Fall 2023-2024

CS 403/534 - Distributed Systems

Programming Assignment (PA) 2 – Report

Meriç Mert Bulca – 28198

Problem Definition:

In this project, I implemented a MapReduce framework using ZeroMQ socket communication. The primary objective was to develop an application using that could efficiently process a large dataset containing paper citations, calculate the number of citations for each paper, and detect cyclic citations using the framework.

Approach:

The implementation follows the steps outlined in the project document:

Implementation of MapReduce class and its sub-classes: I defined an abstract class "MapReduce" with its two concrete sub-classes "FindCitations" and "FindCyclicReferences" and implemented their methods to perform the operations.

Initialization: I initialize the MapReduce class with the number of worker processes.

Data Processing: The start method of MapReduce initializes all processes, reads the input file, prepares a list, and sends it to the producer process which divides the data into chunks. This producer, then, distributes them to the consumer processes concurrently via ZeroMQ Pipeline.

Map Operation: Each worker process performs the respective map operation against its own data chunk.

Partial Results: Partial results resulting from map operations, in the form of dictionaries, are sent to the ResultCollector process via ZeroMQ sockets.

Reduce Operation: The ResultCollector collects partial results, combines them, and performs the reduce operation to obtain the final result.