BIG DATA ANALYSIS

Hadoop MapReduce for Climate Data Analytics

NAME - SAHIL
ROLL NO. -107121086
ELECTRICAL AND ELECTRONICS ENGINEERING

Task 3

(c) Approach

I have implemented a multi-stage MapReduce job where the output of the first MapReduce class serves as the input to the second MapReduce class. This is a common pattern in MapReduce workflows, often referred to as "chaining MapReduce jobs."

First MapReduce Job (First_class):

1. **Mapper (First_TemperatureMapper):**
   - Reads CSV input data.
   - Filters TMAX and TMIN records.
   - Emits key-value pairs with date+station as key and temperature type + value as value.
2. **Reducer (First_TemperatureReducer):**
   - Groups records by date+station.
   - Calculates temperature difference (TMAX - TMIN).
   - Normalizes temperatures if necessary.
   - Writes date+station as key and temperature difference as value.

Second MapReduce Job (Second_class):

1. **Mapper (Second_TemperatureMapper):**
   - Reads output of the first job.
   - Splits records into tokens using whitespace.
   - Emits key-value pairs with date as key and temperature difference as value.
2. **Reducer (Second_TemperatureReducer):**
   - Groups records by date.
   - Calculates the average temperature difference, ignoring zero differences.
   - Writes date as key and average temperature difference as value.

Execution Steps:

1. **Run First_class:**
   - Input: CSV data
   - Output: Key-value pairs with date+station as key and temperature difference as value.
2. **Run Second_class:**
   - Input: Output of First_class
   - Output: Key-value pairs with date as key and average temperature difference as value.

Plot of Output