Департамент образования и науки города Москвы
Государственное автономное образовательное учреждение
высшего образования города Москвы
«Московский городской педагогический университет»
Институт цифрового образования
Департамент информатики управления и технологий

Мареев Георгий Александрович БД-241м

Практическая работа 1.1 Создание и управление базой данных на HDFS
Вариант 13

Направление подготовки/специальность
38.04.05 - Бизнес-информатика
Бизнес-аналитика и большие данные
(очная форма обучения)

Руководитель дисциплины:
Босенко Т.М., доцент департамента
информатики, управления и технологий,
кандидат технических наук

Москва
2025

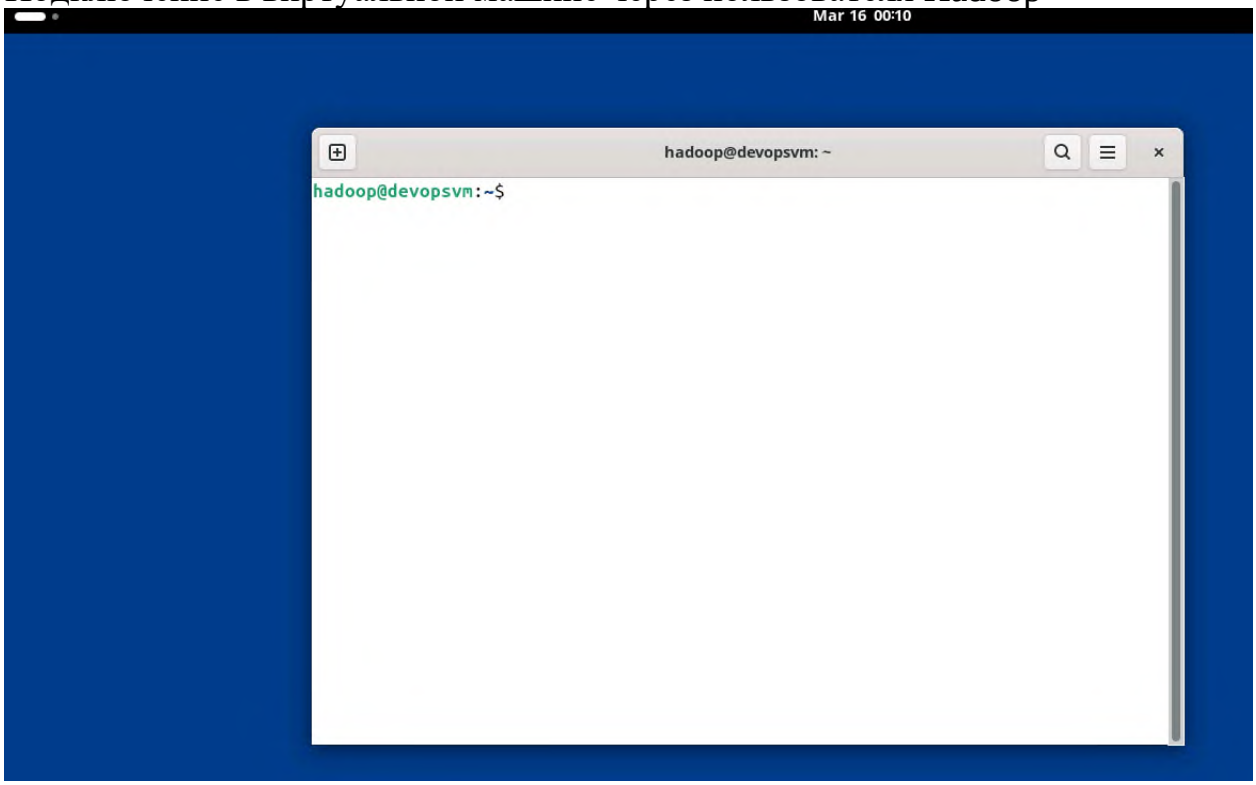
Введение

Цель: изучить основные операции и функциональные возможности системы, что позволит понять принципы работы с данными и распределенными вычислениями.

