МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Систем обработки информации и управления»

ОТЧЕТ

**Лабораторная работа №\_\_2\_\_**

по дисциплине«Технология обработки больших данных»

Тема: «Обработка и анализ больших данных на примере статистики сообщений пользователей»

ИСПОЛНИТЕЛЬ: \_\_Журавлев Н. В.\_\_

ФИО

группа ИУ5-44М \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

подпись

"19"\_\_марта\_\_\_2025 г.

ПРЕПОДАВАТЕЛЬ: \_\_\_\_Сухобоков А.В.\_\_\_\_\_

ФИО

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

подпись

"19"\_марта\_\_\_2025 г.

Москва - 2024

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Описание задачи и алгоритма

Этап 1. Краткое изучение Apache Hadoop и технологии MapReduce (теоретический этап)

Изучите темы:

- Основы Apache Hadoop и экосистемы

- Что такое Hadoop Distributed File System (HDFS)

- Основы подхода MapReduce для обработки больших данных

Этап 2. Установка и настройка Hadoop на локальном ПК

- Скачайте Apache Hadoop на официальном сайте.

- Настройте Hadoop в pseudo-distributed режиме на локальном ПК (Linux или Windows через WSL).

- Убедитесь, что Hadoop успешно установлен и работает.

Этап 3. Подготовка и загрузка исходных данных в HDFS

Используйте файл (или набор файлов), созданный ранее в лабораторной работе №1. Копирование и генерация новых файлов не требуется — необходимо переиспользовать тот же большой CSV-файл (~500МБ–1ГБ).

Этап 4. Реализация задачи обработки с помощью MapReduce

Реализуйте ту же задачу, что и в лабораторной №1, используя технологию MapReduce:

- (A) Анализ активности пользователей (топ-10)

или

- (B) Средняя длина сообщений по дням недели

или

- (C) Подсчёт частоты популярных слов

Используйте язык Python (Hadoop Streaming) или Java для реализации программы обработки.

Этап 5. Запуск MapReduce-задачи

Запустите написанную вами программу в Hadoop, используя Hadoop Streaming или jar-файл Java-программы, проверьте результат выполнения.

Этап 6. Сравнительный анализ решений

- Оцените время работы, загрузку процессора и оперативной памяти при использовании оригинального решения (лабораторная №1) и Hadoop-MapReduce (лабораторная №2).

- Проанализируйте преимущества и недостатки обоих подходов по следующим критериям: простота разработки, масштабируемость, время выполнения, использование вычислительных ресурсов.

# Введение

HDFS (Hadoop Distributed File System) — распределенная файловая система Hadoop для хранения файлов больших размеров с возможностью потокового доступа к информации, поблочно распределённой по узлам вычислительного кластера, который может состоять из произвольного аппаратного обеспечения. Hadoop Distributed File System, как и любая файловая система – это иерархия каталогов с вложенными в них подкаталогами и файлами.

Для запуска необходимых серверных процессов выполняются команды:

sbin/start-dfs.sh

sbin/start-yarn.sh

Затем в hdfs перемещаются файлы для работы:

hdfs dfs -mkdir /lab3

hdfs dfs -put bigdata.csv /lab3

hdfs dfs -put mapper.py /

hdfs dfs -put reducer.py /

Для запуска используется команда:

/home/name/hadoop/hadoop-3.4.1/bin/hadoop jar $HADOOP\_HOME/share/hadoop/tools/lib/hadoop-streaming-\*.jar \

-files mapper.py,reducer.py \

-input /lab3\_input -output /output

Реализация представлена на листинг 1 и листинг 2.

