

Департамент образования города Москвы

**Государственное автономное образовательное учреждение  
высшего образования города Москвы  
«Московский городской педагогический университет»**

Институт цифрового образования  
Департамент информатики, управления и технологий

**ПРАКТИЧЕСКАЯ РАБОТА №5.1**

по дисциплине «Инструменты для хранения и обработки больших данных»  
Направление подготовки 38.03.05 – бизнес-информатика  
Профиль подготовки «Аналитика данных и эффективное управление»  
(очная форма обучения)

**Выполнила:**  
Студентка группы АДЭУ-221  
Вознесенская В. Е.

**Проверил:**  
Босенко Т. М., доцент

Москва  
2025

## Запуск Hadoop

```
hadoop@devopsvm:~/Desktop/lb5/BigDataAnalytic$ docker exec -it -u root hadoop-cluster /bin/bash
root@hadoop:/opt# jps
786 NodeManager
149 NameNode
684 ResourceManager
252 DataNode
430 SecondaryNameNode
1742 Jps
root@hadoop:/opt#
```

## Доступность веб-интерфейсов:

The screenshot shows two browser windows side-by-side. The left window is titled 'localhost:8707/dshealth.html#tab-overview' and displays the 'Overview' page for the Hadoop cluster. It includes sections for 'Started', 'Version', 'Compiled', 'Cluster ID', and 'Block Pool ID'. The right window is titled 'localhost:8088/cluster' and displays the 'Cluster Metrics' page for the Hadoop cluster. It includes sections for 'Cluster Metrics', 'Scheduler Metrics', and a table for 'Capacity Scheduler'.

## Загрузка данных в HDFS:

```
hadoop@devopsvm:~/Desktop/lb5/BigDataAnalytic$ docker cp ~/Downloads/stock_data_lb5.csv hadoop-cluster:~/opt/hadoop/stock_data_lb5.csv
Successfully copied 95.4MB to hadoop-cluster:/opt/hadoop/stock_data_lb5.csv
hadoop@devopsvm:~/Desktop/lb5/BigDataAnalytic$ docker exec -it -u root hadoop-cluster /bin/bash
hadoop@hadoop:~/opt$ ls /opt/hadoop/
LICENSE-binary NOTICE-binary README.txt data include libexec logs share
LICENSE.txt NOTICE.txt bin etc lib licenses-binary sbin stock_data_lb5.csv
hadoop@hadoop:~/opt$ hdfs dfs -mkdir -p /user/hadoop
hadoop@hadoop:~/opt$ hdfs dfs -put /opt/hadoop/stock_data_lb5.csv /user/hadoop/
hadoop@hadoop:~/opt$ hdfs dfs -ls /user/hadoop
Found 1 items
-rw-r--r-- 1 hadoop supergroup 95364952 2025-11-25 22:33 /user/hadoop/stock_data_lb5.csv
hadoop@hadoop:~/opt$
```

## Проверка:

```
hadoop@devopsvm:~/Desktop/lb5/BigDataAnalytic$ docker exec -it hadoop-cluster /bin/bash
root@hadoop:/opt# hdfs dfs -ls /user/hadoop/
Found 1 items
-rw-r--r-- 1 hadoop supergroup 95364952 2025-11-25 22:33 /user/hadoop/stock_data_lb5.csv
root@hadoop:/opt#
root@hadoop:/opt#
```

