

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

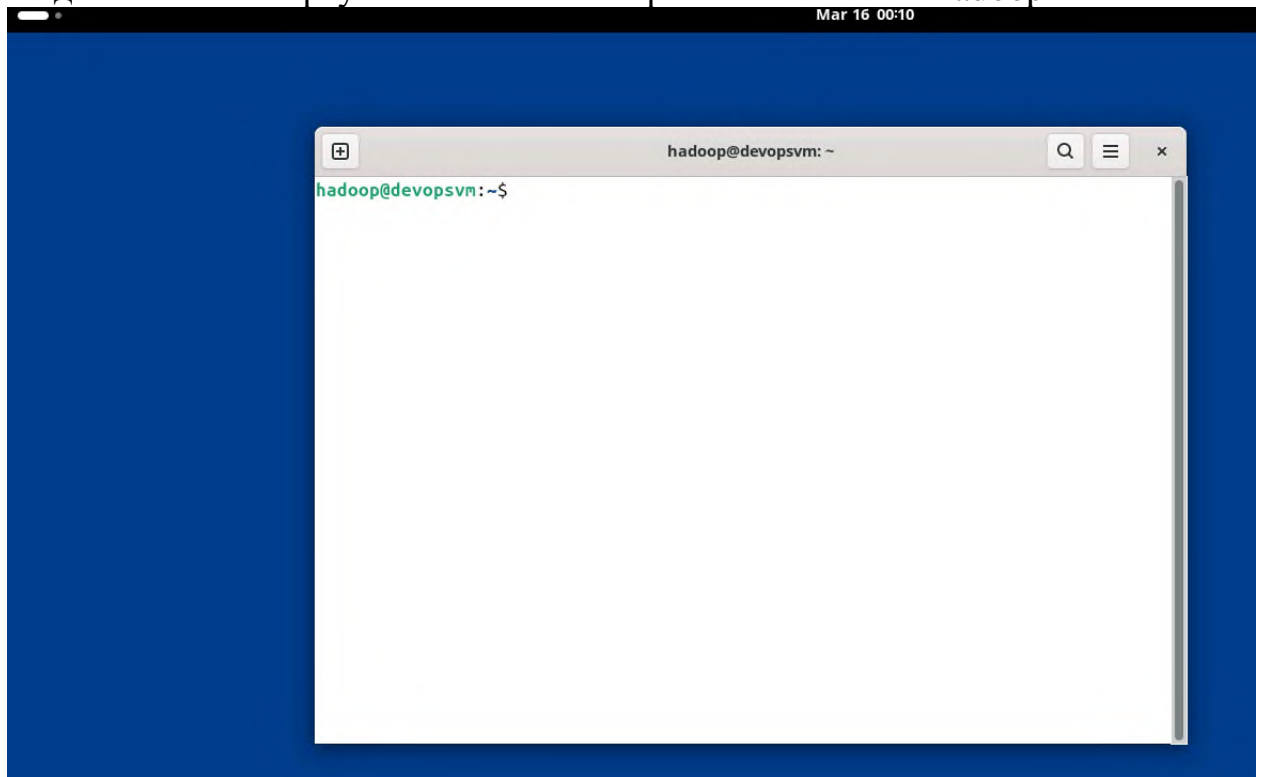
Введение

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

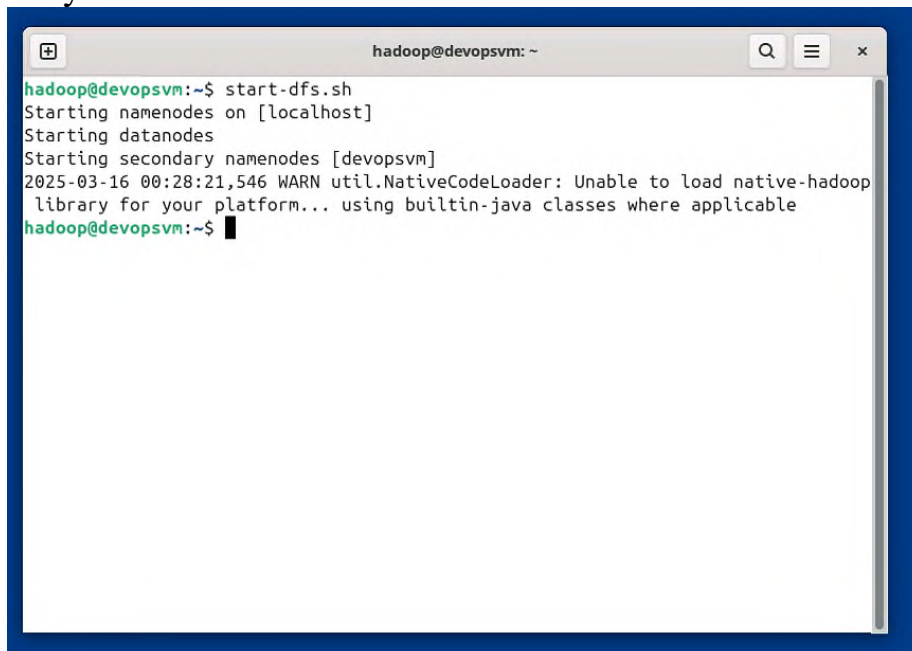
Основная часть

Шаг 1. Запуск Hadoop

Подключение в виртуальной машине через пользователя Hadoop

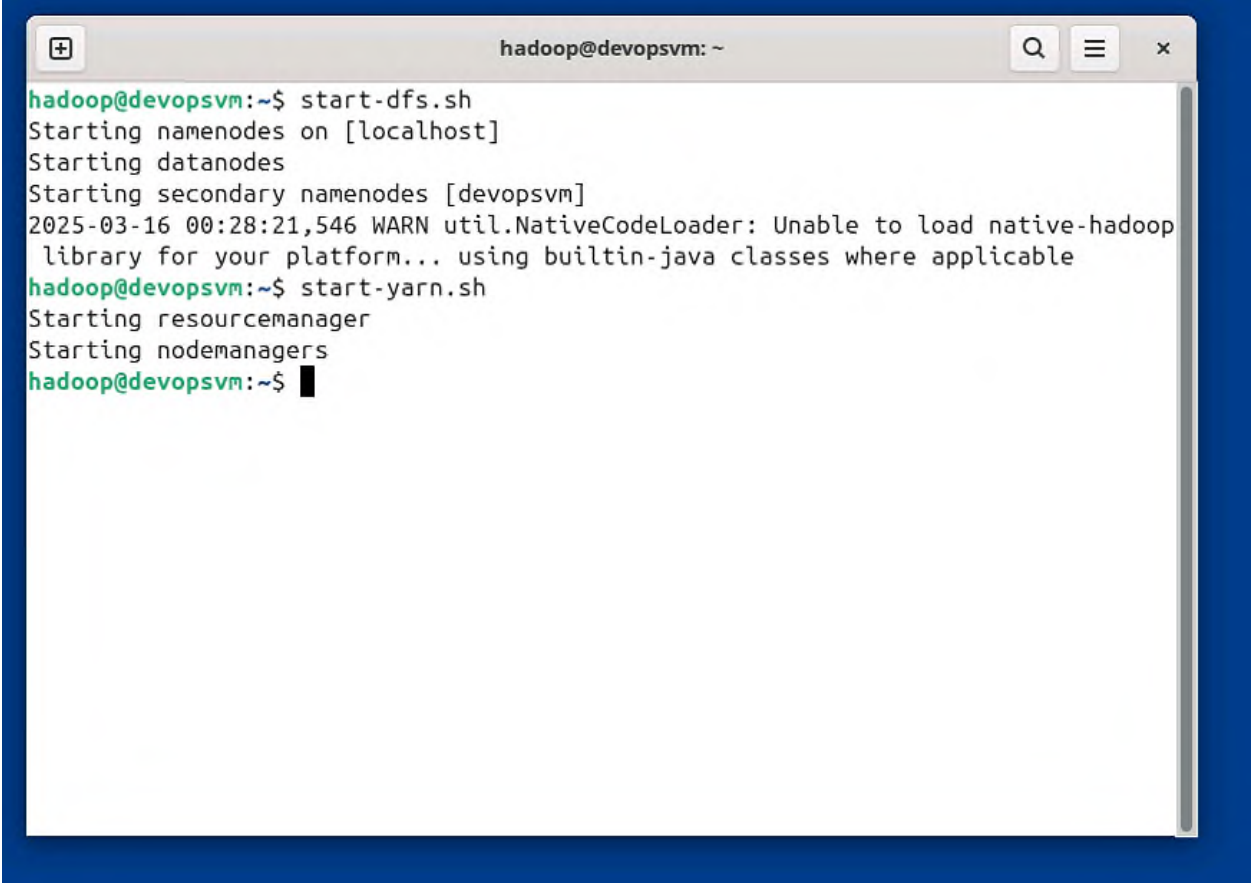


Запуск HDFS



start-dfs.sh

Запуск YARN

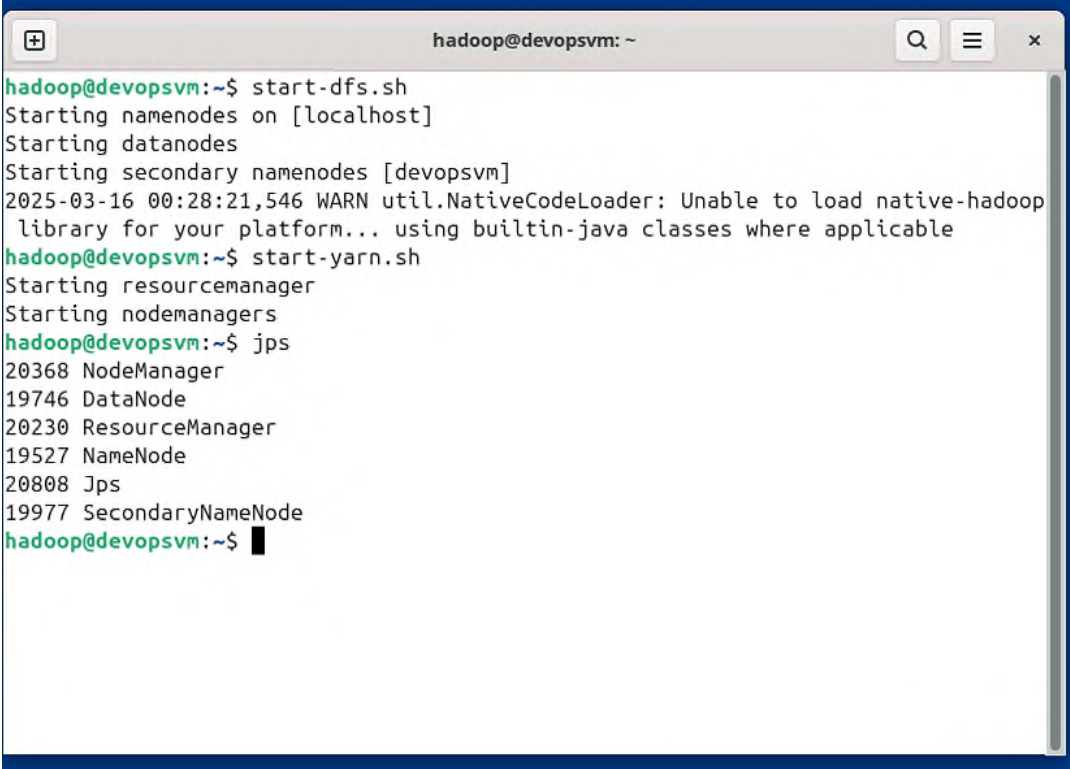


```
hadoop@devopsvm: ~  
hadoop@devopsvm:~$ start-dfs.sh  
Starting namenodes on [localhost]  
Starting datanodes  
Starting secondary namenodes [devopsvm]  
2025-03-16 00:28:21,546 WARN util.NativeCodeLoader: Unable to load native-hadoop  
library for your platform... using builtin-java classes where applicable  
hadoop@devopsvm:~$ start-yarn.sh  
Starting resourcemanager  
Starting nodemanagers  
hadoop@devopsvm:~$
```

start-yarn.sh

Шаг 2. Проверка запущенных служб

jps



```
hadoop@devopsvm: ~  
hadoop@devopsvm:~$ start-dfs.sh  
Starting namenodes on [localhost]  
Starting datanodes  
Starting secondary namenodes [devopsvm]  
2025-03-16 00:28:21,546 WARN util.NativeCodeLoader: Unable to load native-hadoop  
library for your platform... using builtin-java classes where applicable  
hadoop@devopsvm:~$ start-yarn.sh  
Starting resourcemanager  
Starting nodemanagers  
