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_3a Answers

Task #3a:-

The approach of Question:

In the given MapReduce program:

Mapper Phase (TokenizerMapper):

- Reads a CSV file, extracting relevant temperature data (TMAX, TMIN) for the Central Park station (USW00094728).
- Emits key-value pairs with the date as the key and a concatenated string containing maximum and minimum temperatures as the value.

Combiner Phase (TemperatureDifferenceCombiner):

- Gathers temperature data from the Mapper phase.
- Finds the maximum and minimum temperatures for each date, combining these values into a single string per date.

Reducer Phase (TemperatureDifferenceReducer):

- Receives the combined temperature data per date from the Combiner.
- Calculate the temperature difference (TMAX - TMIN) for each date.
- Emits the date as the key and the temperature difference as the value.

Main Method:

- Configures and initiates the MapReduce job, setting input/output paths and types.
- Specifies the Mapper, Combiner, and Reducer classes to execute different tasks.
- Launches the job, waiting for its completion and exiting accordingly.

Overall, the program focuses on processing weather data for the Central Park station, calculating temperature differences between TMAX and TMIN records for each date using MapReduce's distributed computing paradigm.

What is wrong with this approach? Can you propose an alternative solution?

Task_3a suffers from inefficiencies in data processing and unnecessary computation. It lacks filtering at the Mapper phase, emitting all records to the Reducer irrespective of their relevance. This leads to increased data transfer between Map and Reduce phases, higher network overhead, and heavier Reducer computation. Additionally, the Combiner in Task_3a performs partial aggregation but doesn't discard redundant data.

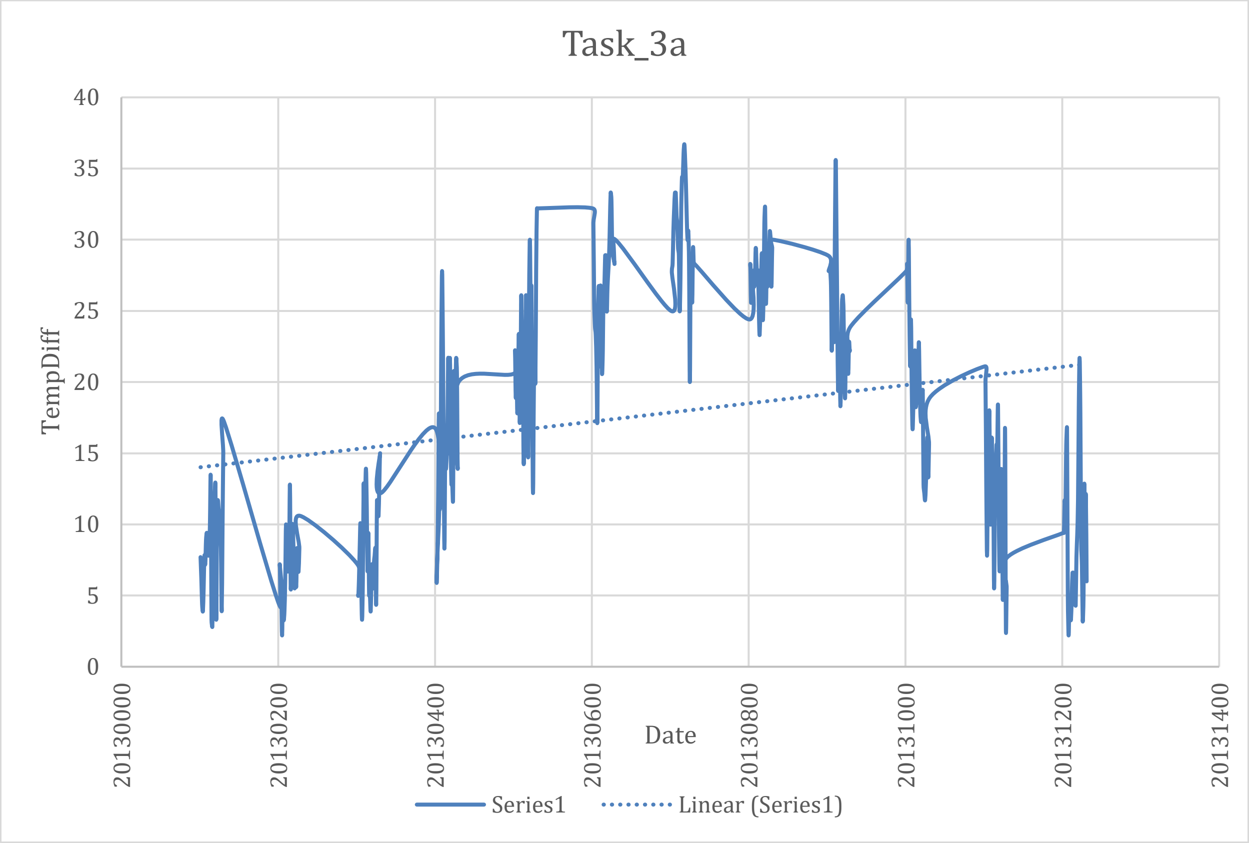
Further possible modifications:-

The enhanced modification for Task_3a involves optimizing the Mapper to filter only essential TMAX and TMIN records and perform temperature difference calculations. By refining the Mapper logic to emit pre-calculated temperature differences for each date, it minimizes Reducer complexity. This modification significantly reduces redundant data transmission, alleviates network congestion, and simplifies Reducer tasks by processing data earlier in the Map phase.

This modification has been implemented as Task_3b, refining the Mapper to efficiently pair TMAX and TMIN records, compute temperature differences, and emit only necessary data to the Reducer. Task_3b's enhancement optimizes the MapReduce process, minimizing redundant computations and data transfer, thereby enhancing overall efficiency and scalability compared to the initial Task_3a approach.

NOTE:- The code files are attached in the form of jar files which can be seen using Eclipse IDE.

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.

Thank You!