

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

Кузьмина Дарья Юрьевна БД-241м

Инструменты хранения и анализа больших данных

Лабораторная работа 1.1. Создание и управление базой данных на HDFS.
Введение в большие данные и их хранение. Инструменты
обработки больших данных (Hadoop)
Вариант 11

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,
доктор экономических наук

Москва
2025

Содержание

| | |
|-----------------------------|-----------|
| Введение | 2 |
| Основная часть | 2 |
| Заключение | 12 |

Введение

Цель

изучить основные операции и функциональные возможности системы, что позволит

понять принципы работы с данными и распределенными вычислениями.

Задачи

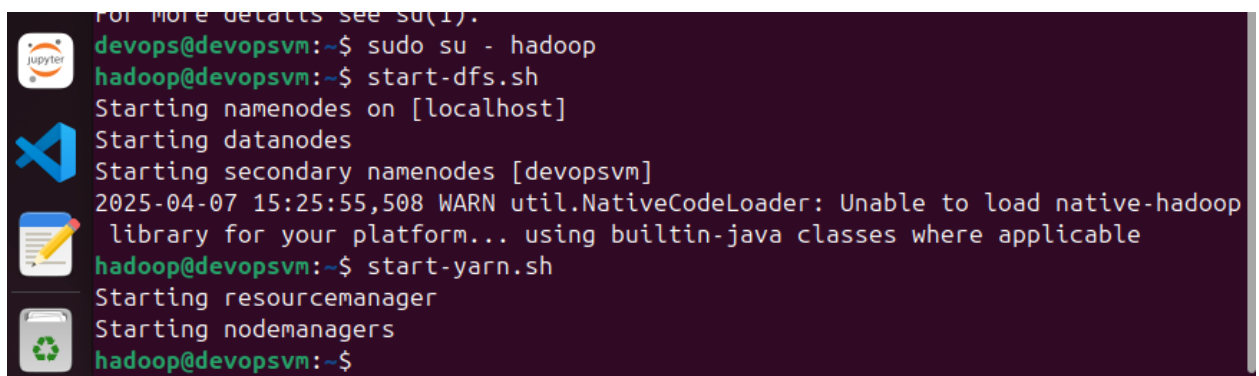
Действия, которые требуется выполнить:

1. Подключение к Hadoop и загрузка данных.
 - Подключиться к HDFS и убедиться, что файл доступен по пути `hdfs://localhost:9000/user01/hadoop/economic_data/БАШ_ФАЙЛ.csv`
 - Использовать PySpark или Pandas для загрузки данных из HDFS в DataFrame, который можно будет использовать для анализа.
2. Исследование и очистка данных.
 - Проверить структуру данных и типы столбцов (например, с помощью `printSchema()` для PySpark или `describe()` для Pandas).
 - Убедиться, что все данные корректны, и преобразовать необходимые столбцы в числовые форматы, если они изначально представлены в виде строк.
 - Проверить данные на наличие пропущенных или некорректных значений, удалить или заполнить такие значения в зависимости от ситуации.
3. Анализ данных.
 - Провести базовый статистический анализ данных:
 - Вычислить средние значения, медианы, минимумы и максимумы для экономических параметров.
 - Проанализировать и выявить тенденции.

- Построить временные ряды, чтобы понять, как изменялась их экономика с течением времени.
4. Визуализация данных.
- Построить графики (например, графики временных рядов).
 - Построить диаграммы для сравнения экономических показателей.
5. Сохранение и экспорт результатов.
- Сохранить результаты анализа и визуализации в формате CSV или изображений.
 - Сохранить обработанные данные (например, данные только для отдельных стран) обратно в HDFS, чтобы другие команды могли использовать их для дальнейшего анализа.
 - Создать отчет, включающий ключевые выводы и визуализации, для представления результатов анализа заинтересованным сторонам.
6. Автоматизация процесса (опционально).
- Создать скрипт или Jupyter Notebook, который автоматизирует процесс загрузки, анализа и визуализации данных для упрощения дальнейших исследований и повторного использования кода.