hadoop@devopsvm:~$ jps  
20368 NodeManager  
19746 DataNode  
20230 ResourceManager  
19527 NameNode  
20808 Jps  
19977 SecondaryNameNode  
hadoop@devopsvm:~$
```

Проверка доступности системы

Namenode information

localhost:9870/dfshealth.html#tab-overview

Relaunch to update

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities

Overview

'localhost:9000' (active)

Started:	Sun Mar 16 00:28:10 +0300 2025
Version:	3.3.5, r706d88266abcee09ed78fbaa0ad5f74d818ab0e9
Compiled:	Wed Mar 15 18:56:00 +0300 2023 by stevel from branch-3.3.5
Cluster ID:	CID-60a52b68-6139-4947-8731-3c039547a32e
Block Pool ID:	BP-1830111676-127.0.1.1-1724666841903

Summary

Security is off.
Safemode is off.
17 files and directories, 5 blocks (5 replicated blocks, 0 erasure coded block groups) = 22 total filesystem object(s).
Heap Memory used 106.47 MB of 182 MB Heap Memory. Max Heap Memory is 1.94 GB.
Non Heap Memory used 54.62 MB of 57.5 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	34.15 GB
Configured Remote Capacity:	0 B
DFS Used:	456 KB (0%)
Non DFS Used:	17.34 GB
DFS Remaining:	15.05 GB (44.06%)

localhost:9870

Каталоги файловой системы

Browsing HDFS

localhost:9870/explorer.html#

Relaunch to update

Hadoop

Overview

Datanodes

Datanode Volume Failures

Snapshot

Startup Progress

Utilities

Browse Directory

/

Go!