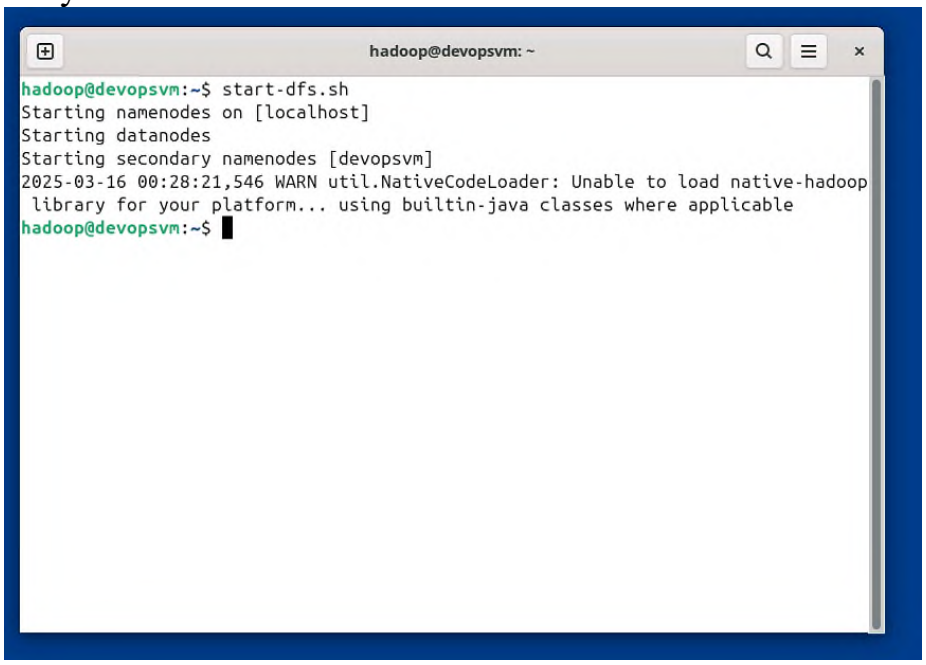
Основная часть

Шаг 1. Запуск Hadoop

Подключение в виртуальной машине через пользователя Hadoop

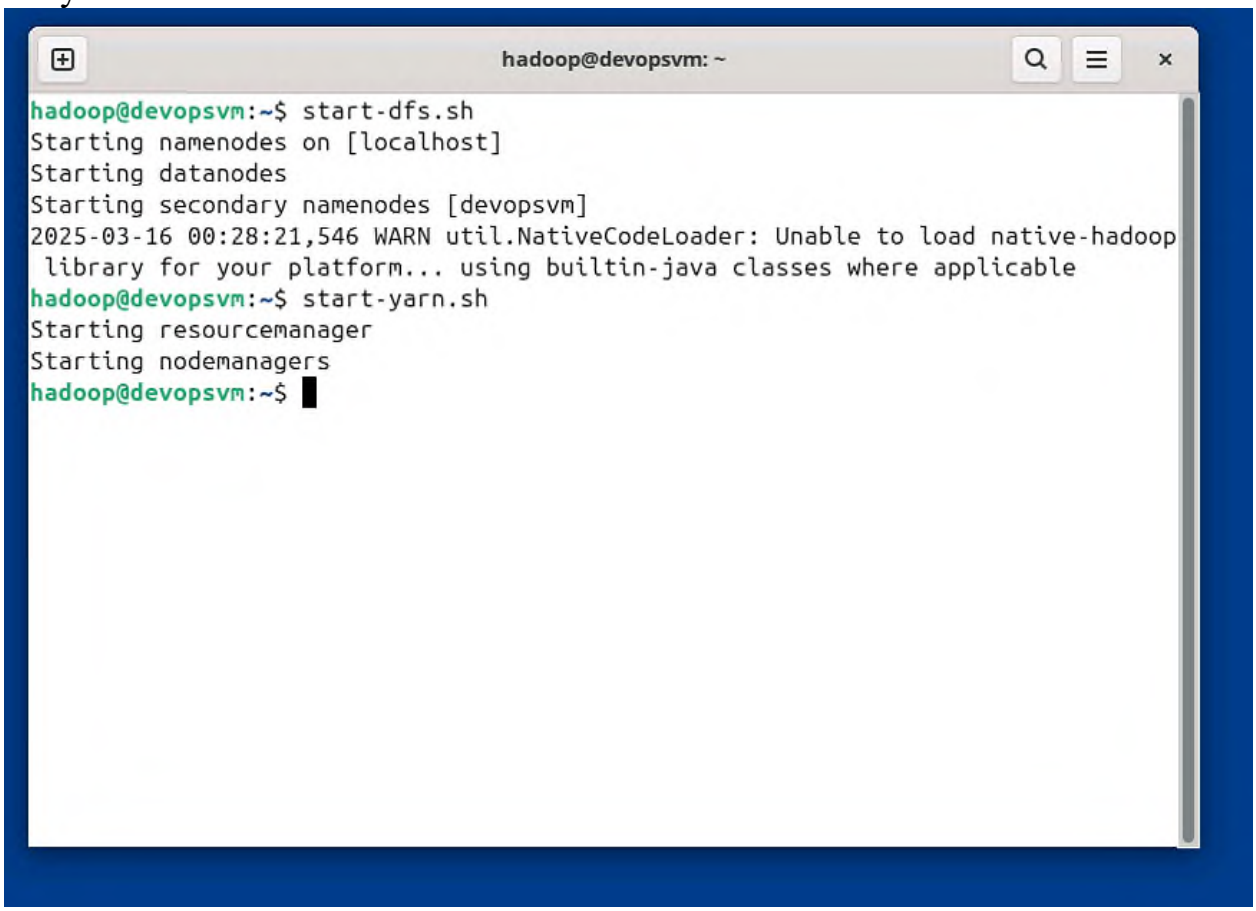


Запуск HDFS



start-dfs.sh

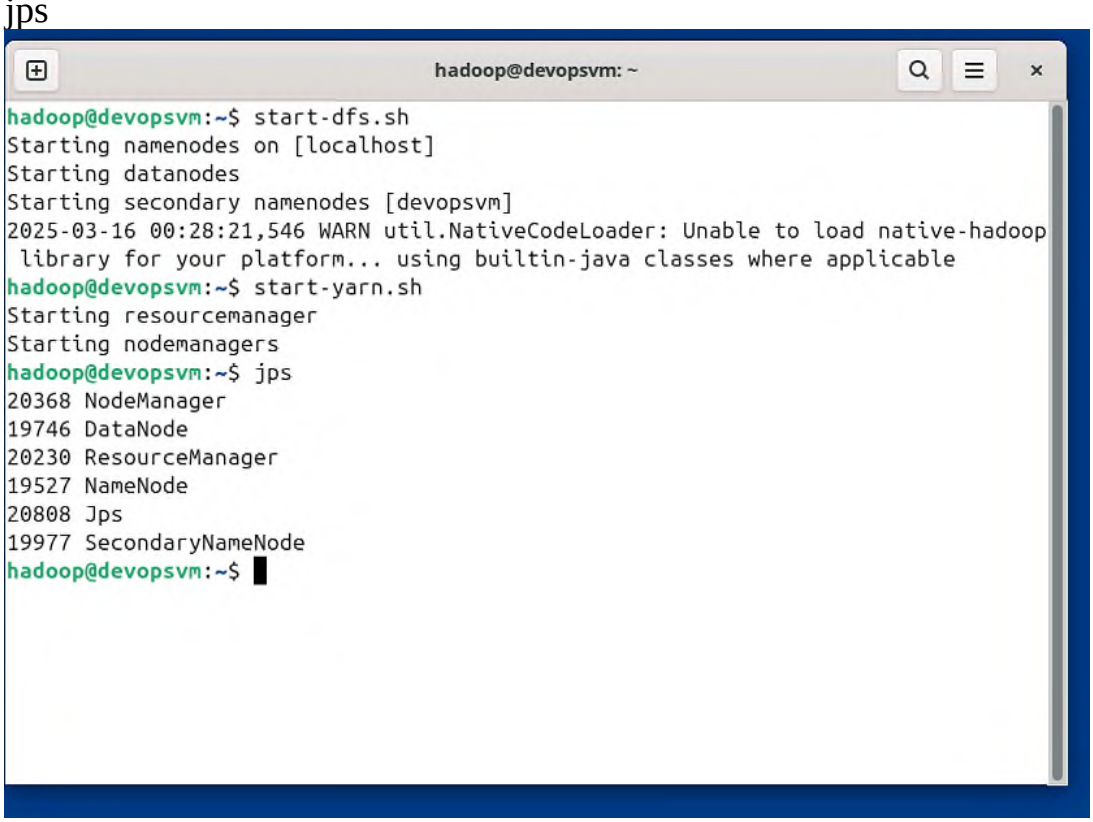
Запуск YARN



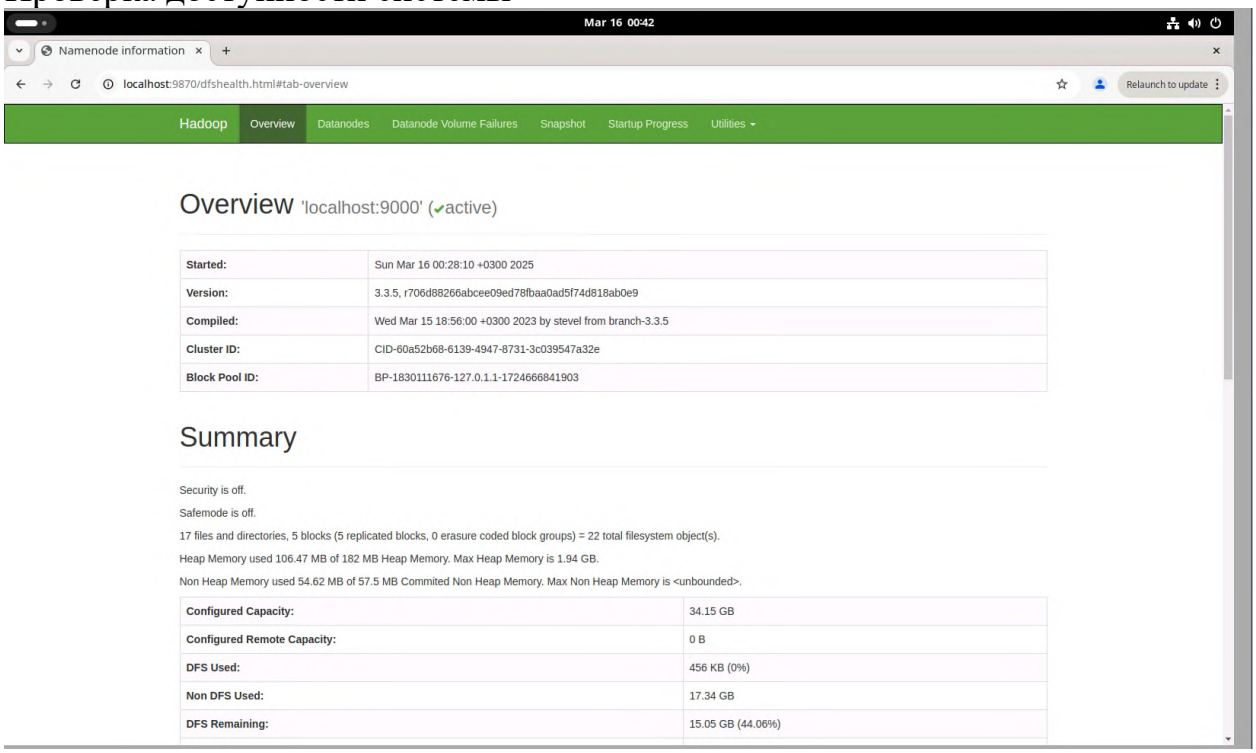
start-yarn.sh

Шаг 2. Проверка запущенных служб

jps

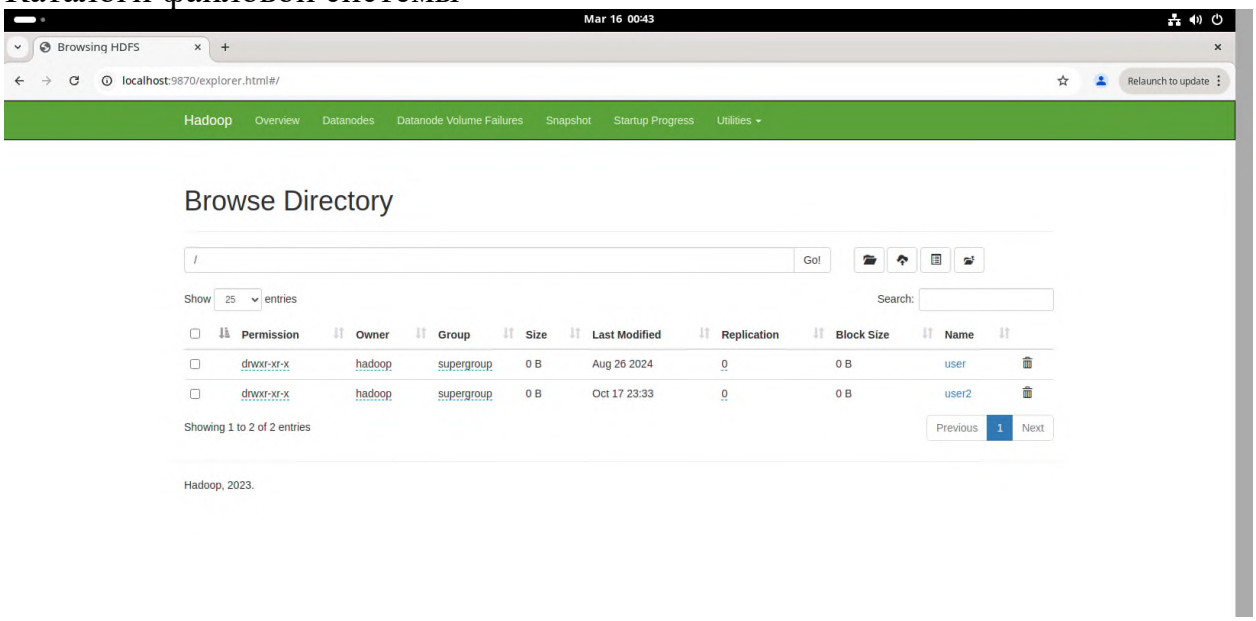


Проверка доступности системы

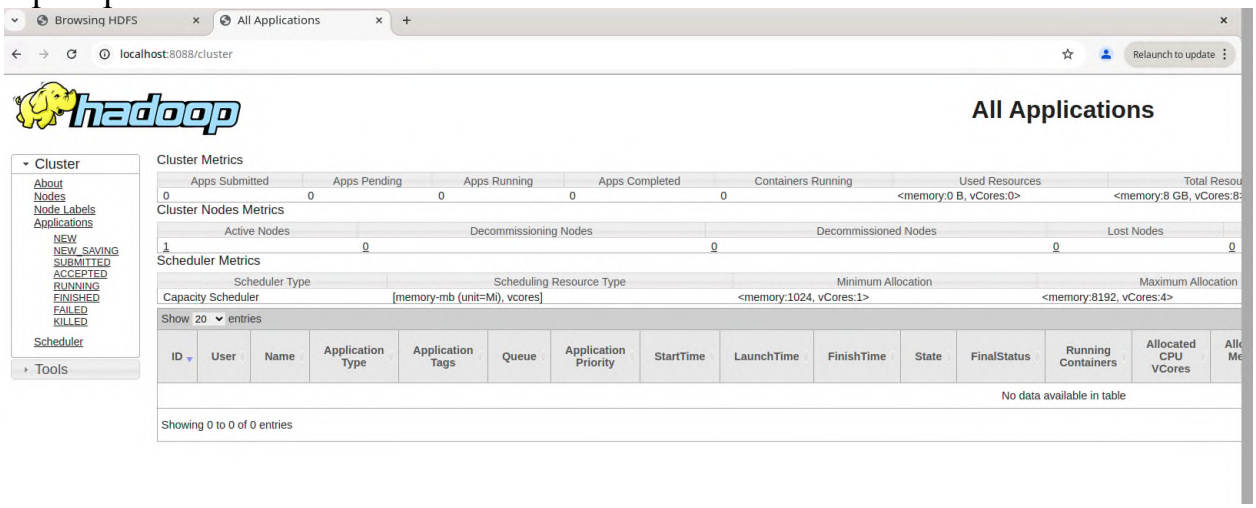


localhost:9870

Каталоги файловой системы

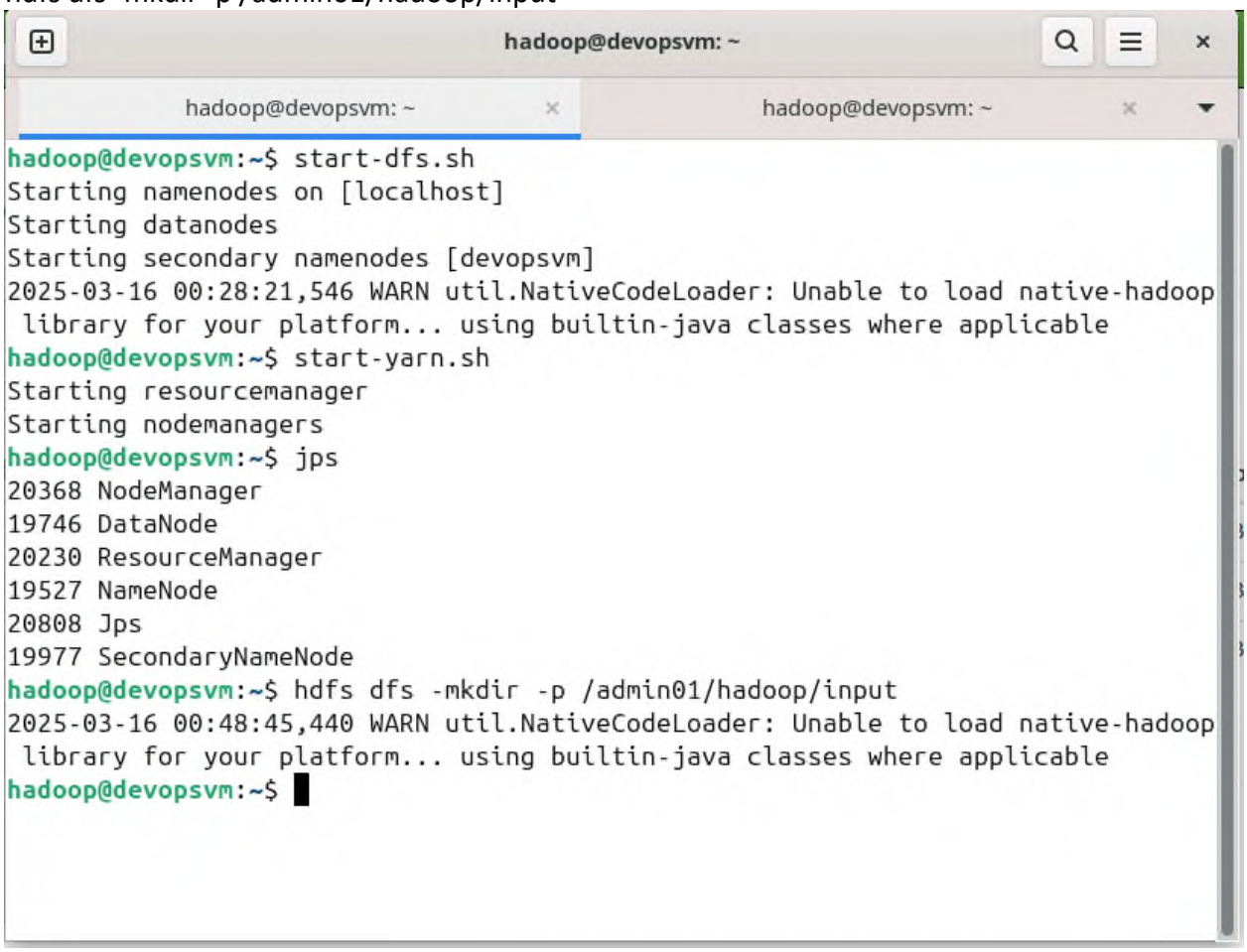


Проверка YARN



Шаг 3. Подготовка рабочего пространства
Создание директории в HDFS

hdfs dfs -mkdir -p /admin01/hadoop/input



Browse Directory

/ Go!

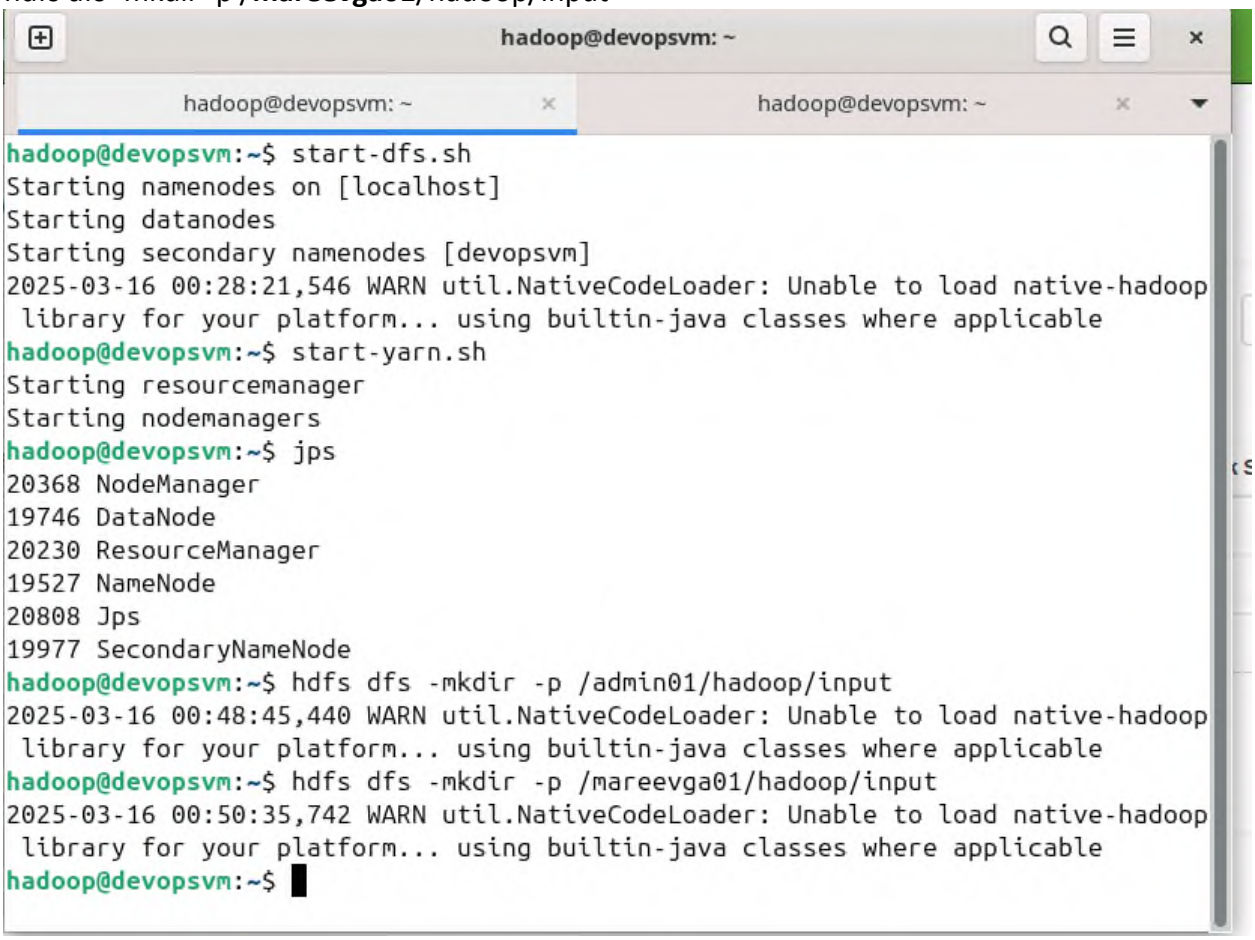
Show 25 entries Search:

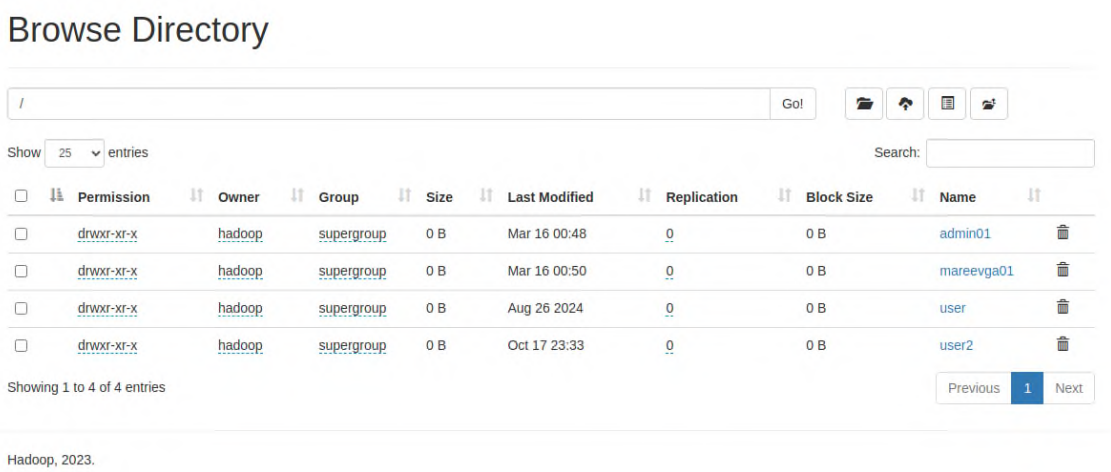
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	hadoop	supergroup	0 B	Mar 16 00:48	0	0 B	admin01
drwxr-xr-x	hadoop	supergroup	0 B	Aug 26 2024	0	0 B	user
drwxr-xr-x	hadoop	supergroup	0 B	Oct 17 23:33	0	0 B	user2

Showing 1 to 3 of 3 entries Previous 1 Next

Hadoop, 2023.

hdfs dfs -mkdir -p /**mareevga01**/hadoop/input



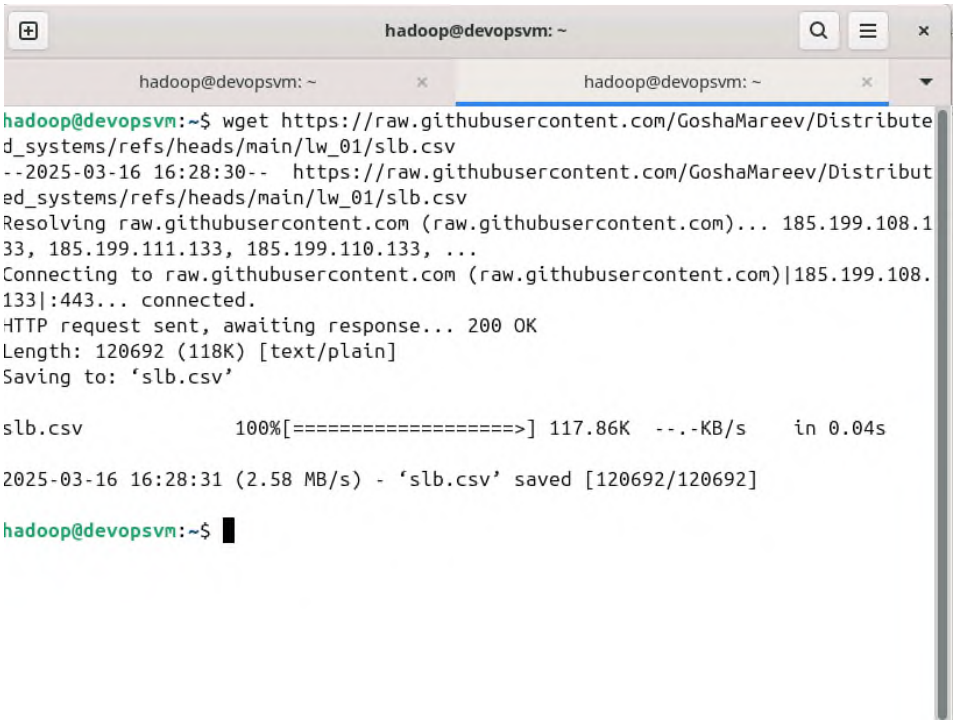


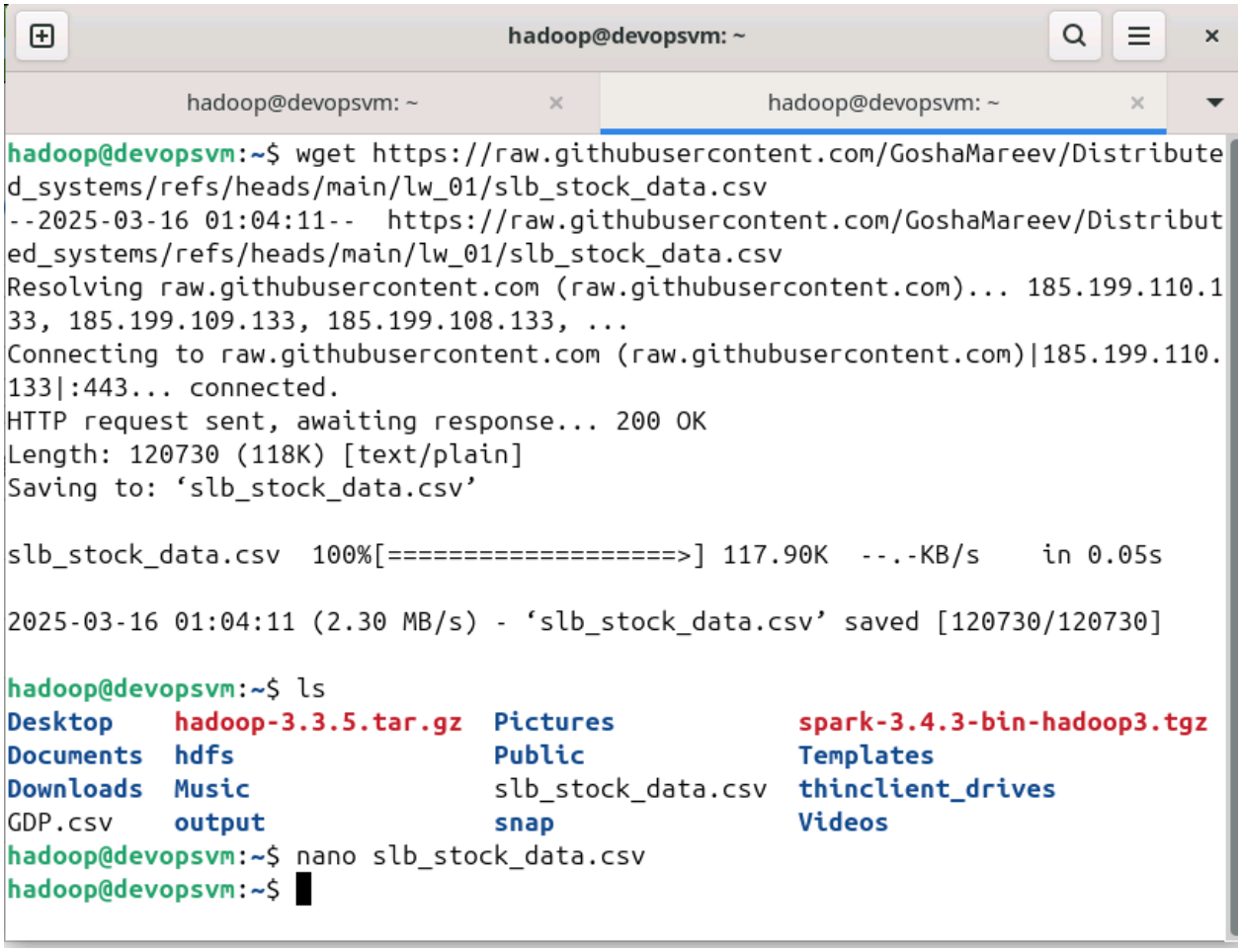
Шаг 4. Загрузка и подготовка данных

Скачивание файла с данными

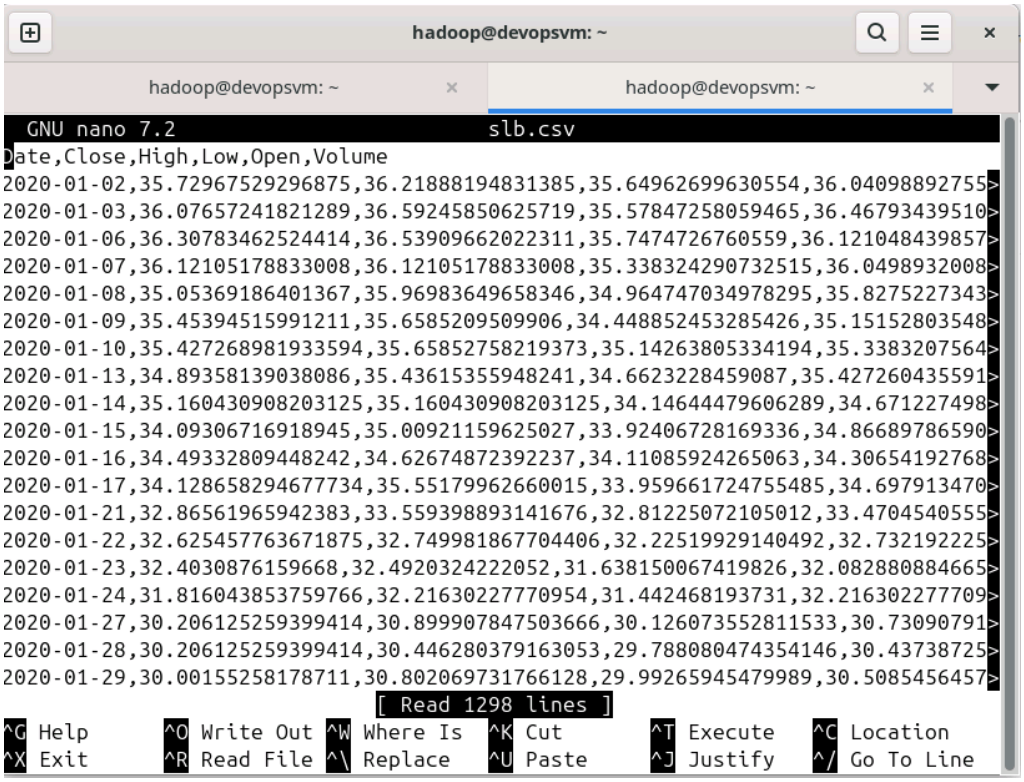
*файл с данными был немного преобразован: сделан заголовок и изменено имя файла для простоты (все изменения отражены в репозитории) https://github.com/GoshaMareev/Distributed_systems/tree/main/lw_01

wget https://raw.githubusercontent.com/GoshaMareev/Distributed_systems/refs/heads/main/lw_01/slb.csv



