producing key-value pairs for each record, the Mapper sets the foundation for efficient joining of datasets.

✓ REDUCER CLASS:

The Reducer class takes over after the shuffle and sort phase, where all records with the same player_id are grouped together. It is responsible for merging the records

from the two datasets based on the shared player_id. To achieve this, the Reducer identifies the origin of each record (whether from player_info or performance_info) by analyzing the number of fields in the value. Once identified, it combines the information into a unified record, ensuring that all details about a player are consolidated into a single output line. The Reducer outputs these merged records in a tab-separated format, ready for further analysis. By handling dataset differentiation and integration, the Reducer ensures that the final output is both comprehensive and logically ordered.

✓ DRIVER CLASS:

The Driver class acts as the controller and coordinator for the entire MapReduce program. It is responsible for configuring the job, specifying the Mapper and Reducer classes, and defining the input and output formats. The Driver also sets the paths for the input datasets and output directory, ensuring seamless data flow. Additionally, it includes error handling mechanisms to validate input arguments and manage exceptions during execution. Once the job is configured, the Driver submits it to the Hadoop cluster for execution and monitors its progress until completion. The Driver's role is crucial in orchestrating the interaction between the Mapper and Reducer, ensuring that the program runs smoothly from start to finish.

✧ PROJECT SETUP AND EXECUTION:

✓ DATASET PREPARATION:

The Cricket Dataset Join Program relies on two primary datasets: player_info and performance_info.

- **player_info Dataset:** This dataset includes player demographic information such as player_id, name, country, and role. Each record provides critical details necessary for identifying and categorizing players.

```
1       Virat Kohli     India   Batsman
2       Shane Warne     Australia       Bowler
3       AB de Villiers  South Africa    All-Rounder
4       Joe Root        England Batsman
5       Muttiah Muralitharan    Sri Lanka       Bowler
6       Steve Smith     Australia       Batsman
7       Ben Stokes      England All-Rounder
8       Kagiso Rabada   South Africa    Bowler
9       David Warner    Australia       Batsman
10      Jofra Archer    England Bowler
11      Faf du Plessis  South Africa    Batsman
12      Dale Steyn      South Africa    Bowler
13      KL Rahul        India   Batsman
14      Rishabh Pant    India   Wicketkeeper
15      Mitchell Starc  Australia       Bowler
```

- **performance_info Dataset:** This dataset captures players' performance statistics, including player_id, matches, runs, wickets, and averages. These metrics are essential for analyzing player performance and contributions in cricket.

1	250	12000	5	52.5
2	194	3000	700	45.2
3	174	9000	50	50.1
4	130	8000	2	48.5
5	170	2500	800	44.5
6	120	5000	10	45.0
7	150	5500	120	47.6
8	90	2500	450	42.8
9	180	7000	10	49.0
10	85	1500	50	43.0
11	100	4000	10	46.5
12	160	3500	500	48.0
13	140	6000	5	49.5
14	90	2500	1	45.0
15	110	3500	120	47.0

Both datasets were carefully formatted to ensure consistency, with player_id acting as the common key to join the two. Each record is stored in a tab-separated format, which is optimal for processing with Hadoop's TextInputFormat.

To facilitate processing with the Hadoop MapReduce framework, the datasets were placed within the Hadoop directory structure at /MDS2024/LAB4. This setup ensured that the files were easily accessible for the program, allowing seamless integration and efficient execution of the MapReduce process.

```
hadoop@Ubuntu22:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [Ubuntu22]
hadoop@Ubuntu22:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hadoop@Ubuntu22:~$ jps
6433 ResourceManager
5957 DataNode
6184 SecondaryNameNode
7209 Jps
6555 NodeManager
5837 NameNode
```

```
hadoop@Ubuntu22:~$ hadoop fs -ls /MDS2024/LAB4/
Found 2 items
-rw-r--r--   3 hadoop supergroup        291 2024-11-29 20:24 /MDS2024/LAB4/performance_info
-rw-r--r--   3 hadoop supergroup        502 2024-11-29 20:24 /MDS2024/LAB4/player_info
```