Основная часть

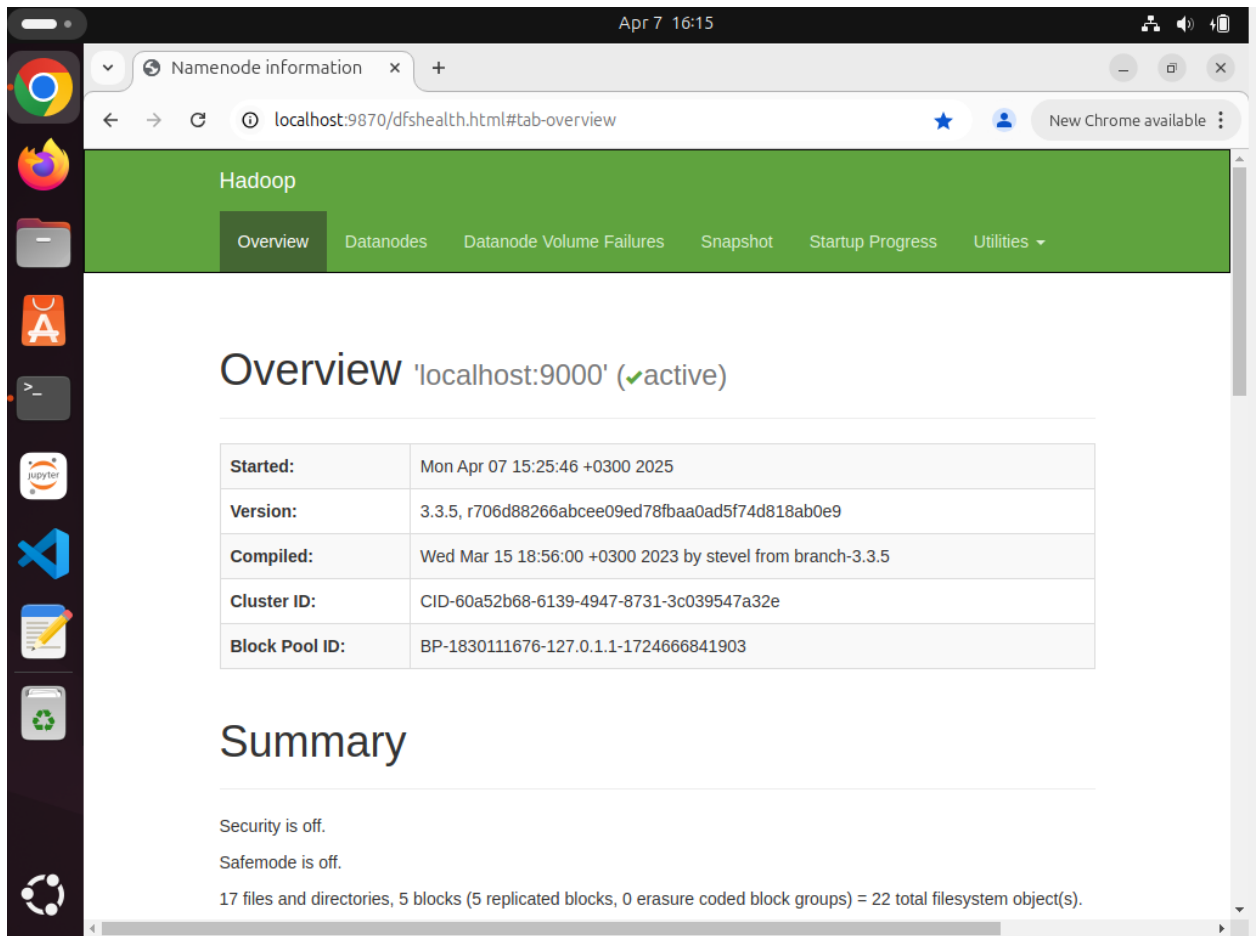
Задача 1. Запускаем yarm и dfs.



```
For more details see su(1).
devops@devopsvm:~$ sudo su - hadoop
hadoop@devopsvm:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [devopsvm]
2025-04-07 15:25:55,508 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:~$
```

```
hadoop@devopsvm:~$ jps
4112 ResourceManager
4244 NodeManager
3399 NameNode
6471 Jps
3800 SecondaryNameNode
3594 DataNode
```

Проверяем запущенные ресурсы.