Show

25

entries

Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	Aug 26 2024	0	0 B	user	<div></div>
<input type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	Oct 17 23:33	0	0 B	user2	<div></div>

Showing 1 to 2 of 2 entries

Previous

1

Next

Hadoop, 2023.

Проверка YARN

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources
0	0	0	0	0	<memory:0 B, vCores:0>	<memory:8 GB, vCores:8>

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes
1	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[memory-mb (unit=M), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>

Showing 0 to 0 of 0 entries

Шаг 3. Подготовка рабочего пространства Создание директории в HDFS

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

```
hadoop@devopsvm:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [devopsvm]
2025-03-16 00:28:21,546 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
hadoop@devopsvm:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hadoop@devopsvm:~$ jps
20368 NodeManager
19746 DataNode
20230 ResourceManager
19527 NameNode
20808 Jps
19977 SecondaryNameNode
hadoop@devopsvm:~$ hdfs dfs -mkdir -p /admin01/hadoop/input
2025-03-16 00:48:45,440 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
hadoop@devopsvm:~$
```

Browse Directory

Show 25 entries

Search:

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

Showing 1 to 3 of 3 entries

Hadoop, 2023.

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

```
hadoop@devopsvm: ~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [devopsvm]
2025-03-16 00:28:21,546 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
hadoop@devopsvm:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hadoop@devopsvm:~$ jps
20368 NodeManager
19746 DataNode
20230 ResourceManager
19527 NameNode
20808 Jps
19977 SecondaryNameNode
hadoop@devopsvm:~$ hdfs dfs -mkdir -p /admin01/hadoop/input
2025-03-16 00:48:45,440 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
hadoop@devopsvm:~$ hdfs dfs -mkdir -p /mareevga01/hadoop/input
2025-03-16 00:50:35,742 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
hadoop@devopsvm:~$
```


Browse Directory

Show 25 entries

Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
<input type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	Mar 16 00:48	0	0 B	admin01	
<input type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	Mar 16 00:50	0	0 B	mareevga01	
<input type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	Aug 26 2024	0	0 B	user	
<input type="checkbox"/>	drwxr-xr-x	hadoop	supergroup	0 B	Oct 17 23:33	0	0 B	user2	

Showing 1 to 4 of 4 entries

Previous

1

Next

Hadoop, 2023.

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: ~$ wget https://raw.githubusercontent.com/GoshaMareev/Distributed_systems/refs/heads/main/lw_01/slb.csv  
--2025-03-16 16:28:30-- https://raw.githubusercontent.com/GoshaMareev/Distributed_systems/refs/heads/main/lw_01/slb.csv  
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.108.133, 185.199.111.133, 185.199.110.133, ...  
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.108.133|:443... connected.  
HTTP request sent, awaiting response... 200 OK  
Length: 120692 (118K) [text/plain]  
Saving to: 'slb.csv'  
  
slb.csv          100%[=====] 117.86K  --.-KB/s    in 0.04s  
  
2025-03-16 16:28:31 (2.58 MB/s) - 'slb.csv' saved [120692/120692]  
  
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

```
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
```

The image shows a terminal window and an HDFS Explorer web interface. The terminal window, titled 'hadoop@devopsvm: ~', shows the process of downloading a CSV file from a GitHub repository. It displays the file path, the resolution of the domain, the connection status, the file size (120730 bytes), and the download progress (100%). The file is saved as 'slb_stock_data.csv'. Below the download progress, the terminal shows the command 'ls' and the output of the 'ls' command, which lists the contents of the current directory. The output includes 'Desktop', 'Documents', 'Downloads', 'GDP.csv', 'hadoop-3.3.5.tar.gz', 'hdfs', 'Music', 'output', 'Pictures', 'Public', 'slb_stock_data.csv', 'snap', 'spark-3.4.3-bin-hadoop3.tgz', 'Templates', 'thinclient_drives', and 'Videos'. The terminal also shows the command 'nano slb_stock_data.csv' and the command 'hdfs dfs -mkdir -p /mareevga01/hadoop/input/economic_data'. The HDFS Explorer web interface, titled 'Browse Directory', shows the path '/mareevga01/hadoop/input/economic_data' and a table of entries. The table is empty, showing 'No data available in table'.

```
hadoop@devopsvm: ~  
--2025-03-16 01:04:11-- https://raw.githubusercontent.com/GoshaMareev/Distribut  
ed_systems/refs/heads/main/lw_01/slb_stock_data.csv  
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.110.1  
33, 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:~$ 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:~$
```

Browsing HDFS x All Applications x +
localhost:9870/explorer.html#/mareevga01/hadoop/input/economic_data
Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities +

Browse Directory

/mareevga01/hadoop/input/economic_data Go!

Show 25 entries Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
No data available in table								

Showing 0 to 0 of 0 entries Previous Next

Hadoop, 2023.

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
Content-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
IDP.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

