Assignment No. 10

Title:
Finding the Coolest/Hottest Year using MapReduce

Problem Statement:
Design and develop a distributed application to find the coolest/hottest year from the available weather data. Use weather data from the Internet and process it using MapReduce.

Theory:

MapReduce for Weather Data Analysis:

- MapReduce is a programming model designed for processing large-scale datasets in a distributed computing environment.
- It consists of two main functions:
  - Map Function: Processes input data and generates key-value pairs.
  - Reduce Function: Aggregates the key-value pairs to compute the final result.
- This model enables parallel processing, scalability, and fault tolerance, making it ideal for analysing large weather datasets.

Process of Finding Coolest/Hottest Year using MapReduce:

1. **Data Collection:** Obtain historical weather data from online sources such as NOAA (National Oceanic and Atmospheric Administration).
2. **Preprocessing:** Extract relevant data such as year, temperature readings, and filter out any missing or corrupt values.
3. **Map Phase:**
   - Parse the dataset and extract temperature readings with their corresponding years.
   - Emit key-value pairs in the format (year, temperature).
4. **Shuffle and Sort:** Group temperature records by year.
5. **Reduce Phase:**
   - Compute the minimum and maximum temperature for each year.
   - Identify the coolest and hottest years based on computed values.
6. **Output:** Display the results showing the coolest and hottest years along with their respective temperature values.

Features of MapReduce:

- Parallel Processing: Distributes computations across multiple nodes to improve efficiency.
- Scalability: Capable of handling massive datasets across distributed systems.
- Fault Tolerance: Automatically recovers from failures using replication mechanisms.
- Flexibility: Can be applied to various data processing tasks, including weather data analysis.

Conclusion:

Thus, we have successfully designed and developed a distributed application to determine the coolest and hottest years from weather data using MapReduce.