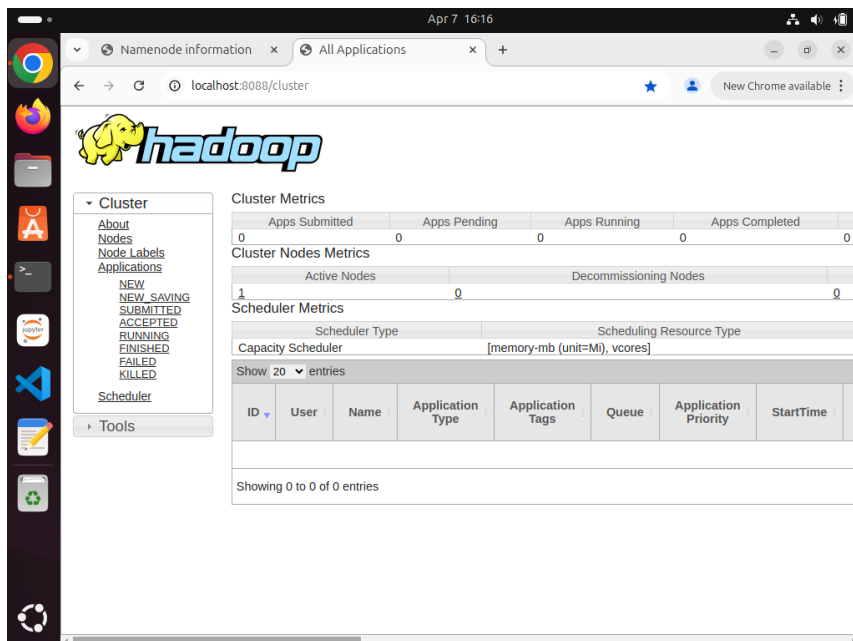


The screenshot shows a web browser window with the URL `localhost:9870/dfshealth.html#tab-overview`. The page title is "Hadoop" and the active tab is "Overview". The main heading is "Overview 'localhost:9000' (✓active)". Below this is a table with the following information:

| | |
|----------------|--|
| Started: | Mon Apr 07 15:25:46 +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 |

Below the table is a "Summary" section with the following text:

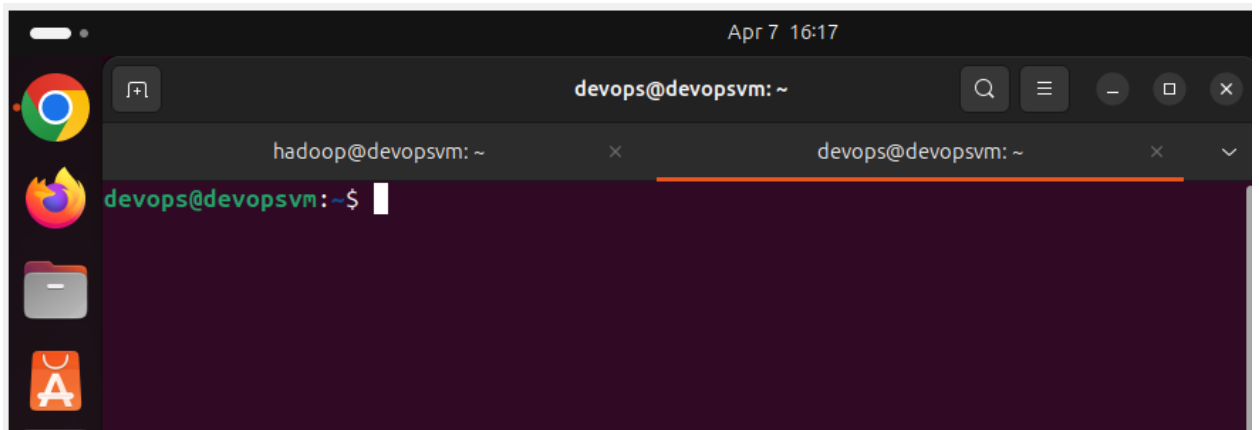
Security is off.
Safemode is off.
17 files and directories, 5 blocks (5 replicated blocks, 0 erasure coded block groups) = 22 total filesystem object(s).



Проверяем что поднялась файловая система и yarn.

Задача 2: Получим данные из источника, трансформируем и посмотрим результат.

Открываем два окна для администрирования и для работы.



Создаем своего пользователя.