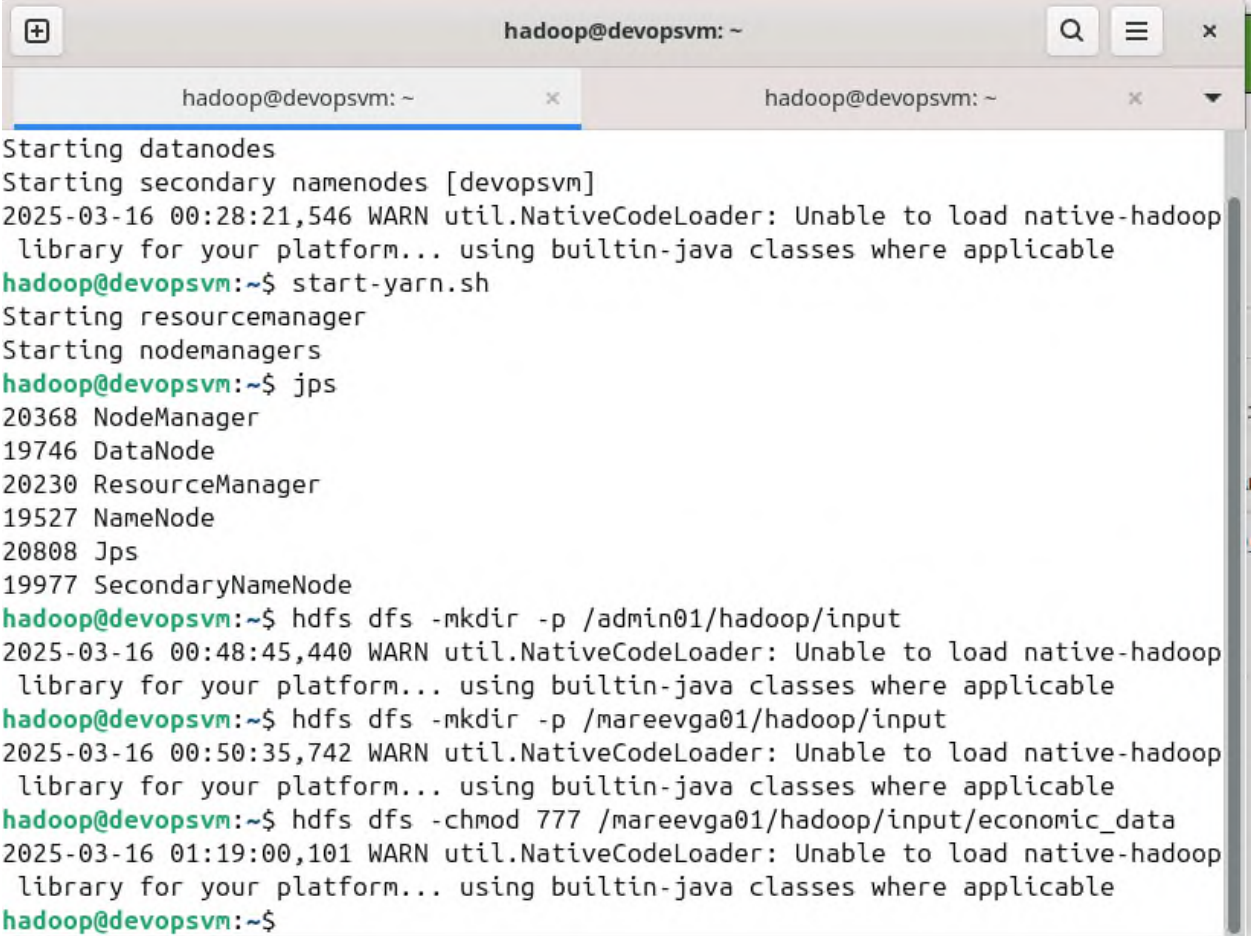
1

Next

Hadoop, 2023.

Установка прав доступа

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



```
hadoop@devopsvm: ~  
Starting datanodes  
Starting secondary namenodes [devopsvm]  
2025-03-16 00:28:21,546 WARN util.NativeCodeLoader: Unable to load native-hadoop  
library for your platform... using builtin-java classes where applicable  
hadoop@devopsvm:~$ start-yarn.sh  
Starting resourcemanager  
Starting nodemanagers  
hadoop@devopsvm:~$ jps  
20368 NodeManager  
19746 DataNode  
20230 ResourceManager  
19527 NameNode  
20808 Jps  
19977 SecondaryNameNode  
hadoop@devopsvm:~$ hdfs dfs -mkdir -p /admin01/hadoop/input  
2025-03-16 00:48:45,440 WARN util.NativeCodeLoader: Unable to load native-hadoop  
library for your platform... using builtin-java classes where applicable  
hadoop@devopsvm:~$ hdfs dfs -mkdir -p /mareevga01/hadoop/input  
2025-03-16 00:50:35,742 WARN util.NativeCodeLoader: Unable to load native-hadoop  
library for your platform... using builtin-java classes where applicable  
hadoop@devopsvm:~$ hdfs dfs -chmod 777 /mareevga01/hadoop/input/economic_data  
2025-03-16 01:19:00,101 WARN util.NativeCodeLoader: Unable to load native-hadoop  
library for your platform... using builtin-java classes where applicable  
hadoop@devopsvm:~$
```

Шаг 5. Обработка данных с помощью Spark

Spark-shell

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:

```
hadoop@devopsvm: ~  
hadoop@devopsvm:~$ spark-shell -i /home/hadoop/scripts/script.scala  
25/03/16 17:57:12 WARN Utils: Your hostname, devopsvm resolves to a loopback address: 127.0.1.1; using 192.168.50.8 instead (on interface enp0s3)  
25/03/16 17:57:12 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address  
Setting default log level to "WARN".  
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).  
25/03/16 17:57:18 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable  
25/03/16 17:57:19 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.  
Spark context Web UI available at http://192.168.50.8:4041  
Spark context available as 'sc' (master = local[*], app id = local-1742137039607).  
Spark session available as 'spark'.  
Средние значения:  
+-----+-----+-----+-----+-----+  
| avg_Close | avg_High | avg_Low | avg_Open | avg_Volume |  
+-----+-----+-----+-----+-----+  
| 36.485162448956586 | 37.06012920306286 | 35.912926763001586 | 36.49513578895775 | 1.2569641788743254E7 |  
+-----+-----+-----+-----+-----+  
Данные за последние 3 года:  
+-----+-----+-----+-----+-----+  
| Date | Close | High | Low | Open | Volume |  
+-----+-----+-----+-----+-----+  
| 2022-03-16 | 35.58138656616211 | 37.460023607495884 | 35.31837494378864 | 36.98096915535493 | 2.74992E7 |  
| 2022-03-17 | 37.11247253417969 | 37.42244936059153 | 35.98529319541303 | 36.39859204741309 | 3.79863E7 |  
| 2022-03-18 | 37.31912612915039 | 37.84514581510905 | 36.76492803289222 | 37.05611807777904 | 2.82676E7 |  
| 2022-03-21 | 38.69052505493164 | 38.97232345205674 | 38.06118342330731 | 38.2020790386544 | 1.92837E7 |  
| 2022-03-22 | 38.699920654296875 | 39.46076630156078 | 38.389943847474164 | 38.699920654296875 | 1.54671E7 |  
| 2022-03-23 | 39.83649444580078 | 40.26858006333437 | 39.44198242541327 | 39.56409124131487 | 1.99675E7 |  
| 2022-03-24 | 40.052547454833984 | 41.06701333684163 | 39.742574157475886 | 39.91165180447716 | 1.62814E7 |  
| 2022-03-25 | 41.02943801879883 | 41.04822482137275 | 39.70499889467511 | 39.892863337198115 | 1.22529E7 |  
| 2022-03-28 | 39.19776153564453 | 40.05254124826423 | 38.925358303996894 | 40.01496764737163 | 1.39842E7 |  
| 2022-03-29 | 40.043155670166016 | 40.2028381231841 | 37.7512158355716 | 38.03301430148926 | 1.32075E7 |  
| 2022-03-30 | 39.64862823486328 | 40.81338461075769 | 39.40440703098388 | 40.231008214418125 | 1.02017E7 |  
| 2022-03-31 | 38.80324935913086 | 39.83649829805685 | 38.737497349374614 | 39.16018781688883 | 1.16187E7 |  
| 2022-04-01 | 39.16018781688883 | 39.89285374421457 | 38.681132143699315 | 38.6999189417556 | 1.22106E7 |  
| 2022-04-04 | 39.028682708740234 | 39.70499317104377 | 38.26783702186445 | 39.376229548217026 | 1.06182E7 |  
| 2022-04-05 | 38.004825592041016 | 39.78953217260422 | 37.92968197542784 | 38.95353928219428 | 1.04739E7 |  
| 2022-04-06 | 38.3054084777832 | 38.83142814414283 | 37.873322814281984 | 38.483881291606274 | 8726600.0 |  
| 2022-04-07 | 38.48388671875 | 38.944150856409024 | 37.32852519477675 | 38.521456741975804 | 8895800.0 |  
| 2022-04-08 | 39.9680061340332 | 40.12769216098632 | 38.61538862575426 | 38.82203986655848 | 1.05264E7 |  
| 2022-04-11 | 38.8220329284668 | 40.05253779424991 | 38.76567253198826 | 40.05253779424991 | 7999000.0 |  
| 2022-04-12 | 39.0098876953125 | 40.30614794739378 | 38.79384491716811 | 39.44197683481622 | 8112800.0 |  
+-----+-----+-----+-----+-----+  
only showing top 20 rows
```


