**Report**  
  
**Part 1a: Setting up Hadoop in Docker:**

Used to run the WordCount example program: bin/hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-3.2.1.jar wordcount input/ output/

Used to concatenate and display the contents of all files in the "output" directory in HDFS: hdfs dfs -cat output/\*

**Part 2: Developing a Hadoop program (N-Gram):**

To transfer required file to HDFS directory /part2: hdfs dfs -put ngram\_input.txt /part2 hdfs dfs -put ngram\_mapper.py /part2 hdfs dfs -put ngram\_reducer.py /part2

Ran a Hadoop streaming job using custom Python scripts to process n-gram frequencies in the input file: bin/hadoop jar share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar

-files ngram\_mapper.py,ngram\_reducer.py

-mapper "python3 ngram\_mapper.py"

-reducer "python3 ngram\_reducer.py"

-input /ngram\_input.txt

-output /output\_ngram

hdfs dfs -cat /output\_ngram/\*

**Part 3: Developing a Hadoop program to analyze real logs**

To transfer required file to HDFS directory /part3: hdfs dfs -put access\_log /part3 hdfs dfs -put part3\_mapper.py /part3 hdfs dfs -put part3\_reducer.py /part3

bin/hadoop jar share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar

-file part3\_mapper.py,part3\_reducer.py

-mapper "python3 part3\_mapper.py"

-reducer "python3 part3\_reducer.py"

-input /access\_log

-output /output\_part3final

hdfs dfs -cat /output\_part3final/\*

bin/hadoop jar share/hadoop/tools/lib/hadoop-streaming-3.2.1.jar

-file part3\_q4910\_mapper.py,part3\_q4910\_reducer.py

-mapper "python3 part3\_q4910\_mapper.py"

-reducer "python3 part3\_q4910\_reducer.py"

-input /access\_log

-output /output\_part3q4910

hdfs dfs -cat /output\_part3q4910/\*

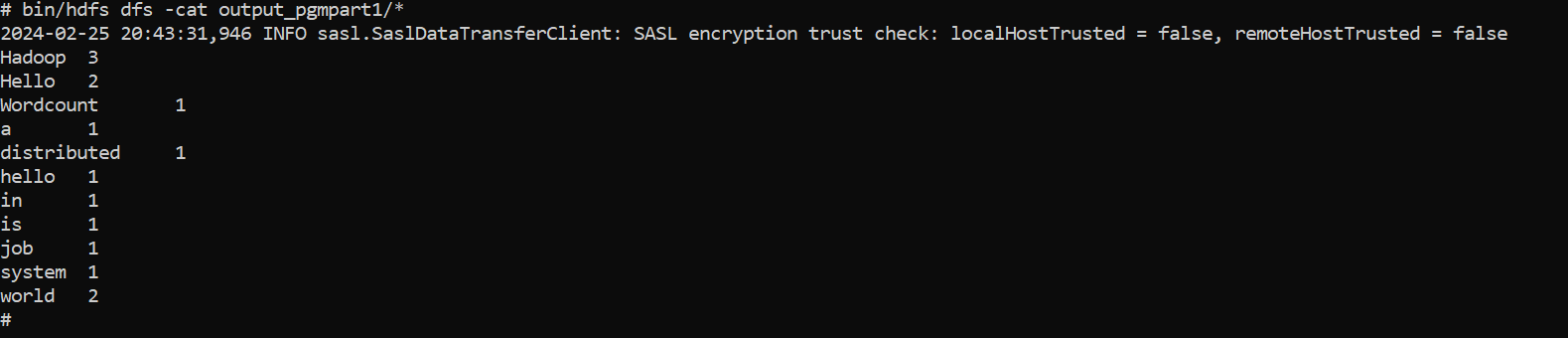
**Part 1: Setting up Hadoop in Docker:**

We proceeded with constructing the Hadoop cluster by adhering to the prescribed sequence of preparatory actions for Hadoop installation and execution. Subsequently, we initiated the cluster and executed the default wordcount program, an integral component of the Hadoop package.