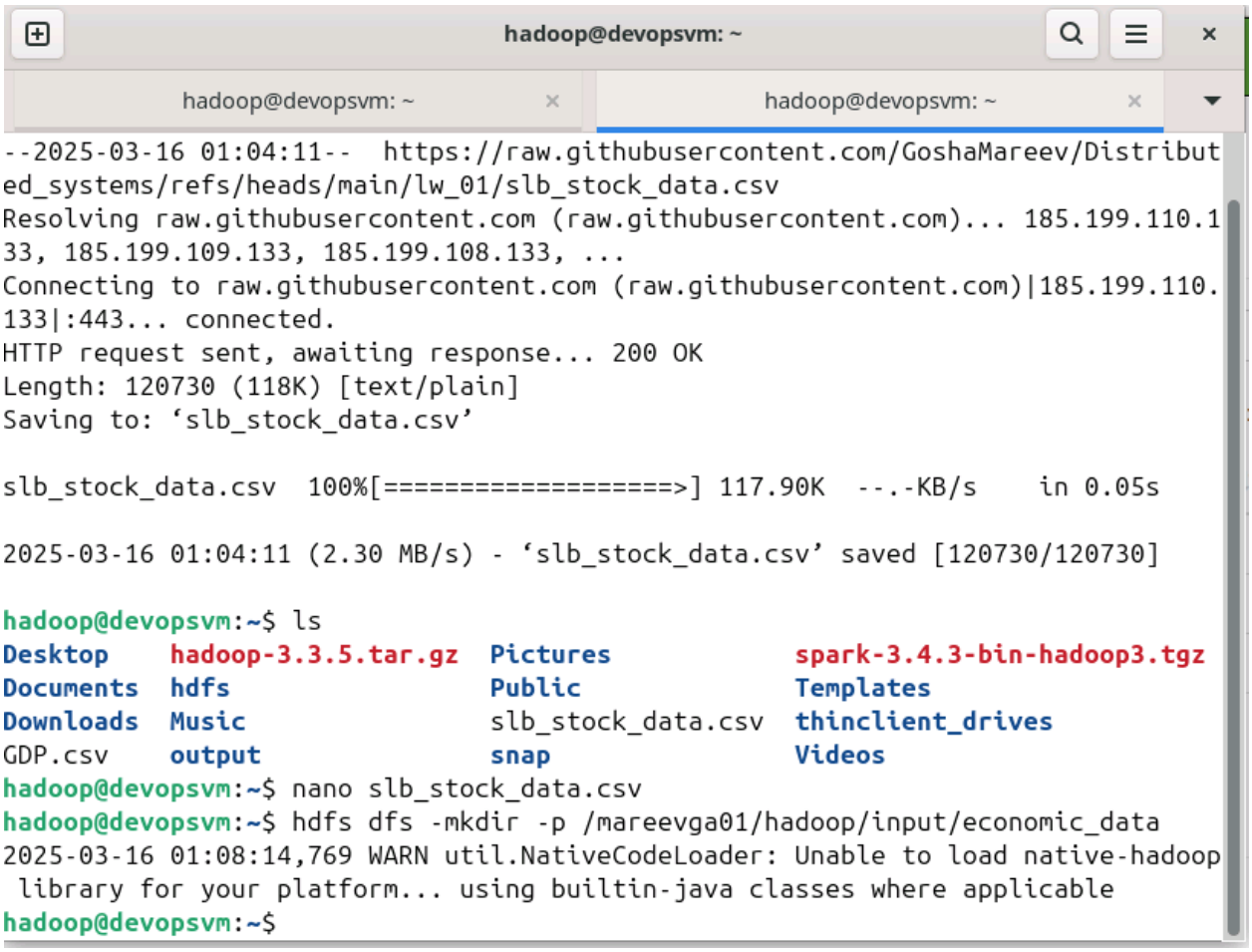

hadoop@devopsvm: ~
hadoop@devopsvm: ~
hadoop@devopsvm: ~
hadoop@devopsvm:~$ wget https://raw.githubusercontent.com/GoshaMareev/Distributed_systems/refs/heads/main/lw_01/slb_stock_data.csv
--2025-03-16 01:04:11-- https://raw.githubusercontent.com/GoshaMareev/Distributed_systems/refs/heads/main/lw_01/slb_stock_data.csv
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.110.133, 185.199.109.133, 185.199.108.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.110.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 120730 (118K) [text/plain]
Saving to: 'slb_stock_data.csv'
slb_stock_data.csv 100%[==========>] 117.90K --.-KB/s in 0.05s
2025-03-16 01:04:11 (2.30 MB/s) - 'slb_stock_data.csv' saved [120730/120730]
hadoop@devopsvm:~$ ls
Desktop hadoop-3.3.5.tar.gz Pictures spark-3.4.3-bin-hadoop3.tgz
Documents hdfs Public Templates
Downloads Music slb_stock_data.csv thinclient_drives
GDP.csv output snap Videos
hadoop@devopsvm:~$ nano slb_stock_data.csv
hadoop@devopsvm:~$

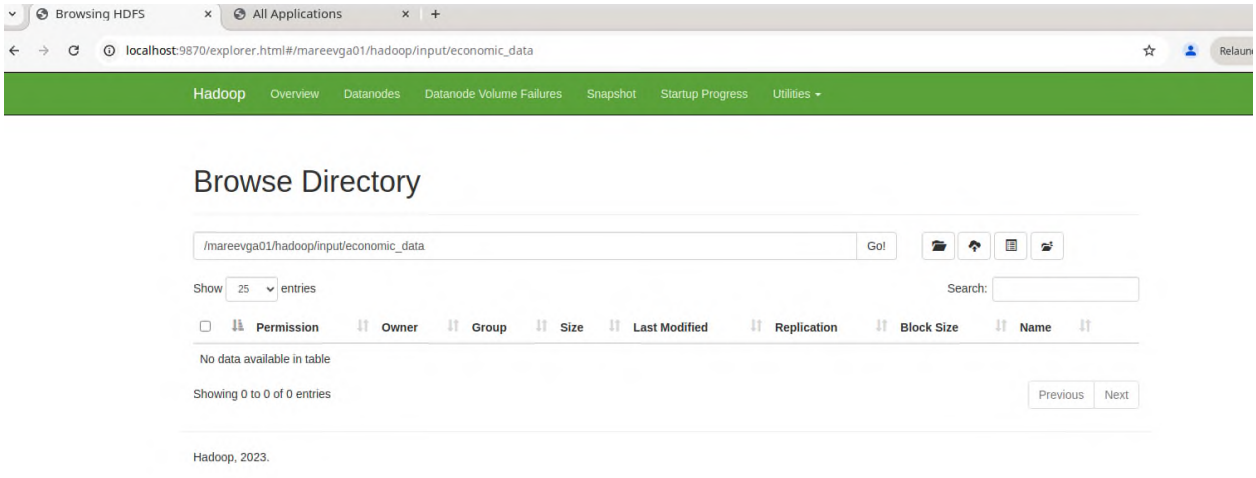
Nano slb.csv


hadoop@devopsvm: ~
hadoop@devopsvm: ~
hadoop@devopsvm: ~
GNU nano 7.2 slb.csv
Date,Close,High,Low,Open,Volume
2020-01-02,35.72967529296875,36.21888194831385,35.64962699630554,36.04098892755
2020-01-03,36.07657241821289,36.59245850625719,35.57847258059465,36.46793439510
2020-01-06,36.30783462524414,36.53909662022311,35.7474726760559,36.121048439857
2020-01-07,36.12105178833008,36.12105178833008,35.338324290732515,36.0498932008
2020-01-08,35.05369186401367,35.96983649658346,34.964747034978295,35.8275227343
2020-01-09,35.45394515991211,35.6585209509906,34.448852453285426,35.15152803548
2020-01-10,35.427268981933594,35.65852758219373,35.14263805334194,35.3383207564
2020-01-13,34.89358139038086,35.43615355948241,34.6623228459087,35.427260435591
2020-01-14,35.160430908203125,35.160430908203125,34.14644479606289,34.671227498
2020-01-15,34.09306716918945,35.00921159625027,33.92406728169336,34.86689786590
2020-01-16,34.49332809448242,34.62674872392237,34.11085924265063,34.30654192768
2020-01-17,34.128658294677734,35.55179962660015,33.959661724755485,34.697913470
2020-01-21,32.86561965942383,33.559398893141676,32.81225072105012,33.4704540555
2020-01-22,32.625457763671875,32.749981867704406,32.22519929140492,32.732192225
2020-01-23,32.4030876159668,32.4920324222052,31.638150067419826,32.082880884665
2020-01-24,31.816043853759766,32.21630227770954,31.442468193731,32.216302277709
2020-01-27,30.206125259399414,30.899907847503666,30.126073552811533,30.73090791
2020-01-28,30.206125259399414,30.446280379163053,29.788080474354146,30.43738725
2020-01-29,30.00155258178711,30.802069731766128,29.99265945479989,30.5085456457
[Read 1298 lines]
^G Help ^O Write Out ^W Where Is ^K Cut ^T Execute ^C Location
^X Exit ^R Read File ^\ Replace ^U Paste ^J Justify ^/ Go To Line

Создание директории для экономических данных

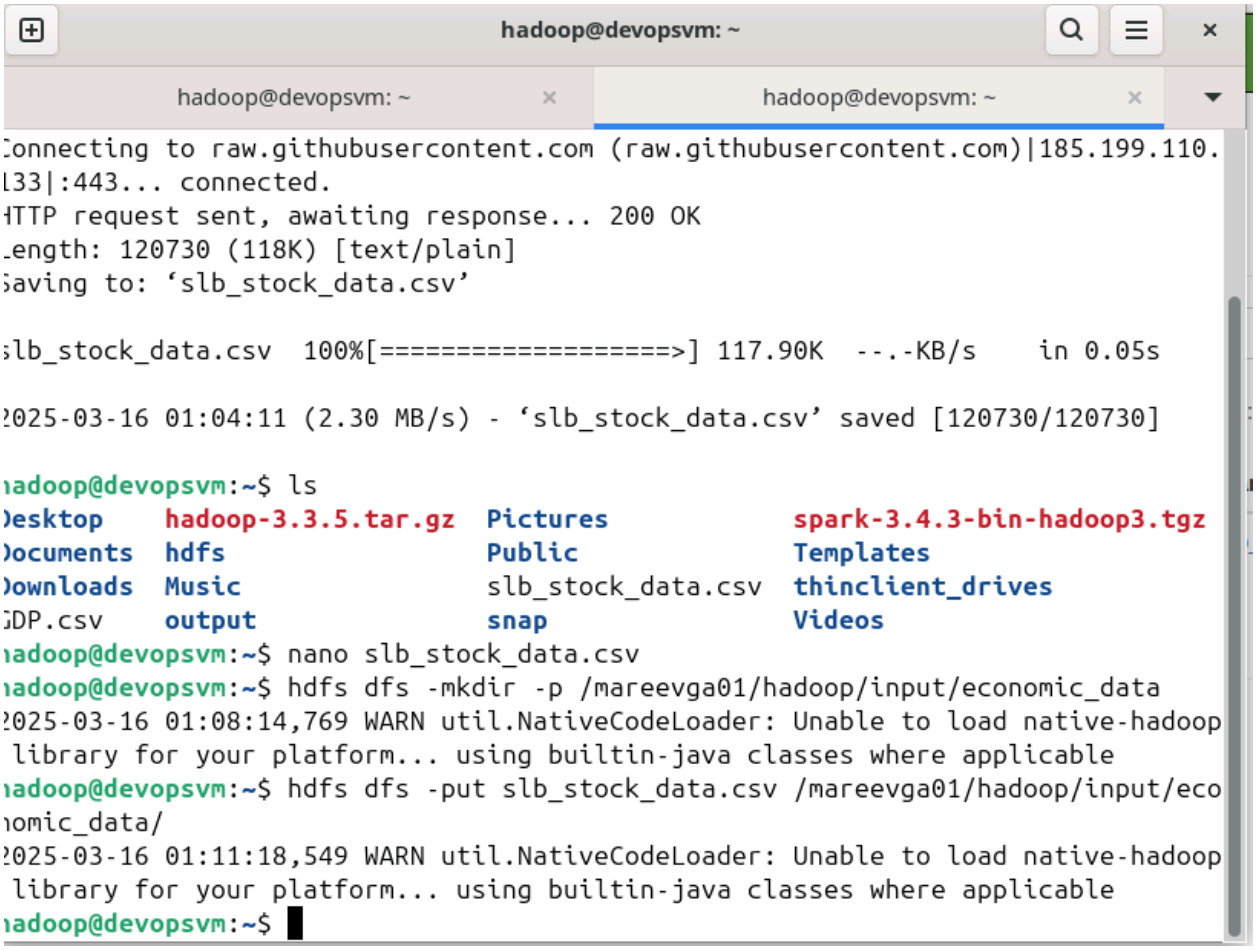
hdfs dfs -mkdir -p /mareevga01/hadoop/input/economic_data

