**Федеральное агентство связи**

**Ордена трудового Красного Знамени**

**Федеральное государственное бюджетное образовательное учреждение высшего образования**

**«Московский технический университет связи и информатики»**



**Практическая работа № 2**

**По дисциплине**

**Введение в большие данные**

Группа: МБД2431

ФИО: Киреев Артём Александрович

**Москва, 2025**

**Цель работы: получить навыки работы с MapReduce и YARN**

Первый запуск

yarn jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-mapreduce-examples.jar pi 5 123456789

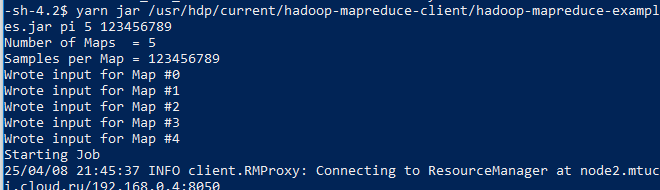




Рисунок 1 - Запуск MapReduce для вычисления числа π

yarn jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-mapreduce-examples.jar pi 5 12345678987

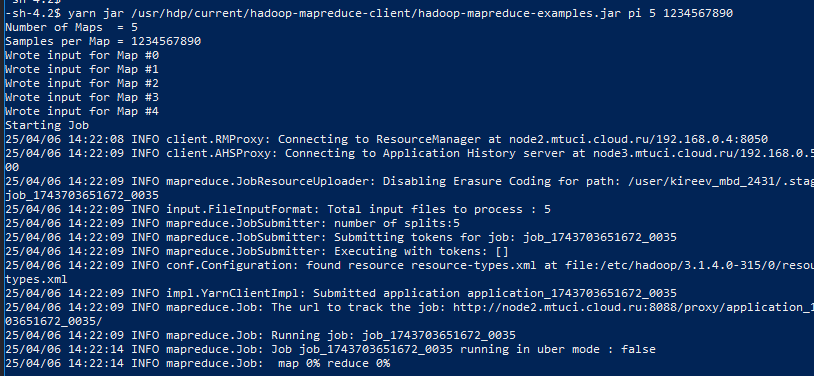
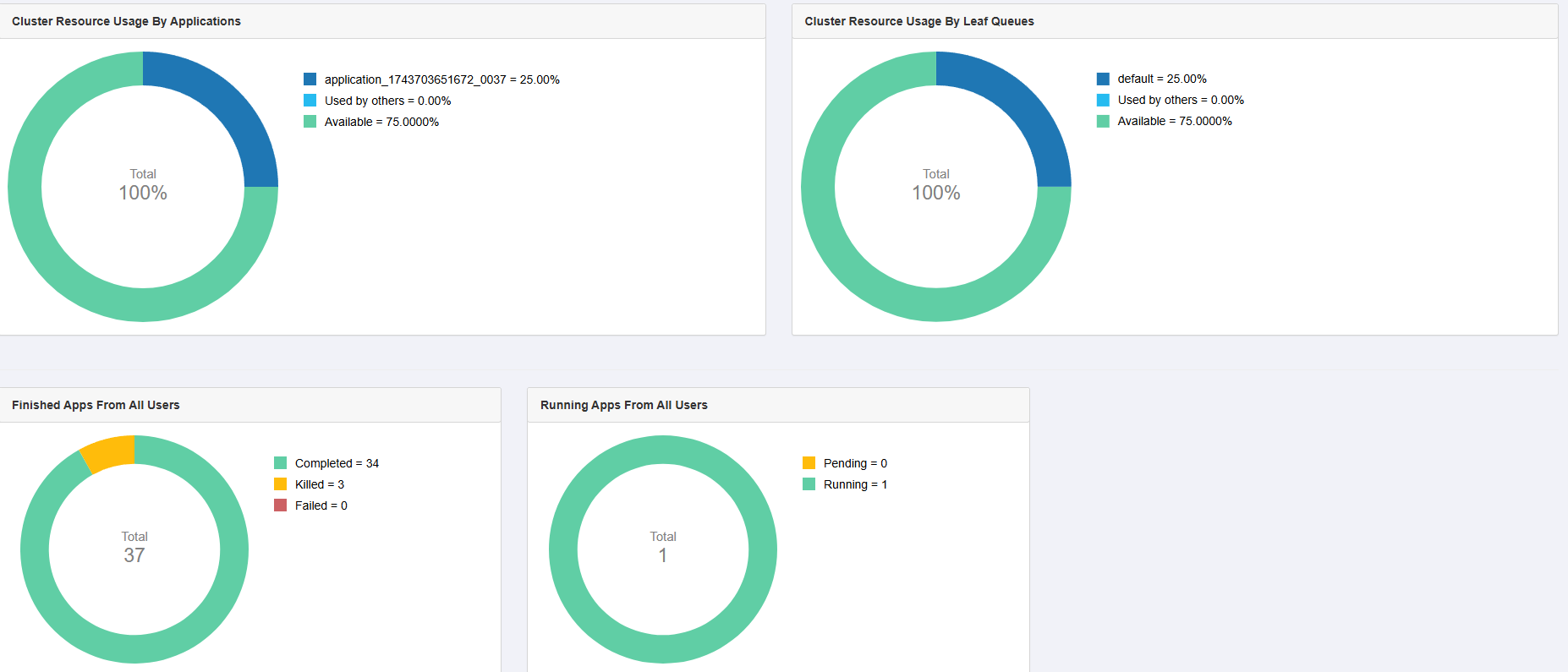




