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

ДИСЦИПЛИНА:

Инструменты для хранения и обработки больших данных

Лабораторная работа №5-1

Тема:

«Развертывание и настройка кластера Hadoop»

Выполнил(а): Морозова Валерия АДЭУ-211

Преподаватель: Босенко Т.М.

Москва

2024

Цель: ознакомление с процессом установки и настройки распределенных систем, таких как Apache(Arenadata) Hadoop. Изучить основные операции и функциональные возможности системы, что позволит понять принципы работы с данными и распределенными вычислениями.

Необходимое ПО:

- Ubuntu 24.04 LTS (22.04, 20.04) или новее.
- Java 8 ил Java11 или новее.
- Apache Spark 3.4.3.
- Python 3.12+.
- pip (менеджер пакетов Python).

Практика на паре

В первую очередь необходимо было войти пользователем hadoop (рисунок 1).

```
devops@devopsvm:~$ sudo su hadoop
[sudo] password for devops:
```

Рисунок 1. Вход через пользователя hadoop.

```
hadoop@devopsvm:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [devopsvm]
2024-10-25 11:20:11,515 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
```

Рисунок 2. Запуск Hadoop

Чтобы получить доступ к веб-интерфейсу HDFS, нужно ввести http://localhost:9870 в адресной строке.