1.Created a small Ubuntu Docker image as a foundational step for virtualized environment setup.

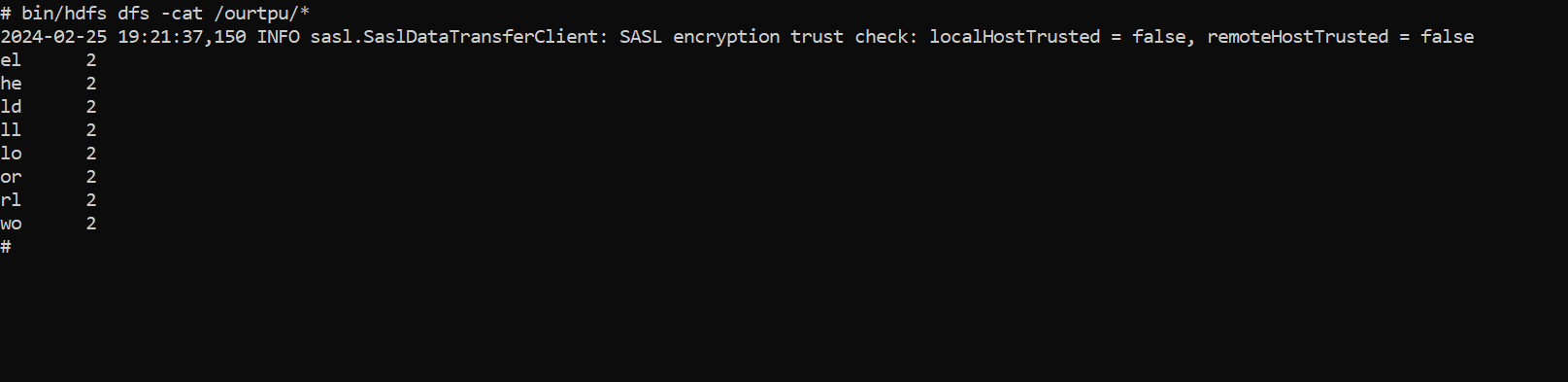
2.Developed a Hadoop Docker image based on the Ubuntu image, implementing basic Hadoop functionalities for local task execution using provided bash files.

3.Tested the Hadoop Docker image by successfully running a Wordcount job locally, including Dockerfile, support files, and a bootstrap script for service startup in the submission.



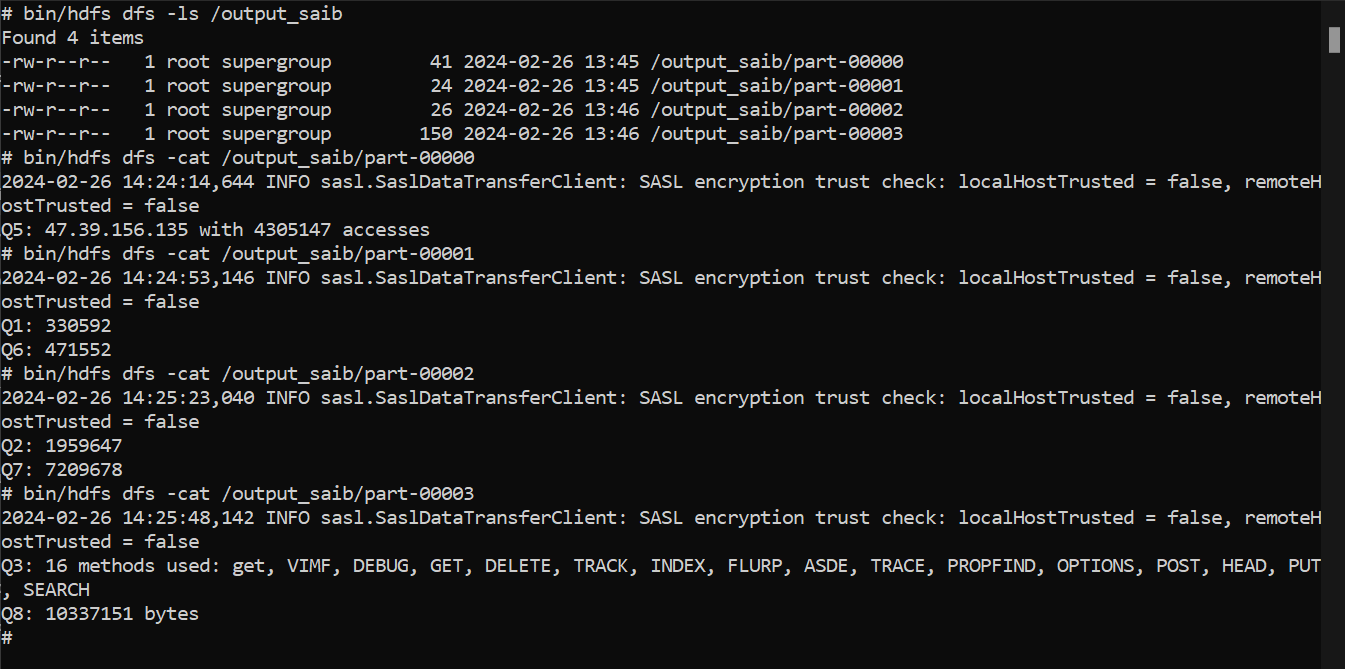
**Part 2: Developing a Hadoop program (N-Gram)**