Рисунок 2 - Второй запуск (с увеличенным количеством точек)

Войди в Ambari под пользователем monitor/monitor и перейди в раздел Services -> YARN -> Quick Links -> ResourceManager UI



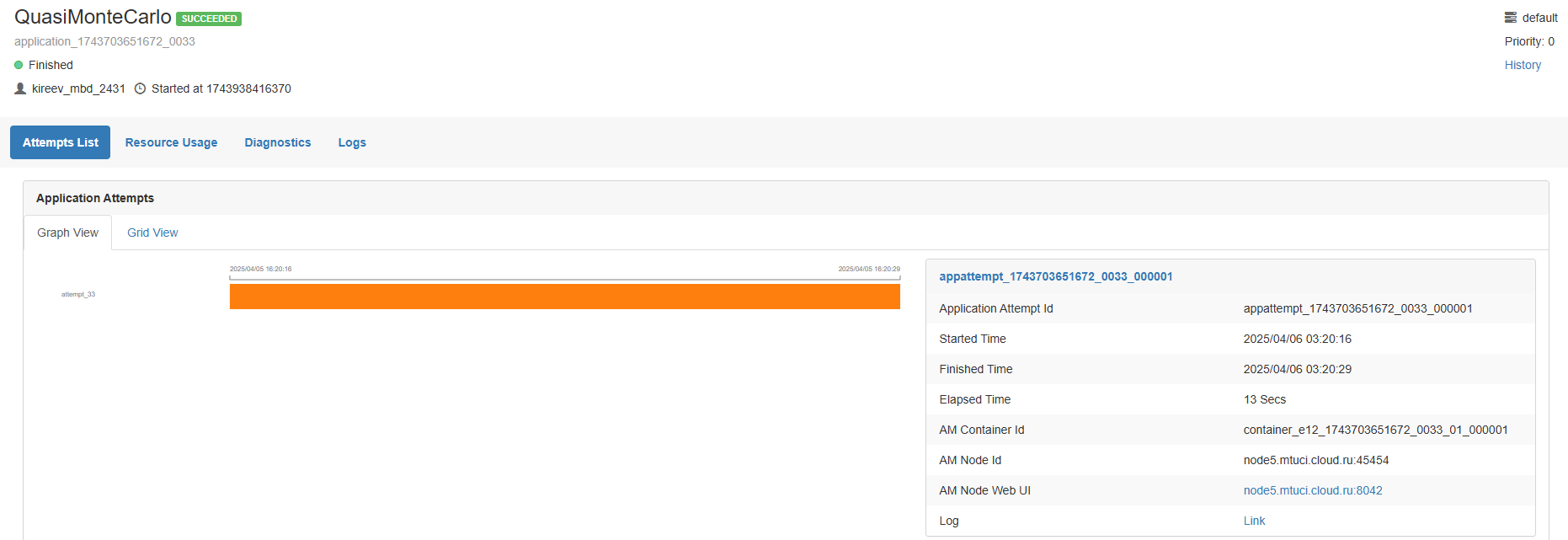


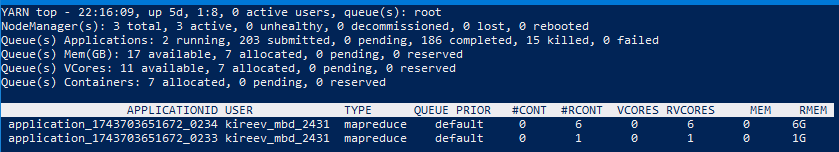


Рисунок 3 - Мониторинг через Ambari

Использование yarn top и принудительное завершение задачи

yarn top

yarn app -kill application\_1743703651672\_0234



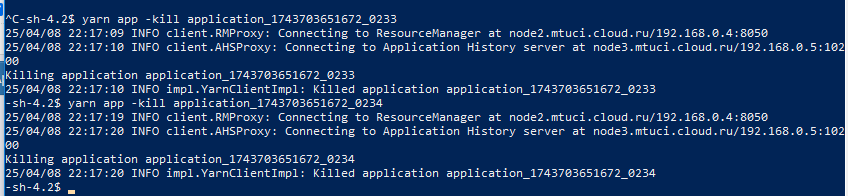


Рисунок 4 - Остановка задачи

**Hadoop Streaming с удалением стоп-слов**

Подготовка входных данных и стоп-слов

mapper\_length.py

#!/usr/bin/env python3

import sys

import re

for line in sys.stdin:

    words = re.findall(r'\b\w+\b', line.lower())

    for word in words:

        if 6 <= len(word) <= 9:

            print(f"{word}\t1")

mapper\_stopwords.py

#!/usr/bin/env python3

import sys

import re

stopwords = set([

'i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd",

'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers',

'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which',

'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been',

'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if',

'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between',

'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out',

'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why',

'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not',

'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't",

'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn',

"couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't",

'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't",

'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"

])

for line in sys.stdin:

words = re.findall(r'\b\w+\b', line.lower())

filtered = [word for word in words if word not in stopwords]

for word in filtered:

print(word)

reducer\_sum.py

#!/usr/bin/env python3

import sys

current\_word = None

current\_count = 0

for line in sys.stdin:

    word, count = line.strip().split("\t")

    count = int(count)

    if current\_word == word:

        current\_count += count

    else:

        if current\_word:

            print(f"{current\_word}\t{current\_count}")

        current\_word = word

        current\_count = count

if current\_word == word:

    print(f"{current\_word}\t{current\_count}")

Делаем файлы исполняемыми и закидываем их в HDFS

hdfs dfs -mkdir -p /data/kireev\_mbd\_2431/

hdfs dfs -chmod 777 /data/kireev\_mbd\_2431/

unzip hadoop\_streaming\_files.zip

chmod +x \*.py

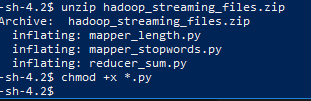


Рисунок 5 - Загрузка файлов

Запуск Hadoop Streaming удаление стоп слов

hadoop jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar \

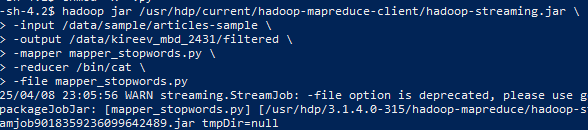
-input /data/sample/articles-sample \

-output /data/kireev\_mbd\_2431/filtered \

-mapper mapper\_stopwords.py \

-reducer /bin/cat \

-file mapper\_stopwords.py



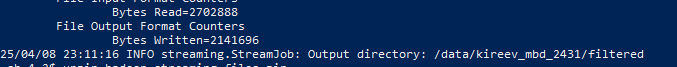


Рисунок 6 - Запуск команды

hdfs dfs -ls /data/kireev\_mbd\_2431/filtered

hdfs dfs -cat /data/kireev\_mbd\_2431/filtered/part-00000

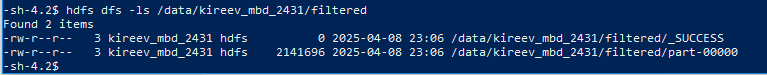


Рисунок 7 - Просмотр результатов

**Сортировка по убыванию количества слов**

hadoop jar /usr/hdp/current/hadoop-mapreduce-client/hadoop-streaming.jar \

-input /data/sample/articles-sample \

-output /data/kireev\_mbd\_2431/wordcount\_6to9 \

-mapper mapper\_length.py \

-reducer reducer\_sum.py \

-file mapper\_length.py \

-file reducer\_sum.py

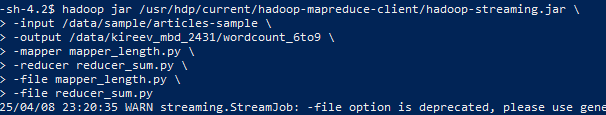


Рисунок 8 - Запуск команды

hdfs dfs -cat /data/kireev\_mbd\_2431/wordcount\_6to9/\* | sort -k2 -nr | head -n 10



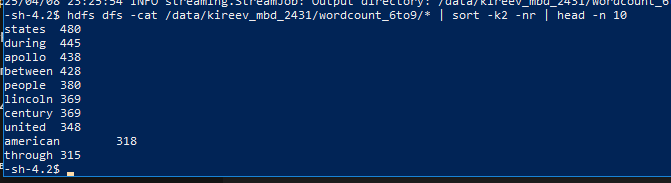


Рисунок 9 - Результат сортировки