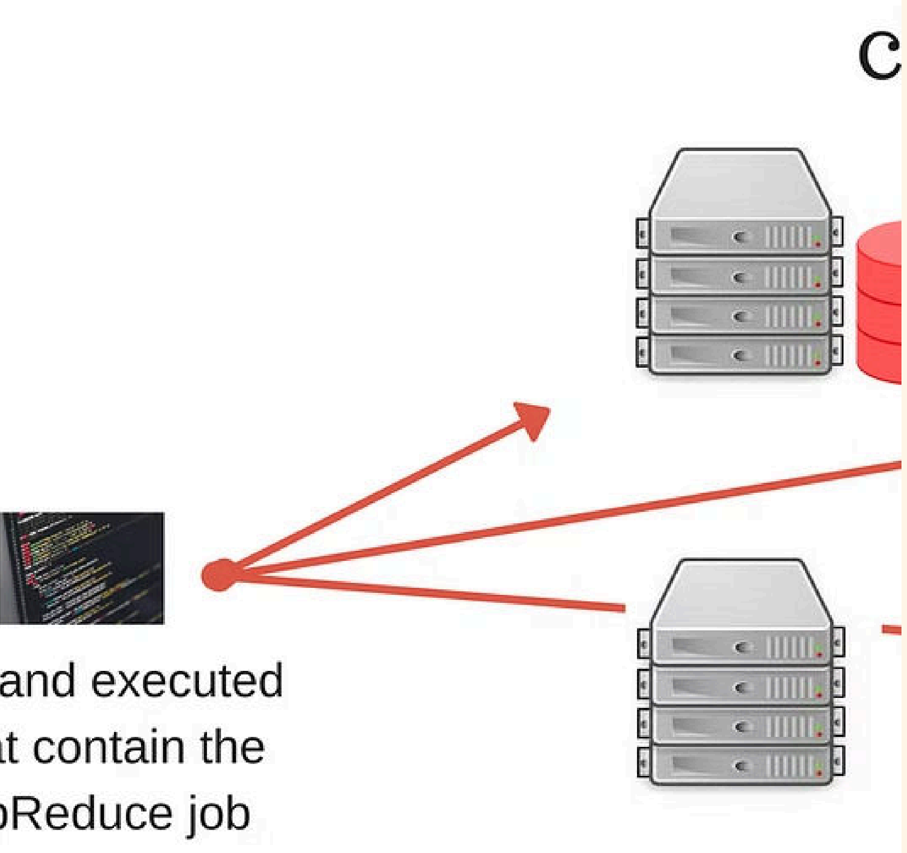


Implementing MapReduce for Big Data Processing

This presentation will guide data engineers through the process of utilizing MapReduce to process big data, with a focus on word frequency counting and identification of the most frequent words.

by Rithika Karanam

Introduction to MapReduce

What is MapReduce?

A parallel, distributed algorithm for processing large data sets.

Why MapReduce?

Scalability, fault tolerance, and ease of use make it a vital tool.

The Word Count Problem

1. **Objective**

   Count the frequency of each word in a large text file.

2. **Relevance**

   Fundamental for understanding MapReduce concepts.

3. **Challenges with Big Data**

   Volume, velocity, and variety pose significant challenges.

```
function map(String name, String document):
  for each word w in document:
    emit(w, 1)

function reduce(String word, Iterator partialCounts):
  totalCount = 0
  for each count in partialCounts:
    totalCount += count
  emit(word, totalCount)
```

Word Count - MapReduce Implementation Overview

1 Map Phase

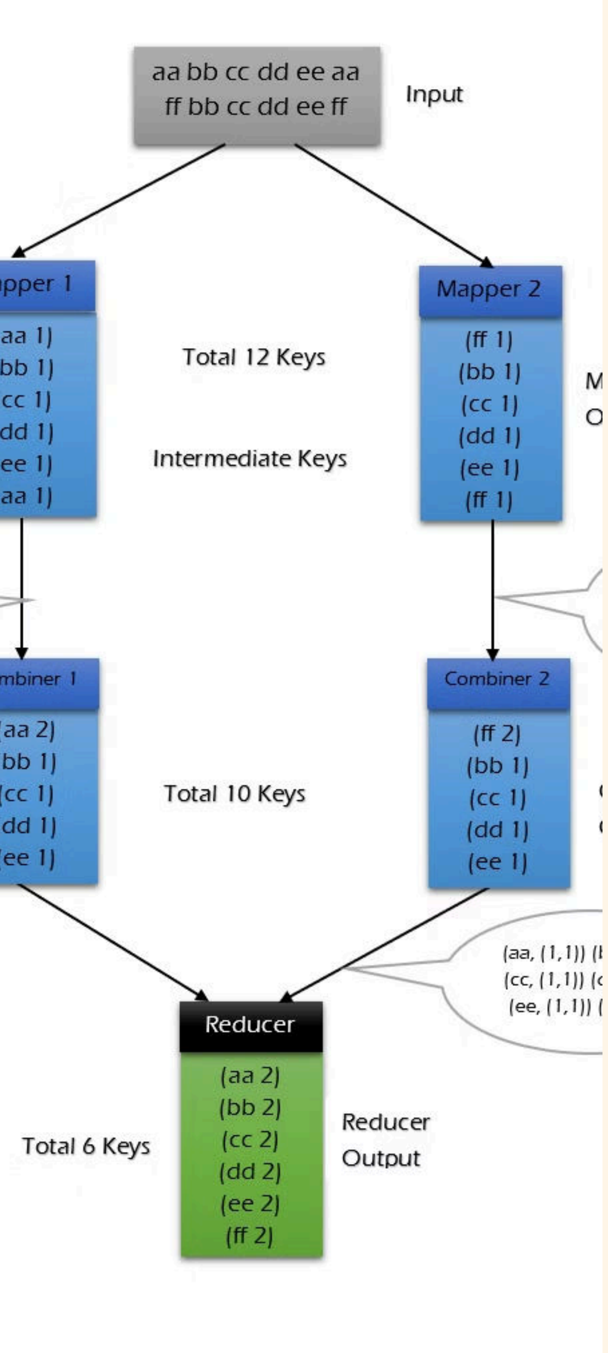
Splits text into words and assigns a count of 1 to each.