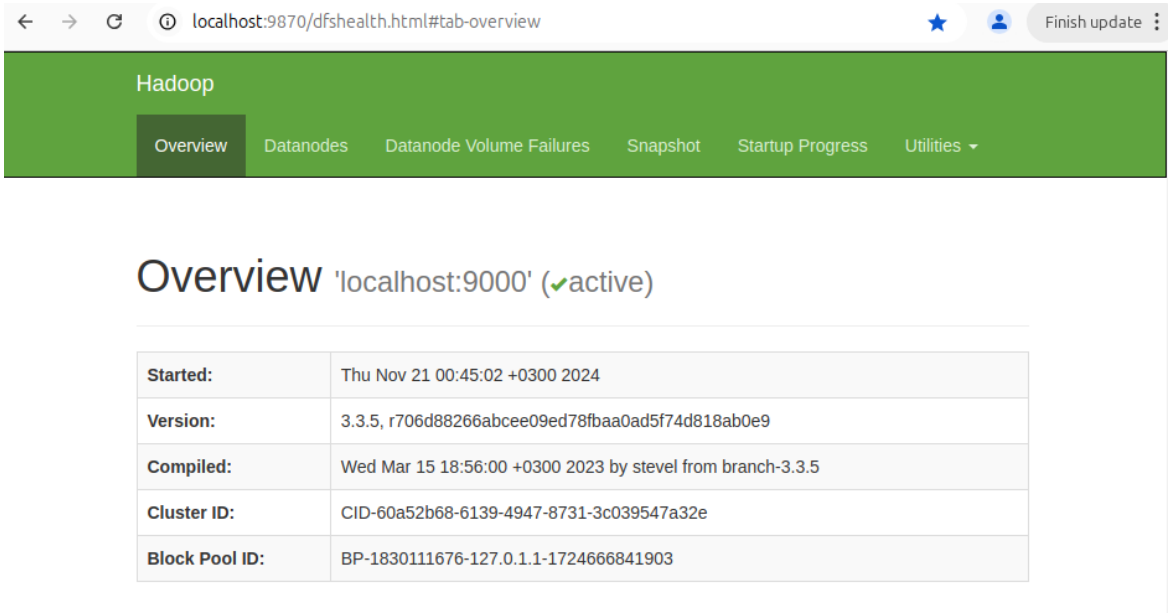


Рисунок 3. Успешный запуск Hadoop в браузере

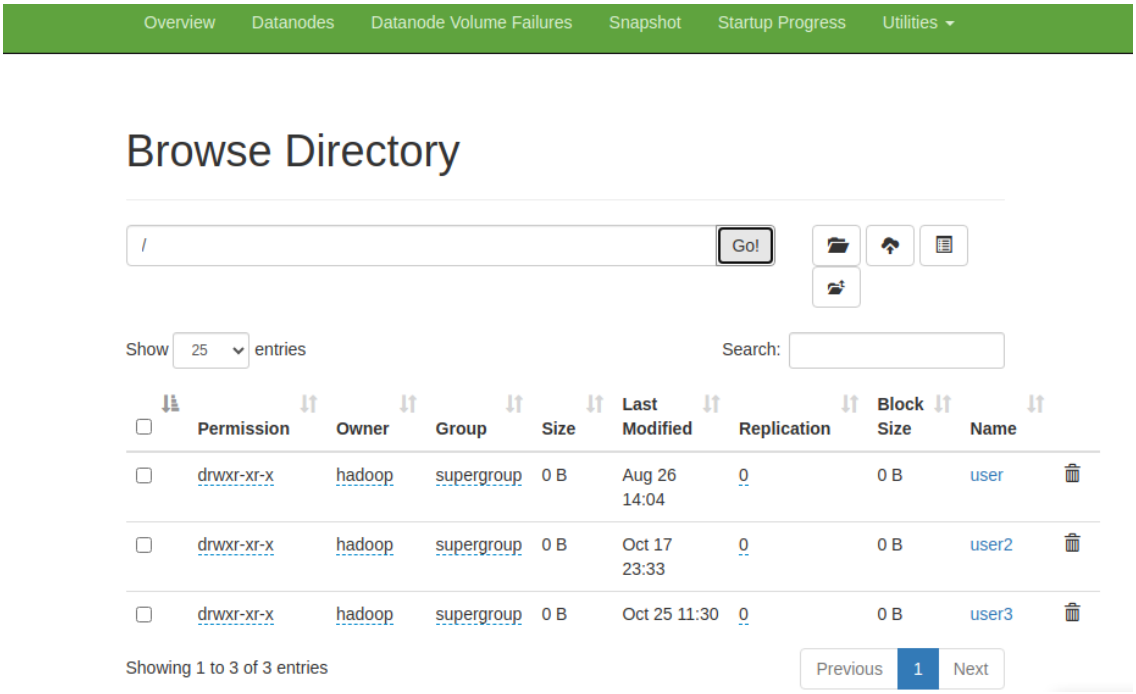


Рисунок 4. Просмотр директория в браузере

На веб-интерфейсе YARN можно увидеть информацию о запущенных приложениях, статусе узлов и других метриках кластера. Это удобный способ управления и мониторинга задач. http://localhost:8088