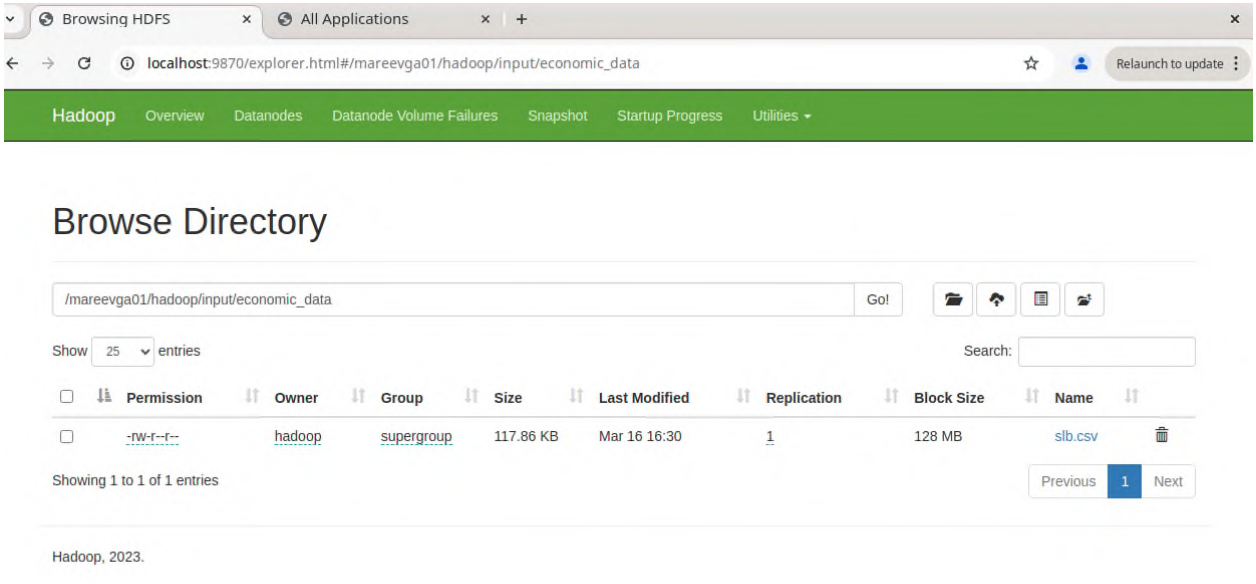




Загрузка данных в HDFS

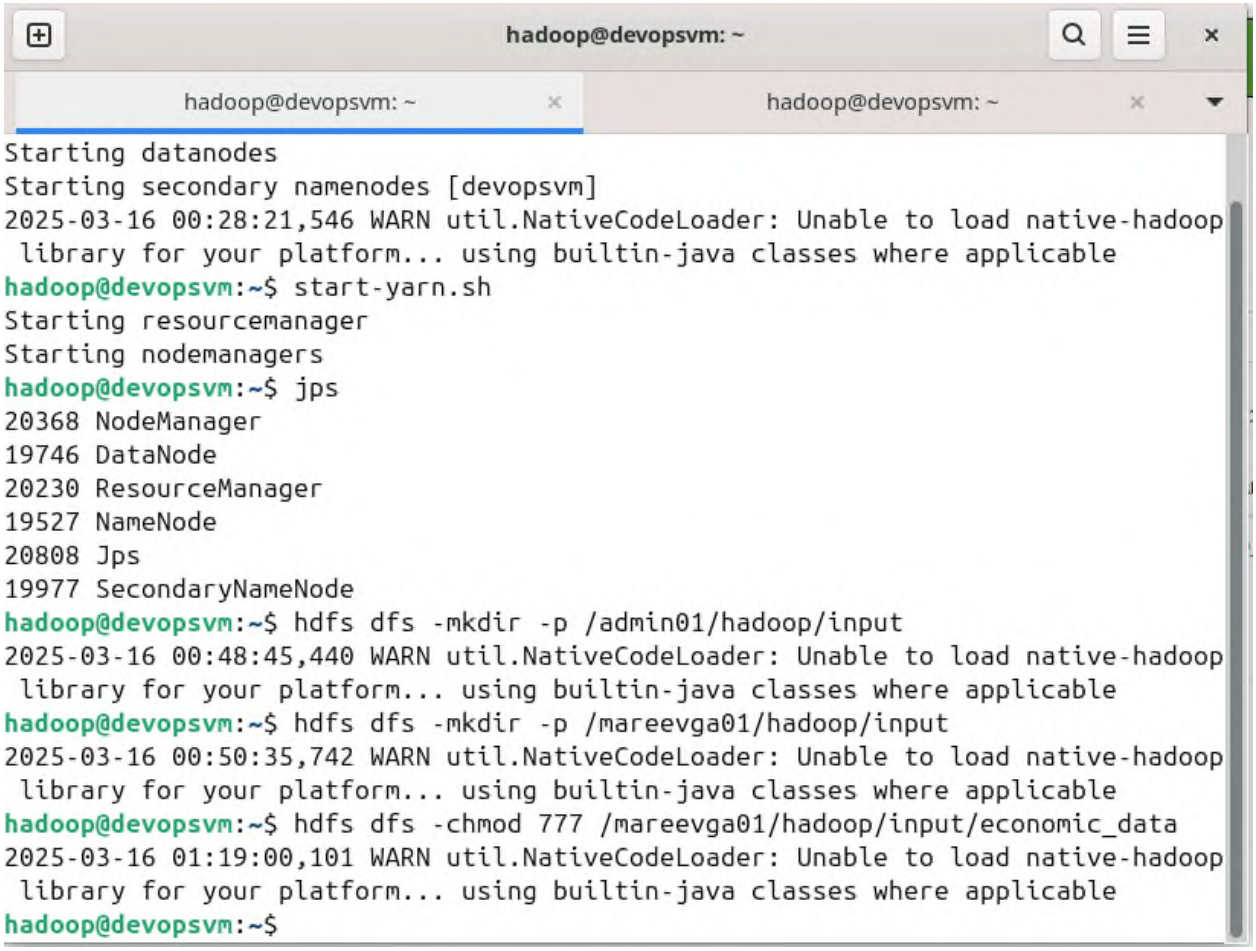
hdfs dfs -put slb.csv /mareevga01/hadoop/input/economic_data/


hadoop@devopsvm: ~
hadoop@devopsvm: ~
hadoop@devopsvm: ~
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.110.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 120730 (118K) [text/plain]
Saving to: 'slb_stock_data.csv'
slb_stock_data.csv 100%[==========>] 117.90K --.-KB/s in 0.05s
2025-03-16 01:04:11 (2.30 MB/s) - 'slb_stock_data.csv' saved [120730/120730]
hadoop@devopsvm:~$ ls
Desktop hadoop-3.3.5.tar.gz Pictures spark-3.4.3-bin-hadoop3.tgz
Documents hdfs Public Templates
Downloads Music slb_stock_data.csv thinclient_drives
GDP.csv output snap Videos
hadoop@devopsvm:~$ nano slb_stock_data.csv
hadoop@devopsvm:~$ hdfs dfs -mkdir -p /mareevga01/hadoop/input/economic_data
2025-03-16 01:08:14,769 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hadoop@devopsvm:~$ hdfs dfs -put slb_stock_data.csv /mareevga01/hadoop/input/economic_data/
2025-03-16 01:11:18,549 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hadoop@devopsvm:~$


Browsing HDFS
All Applications
localhost:9870/explorer.html#/mareevga01/hadoop/input/economic_data
Relaunch to update
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Browse Directory
/mareevga01/hadoop/input/economic_data
Go!
Show 25 entries
Search:
Permission
Owner
Group
Size
Last Modified
Replication
Block Size
Name
-rw-r--r--
hadoop
supergroup
117.86 KB
Mar 16 16:30
1
128 MB
slb.csv
Showing 1 to 1 of 1 entries
Previous
1
Next
Hadoop, 2023.

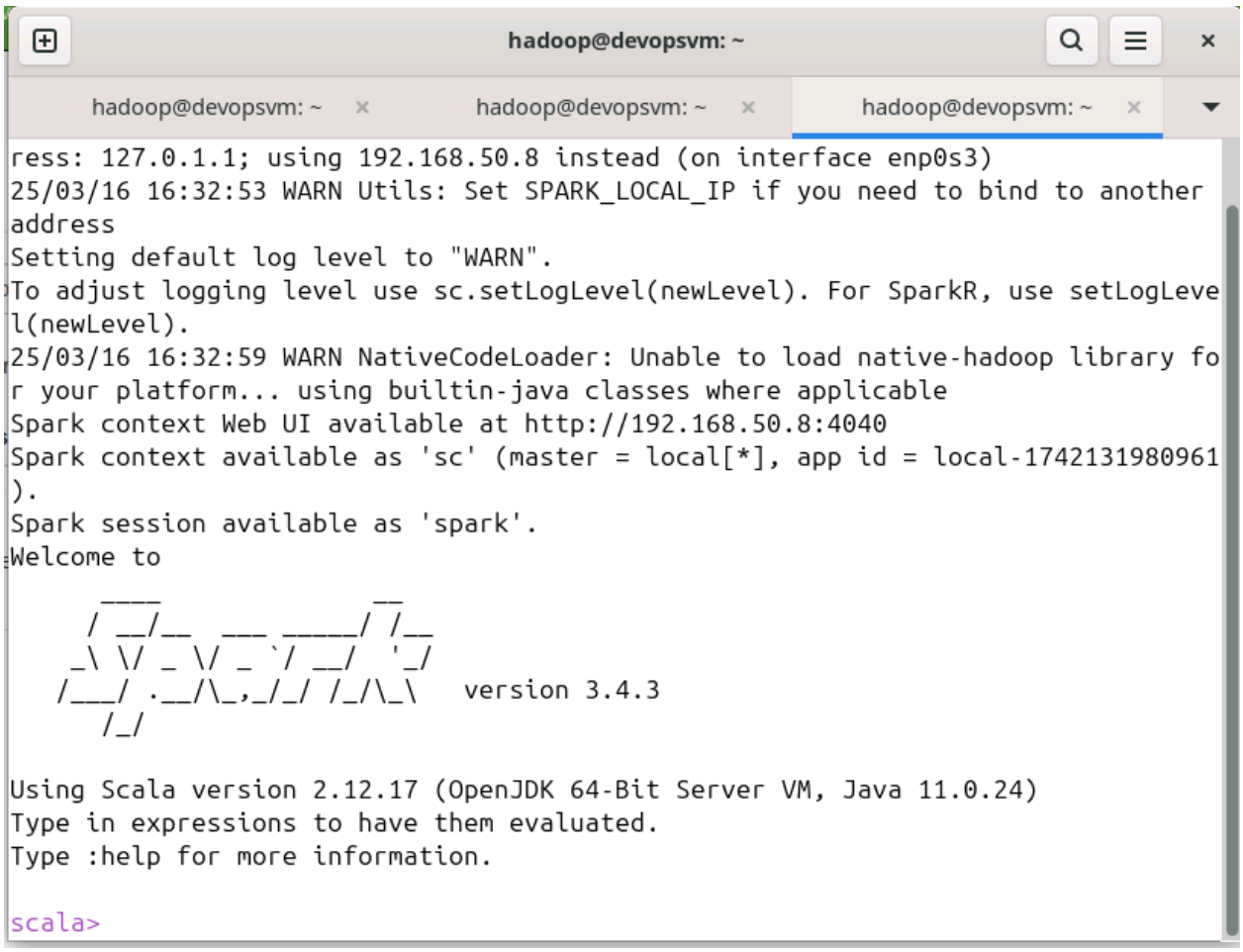
Установка прав доступа

hdfs dfs -chmod 777 /mareevga01/hadoop/input/economic_data



Шаг 5. Обработка данных с помощью Spark

Spark-shell



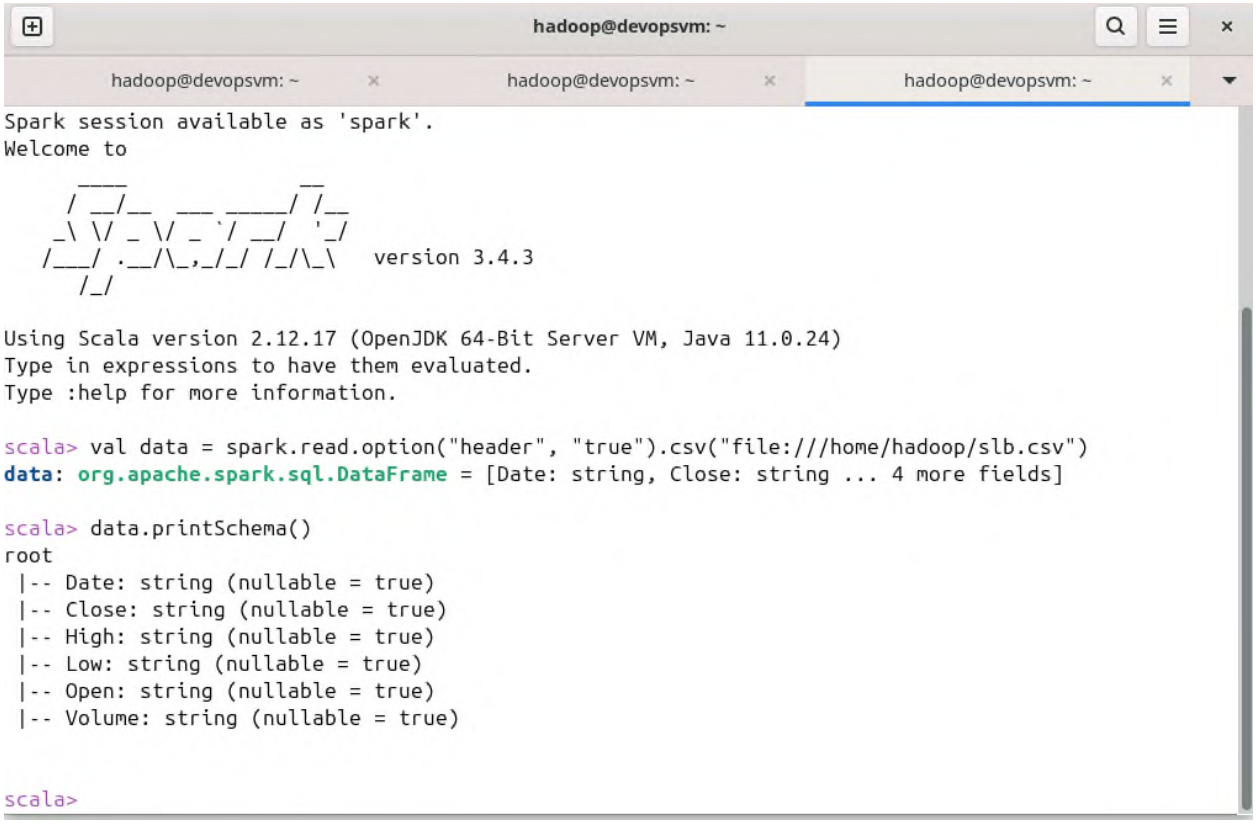
Spark Shell:

Загрузка данных из HDFS с правильным URI

```
val data = spark.read.option("header", "true").csv("file:///home/hadoop/slb.csv")
```