✓ **ECLIPSE SETUP:**

To develop the Cricket Join Program, the Eclipse IDE was utilized as the primary platform for Java development. Known for its reliability and comprehensive features, Eclipse provided a robust environment for coding, testing, and debugging the Hadoop MapReduce components. The development process involved the following steps:

- A new Java project named CricketJoin was created within Eclipse. This project served as the foundational framework for implementing the MapReduce logic required for joining cricket datasets.
- **Main Class and MapReduce Components:** Within the CricketJoin project, the main class was implemented to orchestrate the Mapper, Reducer, and Driver components. Each component played a specific role in the program:

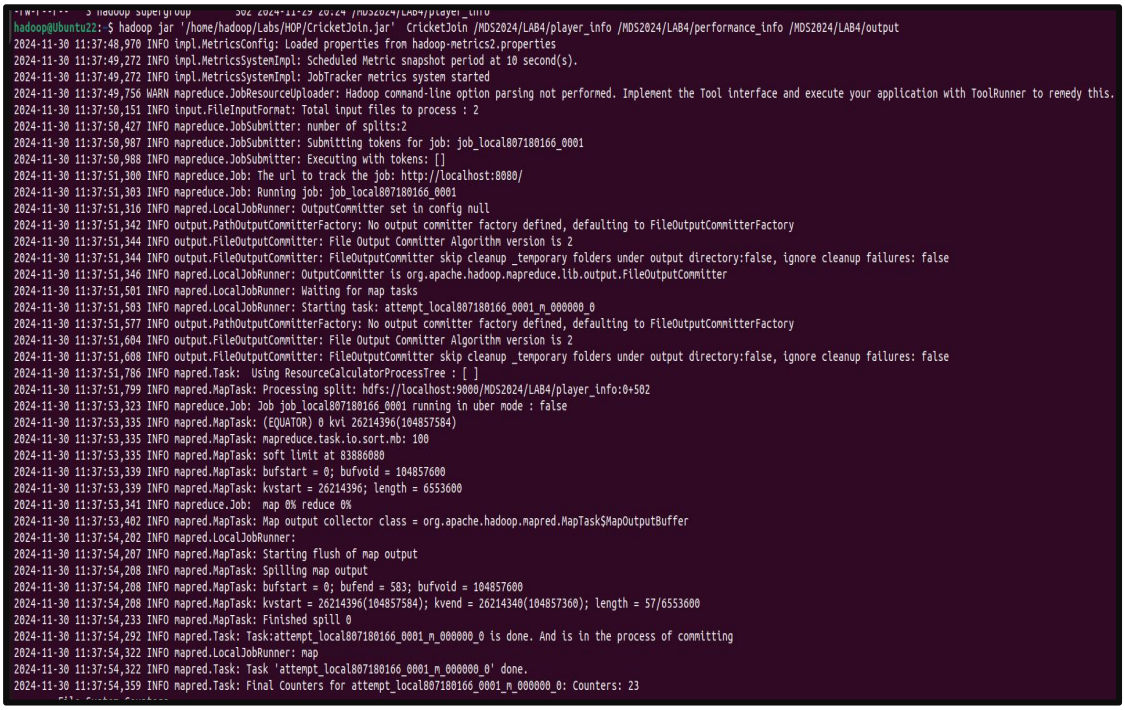
  - The CricketMapper class was designed to process input records from both datasets (player_info and performance_info).It extracted the player_id field from each line and emitted it as the key, along with the full record as the value.This approach enabled the Reducer to distinguish between the datasets based on their structure and join the relevant data efficiently.
  - The CricketReducer class received grouped records based on the common key player_id.It identified the source dataset of each record using the number of fields and merged the corresponding player demographic information with performance statistics.The Reducer then wrote the unified records to the output, ensuring all player details and performance metrics were consolidated into a single, structured line.

- After the code was developed and thoroughly tested, the necessary Hadoop library JAR files were added to the project to ensure Hadoop-specific functionalities during execution. The project was then exported as a JAR file named CricketJoin.jar, making it ready for deployment within the Hadoop environment.