Качаем файлы.

```
hadoop@devopsvm:~$ wget https://raw.githubusercontent.com/BosenkoTM/Distributed_systems/main/practice/2024/lw_01/GDP.csv
--2025-04-07 18:01:28-- https://raw.githubusercontent.com/BosenkoTM/Distributed_systems/main/practice/2024/lw_01/GDP.csv
Resolving raw.githubusercontent.com (raw.githubusercontent.com)... 185.199.109.133, 185.199.110.133, 185.199.111.133, ...
Connecting to raw.githubusercontent.com (raw.githubusercontent.com)|185.199.109.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 30268 (30K) [text/plain]
Saving to: 'GDP.csv.1'

GDP.csv.1          100%[=====] 29.56K  --.-KB/s   in 0.005s

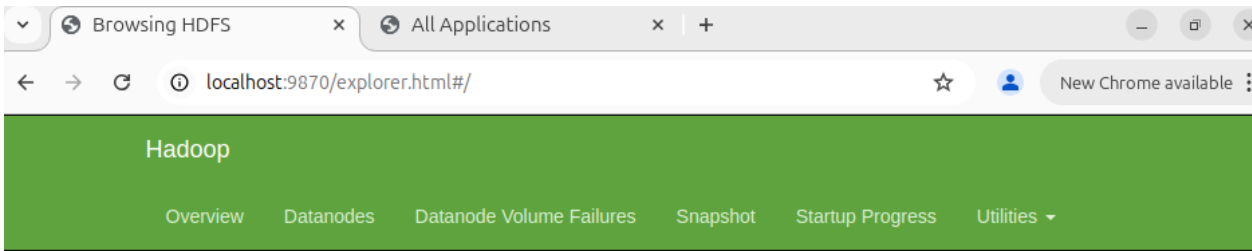
2025-04-07 18:01:29 (5.39 MB/s) - 'GDP.csv.1' saved [30268/30268]

hadoop@devopsvm:~$
```

Смотрим данные.

```
GNU nano 7.2 GDP.csv
Country,Year,GDP,Urban_population,Industry,Business,Mining,Manufacturing,Electr>
Austria,2010,35390,57.4,24,25.2,18.3,24.4,23.6,12.2,9.9,27.5,7.3,9.9,21.2,30.3,>
Austria,2015,36140,57.72,21.8,23.4,13.7,22.7,17.6,9.3,8.2,23.3,11.6,6.4,22.4,30>
Austria,2016,36390,57.91,20.8,22.3,14.4,21.9,13.2,8.2,8.3,23.3,14.5,5.9,20.9,27>
Austria,2017,36980,58.09,20.7,22.3,10.9,21.7,13,8.4,8.3,23.2,12.4,5.7,20.6,28.4>
Austria,2018,37690,58.3,20.4,22,7.9,21.4,14.4,8.1,8.3,23.2,11.7,5.4,20.7,28.2,2>
Austria,2019,38090,58.52,19.9,21.6,8,20.9,13.6,8.3,8.2,22.7,11.9,4.9,20.3,28,28>
Austria,2020,35480,58.75,18.9,20.7,6.1,20.1,12.9,7.7,8,21.8,12.1,5,20.1,27.2,27>
Austria,2021,36950,59,18.8,20.7,6.9,20.2,12.5,7.2,7.8,21.9,10.9,4.5,19.7,27.2,2>
Belgium,2010,33330,97.65,10.2,15.1,-13.4,11.5,29.6,-0.5,-1.6,17.1,4.1,8.7,15.2,>
Belgium,2011,33460,97.7,9.4,14.1,-10.4,10.1,27.7,-1,-2.8,15.6,4.5,7.9,14.4,21.6>
Belgium,2012,33490,97.74,8.3,13,-7.1,8.9,26,-1.2,-4,14.6,4.9,7,13.5,21.2,10.8,1>
Belgium,2013,33490,97.79,7.5,12.1,-4.1,7.6,24.2,-1.6,-5.3,13.3,5.3,6.2,12.7,20.>
Belgium,2014,33870,97.83,6.6,11,-1,6.4,22.4,-2,-6.6,12.1,5.6,5.4,11.8,20.2,3.6,>
Belgium,2015,34360,97.88,6.4,10.8,-1.5,7.4,18.5,1.6,-3.1,11.6,5.3,4.9,11.8,17,5>
Belgium,2016,34620,97.92,6,10.4,-2.1,8.6,14.1,5,0.4,11.1,4.7,4.4,11.6,13.5,6.9,>
Belgium,2017,35050,97.96,5.8,10.2,-2.6,9.6,10,8.4,3.8,10.6,4.1,3.7,11.4,10.2,8.>
Belgium,2018,35510,98,5.8,10.1,-3.1,10.7,6.1,11.8,7.7,10.2,3.6,3.3,11.2,6.9,10.>
Belgium,2019,36110,98.04,5.8,9.9,-3.3,10.4,5.9,11.8,7.2,10,3.4,3.1,11,6.6,9.9,5>
Belgium,2020,34010,98.08,5.3,9.5,-3.5,10.2,5.7,11.6,6.8,9.9,3.1,2.8,10.8,6.4,9.>
[ Read 268 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
```

Создали каталог.