Листинг 1. Листинг mapper

|  |
| --- |
| #!/usr/bin/python3  """mapper.py"""  import sys  local\_result = defaultdict(list)  for row in sys.stdin:  msg = row[3].replace(".", "")  # Получить все слова  all\_words = msg.split()  # Занести все полученные слова в результат  for w in all\_words:  print(local\_result[w].append(msg.count(w))) |

Листинг 2. Листинг reducer

|  |
| --- |
| #!/usr/bin/python3  """reducer.py"""  from operator import itemgetter  import sys  res = defaultdict(list)  for local\_result in sys.stdin:  for word, count\_repeat in local\_result.items():  res[word].extend(count\_repeat)  result\_reduce = res  sorted\_repeated\_word = {word: max(count\_repeat) for word, count\_repeat in result\_reduce.items()}  # Сортируем дни по убыванию  sorted\_days = sorted(sorted\_repeated\_word.items(), key=lambda item: item[1], reverse=True)  # Вывод результата  print("Слова и их повторами:")  max\_word = ""  max\_count = 0  i = 0  for word, count\_repeat in sorted\_days:  if word not in STOP\_WORD:  i += 1  print(f"{word}: Повторов {count\_repeat:.2f}")  if i == 20:  break  if i == 1:  max\_word = word  max\_count = count\_repeat  print(f"\nСамое часто употребляемое слово: {max\_word} ({max\_count:.2f} раз)") |

# Полученные результаты

В результате выполнения получились результаты, представленные в таблице 1.

Таблица 1. Результат выполнения на маленьких даннных

|  |
| --- |
| 2025-03-22 13:03:22,282 INFO mapreduce.Job: Running job: job\_1742635527294\_0002  2025-03-22 13:03:27,360 INFO mapreduce.Job: Job job\_1742635527294\_0002 running in uber mode : false  2025-03-22 13:03:27,361 INFO mapreduce.Job: map 0% reduce 0%  2025-03-22 13:03:32,487 INFO mapreduce.Job: map 100% reduce 0%  2025-03-22 13:03:38,532 INFO mapreduce.Job: map 100% reduce 100%  2025-03-22 13:03:38,543 INFO mapreduce.Job: Job job\_1742635527294\_0002 completed successfully  2025-03-22 13:03:38,626 INFO mapreduce.Job: Counters: 54  File System Counters  FILE: Number of bytes read=124029224  FILE: Number of bytes written=248990961  FILE: Number of read operations=0  FILE: Number of large read operations=0  FILE: Number of write operations=0  HDFS: Number of bytes read=114362948  HDFS: Number of bytes written=122386981  HDFS: Number of read operations=11  HDFS: Number of large read operations=0  HDFS: Number of write operations=2  HDFS: Number of bytes read erasure-coded=0  Job Counters  Launched map tasks=2  Launched reduce tasks=1  Data-local map tasks=2  Total time spent by all maps in occupied slots (ms)=6101  Total time spent by all reduces in occupied slots (ms)=3195  Total time spent by all map tasks (ms)=6101  Total time spent by all reduce tasks (ms)=3195  Total vcore-milliseconds taken by all map tasks=6101  Total vcore-milliseconds taken by all reduce tasks=3195  Total megabyte-milliseconds taken by all map tasks=6247424  Total megabyte-milliseconds taken by all reduce tasks=3271680  Map-Reduce Framework  Map input records=1000000  Map output records=1000000  Map output bytes=121688893  Map output materialized bytes=124029230  Input split bytes=188  Combine input records=0  Combine output records=0  Reduce input groups=1000000  Reduce shuffle bytes=124029230  Reduce input records=1000000  Reduce output records=1000000  Spilled Records=2000000  Shuffled Maps =2  Failed Shuffles=0  Merged Map outputs=2  GC time elapsed (ms)=721  CPU time spent (ms)=9110  Physical memory (bytes) snapshot=1573617664  Virtual memory (bytes) snapshot=7689981952  Total committed heap usage (bytes)=1622147072  Peak Map Physical memory (bytes)=483336192  Peak Map Virtual memory (bytes)=2560425984  Peak Reduce Physical memory (bytes)=609345536  Peak Reduce Virtual memory (bytes)=2570432512  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=114362760  File Output Format Counters  Bytes Written=122386981  2025-03-22 13:03:38,627 INFO streaming.StreamJob: Output directory: /output1 |

