CSOE18 – Big Data Analytics

Name:-	Kshitij Kanade
Roll No:-	107121045
Date:-	02/12/2023
Batch:-	2025
Dept:-	EEE Sec A

Assignment - 2

Hadoop MapReduce for
Climate Data Analytics

Task_3c Answers

Task #3c:-

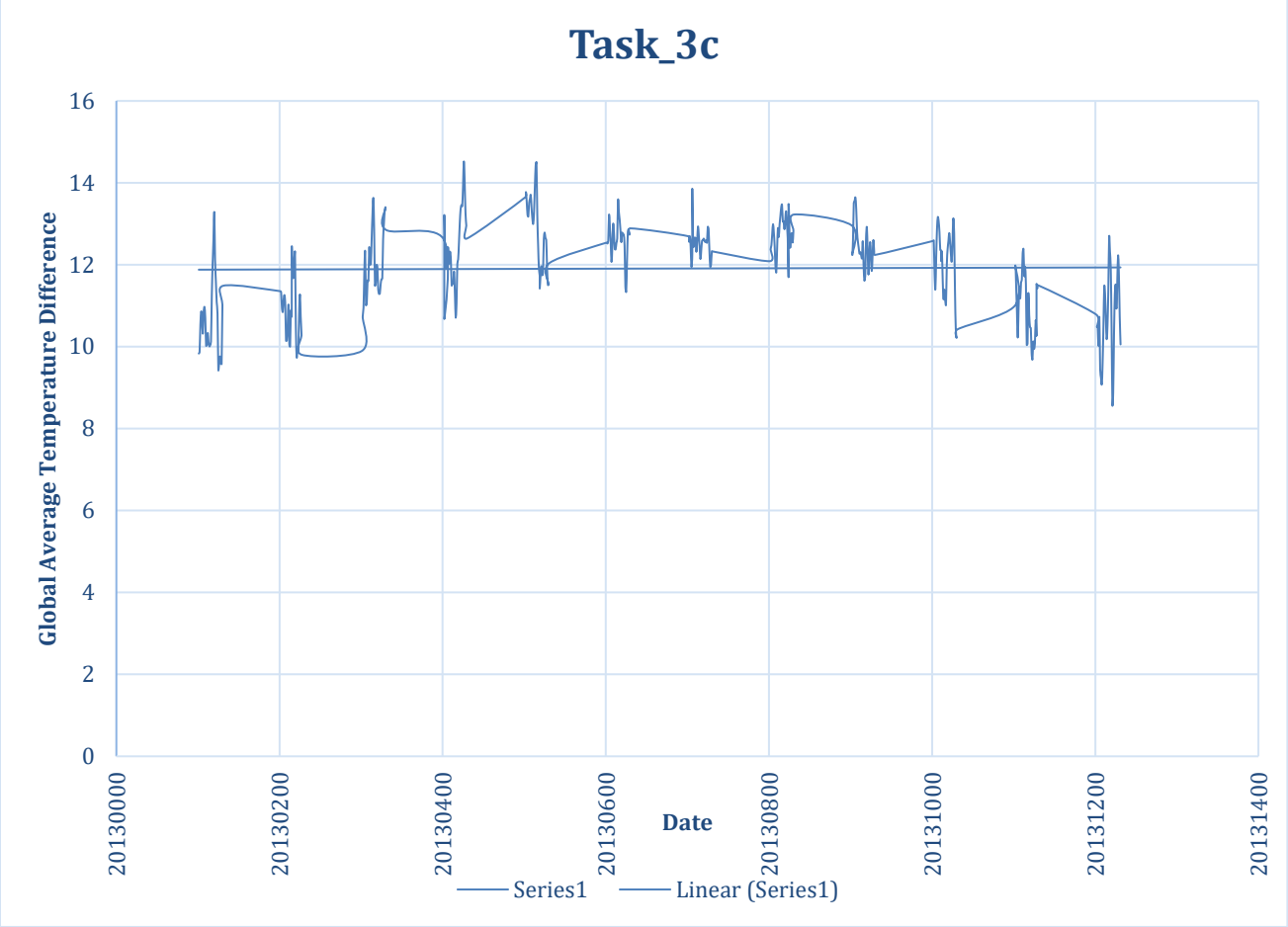
The approach of Question:

This Hadoop MapReduce code calculates the average temperature difference between TMAX and corresponding TMIN records per date. The Mapper extracts relevant data (station code, date, record type, and adjusted temperature) and emits TMAX/TMIN records grouped by date. The Combiner receives these records, identifies TMAX values, looks for corresponding TMIN values for the same date and station, calculates temperature differences, and emits intermediate average differences per date. Finally, the Reducer aggregates these intermediate values per date, computes the overall average temperature difference, and produces the final output with the date as the key and the average temperature difference as the value..

Further possible modifications:-

To enhance this MapReduce job, you might consider a few adjustments. First, optimizing the computation process by using a custom Reducer instead of the default Reducer class could refine the aggregation logic and improve performance. Second, incorporating a secondary sort mechanism within the Reducer to sort temperature values before computing differences could enhance accuracy. Third, implementing a partitioner based on date ranges might balance the workload among reducers for better parallel processing. Additionally, leveraging combiner optimization to pre-aggregate data before it reaches the Reducer can reduce network traffic and improve overall efficiency. Furthermore, considering error handling mechanisms and input data validation would fortify the code against potential issues. Moreover, exploring alternative data structures or compression techniques for intermediate data could reduce storage overhead. Furthermore, enabling speculative execution and tuning Hadoop configuration parameters based on cluster resources and workload characteristics could optimize job execution. Lastly, integrating a comprehensive logging mechanism for better monitoring and debugging and exploring options to handle skewed data distribution across reducers would further refine the MapReduce job.

PLOT:-



References:-

1. The Word Count program on MapReduce demonstrated in class.
2. Regular expressions application in Java.
3. Verbal Discussion with Classmates Rohit Meena, Harshit Malik, Atharv Bhavey, Akanksh Muvva, Vinay.
4. StackOverflow, GeeksForGeeks, etc online platforms.