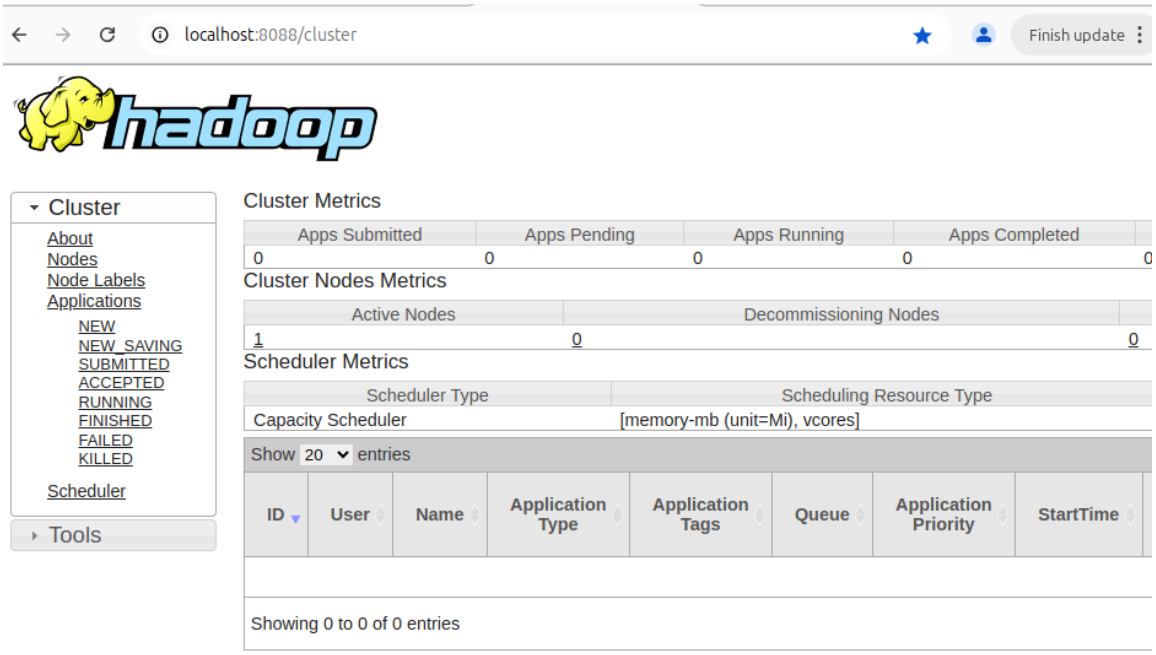


Рисунок 5. Кластеры Hadoop

Теперь, когда все установлено и проверено, нужно создать директорию user 4.

```
hadoop@devopsvm:~$ hdfs dfs -mkdir /user4
2024-10-25 11:29:16,617 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
hadoop@devopsvm:~$ hdfs dfs -mkdir /user4/hadoop
2024-10-25 11:30:55,331 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
hadoop@devopsvm:~$ hdfs dfs -mkdir /user4/hadoop/input
2024-10-25 11:31:07,202 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
```

Рисунок 6. Создание директории user4 в HDFS

```
hadoop@devopsvm:~$ wget https://raw.githubusercontent.com/BosenkoTM/Distributed_
systems/refs/heads/main/practice/2024/lw_01/GDP.csv
--2024-10-25 11:37:44--  https://raw.githubusercontent.com/BosenkoTM/Distributed
_systems/refs/heads/main/practice/2024/lw_01/GDP.csv
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.111.1
33, 185.199.110.133, 185.199.109.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.111.
133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 30268 (30K) [text/plain]
Saving to: ‘GDP.csv’

GDP.csv             100%[===================>]  29.56K  --.-KB/s    in 0.004s

2024-10-25 11:37:44 (7.90 MB/s) - ‘GDP.csv’ saved [30268/30268]
```

Рисунок 8. Загрузка файла, опубликованного на Github

```
hadoop@devopsvm:~$ hdfs dfs -mkdir /user4/hadoop/economic_data
2024-10-25 11:39:54,265 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
hadoop@devopsvm:~$ hdfs dfs -put GDP.csv /user4/hadoop/economic_data/
2024-10-25 11:40:17,899 WARN util.NativeCodeLoader: Unable to load native-hadoop
 library for your platform... using builtin-java classes where applicable
```

Рисунок 9. Загрузка данных в HDFS

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	29.56 KB	Oct 25 11:40	1	128 MB	GDP.csv
drwxr-xr-x	devops	supergroup	0 B	Oct 25 11:55	0	0 B	Italy_data.csv

Showing 1 to 2 of 2 entries

Previous 1 Next

Рисунок 10. Загрузка данных в HDFS выполнена успешно

```
hadoop@devopsvm:~$ hdfs dfs -chmod 777 /user4/hadoop/economic_data
2024-10-25 11:43:23,335 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java
classes where applicable
```

Рисунок 11. Предоставление другим пользователям доступ на изменение данных

Чтобы остановить Hadoop 3 в Ubuntu, выполнила следующие шаги:

```
hadoop@devopsvm:~$ jps
28704 Jps
27909 ResourceManager
18889 DataNode
19113 SecondaryNameNode
18684 NameNode
19614 NodeManager
```

Рисунок 12. Проверка, что все процессы Hadoop запущены

```
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]
2024-11-21 03:55:51,716 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
29684 Jps
```