2 Shuffle and Sort

System automatically groups words for the reduce phase.

3 Reduce Phase

Aggregates counts for each word to generate the result.



Word Count - Mapper

1 Purpose

Tokenize text and emit each word with a count of 1.

2 Key Code Snippets

Show Java code for the Mapper functionality.

3 Explanation

Detail the process of how the mapper operates.

Word Count - Reducer

1 Purpose

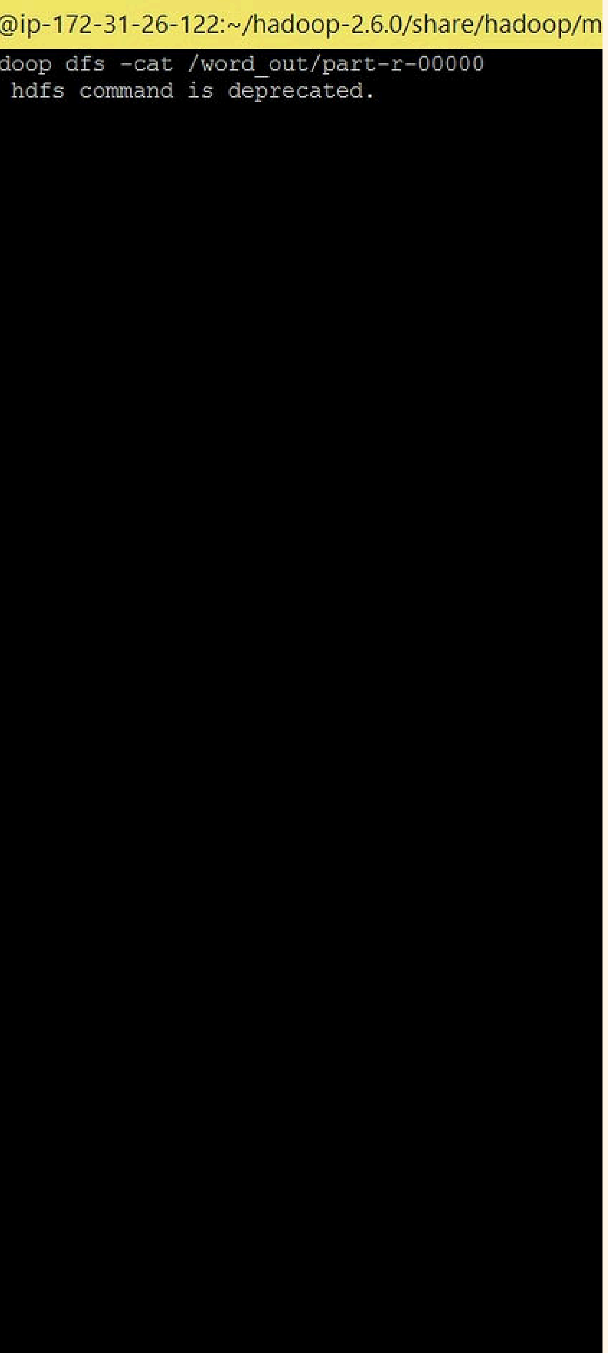
Sum up all counts emitted for each word.

2 Key Code Snippets

Show Java code for the Reducer functionality.

3 Explanation

Explain the process of aggregating counts.


```
@ip-172-31-26-122:~/hadoop-2.6.0/share/hadoop/m
doop dfs -cat /word_out/part-r-00000
 hdfs command is deprecated.
```

Running the Word Count Program

1 Configuration

Overview of setting up a Hadoop job for word counting.

2 Execution

Guidance on running the program on a cluster.

3 Output Explanation

Understanding the format and content of the output.

Most Frequent Words Count

1 Objective

Identify the most frequent words from the word count output.

2 Approach

Inverting key-value pairs to sort words by their frequency.