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:

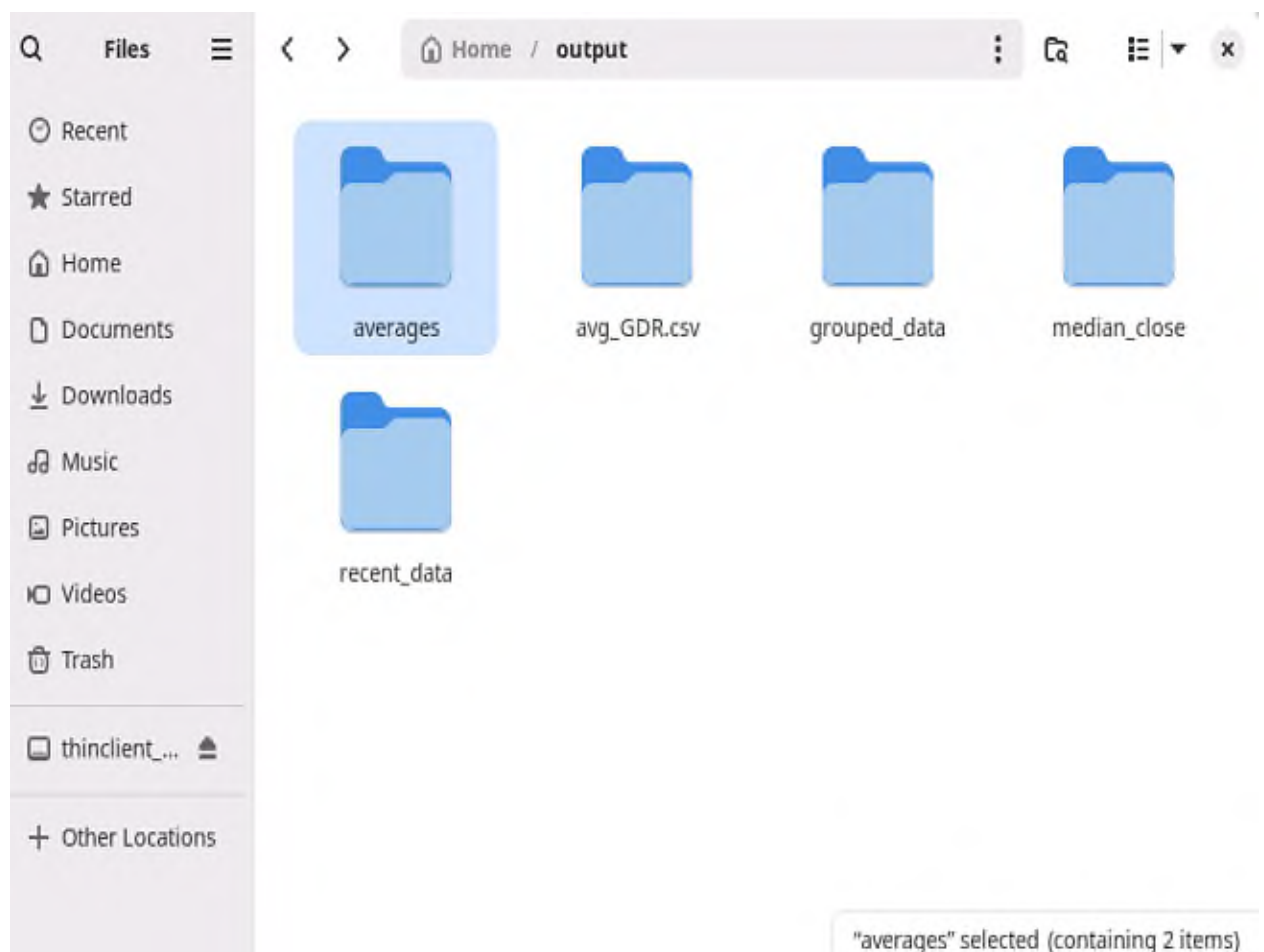
`file:///home/hadoop/output/averages`

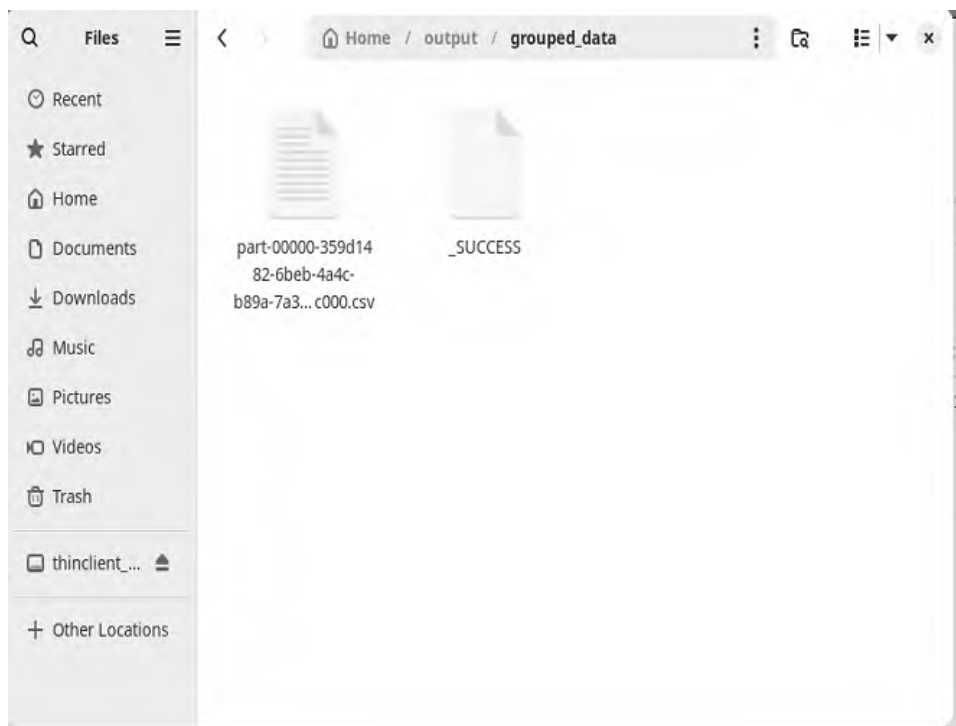