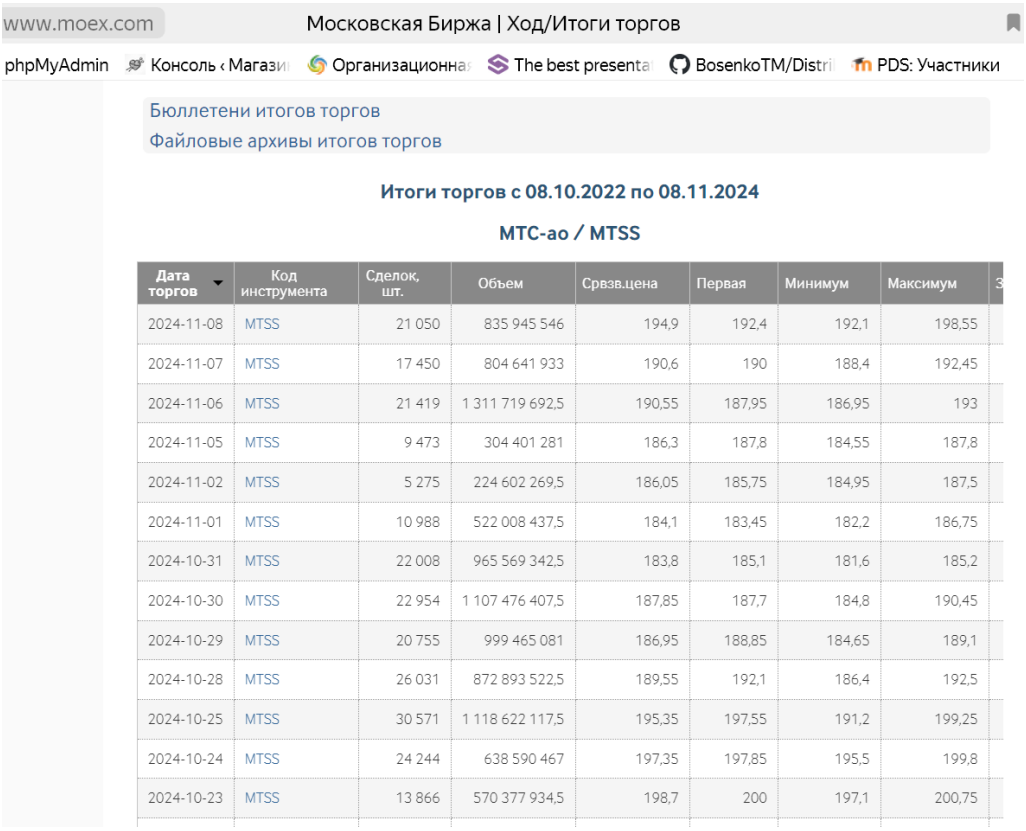
Рисунок 13. Остановка Hadoop 3 в Ubuntu и проверка

Индивидуальное задание

Вариант 10. Настройка Apache Hadoop.

Данные: Исторические данные по акциям МТС (MTSS) с сайта Московской биржи (moex.com) (рисунок 11)

Операции: Фильтрация данных за последние 2 года, расчет максимальной цены закрытия, тренд анализа.



www.moex.com Московская Биржа | Ход/Итоги торгов

phpMyAdmin Консоль ‹ Магази Организационна The best presenta BosenkoTM/Distri PDS: Участники