Таблица 2. Результат выполнения на средних данных

|  |
| --- |
| 2025-03-22 13:06:34,564 INFO mapreduce.Job: Running job: job\_1742635527294\_0003  2025-03-22 13:06:39,644 INFO mapreduce.Job: Job job\_1742635527294\_0003 running in uber mode : false  2025-03-22 13:06:39,645 INFO mapreduce.Job: map 0% reduce 0%  2025-03-22 13:06:46,843 INFO mapreduce.Job: map 20% reduce 0%  2025-03-22 13:06:49,867 INFO mapreduce.Job: map 60% reduce 0%  2025-03-22 13:06:50,877 INFO mapreduce.Job: map 80% reduce 0%  2025-03-22 13:06:51,884 INFO mapreduce.Job: map 100% reduce 0%  2025-03-22 13:07:00,212 INFO mapreduce.Job: map 100% reduce 100%  2025-03-22 13:07:03,241 INFO mapreduce.Job: Job job\_1742635527294\_0003 completed successfully  2025-03-22 13:07:03,315 INFO mapreduce.Job: Counters: 55  File System Counters  FILE: Number of bytes read=1202593058  FILE: Number of bytes written=1824795942  FILE: Number of read operations=0  FILE: Number of large read operations=0  FILE: Number of write operations=0  HDFS: Number of bytes read=572001215  HDFS: Number of bytes written=616013363  HDFS: Number of read operations=20  HDFS: Number of large read operations=0  HDFS: Number of write operations=2  HDFS: Number of bytes read erasure-coded=0  Job Counters  Killed map tasks=1  Launched map tasks=5  Launched reduce tasks=1  Data-local map tasks=5  Total time spent by all maps in occupied slots (ms)=39282  Total time spent by all reduces in occupied slots (ms)=8693  Total time spent by all map tasks (ms)=39282  Total time spent by all reduce tasks (ms)=8693  Total vcore-milliseconds taken by all map tasks=39282  Total vcore-milliseconds taken by all reduce tasks=8693  Total megabyte-milliseconds taken by all map tasks=40224768  Total megabyte-milliseconds taken by all reduce tasks=8901632  Map-Reduce Framework  Map input records=5000000  Map output records=5000000  Map output bytes=608635793  Map output materialized bytes=620337743  Input split bytes=440  Combine input records=0  Combine output records=0  Reduce input groups=5000000  Reduce shuffle bytes=620337743  Reduce input records=5000000  Reduce output records=5000000  Spilled Records=14692969  Shuffled Maps =5  Failed Shuffles=0  Merged Map outputs=5  GC time elapsed (ms)=2504  CPU time spent (ms)=37010  Physical memory (bytes) snapshot=2423894016  Virtual memory (bytes) snapshot=15423762432  Total committed heap usage (bytes)=2398617600  Peak Map Physical memory (bytes)=556048384  Peak Map Virtual memory (bytes)=2568237056  Peak Reduce Physical memory (bytes)=462323712  Peak Reduce Virtual memory (bytes)=2596925440  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=572000775  File Output Format Counters  Bytes Written=616013363  2025-03-22 13:07:03,316 INFO streaming.StreamJob: Output directory: output2 |