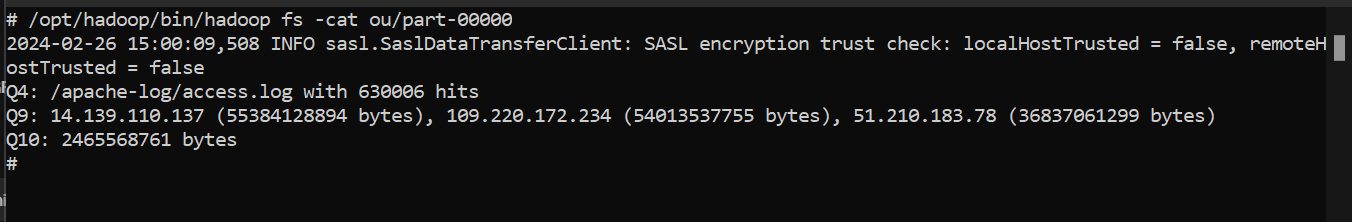
Developed a Hadoop program in Java to compute n-gram frequencies from a given text file. Implemented Mapper and Reducer classes to split the text into n-grams and aggregate the counts.



**Part 3: Developing a Hadoop program to analyze real logs**

Developed MapReduce programs in Hadoop to analyze real anonymous logs stored in an access\_log file. Focused on answering specific questions based on the log data, disregarding the last two fields of the log entries. Implemented solutions without utilizing information from the ignored fields for the problems presented.





**Problems:**

1.How many hits were made to the website directory “/images/smilies/” (including subdirectories and files)?

Sol:330592

2.How many hits were made from the IP: 96.32.128.5?

Sol:1959647

3.How many HTTP request methods are used in this file? What are they?

Sol:16 methods used: get, VIMF, DEBUG, GET, DELETE, TRACK, INDEX, FLURP, ASDE, TRACE, PROPFIND, OPTIONS, POST, HEAD, PUT, SEARCH

4.Which path in the website has been hit most? How many hits were made to the path?

Sol: /apache-log/access.log with 630006 hits

5.Which IP accesses the website most? How many accesses were made by it?

Sol: 47.39.156.135 WITH 4305147 accesses

6.How many POST request were made?

Sol:471552

7.How many requests received a 404 status code?

Sol:7209678

8.How much data was requested on 19/Dec/2020?

Sol:10337151 bytes

9.List 3 IPs that access the most, and what is the total data flow size of each IP?

Sol:14.139.110.137 (55384128894 bytes), 109.220.172.234 (54013537755 bytes), 51.210.183.78

10.How much data(in bytes) was successfully(with status code 200) requested on 16/Jan/2022?

Sol:2465568761 bytes