Бюллетени итогов торгов

Файловые архивы итогов торгов

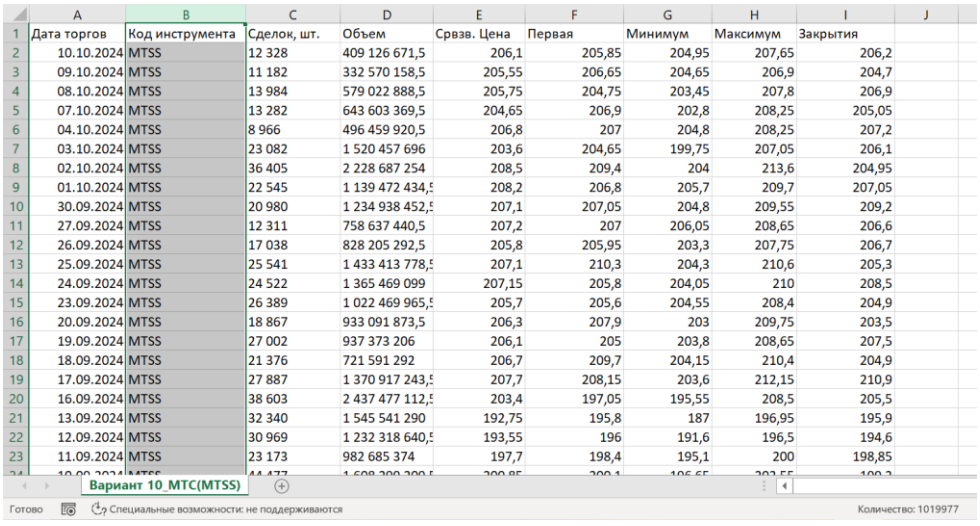
Итоги торгов с 08.10.2022 по 08.11.2024

МТС-ао / MTSS

Дата торгов	Код инструмента	Сделок, шт.	Объем	Срвзв.цена	Первая	Минимум	Максимум
2024-11-08	MTSS	21 050	835 945 546	194,9	192,4	192,1	198,55
2024-11-07	MTSS	17 450	804 641 933	190,6	190	188,4	192,45
2024-11-06	MTSS	21 419	1 311 719 692,5	190,55	187,95	186,95	193
2024-11-05	MTSS	9 473	304 401 281	186,3	187,8	184,55	187,8
2024-11-02	MTSS	5 275	224 602 269,5	186,05	185,75	184,95	187,5
2024-11-01	MTSS	10 988	522 008 437,5	184,1	183,45	182,2	186,75
2024-10-31	MTSS	22 008	965 569 342,5	183,8	185,1	181,6	185,2
2024-10-30	MTSS	22 954	1 107 476 407,5	187,85	187,7	184,8	190,45
2024-10-29	MTSS	20 755	999 465 081	186,95	188,85	184,65	189,1
2024-10-28	MTSS	26 031	872 893 522,5	189,55	192,1	186,4	192,5
2024-10-25	MTSS	30 571	1 118 622 117,5	195,35	197,55	191,2	199,25
2024-10-24	MTSS	24 244	638 590 467	197,35	197,85	195,5	199,8
2024-10-23	MTSS	13 866	570 377 934,5	198,7	200	197,1	200,75

Рисунок 14. Данные по акциям МТС (MTSS) с сайта

Данные были выгружены с сайта Московской биржи и сохранены в файле формата csv, количество строк 1 019 977, размер файла 62 634 КБ.



	A	B	C	D	E	F	G	H	I	J
1	Дата торгов	Код инструмента	Сделок, шт.	Объем	Срвзв. Цена	Первая	Минимум	Максимум	Закрытия	
2	10.10.2024	MTSS	12 328	409 126 671,5	206,1	205,85	204,95	207,65	206,2	
3	09.10.2024	MTSS	11 182	332 570 158,5	205,55	206,65	204,65	206,9	204,7	
4	08.10.2024	MTSS	13 984	579 022 888,5	205,75	204,75	203,45	207,8	206,9	
5	07.10.2024	MTSS	13 282	643 603 369,5	204,65	206,9	202,8	208,25	205,05	
6	04.10.2024	MTSS	8 966	496 459 920,5	206,8	207	204,8	208,25	207,2	
7	03.10.2024	MTSS	23 082	1 520 457 696	203,6	204,65	199,75	207,05	206,1	
8	02.10.2024	MTSS	36 405	2 228 687 254	208,5	209,4	204	213,6	204,95	
9	01.10.2024	MTSS	22 545	1 139 472 434,5	208,2	206,8	205,7	209,7	207,05	
10	30.09.2024	MTSS	20 980	1 234 938 452,5	207,1	207,05	204,8	209,55	209,2	
11	27.09.2024	MTSS	12 311	758 637 440,5	207,2	207	206,05	208,65	206,6	
12	26.09.2024	MTSS	17 038	828 205 292,5	205,8	205,95	203,3	207,75	206,7	
13	25.09.2024	MTSS	25 541	1 433 413 778,5	207,1	210,3	204,3	210,6	205,3	
14	24.09.2024	MTSS	24 522	1 365 469 099	207,15	205,8	204,05	210	208,5	
15	23.09.2024	MTSS	26 389	1 022 469 965,5	205,7	205,6	204,55	208,4	204,9	
16	20.09.2024	MTSS	18 867	933 091 873,5	206,3	207,9	203	209,75	203,5	
17	19.09.2024	MTSS	27 002	937 373 206	206,1	205	203,8	208,65	207,5	
18	18.09.2024	MTSS	21 376	721 591 292	206,7	209,7	204,15	210,4	204,9	
19	17.09.2024	MTSS	27 887	1 370 917 243,5	207,7	208,15	203,6	212,15	210,9	
20	16.09.2024	MTSS	38 603	2 437 477 112,5	203,4	197,05	195,55	208,5	205,5	
21	13.09.2024	MTSS	32 340	1 545 541 290	192,75	195,8	187	196,95	195,9	
22	12.09.2024	MTSS	30 969	1 232 318 640,5	193,55	196	191,6	196,5	194,6	
23	11.09.2024	MTSS	23 173	982 685 374	197,7	198,4	195,1	200	198,85	