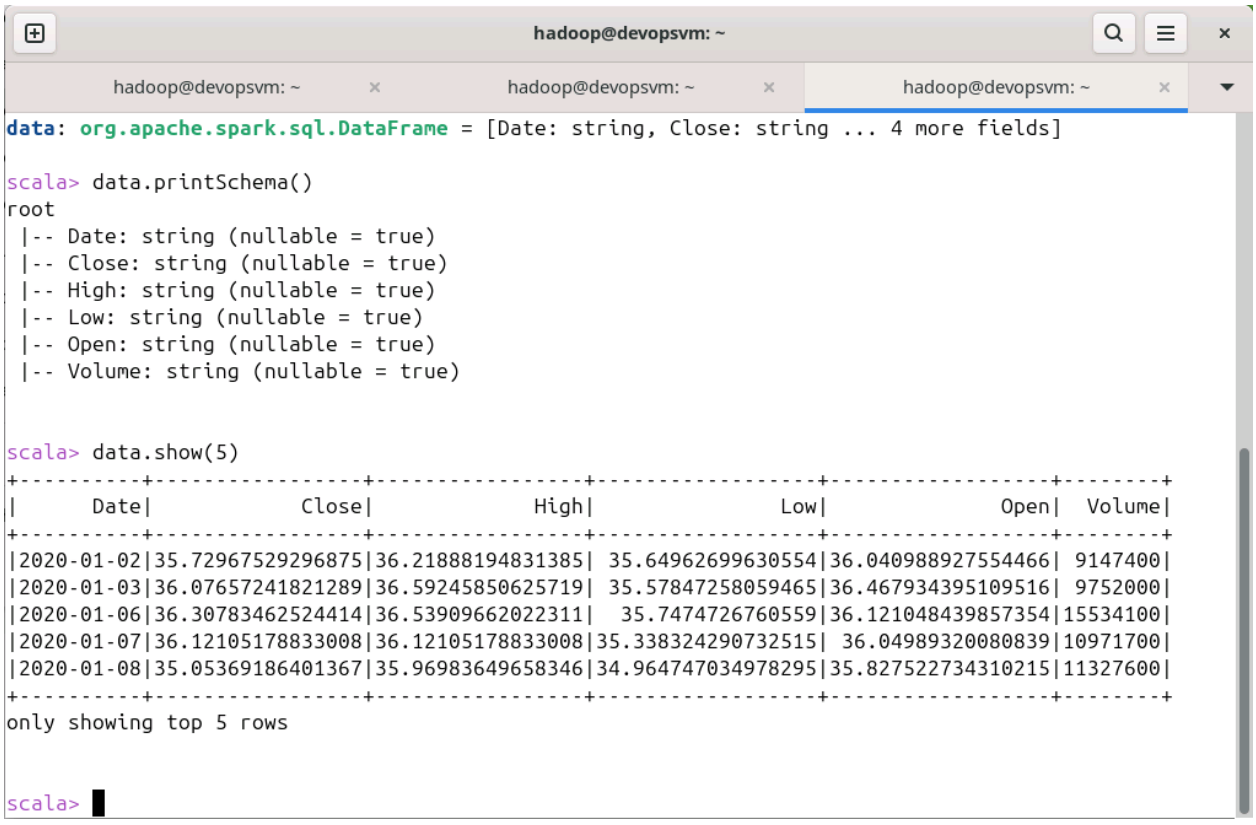
Проверка схемы данных

```
data.printSchema()
```



Первые 5 строк

Data.show(5)



Все столбцы имеют строковый (string) формат, для вычислений нужно будет преобразовать:

Data – в формат даты

Close, High, Low, Open, Volume в числовой формат

Вычисление среднего значения

Фильтрация данных за последние 3 года

Расчет медианной цены закрытия

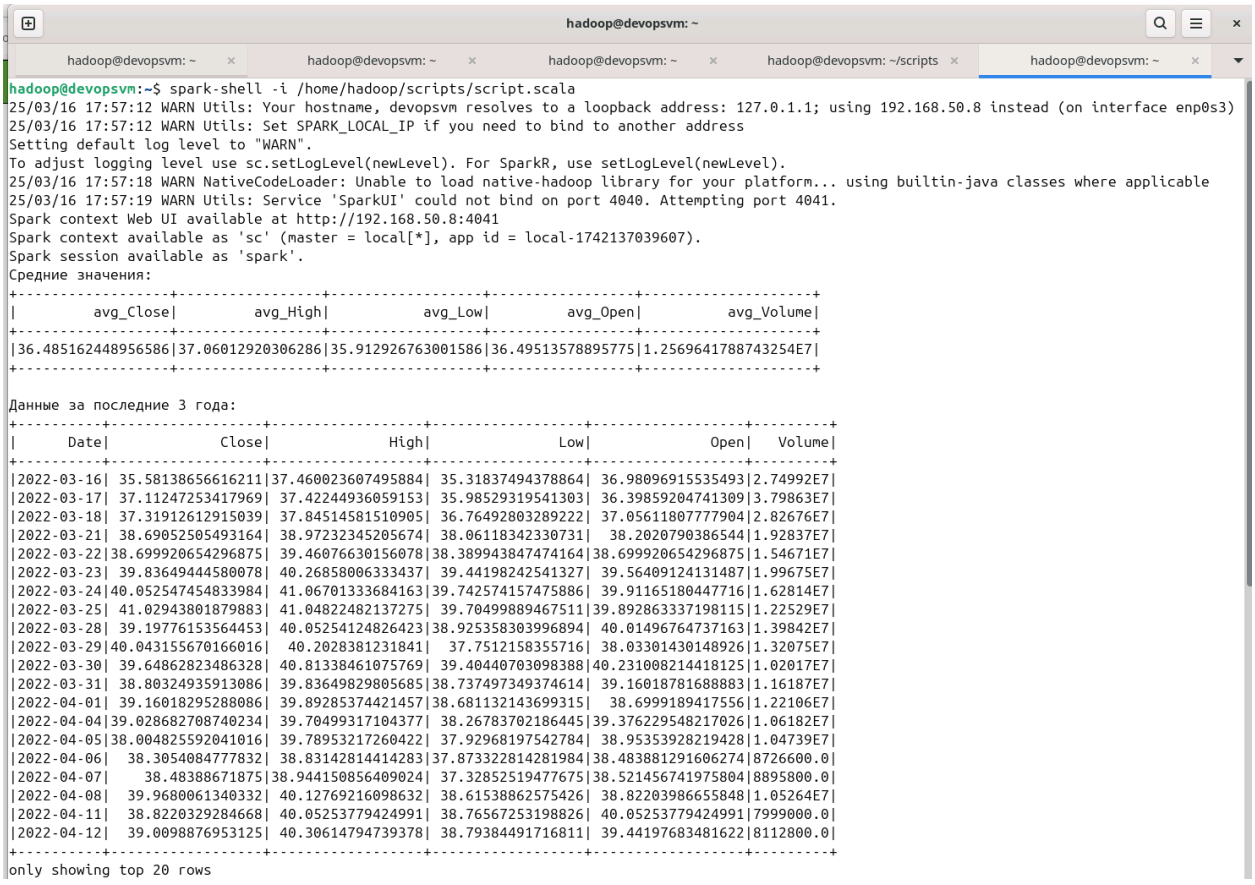
Группировка по кварталам

Сохранение результата в CSV файл

Для удобства была создана директория scripts/ и файл script.scala

Создание директории и файла script.scala с кодом

Результат выполнения:



```
Медианная цена закрытия: 38.074913024902344
Группировка по кварталам:
+----+-------+------------------+
|Year|Quarter|         avg_Close|
+----+-------+------------------+
|2020|      3|  16.9349362552166|
|2024|      3| 43.72580277919769|
|2022|      2|39.226578497117565|
|2022|      3| 34.23342174291611|
|2024|      2|  47.0983515542651|
|2023|      3|55.746014367966424|
|2023|      2| 45.60234648181546|
|2020|      4|17.305005356669426|
|2023|      4| 52.46532997252449|
|2021|      3|26.637318402528763|
|2021|      2|28.453633747403583|
|2024|      1| 49.00227587340308|
|2020|      1| 25.90806208887408|
|2022|      1|36.778632071710405|
|2021|      1|  24.2893645177122|
|2025|      1| 40.83830398168319|
|2020|      2|15.854010763622465|
|2024|      4| 41.50260633230209|
|2022|      4| 47.20831637912326|
|2021|      4|29.415187895298004|
+----+-------+------------------+
only showing top 20 rows
```

Выход из Spark-shell:

```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.4.3
      /_/

Using Scala version 2.12.17 (OpenJDK 64-Bit Server VM, Java 11.0.24)
Type in expressions to have them evaluated.
Type :help for more information.

scala> :q
hadoop@devopsvm:~$
```

Шаг 6. Работа с результатами

cd /home/hadoop

Переход в директорию с результатами

Проверка результатов

После выполнения скрипта результаты будут сохранены в указанных директориях:

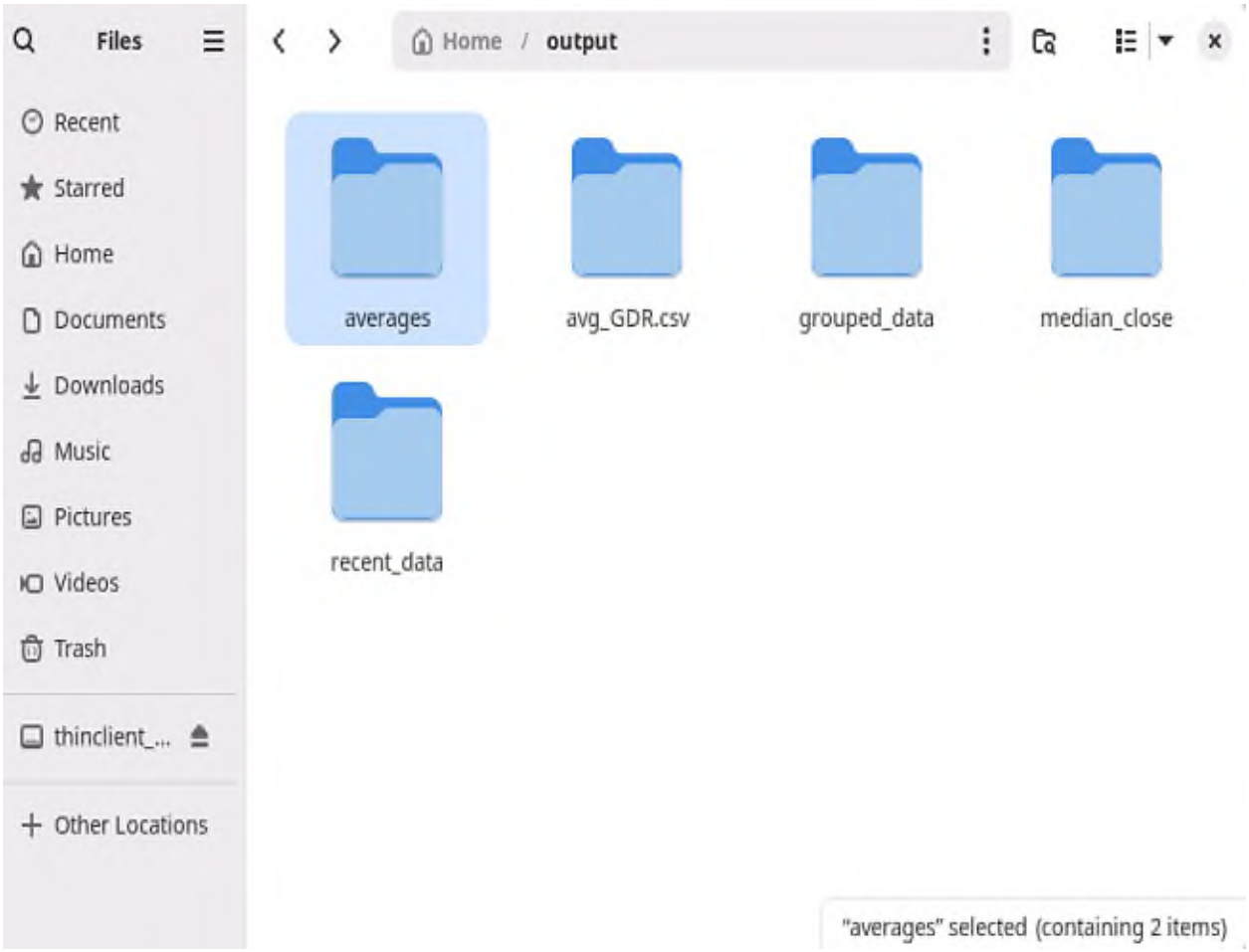
file:///home/hadoop/output/averages

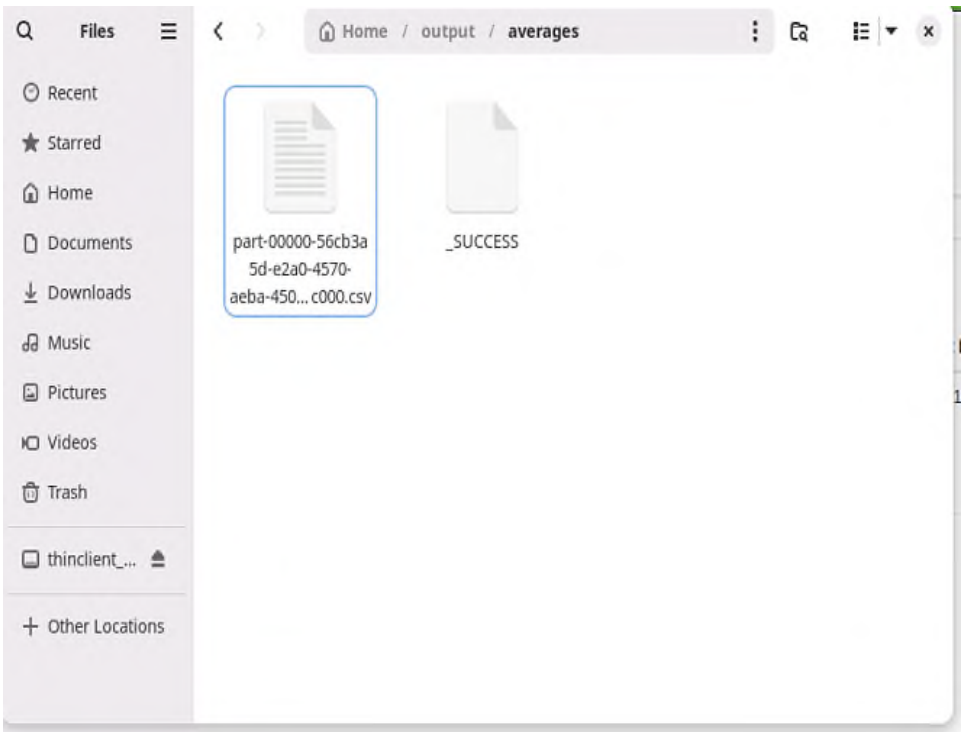
file:///home/hadoop/output/recent_data

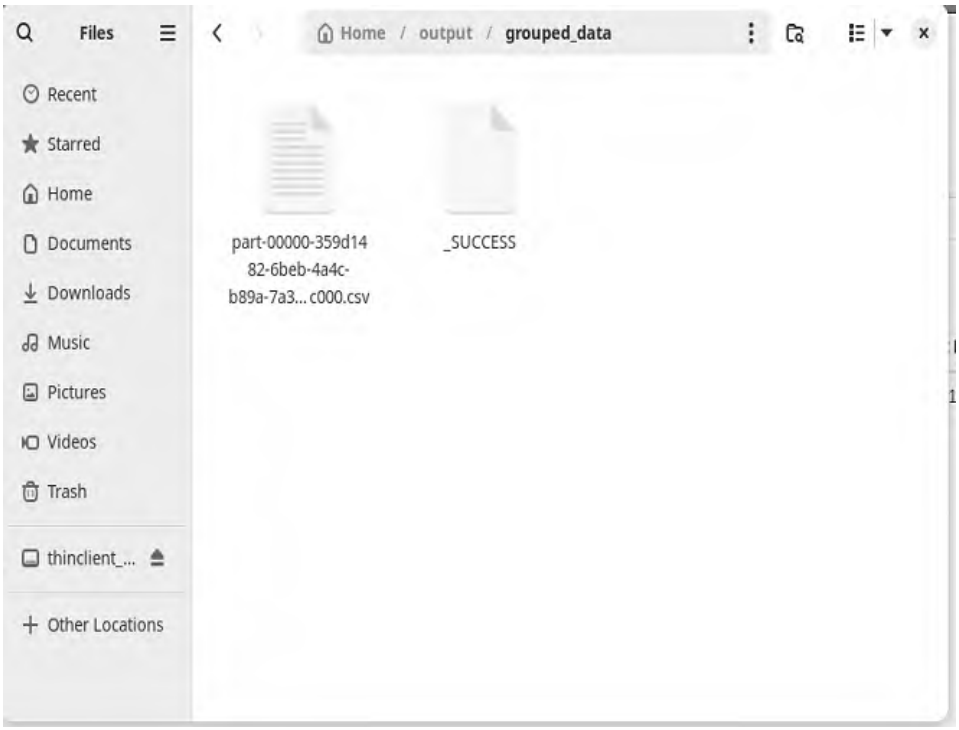
file:///home/hadoop/output/median_close

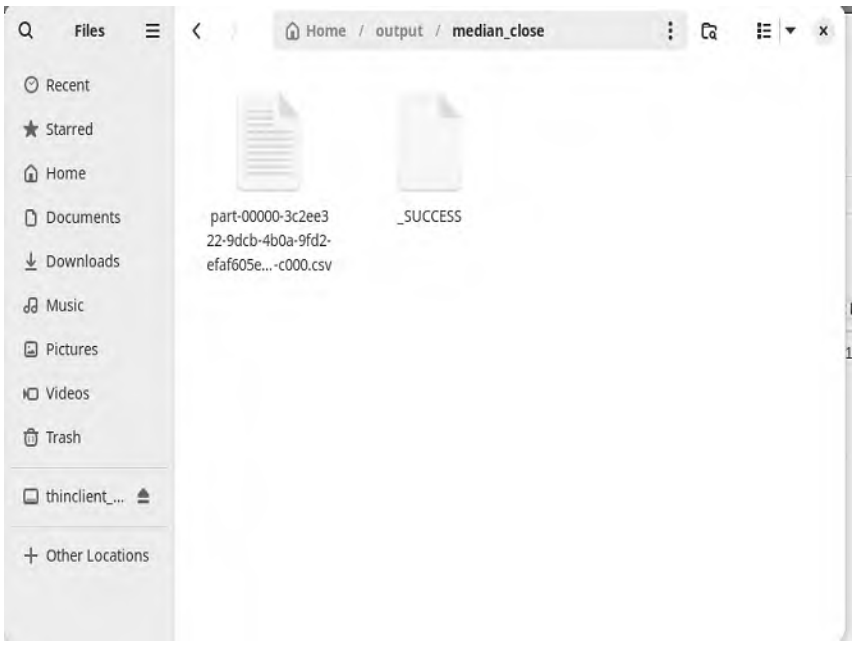
file:///home/hadoop/output/grouped_data

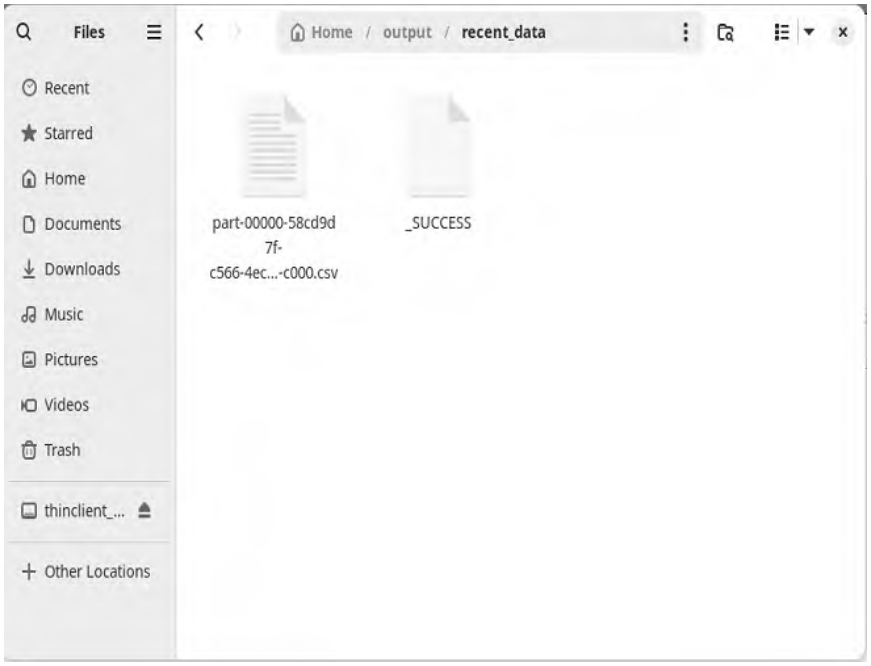
```
hadoop@devopsvm:~$ ls /home/hadoop/output/
averages  avg_GDR.csv  grouped_data  median_close  recent_data
hadoop@devopsvm:~$
```



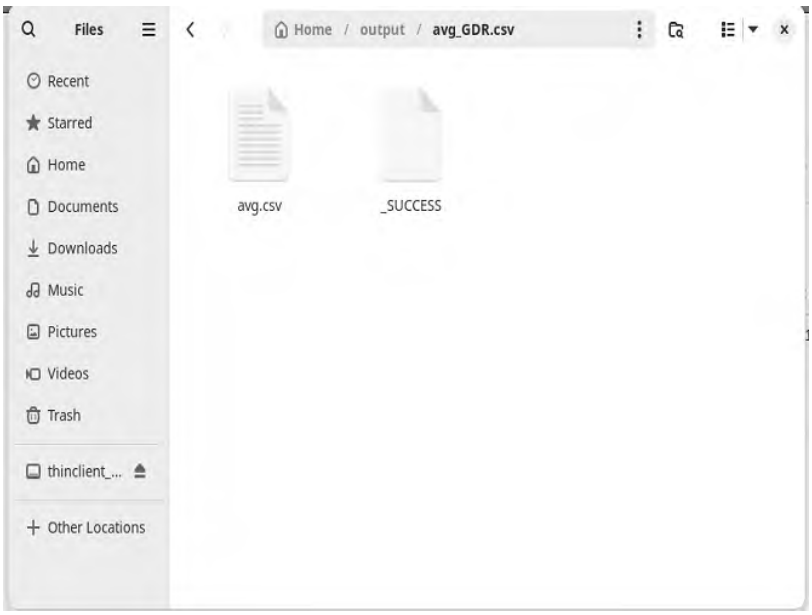

Files
Home / output / averages
Recent
Starred
Home
Documents
Downloads
Music
Pictures
Videos
Trash
thinclient_...
Other Locations
part-00000-56cb3a5d-e2a0-4570-aeba-450... c000.csv
_SUCCESS


Files
Home / output / grouped_data
Recent
Starred
Home
Documents
Downloads
Music
Pictures
Videos
Trash
thinclient_...
Other Locations
part-00000-359d1482-6beb-4a4c-b89a-7a3... c000.csv
_SUCCESS





Файлы из примера



Переименование файлов с результатами

Averages:

mv part-00000-*.csv avg.csv

grouped_data:

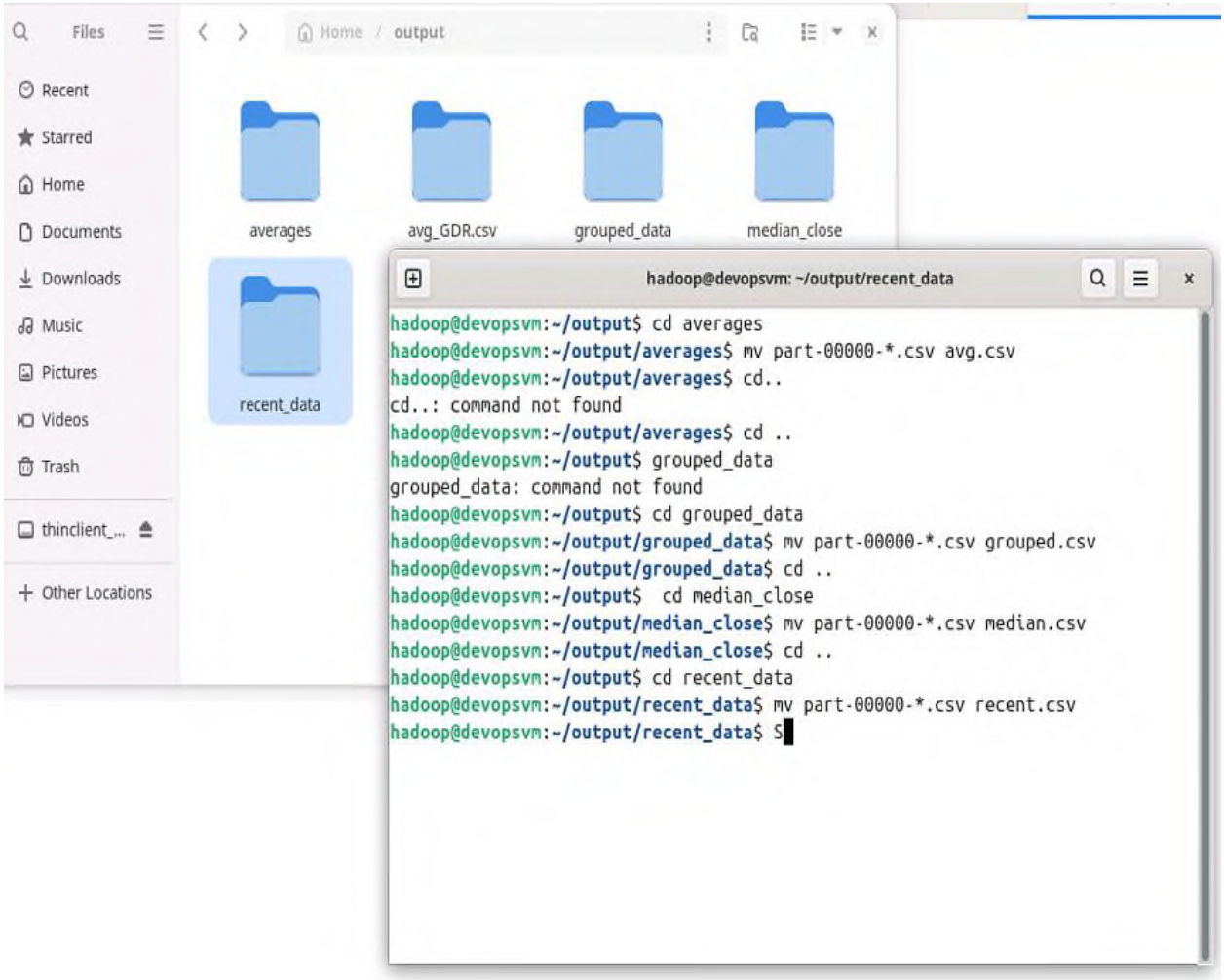
mv part-00000-*.csv grouped.csv

median:

mv part-00000-*.csv median.csv

recent:

mv part-00000-*.csv recent.csv



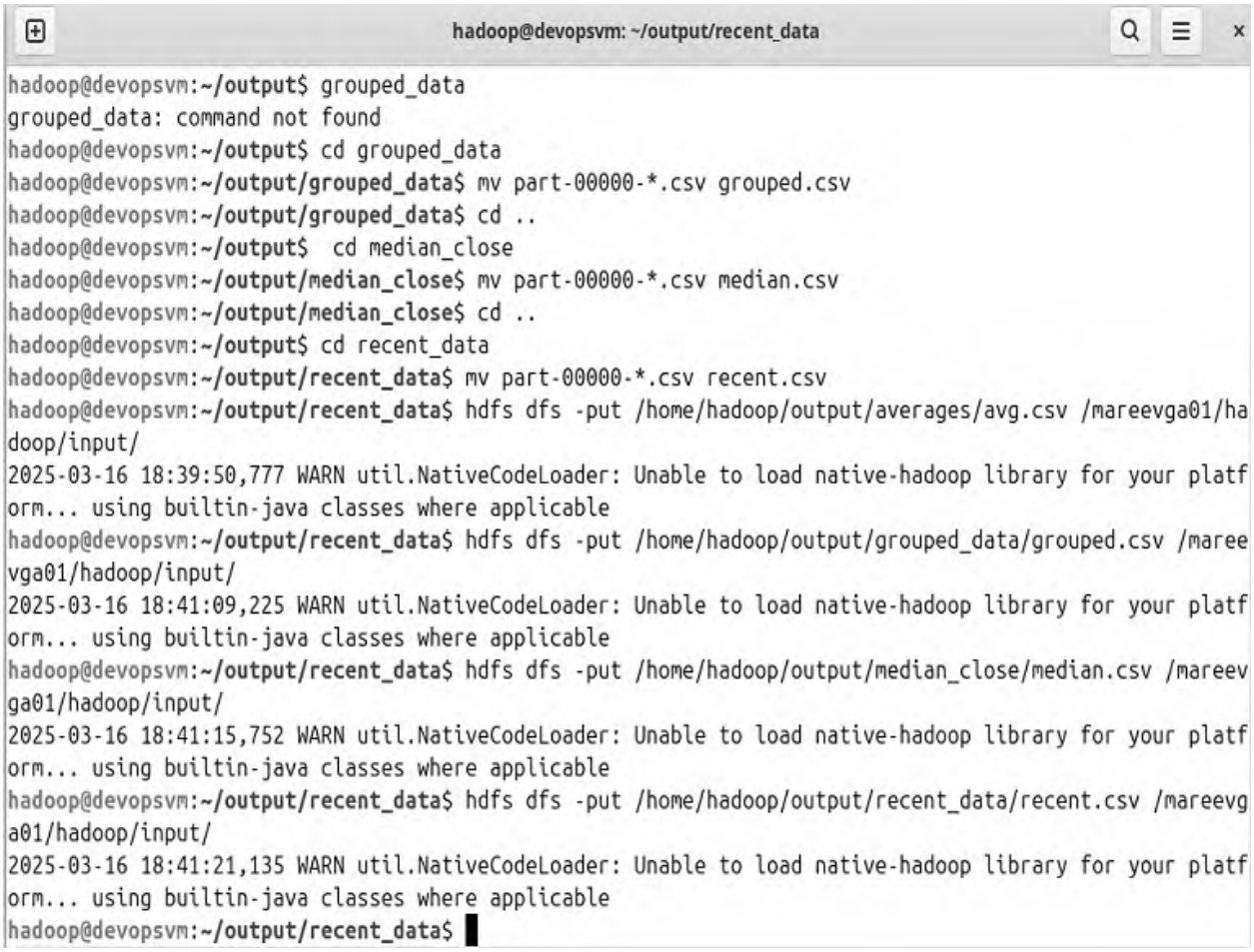
Загрузка результатов в HDFS

hdfs dfs -put /home/hadoop/output/averages/avg.csv /mareevga01/hadoop/input/

hdfs dfs -put /home/hadoop/output/grouped_data/grouped.csv /mareevga01/hadoop/input/

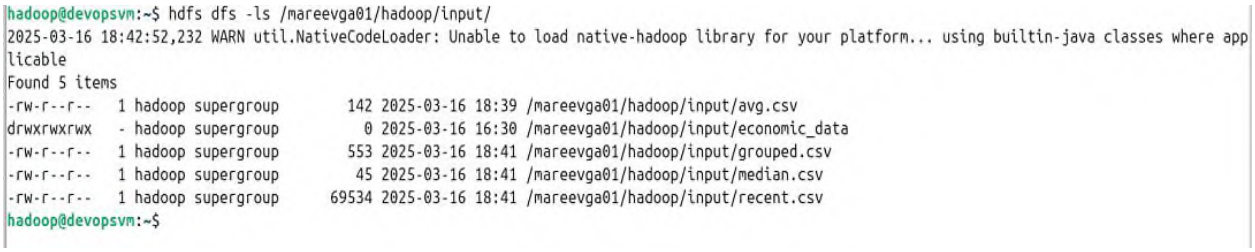
hdfs dfs -put /home/hadoop/output/median_close/median.csv /mareevga01/hadoop/input/

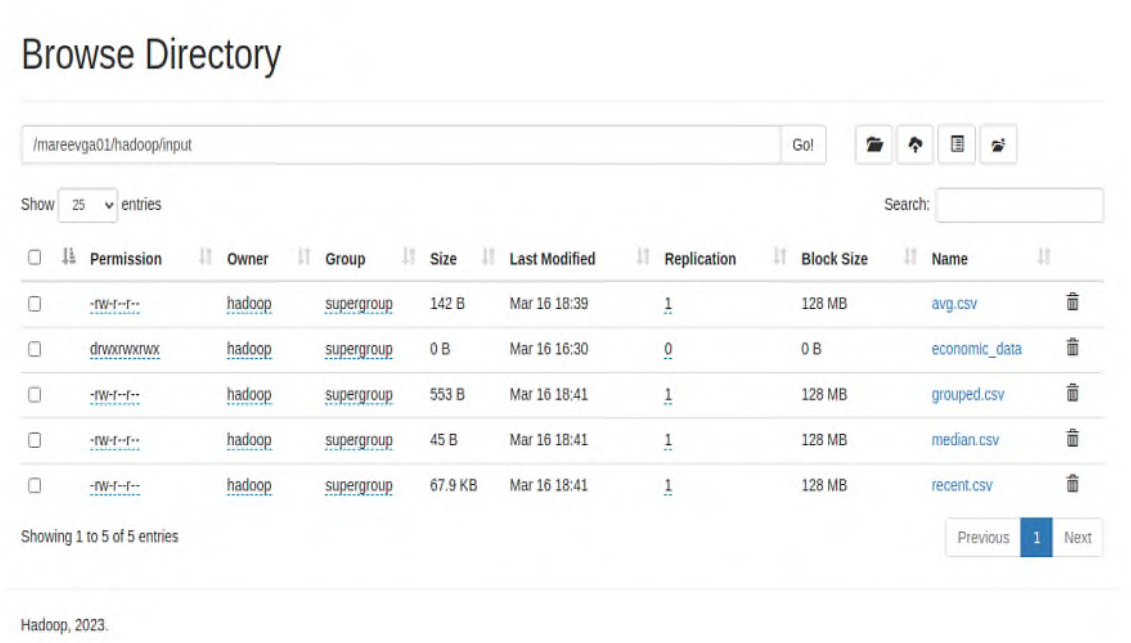
hdfs dfs -put /home/hadoop/output/recent_data/recent.csv /mareevga01/hadoop/input/



Проверка загрузки

hdfs dfs -ls /mareevga01/hadoop/input/





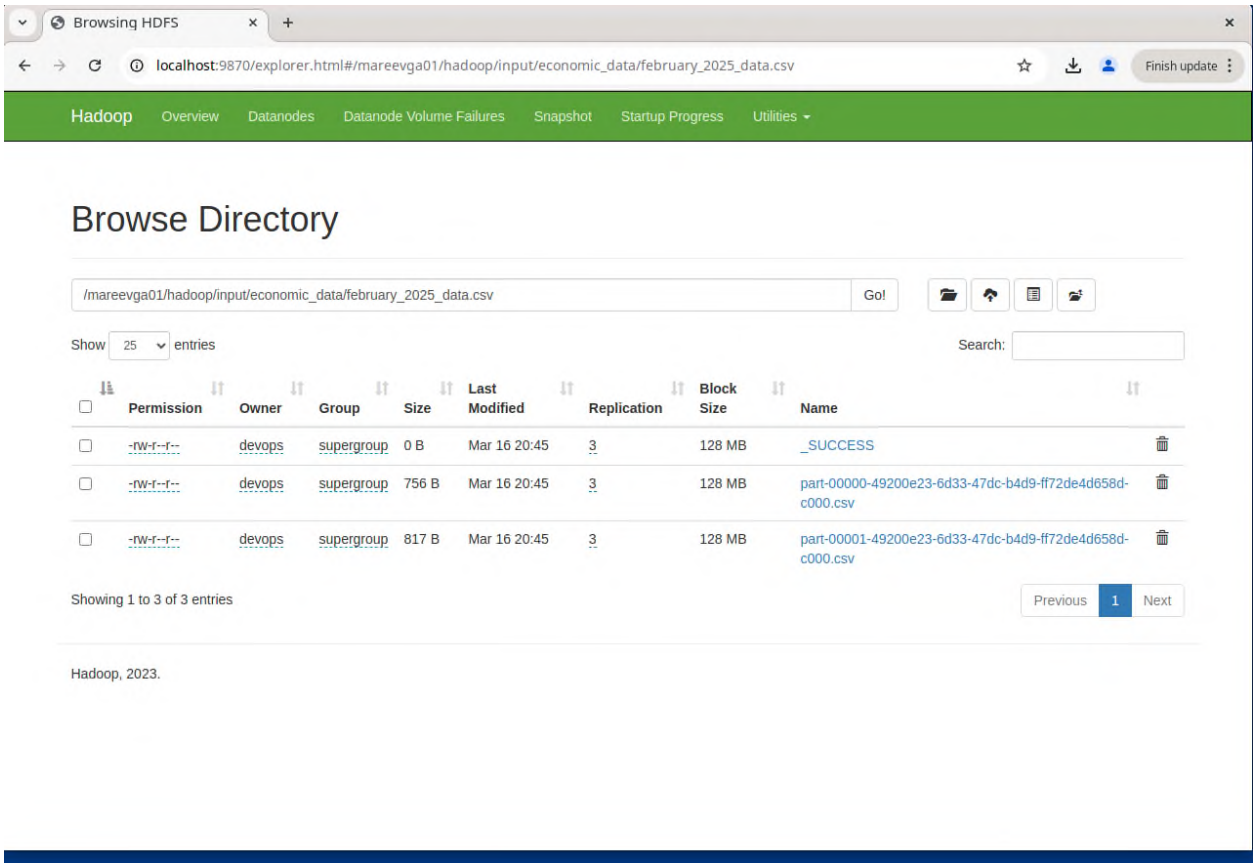
Работы в Jupyter(блокнот work_with_data_2025 приложен):

```
pandas_df = df.toPandas()
pandas_df.head()
```

	Date	Close	High	Low	Open	Volume
0	2020-01-02	35.729675	36.218882	35.649627	36.040989	9147400
1	2020-01-03	36.076572	36.592459	35.578473	36.467934	9752000
2	2020-01-06	36.307835	36.539097	35.747473	36.121048	15534100
3	2020-01-07	36.121052	36.121052	35.338324	36.049893	10971700
4	2020-01-08	35.053692	35.969836	34.964747	35.827523	11327600

Были выгружены данные за февраль 2025 г.

Результат записи



Шаг 7. Завершение работы с Hadoop

```
hadoop@devopsvm:~$ jps
20368 NodeManager
44419 Jps
hadoop@devopsvm:~$ stop-all.sh
WARNING: Stopping all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: Use CTRL-C to abort.
Stopping namenodes on [localhost]
Stopping datanodes
Stopping secondary namenodes [devopsvm]
2025-03-16 20:58:26,288 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
Stopping nodemanagers
Stopping resourcemanager
hadoop@devopsvm:~$ jps
20368 NodeManager
45273 Jps
hadoop@devopsvm:~$
```

Индивидуальное задание. Вариант 13

Поиск медианных значений и квартальная агрегация

- фильтрация данных за последние 3 года,
- расчет медианной цены закрытия,
- группировка по кварталам

Исторические данные по акциям Мечела (MLTR)

https://www.kaggle.com/datasets/svtxvt/moscow-exchange-daily-pricedata), YahooFinance (https://finance.yahoo.com/quote/MTL/history)

Так как ссылки не работали, были скачаны данные по акциям компании Schlumberger Limited (SLB) с YahooFinance за период(с "2020-01-01" по "2025-03-01")

```
import yfinance as yf

ticker = "SLB"

data = yf.download(ticker, start="2020-01-01", end="2025-03-01")

data.to_csv("slb_stock_data.csv")


print(data.head())
```

Результаты вычислений находятся в репозитории в /results:

avg.csv, grouped.csv, median.csv, recent.csv

Заключение

В результате изучения основных операций и функциональных возможностей системы удалось получить понимание принципов работы с данными и распределенными вычислениями. Были рассмотрены ключевые аспекты, такие как управление данными, выполнение операций чтения/записи, обработка больших объемов информации в распределенной среде, а также механизмы параллельных вычислений.