Browse Directory

/

Go!

Show

25

▼

entries

Search:

| <div><div><div></div></div></div> | <div><div><div></div></div></div> Permission | <div><div><div></div></div></div> Owner | <div><div><div></div></div></div> Group | <div><div><div></div></div></div> Size | <div><div><div></div></div></div> Last Modified | <div><div><div></div></div></div> Replication | <div><div><div></div></div></div> Block Size | <div><div><div></div></div></div> Name | |
|-----------------------------------|--|--|--|--|---|---|--|---|---|
| <div><div><div></div></div></div> | <div><div><div></div></div></div> drwxr-xr-x | <div><div><div></div></div></div> hadoop | <div><div><div></div></div></div> supergroup | <div><div><div></div></div></div> 0 B | <div><div><div></div></div></div> Apr 07 17:58 | <div><div><div></div></div></div> 0 | <div><div><div></div></div></div> 0 B | <div><div><div></div></div></div> kuzmina01 | <div><div><div></div></div></div> <div></div> |
| <div><div><div></div></div></div> | <div><div><div></div></div></div> drwxr-xr-x | <div><div><div></div></div></div> hadoop | <div><div><div></div></div></div> supergroup | <div><div><div></div></div></div> 0 B | <div><div><div></div></div></div> Aug 26 2024 | <div><div><div></div></div></div> 0 | <div><div><div></div></div></div> 0 B | <div><div><div></div></div></div> user | <div><div><div></div></div></div> <div></div> |
| <div><div><div></div></div></div> | <div><div><div></div></div></div> drwxr-xr-x | <div><div><div></div></div></div> hadoop | <div><div><div></div></div></div> supergroup | <div><div><div></div></div></div> 0 B | <div><div><div></div></div></div> Oct 17 23:23 | <div><div><div></div></div></div> 0 | <div><div><div></div></div></div> 0 B | <div><div><div></div></div></div> user2 | <div><div><div></div></div></div> <div></div> |

Задача 3: Перегоняем в каталог данные.

Browse Directory

/kuzmina01/hadoop/input/economic_data

Go!

Show

25

▼

entries

Search:

| | <div> <div></div> <div></div> </div> Permission | <div> <div></div> <div></div> </div> Owner | <div> <div></div> <div></div> </div> Group | <div> <div></div> <div></div> </div> Size | <div> <div></div> <div></div> </div> Last Modified | <div> <div></div> <div></div> </div> Replication | <div> <div></div> <div></div> </div> Block Size | <div> <div></div> <div></div> </div> Name |
|--------------------------|---|--|--|---|--|--|---|--|
| <input type="checkbox"/> | -rw-r--r-- | hadoop | supergroup | 29.56 KB | Apr 07 18:08 | 1 | 128 MB | GDP.csv <div> <div></div> </div> |

