**PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE**

**ACADEMIC YEAR: 2023-24**

## **DEPARTMENT of COMPUTER ENGINEERING DEPARTMENT**

**CLASS: B.E. SEMESTER: I**

**SUBJECT: LP-IV**

|  |  |
| --- | --- |
| **ASSIGNMENT NO.** | A4 |
| **TITLE** | **Map Reduce Program** |
| **PROBLEM STATEMENT /DEFINITION** | Write a **Map Reduce** program to count the number of occurrences of each alphabetic character in the given dataset. The count of each letter should be case-insensitive. |
| **OBJECTIVE** | Using Map Reduce processing technique to count the number of occurrences of each alphabetic character in the given document. |
| **OUTCOME** | Students will be able to:   * Implement the Map Reduce program to count the occurrences of each alphabetic character in the given dataset. * Understand the working of the Map Reduce program. |
| **S/W PACKAGES AND**  **HARDWARE APPARATUS USED** | Python 3.9.0,  VS Code,  Intel I5 8 GB RAM,  Windows 10 (64-bit) OS |
| **REFERENCES** | 1. C.J. Rijsbergen, &quot;Information Retrieval&quot;, (http://www.dcs.gla.ac.uk/Keith/Preface.html)  2. W.R. Hersh, ―Information Retrieval: A Health and Biomedical Perspective‖, Springer, 2002.  3. G. Kowalski, M.T. Maybury. &quot;Information storage and Retrieval System&quot; , Springer, 2005 |
| **STEPS** | Refer to theory, algorithm, test input, test output |
| **INSTRUCTIONS FOR**  **WRITING JOURNAL** | 1. Date  2. Assignment no.  3. Problem definition  4. Learning objective  5. Learning Outcome  6. Concepts related Theory  7. Algorithm  8. Test cases  10. Conclusion/Analysis |

**Prerequisites:**

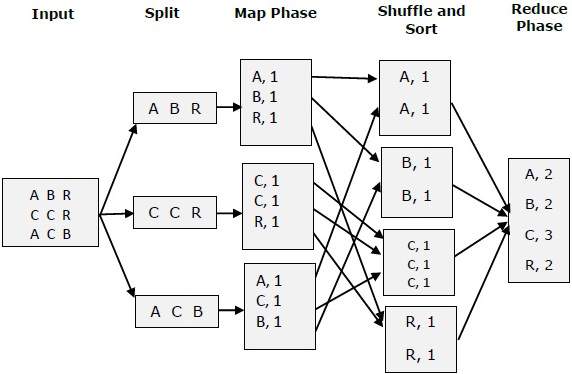
**Concepts related Theory:**

**Map Reduce**

A MapReduce program is composed of a map procedure, which performs filtering and sorting (such as sorting students by first name into queues, one queue for each name), and a reduce method, which performs a summary operation (such as counting the number of students in each queue, yielding name frequencies). The "MapReduce System" (also called "infrastructure" or "framework") orchestrates the processing by marshalling the distributed servers, running the various tasks in parallel, managing all communications and data transfers between the various parts of the system, and providing for redundancy and fault tolerance.

A MapReduce framework (or system) is usually composed of three operations (or steps):

1. **Map:** each worker node applies the map function to the local data, and writes the output to a temporary storage. A master node ensures that only one copy of the redundant input data is processed.
2. **Shuffle:** worker nodes redistribute data based on the output keys (produced by the map function), such that all data belonging to one key is located on the same worker node.
3. **Reduce:** worker nodes now process each group of output data, per key, in parallel.



*Above diagram shows the working process of a MapReduce program.*

**Algorithm:**

* The map takes data in the form of pairs and returns a list of <key, value> pairs. The keys will not be unique in this case.
* Using the output of Map, sort and shuffle are applied by the Hadoop architecture. This sort and shuffle acts on these list of <key, value> pairs and sends out unique keys and a list of values associated with this unique key <key, list(values)>.
* An output of sort and shuffle sent to the reducer phase. The reducer performs a defined function on a list of values for unique keys, and Final output <key, value> will be stored/displayed.

**Conclusion:** Successfully learned and implemented Map Reduce function to count the number of occurrences of each alphabetic character in the given dataset.

**Review Questions**:

Q1. What are the main components of MapReduce Job?

Q2. What is Shuffling and Sorting in MapReduce?

Q3. What is Text Input Format?

Q4. What is a MapReduce Combiner?

Q5. What are the parameters of mappers and reducers?