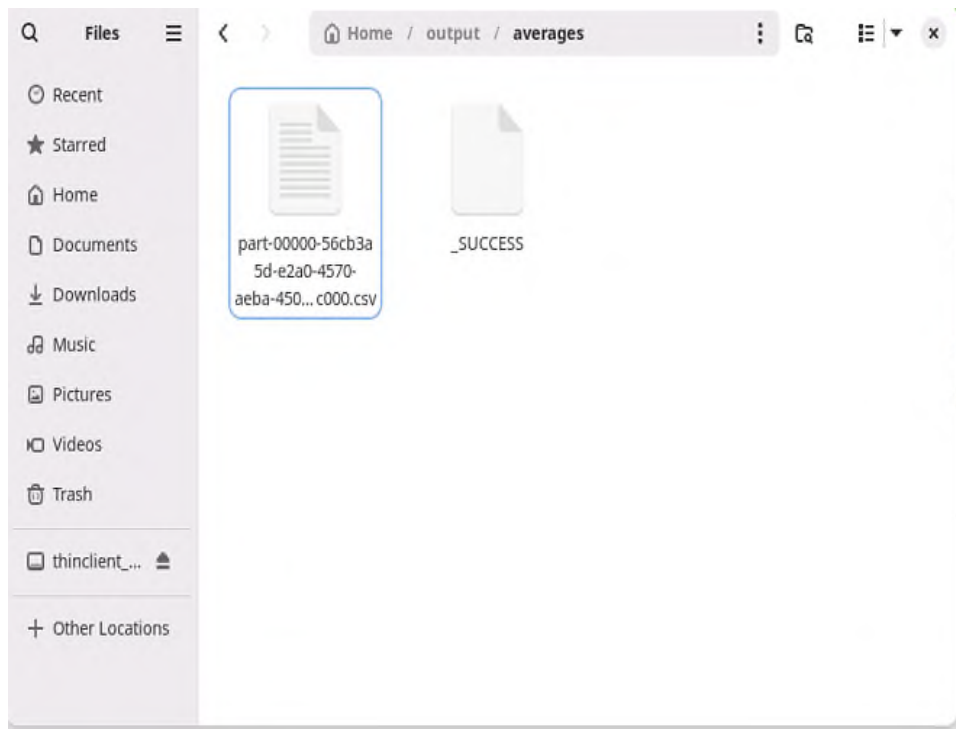
`file:///home/hadoop/output/recent_data`

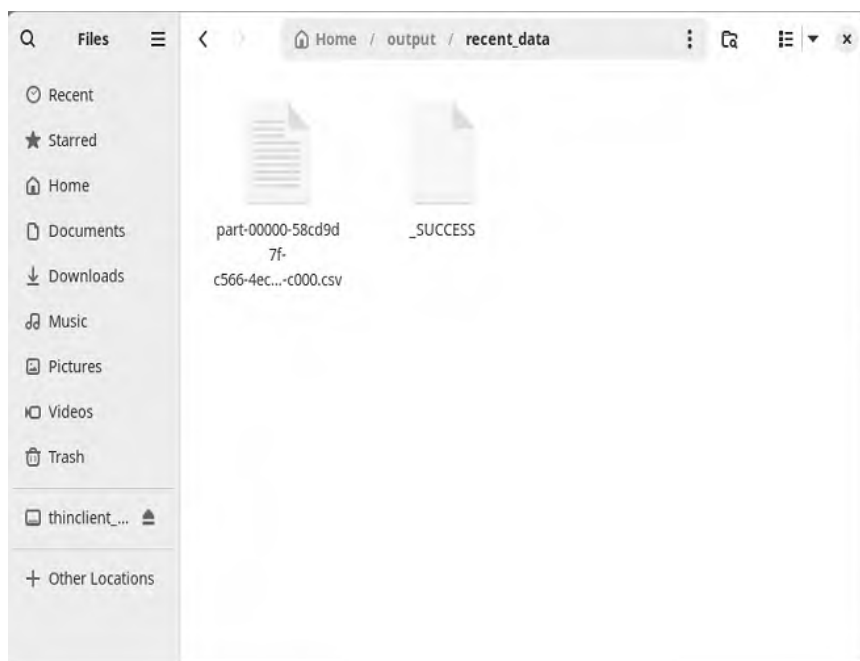
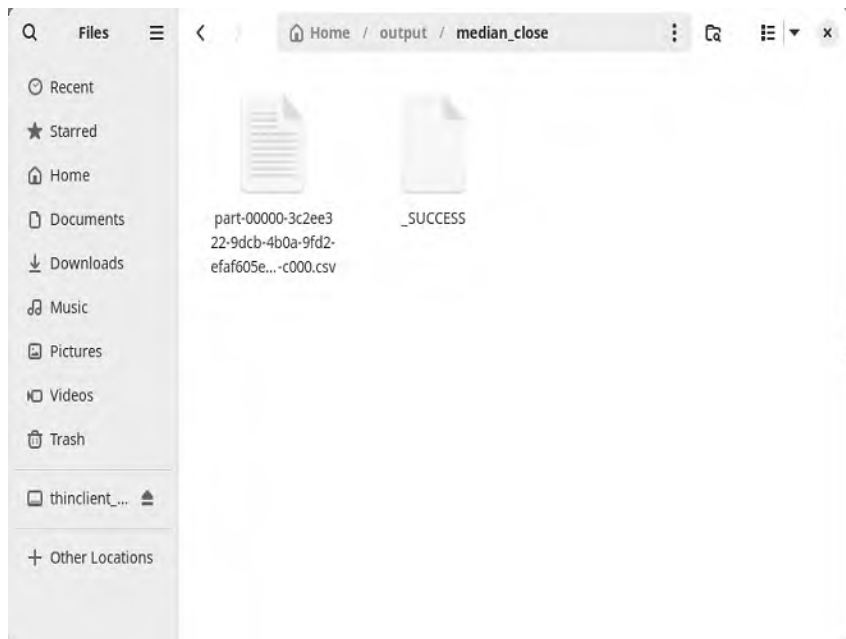
`file:///home/hadoop/output/median_close`