Showing 1 to 1 of 1 entries

Previous

1

Next

Проверяем визуализацию данных.

```
[*]: # Чтение данных из HDFS
file_path = "hdfs://localhost:9000/kuzmina01/hadoop/economic_data/GDP.csv"
df = spark.read.csv(file_path, header=True, inferSchema=True)

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

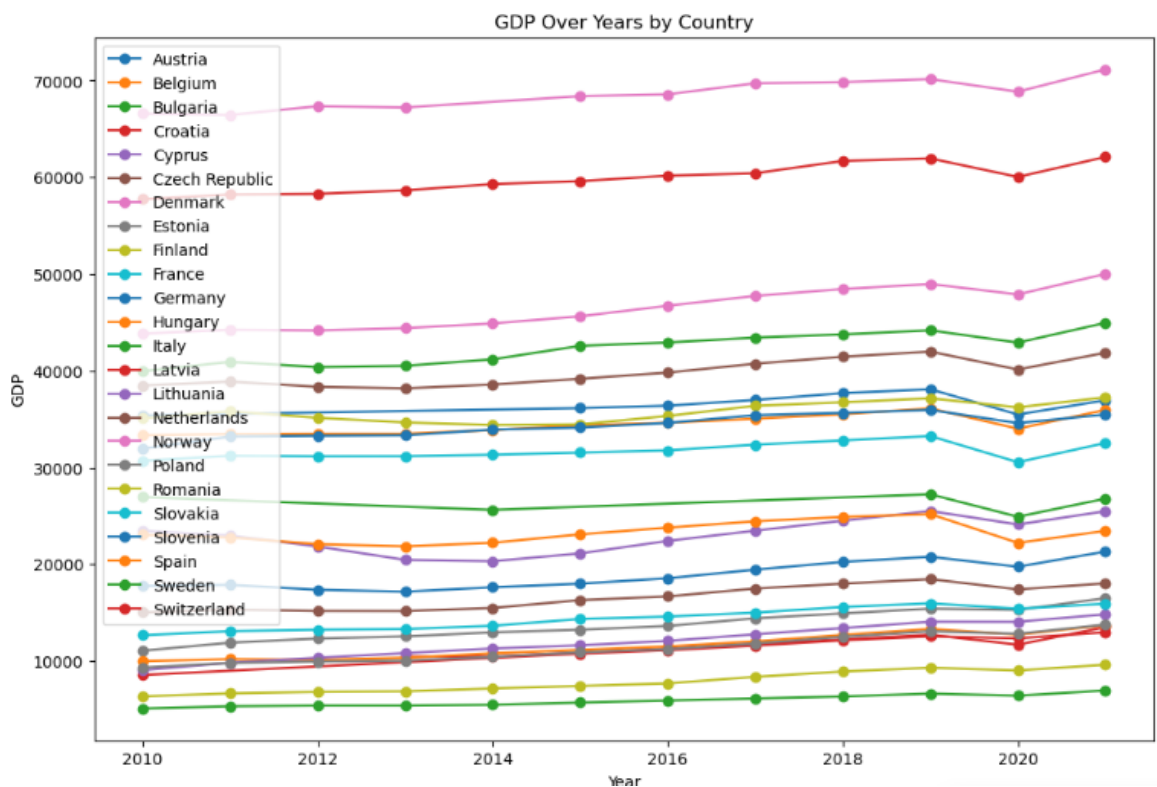
```
[5]: pandas_df = df.toPandas()
pandas_df.head()
```

```
[5]:
```

| | Country | Year | GDP | Urban_population | Industry | Business | Mining | Manufacturing | Electricity_supply | Water_s |
|---|---------|------|-------|------------------|----------|----------|--------|---------------|--------------------|---------|
| 0 | Austria | 2010 | 35390 | 57.40 | 24.0 | 25.2 | 18.3 | 24.4 | 23.6 | |
| 1 | Austria | 2015 | 36140 | 57.72 | 21.8 | 23.4 | 13.7 | 22.7 | 17.6 | |
| 2 | Austria | 2016 | 36390 | 57.91 | 20.8 | 22.3 | 14.4 | 21.9 | 13.2 | |
| 3 | Austria | 2017 | 36980 | 58.09 | 20.7 | 22.3 | 10.9 | 21.7 | 13.0 | |
| 4 | Austria | 2018 | 37690 | 58.30 | 20.4 | 22.0 | 7.9 | 21.4 | 14.4 | |

5 rows x 23 columns

summary_statistics



Индивидуальное задание Вариант 11

Добавляем свои данные с гитхаба.

```
hadoop@devopsvm:~$ curl -L "https://raw.githubusercontent.com/Iezekiss/BDSAD_MGP
U/refs/heads/main/Lab1.1/TATN.csv" | hdfs dfs -put - /kuzmina01/hadoop/input/tat
n_data/TATN.csv
% Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
           Dload  Upload   Total     Spent    Left     Speed
100  3198  100  3198    0     0  6300      0  --:--:-- --:--:-- --:--:--  6332
2025-04-07 19:34:23,921 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
hadoop@devopsvm:~$
```

Проверяем, загрузились ли.

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 | Apr 07 18:08 | 0 | 0 B | economic_data | |
| <input type="checkbox"/> | drwxr-xr-x | hadoop | supergroup | 0 B | Apr 07 19:34 | 0 | 0 B | tatn_data | |

Showing 1 to 2 of 2 entries

Previous

1

Next

Browse Directory

Show 25 entries

Search:

| <input type="checkbox"/> | Permission | Owner | Group | Size | Last Modified | Replication | Block Size | Name | |
|--------------------------|----------------------------|------------------------|----------------------------|---------|---------------|-------------------|------------|--------------------------|--|
| <input type="checkbox"/> | -rw-r--r-- | hadoop | supergroup | 3.12 KB | Apr 07 19:34 | 1 | 128 MB | TATN.csv | |

Showing 1 to 1 of 1 entries

Previous

1

Next

Запускаем в блокноте свои данные.