Таблица 3. Результат выполнения больших данных

|  |
| --- |
| 2025-03-22 13:10:05,324 INFO mapreduce.Job: Running job: job\_1742635527294\_0004  2025-03-22 13:10:11,689 INFO mapreduce.Job: Job job\_1742635527294\_0004 running in uber mode : false  2025-03-22 13:10:11,691 INFO mapreduce.Job: map 0% reduce 0%  2025-03-22 13:10:29,914 INFO mapreduce.Job: map 48% reduce 0%  2025-03-22 13:10:34,053 INFO mapreduce.Job: map 51% reduce 0%  2025-03-22 13:10:35,117 INFO mapreduce.Job: map 67% reduce 0%  2025-03-22 13:10:44,511 INFO mapreduce.Job: map 78% reduce 0%  2025-03-22 13:10:47,528 INFO mapreduce.Job: map 89% reduce 0%  2025-03-22 13:10:50,547 INFO mapreduce.Job: map 100% reduce 0%  2025-03-22 13:10:53,568 INFO mapreduce.Job: map 100% reduce 11%  2025-03-22 13:10:59,598 INFO mapreduce.Job: map 100% reduce 67%  2025-03-22 13:11:05,632 INFO mapreduce.Job: map 100% reduce 82%  2025-03-22 13:11:12,938 INFO mapreduce.Job: map 100% reduce 91%  2025-03-22 13:11:15,953 INFO mapreduce.Job: map 100% reduce 100%  2025-03-22 13:11:17,979 INFO mapreduce.Job: Job job\_1742635527294\_0004 completed successfully  2025-03-22 13:11:18,049 INFO mapreduce.Job: Counters: 55  File System Counters  FILE: Number of bytes read=2481317874  FILE: Number of bytes written=3725085389  FILE: Number of read operations=0  FILE: Number of large read operations=0  FILE: Number of write operations=0  HDFS: Number of bytes read=1143986522  HDFS: Number of bytes written=1234240390  HDFS: Number of read operations=32  HDFS: Number of large read operations=0  HDFS: Number of write operations=2  HDFS: Number of bytes read erasure-coded=0  Job Counters  Killed map tasks=1  Launched map tasks=9  Launched reduce tasks=1  Data-local map tasks=10  Total time spent by all maps in occupied slots (ms)=160066  Total time spent by all reduces in occupied slots (ms)=38744  Total time spent by all map tasks (ms)=160066  Total time spent by all reduce tasks (ms)=38744  Total vcore-milliseconds taken by all map tasks=160066  Total vcore-milliseconds taken by all reduce tasks=38744  Total megabyte-milliseconds taken by all map tasks=163907584  Total megabyte-milliseconds taken by all reduce tasks=39673856  Map-Reduce Framework  Map input records=10000000  Map output records=10000000  Map output bytes=1217255694  Map output materialized bytes=1240658910  Input split bytes=828  Combine input records=0  Combine output records=0  Reduce input groups=10000000  Reduce shuffle bytes=1240658910  Reduce input records=10000000  Reduce output records=10000000  Spilled Records=30000000  Shuffled Maps =9  Failed Shuffles=0  Merged Map outputs=9  GC time elapsed (ms)=3554  CPU time spent (ms)=65970  Physical memory (bytes) snapshot=4712181760  Virtual memory (bytes) snapshot=25670389760  Total committed heap usage (bytes)=4759486464  Peak Map Physical memory (bytes)=545759232  Peak Map Virtual memory (bytes)=2568908800  Peak Reduce Physical memory (bytes)=519884800  Peak Reduce Virtual memory (bytes)=2583207936  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=1143985694  File Output Format Counters  Bytes Written=1234240390 |

# Сравнительный анализ решений

Сравнение с предыдущей работой, представлено в таблице 2.

Таблица 2. Результат выполнения из предыдущей лабораторной

|  |  |  |  |
| --- | --- | --- | --- |
| Размер | 109 МБ | 545 МБ | 1038 МБ |
| Время выполнения (лабораторная 1) | 5.57 с | 31.41 с | 67.24 с |
| Время выполнения (лабораторная 2) | 16.34 с | 28.75 с | 72.23 с |

Время выполнения на большом размере файла незначительно увеличилось, а на маленьких значительно. На среднем значении время выполнения незначительно сократилось.

# Вывод

Были получены навыки установки и использования Apache Hadoop локально, применение технологии MapReduce, а также выбора подхода и инструментов обработки больших данных и умение определить преимущества и недостатки различных реализаций на практике.