

Департамент образования и науки города Москвы

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

Институт цифрового образования

Департамент информатики, управления и технологий

Лабораторная работа 5.1.

Развертывание и настройка Nadoor. Анализ данных с использованием
экосистемы Nadoor.

Выполнил студент группы АДЭУ-221

Джамалова Сабина Шахиновна

Проверил доцент

Босенко Тимур Муртазович

Москва

2025

Оглавление

| | |
|--|----|
| Введение | 3 |
| Запуск проекта..... | 8 |
| Шаг 1. Клонирование и подготовка | 8 |
| Шаг 2. Запуск Docker контейнера | 8 |
| Шаг 3. Подключение к контейнеру | 9 |
| Шаг 4. Проверка компонентов Hadoop | 9 |
| Работа с HDFS | 10 |
| Шаг 1. Создание директорий | 10 |
| Шаг 2. Загрузка данных..... | 10 |
| Шаг 3. Просмотр данных в HDFS..... | 10 |
| Шаг 4. Веб-интерфейсы | 10 |
| Анализ данных | 12 |
| Загрузка данных: | 12 |
| Очистка данных: | 13 |
| MapReduce анализ на очищенных данных | 15 |
| Расширенный анализ факторов риска..... | 16 |
| Выводы: | 21 |

Введение

Цель работы: получить практические навыки развертывания одноузлового кластера Hadoop, освоить базовые операции с распределенной файловой системой HDFS, выполнить загрузку и простейшую обработку данных, а также научиться выгружать результаты для последующего анализа и визуализации во внешней среде.

Вариант задания: 6.

Кейс: Сердечно-сосудистые заболевания

Аналитическая часть: Средний возраст, вес, рост пациентов заболеванием и без (MapReduce)

Источник: <https://www.kaggle.com/datasets/colewelkins/cardiovascular-disease>

Описание датасета:

Демографические переменные

ID

Уникальный идентификатор пациента. Используется для однозначной идентификации записей в датасете.

age

Возраст пациента в днях на момент обследования. Преобразуется в годы для анализа. Диапазон: 18-100 лет (6570-36500 дней).

age_years

Возраст пациента в годах (производная переменная). Рассчитывается как $age / 365$. Критически важный фактор риска сердечно-сосудистых заболеваний.

gender

Пол пациента. Кодировка: 1 - женский, 2 - мужской. Мужской пол является независимым фактором риска.

height

Рост пациента в сантиметрах. Используется вместе с весом для расчета индекса массы тела (BMI). Диапазон: 100-250 см.

weight

Вес пациента в килограммах. Вместе с ростом определяет статус питания пациента. Диапазон: 30-200 кг.

Клинические показатели

ap_hi

Систолическое артериальное давление (верхнее значение) в мм рт.ст.

Показывает давление в артериях в момент сокращения сердца. Норма: 90-120 мм рт.ст.

ap_lo

Диастолическое артериальное давление (нижнее значение) в мм рт.ст.

Показывает давление в артериях в момент расслабления сердца. Норма: 60-80 мм рт.ст.

cholesterol

Уровень холестерина в крови. Категориальная переменная:

- 1: Нормальный уровень (< 200 мг/дл)
- 2: Выше нормального (200-239 мг/дл)
- 3: Значительно выше нормального (≥ 240 мг/дл)

gluc

Уровень глюкозы в крови. Категориальная переменная:

- 1: Нормальный уровень (< 100 мг/дл)
- 2: Выше нормального (100-125 мг/дл)
- 3: Значительно выше нормального (≥ 126 мг/дл)

Факторы образа жизни

smoke

Статус курения. Бинарная переменная:

- 0: Не курит
- 1: Курит (текущий курильщик)

alco

Употребление алкоголя. Бинарная переменная:

- 0: Не употребляет алкоголь
- 1: Употребляет алкоголь

active

Уровень физической активности. Бинарная переменная:

- 0: Не активен (сидячий образ жизни)
- 1: Активен (регулярная физическая активность)

Производные и целевые переменные

bmi

Индекс массы тела (Body Mass Index). Рассчитывается как $\text{weight} / (\text{height}/100)^2$. Классификация:

- <18.5 : Недостаточный вес
- 18.5-24.9: Нормальный вес
- 25-29.9: Избыточный вес
- ≥ 30 : Ожирение

bp_category

Категория артериального давления на основе `ap_hi` и `ap_lo`:

- "Normal": Нормальное ($<120 / <80$)

- "Elevated": Повышенное (120-129/<80)
- "Hypertension Stage 1": Гипертония 1 стадии (130-139/80-89)
- "Hypertension Stage 2": Гипертония 2 стадии ($\geq 140/\geq 90$)
- "Hypertensive Crisis": Гипертонический криз ($>180/>120$)

bp_category_encoded

Закодированная версия bp_category для машинного обучения:

- 0: Normal
- 1: Elevated
- 2: Hypertension Stage 1
- 3: Hypertension Stage 2
- 4: Hypertensive Crisis

cardio

Целевая переменная - наличие сердечно-сосудистого заболевания:

- 0: Отсутствие сердечно-сосудистого заболевания
- 1: Наличие сердечно-сосудистого заболевания

Дополнительные расчетные переменные

risk_factors

Суммарное количество факторов риска у пациента. Рассчитывается как сумма:

- Высокое давление ($\text{bp_category_encoded} \geq 2$)
- Высокий холестерин ($\text{cholesterol} \geq 2$)
- Высокая глюкоза ($\text{gluc} \geq 2$)
- Курение ($\text{smoke} = 1$)

- Ожирение ($\text{bmi} \geq 30$)

age_group

Возрастная группа пациента. Категории:

- 18-29 лет
- 30-39 лет
- 40-49 лет
- 50-59 лет
- 60-69 лет
- 70+ лет

bmi_category

Категория индекса массы тела:

- Недостаточный вес (<18.5)
- Нормальный вес ($18.5-24.9$)
- Избыточный вес ($25-29.9$)
- Ожирение I степени ($30-34.9$)
- Ожирение II+ степени (≥ 35)

Выполнение:

Запуск проекта

Шаг 1. Клонирование и подготовка

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

● hadoop@devopsvm:~$ cd /home/hadoop/Downloads/BigDataAnalytic-main/lw/2025/lw_5_1
● hadoop@devopsvm:~/Downloads/BigDataAnalytic-main/lw/2025/lw_5_1$ ls -la hadoop/
total 36
drwxrwxr-x 2 hadoop hadoop 4096 Oct 29 08:03 .
drwxrwxr-x 5 hadoop hadoop 4096 Oct 29 08:03 ..
-rw-rw-r-- 1 hadoop hadoop 1215 Oct 29 08:03 capacity-scheduler.xml
-rw-rw-r-- 1 hadoop hadoop 1509 Oct 29 08:03 core-site.xml
-rw-rw-r-- 1 hadoop hadoop 2820 Oct 29 08:03 hdfs-site.xml
-rw-rw-r-- 1 hadoop hadoop 1423 Oct 29 08:03 log4j.properties
-rw-rw-r-- 1 hadoop hadoop 2260 Oct 29 08:03 mapred-site.xml
-rw-rw-r-- 1 hadoop hadoop 9 Oct 29 08:03 workers
-rw-rw-r-- 1 hadoop hadoop 3040 Oct 29 08:03 yarn-site.xml
○ hadoop@devopsvm:~/Downloads/BigDataAnalytic-main/lw/2025/lw_5_1$