```
[12]: # Чтение данных из HDFS
file_path = "hdfs://localhost:9000/kuzmina01/hadoop/input/tatn_data/TATN.csv"
df = spark.read.csv(file_path, header=True, inferSchema=True)

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

```
+-----+
|TATN;M;200501;000000;540.6;575.4;508.2;532;102322400|
+-----+
|                                     TATN;M;200601;000...|
|                                     TATN;M;200701;000...|
|                                     TATN;M;200801;000...|
|                                     TATN;M;200901;000...|
|                                     TATN;M;201001;000...|
+-----+
only showing top 5 rows
```

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

```
[13]: TATN;M;200501;000000;540.6;575.4;508.2;532;102322400
0      TATN;M;200601;000000;544.2;599.4;501.2;556.3;1...
1      TATN;M;200701;000000;562.4;579.8;534.2;554.7;5...
2      TATN;M;200801;000000;552.7;616.4;541.5;547.8;6...
3      TATN;M;200901;000000;549.9;557.5;448.7;463.6;1...
4      TATN;M;201001;000000;465.6;486.7;395.5;412.6;1...
```

```
[12]: # Чтение данных из HDFS
file_path = "hdfs://localhost:9000/kuzmina01/hadoop/input/tatn_data/TATN.csv"
df = spark.read.csv(file_path, header=True, inferSchema=True)

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

```
+-----+
|TATN;M;200501;000000;540.6;575.4;508.2;532;102322400|
+-----+
|                                     TATN;M;200601;000...|
|                                     TATN;M;200701;000...|
|                                     TATN;M;200801;000...|
|                                     TATN;M;200901;000...|
|                                     TATN;M;201001;000...|
+-----+
only showing top 5 rows
```

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