## Просмотр первых 10 строк файла:

```
root@hadoop:/opt$ docker exec -it -u hadoop hadoop-cluster /bin/bash
hadoop@hadoop:/opt$ hdfs dfs -cat /user/hadoop/stock_data_lb5.csv | head -n 10

date,company,open,high,low,close,volume
2020-11-19 22:13:11.218360,AAPL,188.57,191.73,187.14,189.78,3405915
2020-11-20 22:13:11.218360,AAPL,192.69,194.0,190.69,191.62,7002713
2020-11-21 22:13:11.218360,AAPL,195.79,198.27,194.96,196.3,12668393
2020-11-22 22:13:11.218360,AAPL,200.0,201.26,198.47,198.79,9291662
2020-11-23 22:13:11.218360,AAPL,203.18,203.43,201.88,202.45,2939593
2020-11-24 22:13:11.218360,AAPL,204.26,206.1,203.83,205.09,14885439
2020-11-25 22:13:11.218360,AAPL,205.12,206.0,202.16,203.88,321190
2020-11-26 22:13:11.218360,AAPL,204.98,206.26,202.2,203.98,13895769
2020-11-27 22:13:11.218360,AAPL,196.49,198.9,195.67,197.42,5159216
cat: Unable to write to output stream.
```

Выгрузка файла из HDFS на локальную файловую систему контейнера,  
далее чтение его в Jupyter

```
root@hadoop:/opt$ docker exec -it -u hadoop hadoop-cluster /bin/bash
hadoop@hadoop:/opt$ hdfs dfs -ls /user/hadoop
Found 1 items
-rw-r--r-- 1 hadoop supergroup 95364952 2025-11-25 22:33 /user/hadoop/stock_data_lb5.csv
hadoop@hadoop:/opt$ mkdir -p /home/hadoop/notebooks/data
hadoop@hadoop:/opt$ hdfs dfs -get /user/hadoop/stock_data_lb5.csv /home/hadoop/notebooks/data/
hadoop@hadoop:/opt$ 
```

Открываем в Jupyter:

```
In [2]:  
import pandas as pd  
import matplotlib.pyplot as plt  
import seaborn as sns  
  
# Загрузка CSV  
file_path = "stock_data_lb5.csv"  
df = pd.read_csv(file_path, parse_dates=['date'])  
  
# Создаем колонку с годом  
df['year'] = df['date'].dt.year  
  
df.head()
```

Out[2]:

	date	company	open	high	low	close	volume	year
0	2020-11-19 22:13:11.218360	AAPL	188.57	191.73	187.14	189.78	3405915	2020
1	2020-11-20 22:13:11.218360	AAPL	192.69	194.00	190.69	191.62	7002713	2020
2	2020-11-21 22:13:11.218360	AAPL	195.79	198.27	194.96	196.30	12668393	2020
3	2020-11-22 22:13:11.218360	AAPL	200.00	201.26	198.47	198.79	9291662	2020
4	2020-11-23 22:13:11.218360	AAPL	203.18	203.43	201.88	202.45	2939593	2020

Группируем по компаниям и году:

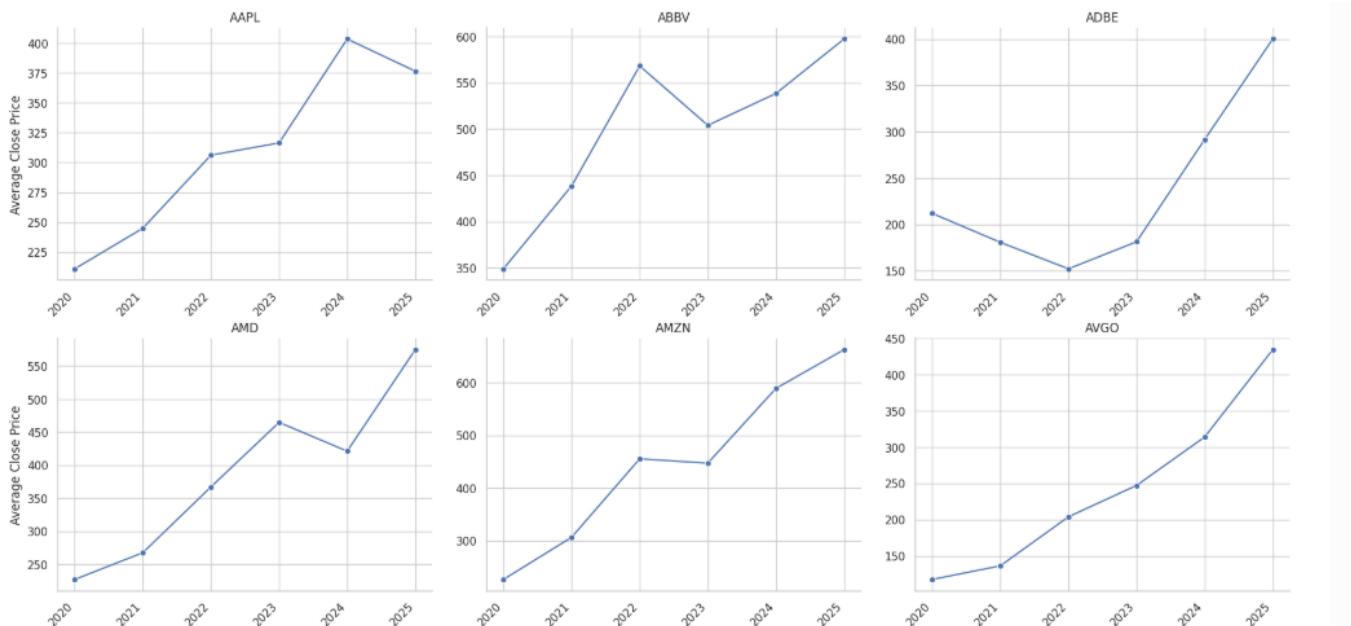
```
In [6]: # Группировка по компании и году
result = df.groupby(['company', 'year']).agg(
    avg_close=('close', 'mean'),
    avg_volume=('volume', 'mean')
).reset_index()

result
```

```
Out[6]:   company  year  avg_close  avg_volume
0      AAPL  2020  211.075581  8.719264e+06
1      AAPL  2021  245.037836  7.564756e+06
2      AAPL  2022  306.210603  7.695494e+06
3      AAPL  2023  316.456493  7.500412e+06
4      AAPL  2024  403.468279  7.736035e+06
...
289     XOM  2021  263.067945  7.618110e+06
290     XOM  2022  203.571151  7.450148e+06
291     XOM  2023  388.654767  7.326581e+06
292     XOM  2024  521.225847  7.608364e+06
293     XOM  2025  513.344174  7.672952e+06
```

294 rows × 4 columns

Расчет средней цены закрытия по компаниям за все годы (полные результаты предоставлены в github)



Выводы по графикам средней цены закрытия по компаниям

Общая

тенденция:

На графиках показана динамика средней цены закрытия акций каждой компании за год (или несколько лет). Для большинства компаний видно устойчивый рост цены, что говорит о положительной динамике на рынке и росте капитализации этих компаний.

Компании с резким ростом:

Некоторые компании, например AMZN, GOOG, META, демонстрируют значительный рост средней цены закрытия, что может свидетельствовать о расширении бизнеса, успешных продуктах или увеличении спроса на их акции.

Компании с падением или нестабильной динамикой:

Есть компании с просадками или волатильной ценой, например NVDA, BTC (если криптовалютные данные включены). Это может быть связано с рыночными колебаниями, сезонностью или внешними факторами. Такие компании требуют повышенного внимания при принятии инвестиционных решений.

Влияние внешних факторов:

Для некоторых компаний заметны резкие скачки или падения в отдельных годах. Это может отражать события на рынке, новости о компании или изменения в экономической ситуации.

Бизнес-инсайты:

Инвестиционная стратегия: компании с устойчивым ростом могут быть интересны для долгосрочных инвестиций.

Риск-менеджмент: для волатильных компаний важно учитывать возможные падения и диверсифицировать портфель.

Сравнительный анализ: визуализация позволяет сравнивать динамику цен разных компаний одновременно, выявлять лидеров и отстающих.

Рекомендации для дальнейшего анализа:

Построить аналогичные графики по объему торгов, чтобы оценить активность рынка и ликвидность акций.

Рассмотреть корреляцию между объемом торгов и динамикой цен.

Выявить компании с ростом цены при низких объемах торгов — это может сигнализировать о спекулятивной активности.