Вариант 10_MTC(MTSS)

Готово Специальные возможности: не поддерживаются Количество: 1019977

Рисунок 15. Файл данных МТС в csv

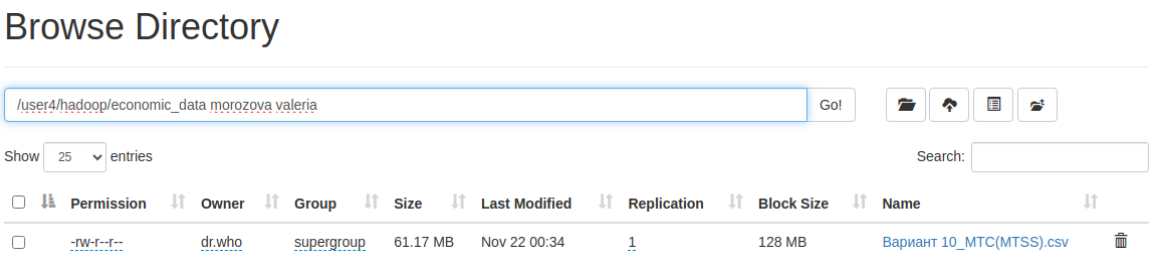


Рисунок 16. Данные МТС успешно загружены в Hadoop

Теперь нужно перейти в Jupyter notebook, чтобы выполнить подключение и операции согласно индивидуальному заданию.