● hadoop@devopsvm:~/Downloads/BigDataAnalytic-main/lw/2025/lw_5_1$ ls -la scripts/
total 28
drwxrwxr-x 2 hadoop hadoop 4096 Oct 29 08:03 .
drwxrwxr-x 5 hadoop hadoop 4096 Oct 29 08:03 ..
-rw-rw-r-- 1 hadoop hadoop 4145 Oct 29 08:03 analyze_pandas.py
-rw-rw-r-- 1 hadoop hadoop 5579 Oct 29 08:03 analyze_spark.py
-rw-rw-r-- 1 hadoop hadoop 643 Oct 29 08:03 start_jupyter.sh
● hadoop@devopsvm:~/Downloads/BigDataAnalytic-main/lw/2025/lw_5_1$ ls -la notebooks/
total 144
drwxrwxr-x 2 hadoop hadoop 4096 Oct 29 08:03 .
drwxrwxr-x 5 hadoop hadoop 4096 Oct 29 08:03 ..
-rw-rw-r-- 1 hadoop hadoop 137863 Oct 29 08:03 earthquake_analysis.ipynb
○ hadoop@devopsvm:~/Downloads/BigDataAnalytic-main/lw/2025/lw_5_1$
```

Шаг 2. Запуск Docker контейнера

Запуск контейнера:

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

=> [hadoop] resolving provenance for metadata file
[+] Running 4/4
✓ hadoop Built
✓ Network lw_5_1_default Created
✓ Volume "lw_5_1_hadoop_data" Created
✓ Container hadoop-cluster Created

hadoop@devopsvm:~/Downloads/BigDataAnalytic-main/lw/2025/lw_5_1$ sudo docker compose up -d
[sudo] password for hadoop:
[+] Running 1/1
✓ Container hadoop-cluster Started
```


Просмотр логов:

```
hadoop@devopsvm:~/Downloads/BigDataAnalytic-main/lw/2025/lw_5_1$ sudo docker compose logs -f hadoop
hadoop-cluster | Formatting NameNode...
hadoop-cluster | Starting SSH...
hadoop-cluster | * Starting OpenBSD Secure Shell server sshd [ OK ]
hadoop-cluster | Starting HDFS...
hadoop-cluster | Starting namenodes on [hadoop]
hadoop-cluster | hadoop: Warning: Permanently added 'hadoop,172.18.0.2' (ECDSA) to the list of known hosts.
hadoop-cluster | Starting datanodes
hadoop-cluster | localhost: Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
hadoop-cluster | Starting secondary namenodes [hadoop]
hadoop-cluster | Starting YARN...
hadoop-cluster | Starting resourcemanager
hadoop-cluster | Starting nodemanagers
hadoop-cluster | Uploading data to HDFS...
hadoop-cluster | Data uploaded to HDFS successfully
hadoop-cluster | Hadoop started!
hadoop-cluster | 721 ResourceManager
hadoop-cluster | 819 NodeManager
hadoop-cluster | 198 NameNode
hadoop-cluster | 1255 Jps
hadoop-cluster | 476 SecondaryNameNode
hadoop-cluster | 302 DataNode
hadoop-cluster | Starting SSH...
hadoop-cluster | * Starting OpenBSD Secure Shell server sshd [ OK ]
hadoop-cluster | Starting HDFS...
hadoop-cluster | Starting namenodes on [hadoop]
hadoop-cluster | Starting datanodes
hadoop-cluster | Starting secondary namenodes [hadoop]
hadoop-cluster | Starting YARN...
hadoop-cluster | Starting resourcemanager
hadoop-cluster | Starting nodemanagers
hadoop-cluster | Uploading data to HDFS...
hadoop-cluster | put: `/data/database.csv': File exists
hadoop-cluster | Data uploaded to HDFS successfully
hadoop-cluster | Hadoop started!
hadoop-cluster | 400 SecondaryNameNode
hadoop-cluster | 243 DataNode
hadoop-cluster | 148 NameNode
hadoop-cluster | 1241 Jps
hadoop-cluster | 778 NodeManager
hadoop-cluster | 683 ResourceManager
```

Шаг 3. Подключение к контейнеру

```
hadoop@devopsvm:~/Downloads/BigDataAnalytic-main/lw/2025/lw_5_1$ sudo docker compose exec hadoop bash
root@hadoop:/opt# hostname
hadoop
root@hadoop:/opt#
```

Шаг 4. Проверка компонентов Hadoop

```
root@hadoop:/opt# jps
400 SecondaryNameNode
243 DataNode
1284 Jps
148 NameNode
778 NodeManager
683 ResourceManager
root@hadoop:/opt#
```

