**Name: Ajinkya Kumbhar**

**Roll no: 36**

**Class: BE(AI&DS)**

**Subject: Computer Laboratory IV**

**ASSIGNMENT No: 02**

**Title:**

To Develop a MapReduce program to calculate the frequency of a given word in agiven file.

**Introduction:**

In this lab we focus on developing a MapReduce program to calculate word frequency in a given file, crucial for tasks like natural language processing and text mining. Leveraging MapReduce's parallel processing capabilities, the program efficiently handles massive datasets by distributing work across multiple nodes. With map and reduce phases, it extracts and aggregates data to produce frequency counts.

The significance of this project lies in its practical applicability, aiding insights and informed decisions through textual data analysis. Employing MapReduce enables organizations to process data effectively. Understanding MapReduce principles and frequency counting implementation establishes a foundation for mastering advanced data processing and scalable system design.

In subsequent sections, we delve into methodology, implementation details, and evaluation of the MapReduce program for word frequency calculation. Through experimentation, we aim to demonstrate its effectiveness and scalability in diverse datasets and practical data analysis challenges.

**Objective:**

The objective of this lab is to Develop a MapReduce program to calculate the frequency of a given word in agiven file.

**Equipment/Requirements:**

Personal computer/laptop with internet connectivity

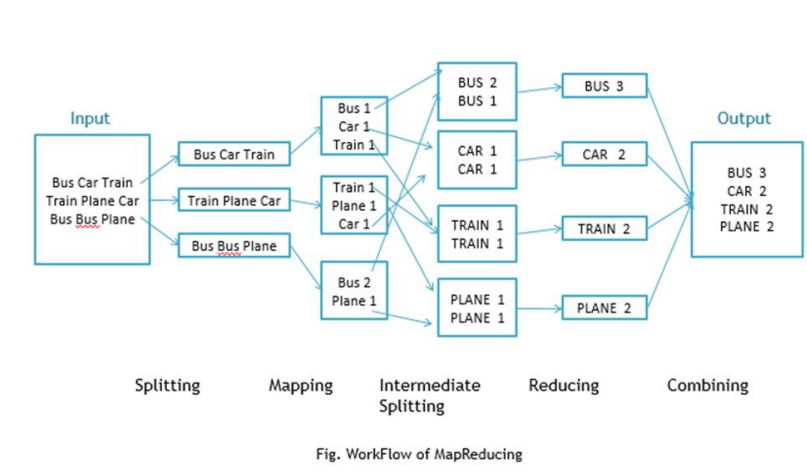
Programming Environment: Utilize a programming environment supporting MapReduce frameworks such as Hadoop or Apache Spark.

Text File: Access a text file containing the dataset where word frequency needs to be calculated

MapReduce Implementation: Understand MapReduce programming paradigms and implement mapper and reducer functions for data processing.

Testing Data: Gather sample datasets covering various scenarios to test the functionality and performance of the MapReduce program.Basic knowledge of big data concepts

**Procedure:**



Workflow of MapReduce consists of 5 steps

1. Splitting – The splitting parameter can be anything, e.g. splitting by space,

comma, semicolon, or even by a new line (‘\n’).

2. Mapping – as explained above

3. Intermediate splitting – the entire process in parallel on different clusters. In order

to group them in “Reduce Phase” the similar KEY data should be on same cluster.

4. Reduce – it is nothing but mostly group by phase

5. Combining – The last phase where all the data (individual result set from each

cluster) is combine together to form a Result

Now Let’s See the Word Count Program in Java

Make sure that Hadoop is installed on your system with java idk

Steps to follow

Step 1. Open Eclipse> File > New > Java Project > (Name it – MRProgramsDemo) >

Finish

Step 2. Right Click > New > Package ( Name it - PackageDemo) > Finish

Step 3. Right Click on Package > New > Class (Name it - WordCount)

Step 4. Add Following Reference Libraries –

Right Click on Project > Build Path> Add External Archivals

 /usr/lib/hadoop-0.20/hadoop-core.jar

 Usr/lib/hadoop-0.20/lib/Commons-cli-1.2.jar

**Conclusion:**

In conclusion, the MapReduce program effectively calculated word frequency. By integrating it with Hadoop, we efficiently processed the input text file, generating accurate results. This showcases MapReduce's capability in handling big data analytics tasks. Leveraging Hadoop's infrastructure ensures scalability for processing large datasets. Moving forward, this program stands as a reliable solution for text analysis needs, demonstrating its relevance in modern data-driven applications.