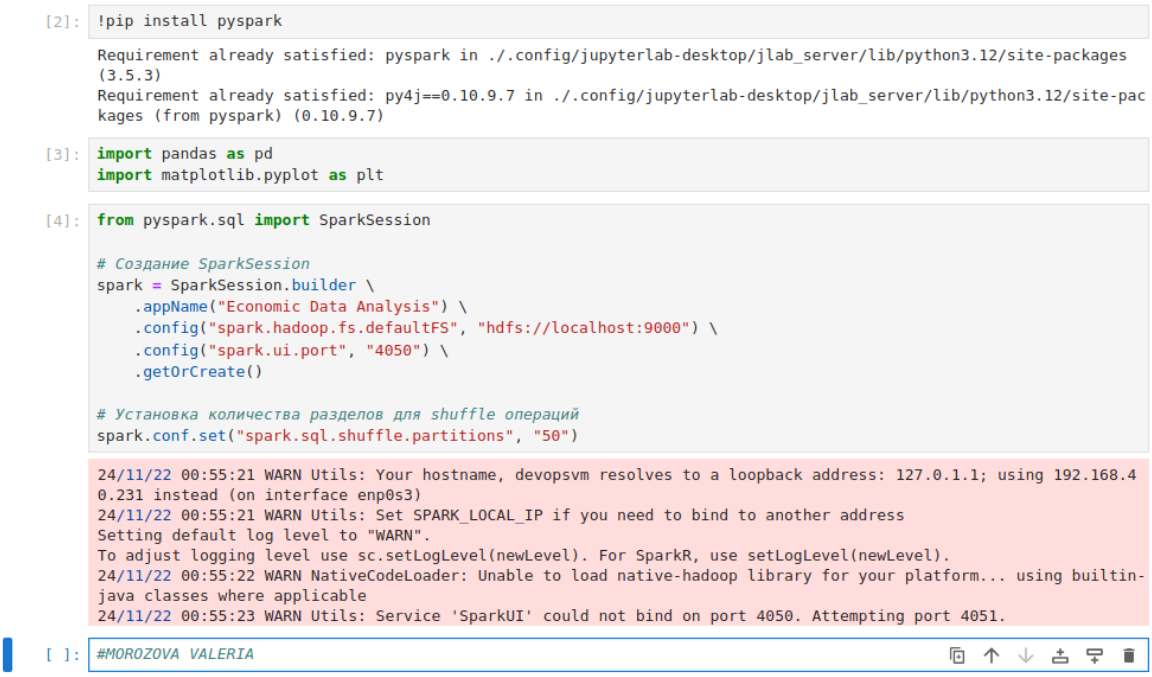


Рисунок 17. Установка модуля и подключение прошли успешно

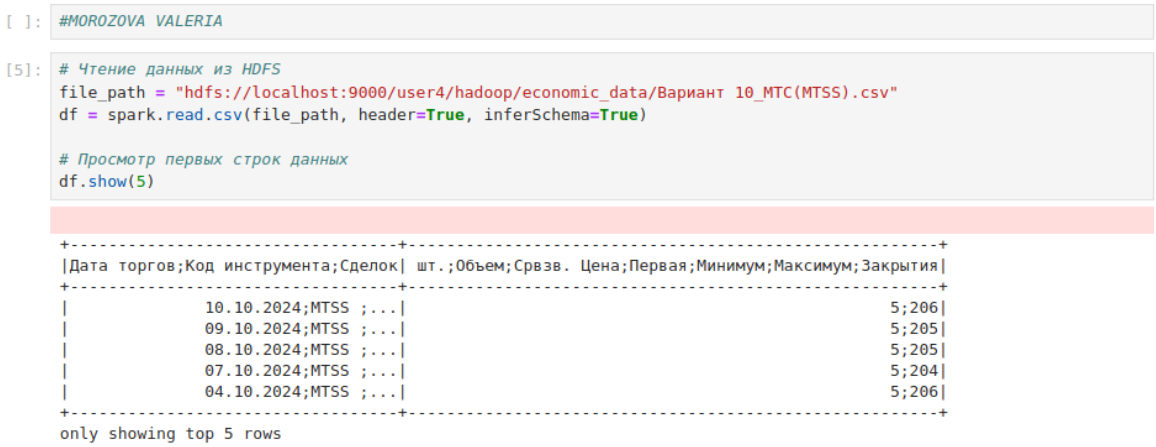


Рисунок 18. Чтение данных из HDFS

Как видно, есть необходимость изменить разделитель для корректного отображения данных (рисунок 19).

При добавлении параметра разделителя, данные читаются корректно.

```
[4]: # Чтение данных из HDFS
file_path = "hdfs://localhost:9000/user4/hadoop/economic_data/Вариант 10_MTC(MTSS).csv"
df = spark.read.csv(file_path, header=True, inferSchema=True, sep=";")

# Просмотр первых строк данных
df.show(5)
```

```
+----------+---------------+----------+-------------+----------+------+-------+--------+--------+
|Дата торгов|Код инструмента|Сделок, шт.|        Объем|Срвзв. Цена|Первая|Минимум|Максимум|Закрытия|
+----------+---------------+----------+-------------+----------+------+-------+--------+--------+
| 10.10.2024|          MTSS |     12 328|409 126 671,5|     206,1|205,85| 204,95|  207,65|   206,2|
| 09.10.2024|          MTSS |     11 182|332 570 158,5|    205,55|206,65| 204,65|   206,9|   204,7|
| 08.10.2024|          MTSS |     13 984|579 022 888,5|    205,75|204,75| 203,45|   207,8|   206,9|
| 07.10.2024|          MTSS |     13 282|643 603 369,5|    204,65|206,90| 202,80|  208,25|  205,05|
| 04.10.2024|          MTSS |      8 966|496 459 920,5|     206,8|207,00| 204,80|  208,25|   207,2|
+----------+---------------+----------+-------------+----------+------+-------+--------+--------+
only showing top 5 rows
```