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

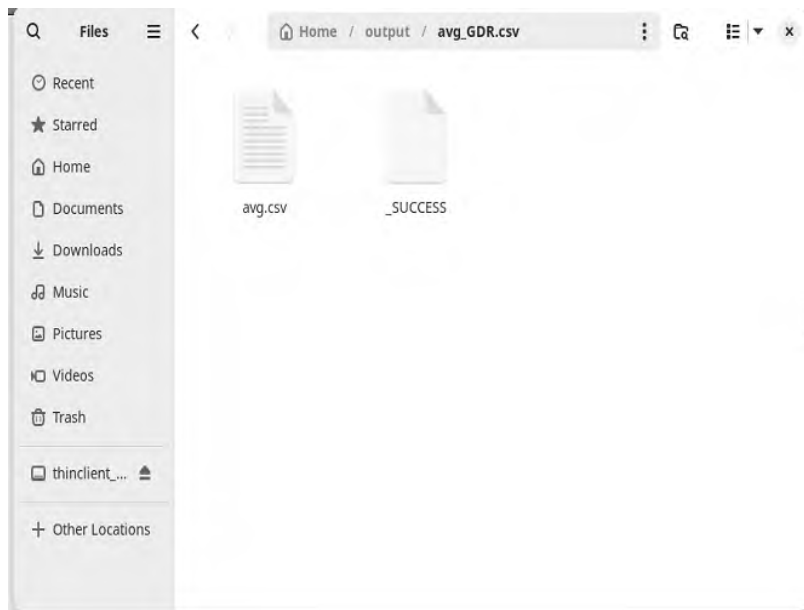
```
hadoop@devopsvm:~$ ls /home/hadoop/output/  
averages  avg_GDR.csv  grouped_data  median_close  recent_data  
hadoop@devopsvm:~$
```







Файлы из примера



Переименование файлов с результатами

Averages:

```
mv part-00000-*.csv avg.csv
```

grouped_data:

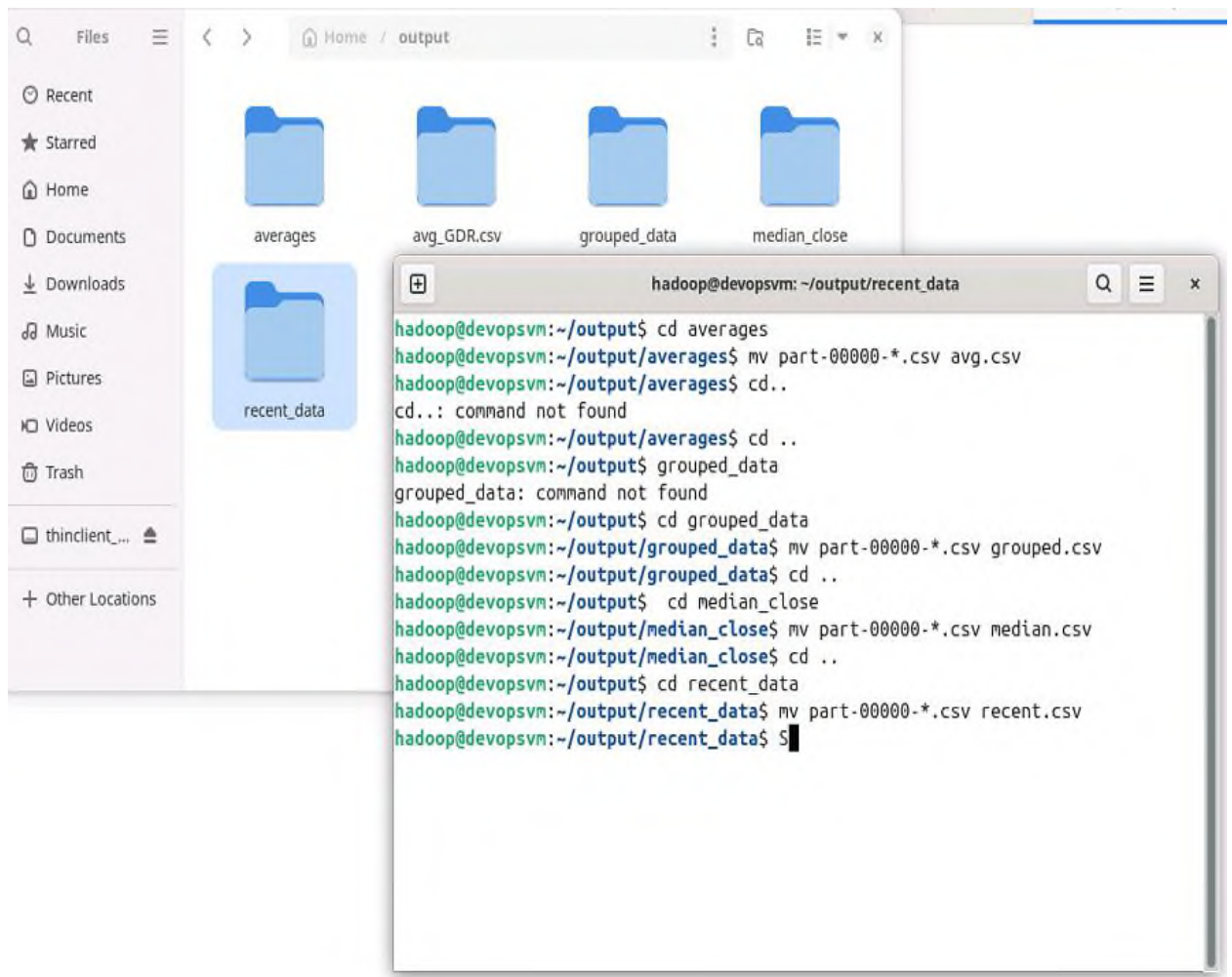
```
mv part-00000-*.csv grouped.csv
```

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/

```
hadoop@devopsvm: ~/output/recent_data
hadoop@devopsvm:~/output$ grouped_data
grouped_data: command not found
hadoop@devopsvm:~/output$ cd grouped_data
hadoop@devopsvm:~/output/grouped_data$ mv part-00000-*.csv grouped.csv
hadoop@devopsvm:~/output/grouped_data$ cd ..
hadoop@devopsvm:~/output$ cd median_close
hadoop@devopsvm:~/output/median_close$ mv part-00000-*.csv median.csv
hadoop@devopsvm:~/output/median_close$ cd ..
hadoop@devopsvm:~/output$ cd recent_data
hadoop@devopsvm:~/output/recent_data$ mv part-00000-*.csv recent.csv
hadoop@devopsvm:~/output/recent_data$ hdfs dfs -put /home/hadoop/output/averages/avg.csv /mareevga01/hadoop/input/
2025-03-16 18:39:50,777 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hadoop@devopsvm:~/output/recent_data$ hdfs dfs -put /home/hadoop/output/grouped_data/grouped.csv /mareevga01/hadoop/input/
2025-03-16 18:41:09,225 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hadoop@devopsvm:~/output/recent_data$ hdfs dfs -put /home/hadoop/output/median_close/median.csv /mareevga01/hadoop/input/
2025-03-16 18:41:15,752 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hadoop@devopsvm:~/output/recent_data$ hdfs dfs -put /home/hadoop/output/recent_data/recent.csv /mareevga01/hadoop/input/
2025-03-16 18:41:21,135 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hadoop@devopsvm:~/output/recent_data$
```

