HW1

1) запуск standalone сервиса Hadoop:

```
cd ./hadoop/
docker compose up -d
cd ..
```

2) берем название сети контейнера с Hadoop для запуска нашего основного докера `docker network ls` -> `hadoop_default`

3) Билдим контейнер `docker build -t hw1 .`:

```
FROM apache/hadoop:3.3.6

COPY /data /data

CMD ["bash"]
```

4) Запсукаем основной контейнер `docker run --rm -it --network hadoop_default hw1`

Задания:

1) Создаем директорию /createme:

```
hdfs dfs -fs hdfs://master:8020 -mkdir -p /createme
```

2) Удаляем директорию /delme:

```
hdfs dfs -fs hdfs://master:8020 -rm -r -f /delme
```

3) Создаем файл с произвольным текстом nonnull.txt:

```
echo "               - " | hdfs dfs -fs hdfs://master:8020 -put -f - /nonnull.txt
hdfs dfs -fs hdfs://master:8020 -cat /nonnull.txt
```

4) Добавляем файл с рыбатекстом и словами 'Innsmouth' (`/data/shadow.txt`):

```
hdfs dfs -fs hdfs://master:8020 -put -f /data/shadow.txt /shadow.txt
```

5) Считаем количество всех слов (токенов) чкркз MR wordcount:

```
hdfs dfs -fs hdfs://master:8020 -rm -r -f /wc_out #

hadoop jar \
  /opt/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.6.jar \
  wordcount \
  hdfs://master:8020/shadow.txt \
  hdfs://master:8020/wc_out
```

Краткий вывод:

```
...
2025-12-15 20:23:49 INFO  Job:1773 - Counters: 36
        File System Counters
                FILE: Number of bytes read=566136
                FILE: Number of bytes written=1851197
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=6524
                HDFS: Number of bytes written=1075
                HDFS: Number of read operations=15
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=28
                Map output records=506
                Map output bytes=5265
                Map output materialized bytes=1557
                Input split bytes=94
                Combine input records=506
                Combine output records=121
                Reduce input groups=121
                Reduce shuffle bytes=1557
                Reduce input records=121
                Reduce output records=121
                Spilled Records=242
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=61
                Total committed heap usage (bytes)=1970274304
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=3262
        File Output Format Counters
                Bytes Written=1075
```

6) Считаем количество вхождения токена ‘Innsmouth’:

```
hdfs dfs -fs hdfs://master:8020 -cat /wc_out/* | awk '$1=="Innsmouth"{print $2}'
```

Вывод: - 3 (тк `Innsmouth`!=`Innsmouth.` или `Innsmouth\`)