Рисунок 19. Чтение данных из HDFS с учетом разделителя

```
[5]: pandas_df = df.toPandas()
pandas_df.head()
```

[5]:

	Дата торгов	Код инструмента	Сделок, шт.	Объем	Срвзв. Цена	Первая	Минимум	Максимум	Закрытия
0	10.10.2024	MTSS	12 328	409 126 671,5	206,1	205,85	204,95	207,65	206,2
1	09.10.2024	MTSS	11 182	332 570 158,5	205,55	206,65	204,65	206,9	204,7
2	08.10.2024	MTSS	13 984	579 022 888,5	205,75	204,75	203,45	207,8	206,9
3	07.10.2024	MTSS	13 282	643 603 369,5	204,65	206,90	202,80	208,25	205,05
4	04.10.2024	MTSS	8 966	496 459 920,5	206,8	207,00	204,80	208,25	207,2

```
[ ]: #MOROZOVA VALERIA
```

Рисунок 20. Перевод в pandas

Прежде чем приступить к выполнению операций, нужно провести анализ, проверить тип данных и получить сводную статистику по каждому столбцу.

```
[ ]: #MOROZOVA VALERIA
```

```
[6]: pandas_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1019976 entries, 0 to 1019975
Data columns (total 9 columns):
 #   Column           Non-Null Count    Dtype 
---  ------           --------------    ----- 
 0   Дата торгов      1019976 non-null  object
 1   Код инструмента  1019976 non-null  object
 2   Сделок, шт.      1019976 non-null  object
 3   Объем            1019976 non-null  object
 4   Срвзв. Цена      1019976 non-null  object
 5   Первая           1019976 non-null  object
 6   Минимум          1019976 non-null  object
 7   Максимум         1019976 non-null  object
 8   Закрытия         1019976 non-null  object
dtypes: object(9)
memory usage: 70.0+ MB
```

Рисунок 21. Информация о типе данных каждого столбца

Все данные были по типу объекты, их нужно изменить.

```
[12]: pandas_df['Дата торгов'] = pd.to_datetime(pandas_df['Дата торгов'], errors='coerce')
      pandas_df['Код инструмента'] = pd.to_numeric(pandas_df['Код инструмента'], errors='coerce')
      pandas_df['Сделок, шт.'] = pd.to_numeric(pandas_df['Сделок, шт.'], errors='coerce')
      pandas_df['Объем'] = pd.to_numeric(pandas_df['Объем'], errors='coerce')
      pandas_df['Срвзв. Цена'] = pd.to_numeric(pandas_df['Срвзв. Цена'], errors='coerce')
      pandas_df['Первая'] = pd.to_numeric(pandas_df['Первая'], errors='coerce')
      pandas_df['Минимум'] = pd.to_numeric(pandas_df['Минимум'], errors='coerce')
      pandas_df['Максимум'] = pd.to_numeric(pandas_df['Максимум'], errors='coerce')
      pandas_df['Закрытия'] = pd.to_numeric(pandas_df['Закрытия'], errors='coerce')

[13]: pandas_df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1019976 entries, 0 to 1019975
Data columns (total 9 columns):
 #   Column           Non-Null Count    Dtype
---  ------           --------------    -----
 0   Дата торгов      401945 non-null   datetime64[ns]
 1   Код инструмента  0 non-null        float64
 2   Сделок, шт.      0 non-null        float64
 3   Объем            0 non-null        float64
 4   Срвзв. Цена      1019499 non-null  float64
 5   Первая           0 non-null        float64
 6   Минимум          0 non-null        float64
 7   Максимум         1019550 non-null  float64
 8   Закрытия         1019540 non-null  float64
dtypes: datetime64[ns](1), float64(8)
memory usage: 70.0 MB

[ ]: #MOROZOVA VALERIA
```

Рисунок 22. Успешное изменение типа данных

Вывод

Hadoop более ориентирован на хранение данных, запросы выполняются с трудом и гораздо медленнее, чем на Spark. Кроме того, нет возможности выполнять операции на локальном компьютере.

1. Подключение было успешно установлено
2. Файлы загружены и прочитаны
3. Переведены в пандас