✧ EXECUTION PROCESS:

- The next step involved executing the program within the Hadoop environment, using the following command:

$ hadoop jar '/home/hadoop/Labs/HOP/CricketJoin.jar' CricketJoin /MDS2024/LAB4/player_info /MDS2024/LAB4/performance_info /MDS2024/LAB4/output



~$ hadoop fs -ls /MDS2024/LAB4/output

~$ hadoop fs -ls /MDS2024/LAB4/output/part-r-00000

```
                Bytes Written=757
hadoop@Ubuntu22:~$ hadoop fs -ls /MDS2024/LAB4/output
Found 2 items
-rw-r--r--   3 hadoop supergroup          0 2024-11-30 12:36 /MDS2024/LAB4/output/_SUCCESS
-rw-r--r--   3 hadoop supergroup        757 2024-11-30 12:36 /MDS2024/LAB4/output/part-r-00000
hadoop@Ubuntu22:~$ hadoop fs -cat /MDS2024/LAB4/output/part-r-00000
1       Virat Kohli     India   Batsman 250     12000   5       52.5
2       Shane Warne     Australia       Bowler  194     3000    700     45.2
3       AB de Villiers  South Africa    All-Rounder     174     9000    50      50.1
4       Joe Root        England Batsman 130     8000    2       48.5
5       Muttiah Muralitharan    Sri Lanka       Bowler  170     2500    800     44.5
6       Steve Smith     Australia       Batsman 120     5000    10      45.0
7       Ben Stokes      England All-Rounder     150     5500    120     47.6
8       Kagiso Rabada   South Africa    Bowler  90      2500    450     42.8
9       David Warner    Australia       Batsman 180     7000    10      49.0
10      Jofra Archer    England Bowler  85      1500    50      43.0
11      Faf du Plessis  South Africa    Batsman 100     4000    10      46.5
12      Dale Steyn      South Africa    Bowler  160     3500    500     48.0
13      KL Rahul        India   Batsman 140     6000    5       49.5
14      Rishabh Pant    India   Wicketkeeper    90      2500    1       45.0
15      Mitchell Starc  Australia       Bowler  110     3500    120     47.0
```

OUTPUT:

1	Virat Kohli	India	Batsman	250	12000	5	52.5
2	Shane Warne	Australia	Bowler	194	3000	700	45.2
3	AB de Villiers	South Africa	All-Rounder	174	9000	50	50.1
4	Joe Root	England	Batsman	130	8000	2	48.5
5	Muttiah Muralitharan	Sri Lanka	Bowler	170	2500	800	44.5
6	Steve Smith	Australia	Batsman	120	5000	10	45
7	Ben Stokes	England	All-Rounder	150	5500	120	47.6
8	Kagiso Rabada	South Africa	Bowler	90	2500	450	42.8
9	David Warner	Australia	Batsman	180	7000	10	49
10	Jofra Archer	England	Bowler	85	1500	50	43
11	Faf du Plessis	South Africa	Batsman	100	4000	10	46.5
12	Dale Steyn	South Africa	Bowler	160	3500	500	48
13	KL Rahul	India	Batsman	140	6000	5	49.5

14	Rishabh Pant	India	Wicketkeeper	90	2500	1	45
15	Mitchell Starc	Australia	Bowler	110	3500	120	47

✧ CONCLUSION:

The Cricket Dataset Join Program demonstrates the power and efficiency of Hadoop's MapReduce framework in processing and integrating large datasets. By combining player demographic data and performance statistics, the program provides a unified view of cricket player profiles, enabling deeper insights and comprehensive analytics. The implementation highlights the scalability of MapReduce, capable of handling vast amounts of data distributed across multiple nodes. The Mapper and Reducer classes work seamlessly to categorize and merge datasets based on the player_id, ensuring accurate and efficient data processing.

The program not only addresses the challenge of joining datasets in a distributed environment but also lays the groundwork for more complex data analysis tasks in cricket and other domains. Its design can be easily adapted to accommodate additional datasets or extended to perform advanced computations and queries. The structured output generated can serve as a valuable resource for data analysts, cricket enthusiasts, and researchers looking to derive actionable insights. Overall, the program is a testament to the robustness of MapReduce for big data analytics, making it a cornerstone for similar applications in the future**.**