```
[13]: TATN;M;200501;000000;540.6;575.4;508.2;532;102322400
0      TATN;M;200601;000000;544.2;599.4;501.2;556.3;1...
1      TATN;M;200701;000000;562.4;579.8;534.2;554.7;5...
2      TATN;M;200801;000000;552.7;616.4;541.5;547.8;6...
3      TATN;M;200901;000000;549.9;557.5;448.7;463.6;1...
4      TATN;M;201001;000000;465.6;486.7;395.5;412.6;1...
```

[3]: # Чтение данных из HDFS
file_path = "hdfs://localhost:9000/kuzmina01/hadoop/input/economic_data/GDP.csv"
df = spark.read.csv(file_path, header=True, inferSchema=True)

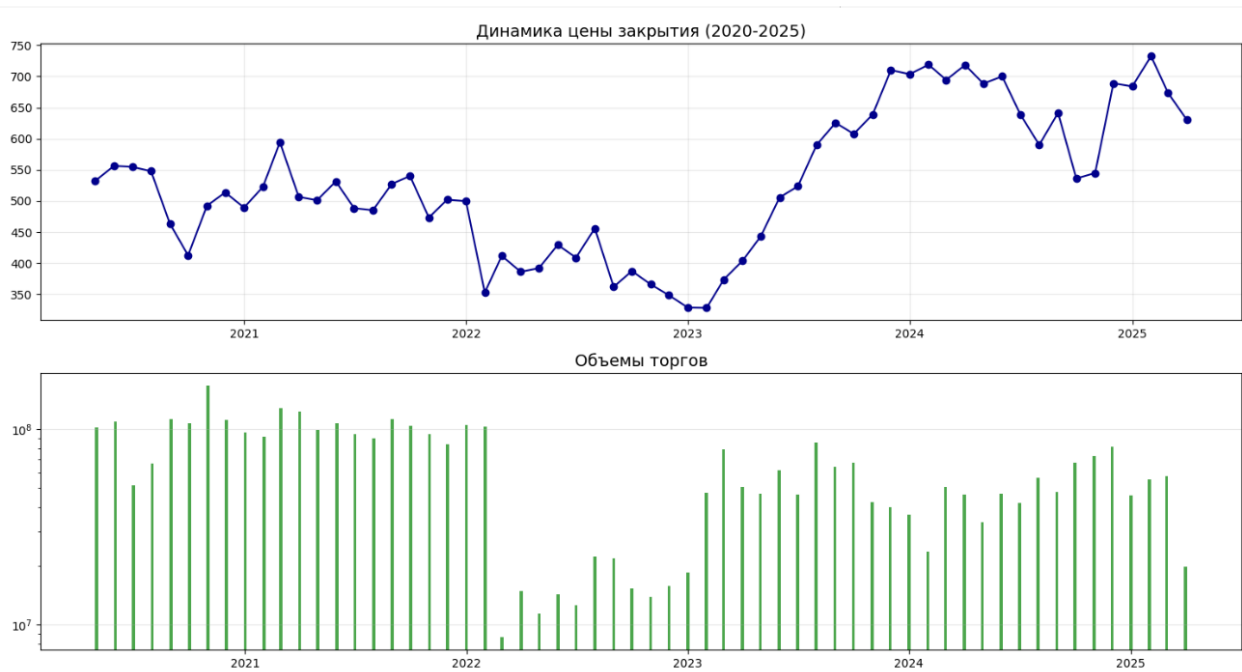
Просмотр первых строк данных
df.show(5)

| Country | Year | GDP | Urban_population | Industry | Business | Mining | Manufacturing | Electricity_supply | Water_supply | Construction | Retail_trade | Transportation | Accommodation | Information | Financial | Real estate | Professional_scientific | Administrative | Education | Human_health | Arts | Other |
|---------|------|-------|------------------|----------|----------|--------|---------------|--------------------|--------------|--------------|--------------|----------------|---------------|-------------|-----------|-------------|-------------------------|----------------|-----------|--------------|------|-------|
| Austria | 2010 | 35390 | 57.4 | 24.0 | 25.2 | 18.3 | 24.4 | 23.6 | 12.2 | 9.9 | 27.5 | 7.3 | 9.9 | 21.2 | 30.3 | 27.0 | 34.0 | 22.5 | 27.8 | 12.0 | 34.0 | 32.0 |
| Austria | 2015 | 36140 | 57.72 | 21.8 | 23.4 | 13.7 | 22.7 | 17.6 | 9.3 | 8.2 | 23.3 | 11.6 | 6.4 | 22.4 | 30.3 | 28.0 | 31.3 | 20.0 | 24.2 | 12.9 | 26.2 | 28.3 |
| Austria | 2016 | 36390 | 57.91 | 20.8 | 22.3 | 14.4 | 21.9 | 13.2 | 8.2 | 8.3 | 23.3 | 14.5 | 5.9 | 20.9 | 27.1 | 28.7 | 30.4 | 17.8 | 24.3 | 14.5 | 20.8 | 27.8 |
| Austria | 2017 | 36980 | 58.09 | 20.7 | 22.3 | 10.9 | 21.7 | 13.0 | 8.4 | 8.3 | 23.2 | 12.4 | 5.7 | 20.6 | 28.4 | 29.0 | 29.4 | 17.4 | 23.7 | 15.0 | 19.1 | 26.9 |
| Austria | 2018 | 37690 | 58.3 | 20.4 | 22.0 | 7.9 | 22.0 | 13.0 | 8.1 | 8.3 | 23.2 | 11.7 | 5.4 | | | | 28.3 | 17.1 | 23.6 | 15.3 | 18.3 | 26.4 |

Would you like to get notified about official Jupyter news?

[Open privacy policy](#) Yes No

Строим графики.



Заключение

Вывод:

В ходе выполнения лабораторной работы были изучены основные принципы работы с распределенной файловой системой Hadoop (HDFS) и инструментами обработки больших данных. Цель работы — освоение базовых операций управления данными и распределенными вычислениями — достигнута. Реализованы следующие задачи:

1. Подключение к HDFS и загрузка данных:

- Успешно настроено окружение Hadoop, запущены HDFS и YARN.
- Данные загружены в HDFS, созданы необходимые каталоги и обеспечен корректный доступ к файлам.

2. Исследование и очистка данных:

- Проведена проверка структуры данных, преобразование типов столбцов и обработка некорректных значений.
- Устранены пропуски и аномалии, что обеспечило готовность данных для анализа.

3. Анализ данных:

- Выполнен базовый статистический анализ (средние значения, минимумы, максимумы).
- Построены временные ряды, выявлены тенденции в динамике экономических показателей.

4. Визуализация:

- Созданы графики временных рядов и диаграммы для наглядного представления результатов.

5. Сохранение результатов:

- Результаты анализа экспортированы в форматы CSV и изображений.
- Обработанные данные сохранены в HDFS для дальнейшего использования.

6. Автоматизация:

- Разработан Jupyter Notebook, автоматизирующий процесс загрузки, анализа и визуализации данных.

Работа с Hadoop и HDFS продемонстрировала их эффективность для хранения и обработки больших объемов данных. Полученные навыки критически важны для задач бизнес-аналитики, где требуется масштабируемость и скорость обработки данных.