Проверка загрузки

`hdfs dfs -ls /mareevga01/hadoop/input/`

```
hadoop@devopsvm:~$ hdfs dfs -ls /mareevga01/hadoop/input/
2025-03-16 18:42:52,232 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 5 items
-rw-r--r-- 1 hadoop supergroup 142 2025-03-16 18:39 /mareevga01/hadoop/input/avg.csv
drwxrwxrwx - hadoop supergroup 0 2025-03-16 16:30 /mareevga01/hadoop/input/economic_data
-rw-r--r-- 1 hadoop supergroup 553 2025-03-16 18:41 /mareevga01/hadoop/input/grouped.csv
-rw-r--r-- 1 hadoop supergroup 45 2025-03-16 18:41 /mareevga01/hadoop/input/median.csv
-rw-r--r-- 1 hadoop supergroup 69534 2025-03-16 18:41 /mareevga01/hadoop/input/recent.csv
hadoop@devopsvm:~$
```


Browse Directory

Show 25 entries

Search:

<input type="checkbox"/>	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	<input type="checkbox"/>
<input type="checkbox"/>	-rw-r--r--	hadoop	supergroup	142 B	Mar 16 18:39	1	128 MB	avg.csv	<input type="checkbox"/>
<input type="checkbox"/>	drwxrwxrwx	hadoop	supergroup	0 B	Mar 16 16:30	0	0 B	economic_data	<input type="checkbox"/>
<input type="checkbox"/>	-rw-r--r--	hadoop	supergroup	553 B	Mar 16 18:41	1	128 MB	grouped.csv	<input type="checkbox"/>
<input type="checkbox"/>	-rw-r--r--	hadoop	supergroup	45 B	Mar 16 18:41	1	128 MB	median.csv	<input type="checkbox"/>
<input type="checkbox"/>	-rw-r--r--	hadoop	supergroup	67.9 KB	Mar 16 18:41	1	128 MB	recent.csv	<input type="checkbox"/>

Showing 1 to 5 of 5 entries

Hadoop, 2023.

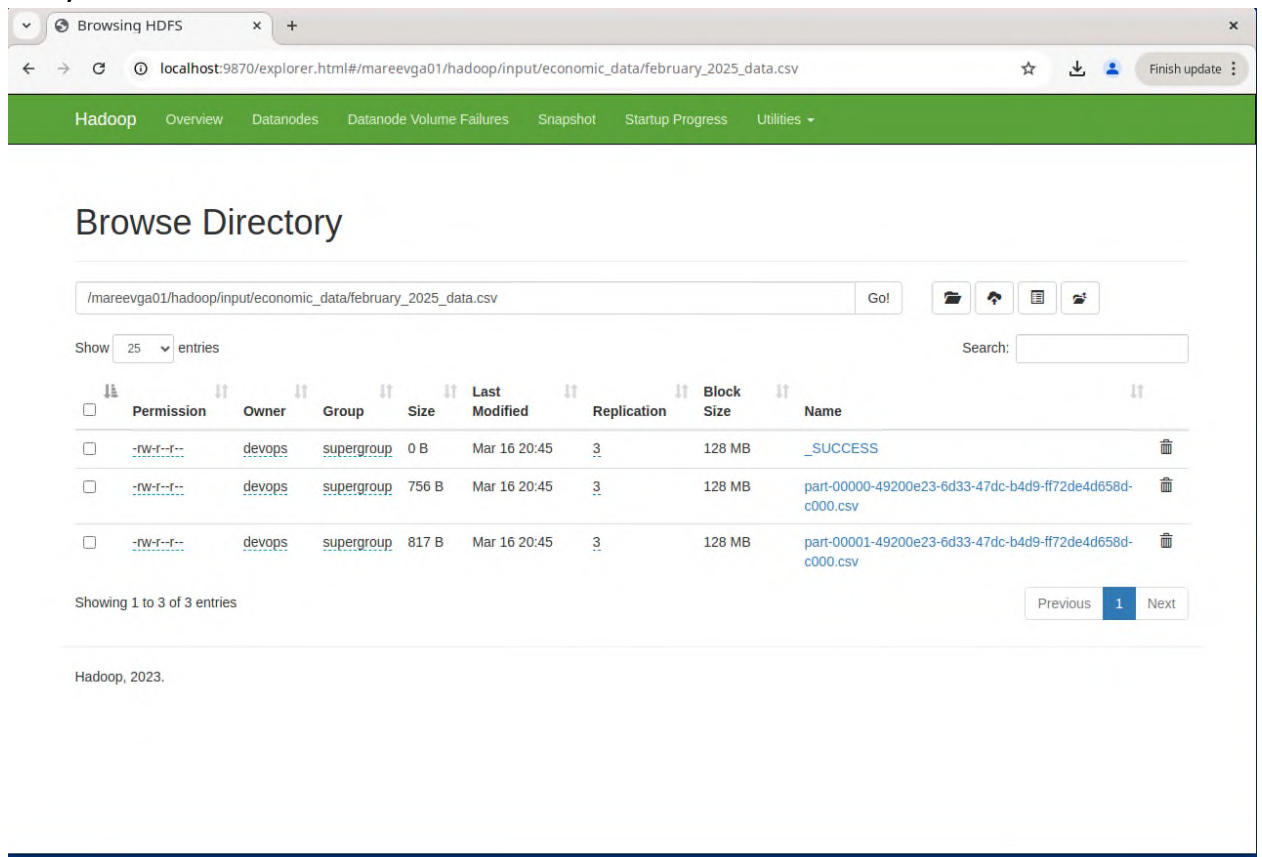
Работы в Jupyter(блокнот work_with_data_2025 приложен):

```
: 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



Browsing HDFS

localhost:9870/explorer.html#/mareevga01/hadoop/input/economic_data/february_2025_data.csv

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Browse Directory

/mareevga01/hadoop/input/economic_data/february_2025_data.csv Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	devops	supergroup	0 B	Mar 16 20:45	3	128 MB	_SUCCESS
-rw-r--r--	devops	supergroup	756 B	Mar 16 20:45	3	128 MB	part-00000-49200e23-6d33-47dc-b4d9-ff72de4d658d-c000.csv
-rw-r--r--	devops	supergroup	817 B	Mar 16 20:45	3	128 MB	part-00001-49200e23-6d33-47dc-b4d9-ff72de4d658d-c000.csv

Showing 1 to 3 of 3 entries

Previous 1 Next

Hadoop, 2023.

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.