Vidyavardhini's College of Engineering and Technology

Department of Artificial Intelligence & Data Science

AY: 2024 - 25

Class:	BE	Semester:	VII
Course Code:		Course Name:	BDA

Name of Student:	BARI ANKIT VINOD
Roll No. :	61
Assignment No.:	2
Title of Assignment:	
Date of Submission:	
Date of Correction:	

Evaluation

Performance Indicator	Max. Marks	Marks Obtained
Completeness	5	4
Demonstrated Knowledge Legibility	3	3
Legibility	2	2
Total	10	9

Performance Indicator	Exceed Expectations (EE)	Meet Expectations (ME)	Below Expectations (BE)
Completeness	5	3-4	1-2
Demonstrated Knowledge Legibility	3	2	1
Legibility	2	1	0

Checked by

Name of Faculty : Ms. Sweety Patil

Signature :

Date : 18/8/25

Assignment No.- 2

DA

A. 1) A reccomendations system company stores customer-item iteractions data as a matrix, where each row representans a customer and each column an item. To calculate personalized scores based on customers preferences, the company wants to perform matrix-vector multiplication at scale using distributed systems. Apply the mapReduce model to solve the matrix-vector multiplication problem. describe how the map and reduce phase would be used to compute the final output efficiently.

→ given, $y = M.v$

∠ That means : $y_i = \sum_{j=1}^{n} M_{ij} \cdot v_j$

- MapReduce - input - matrix rows and vector elements.
  - each matrix entry M_{ij} is paired with the corresponding vector elements v_j.

$$emit(i, M_{ij}, v_j)$$

- efficiency advantage -
- parallelism - each row's computation is independent, so it scales horizontal across many machines.
- locally - vectors v can be broadcast to all mappers since it's small compared to the matrix.
- scalability - hadles very large matrices.

Matrix M : $\begin{bmatrix} 2 & 3 \\ 4 & 5 \end{bmatrix}$, Vector v : $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$

∠ for row 1 : emit(1, 2×1), emit(1, 3×2) → (1,2), (1,6)

∠ for row 2 : emit(2, 4×1), emit(2, 5×2) → (2,4), (2,10)

$\therefore y = \begin{bmatrix} 8 \\ 14 \end{bmatrix}$

Q. 2> A company wants to analyze its website logs to calculate how many times each page was visited in day. The data is stored across multiple servers and processing needs to be distributed due to large volume. Apply the mapReduce would work to solve this problem efficiently. include key steps and any optimizations like use of combinations.

→ Goal : Count how many times each web page visited in a day.

- Data : website log files distributed across multiple servers. each log file typically contains.

  [timestamp] user_id page_url response_code

(i) MapReduce Approach -

- Input - each mapper reads raw log files from distributed storage.
- Processing - extract the page URL from each log entry.
- emit - emit (page_url, 1)

(ii) Combiner -

before data is shuffled across the network, we can combine results locally on each mapper.

Purpose - reduce communication overhead.

Eg. /home : [1, 1, 1, 1] → emit (/home, 4)

(iii) shuffle phase -

- framework groups intermediate key-value pairs by page URL across all servers.
- Eg. /home : [4, 7, 5]
  /contact : [1, 2, 4]

(iv) Reduce phase -

- each reducer gets one key and it's list of counts.
- reducer sums up all values for that page.
- emit final result - emit (page_url, total_visits)
- Eg. /home → 16
  /contact → 7