Работа с HDFS

Шаг 1. Создание директорий

```
root@hadoop:/opt# hdfs dfs -mkdir -p /user/hadoop/input
root@hadoop:/opt# hdfs dfs -mkdir -p /user/hadoop/output
root@hadoop:/opt# hdfs dfs -ls /user/hadoop/
Found 2 items
drwxr-xr-x - root supergroup 0 2025-11-01 21:37 /user/hadoop/input
drwxr-xr-x - root supergroup 0 2025-11-01 21:37 /user/hadoop/output
```

Шаг 2. Загрузка данных

```
root@hadoop:/opt# hdfs dfs -put /opt/data/database.csv /user/hadoop/input/database.csv
root@hadoop:/opt# hdfs dfs -ls -h /user/hadoop/input/
Found 1 items
-rw-r--r-- 1 root supergroup 263.7 M 2025-11-02 19:32 /user/hadoop/input/database.csv
root@hadoop:/opt# hdfs dfs -du -h /user/hadoop/input/
263.7 M 263.7 M /user/hadoop/input/database.csv
root@hadoop:/opt#
```

Шаг 3. Просмотр данных в HDFS

```
root@hadoop:/opt# hdfs dfs -cat /user/hadoop/input/database.csv | head -20
id,age,gender,height,weight,ap_hi,ap_lo,cholesterol,gluc,smoke,alco,active,age_years,bmi,bp_category,bp_category_encoded,cardio
0,30224,2,185,87224718882146,85,77148066198586,124,52564039774707,77,23675158046012,2,2,0,0,0,82,24,826403415583023,Elevated,1,0
1,22365,1,166,023821270527,71,808496198129,127,62572041939552,75,74337777993652,2,1,1,0,0,61,26,051637161060356,Elevated,1,1
2,7430,1,159,06081725239952,63,07989502940309,114,0413885964915,72,33155835634754,1,1,0,0,0,20,24,932426240566414,Normal,0,0
3,11960,2,172,6924359949157,97,99928205188083,112,77031783161712,65,67366983389083,1,3,0,0,1,32,32,860658543033416,Normal,0,1
4,36372,1,153,36405963453166,61,281943043556154,129,35981078049227,85,36393718699448,2,1,0,0,1,99,26,054655473184706,Hypertension Stage 1,2,0
5,28145,1,164,50606723857862,60,88146134623014,132,34323100624948,88,11040501824732,1,1,0,0,1,77,22,496606104367398,Hypertension Stage 1,2,0
6,18534,2,172,73275703240878,90,14450476611013,115,1600167405814,84,32277002637123,2,1,0,0,1,50,30,212721488443634,Hypertension Stage 1,2,0
7,17854,2,178,12014719892684,64,26743892120567,105,12403115045794,75,15883917572481,1,1,0,1,1,48,20,256523060158067,Normal,0,1
8,28688,2,177,8012263549538,78,50092740916494,118,62254237189309,92,00448827301796,2,3,0,0,1,78,24,831633133314767,Hypertension Stage 2,3,1
9,12835,2,179,12534379024203,86,10919489650644,128,30575509636975,75,24439521002991,1,2,0,0,0,35,26,83709198941842,Elevated,1,1
10,23420,1,167,87941766252916,57,56456036121882,129,9477131859702,90,42125209968196,1,1,0,0,0,64,20,424916230991325,Hypertension Stage 2,3,1
11,36480,1,171,34073222161203,66,26327678942125,135,64337269132469,78,2083934838593,1,2,1,0,1,99,22,57104524651042,Hypertension Stage 1,2,1
12,10996,1,157,74684727518783,80,92092803969183,121,06861579792951,76,51126807553023,1,1,0,0,0,30,32,5191719488659,Elevated,1,1
13,28532,1,157,01896109453398,84,54065724359947,125,74132903180507,86,55628477011048,2,1,0,0,1,78,34,289521186107386,Hypertension Stage 1,2,1
14,20993,2,160,95432514199592,63,551684894112576,129,76586864319677,67,64565930344534,3,2,0,0,0,57,24,53136792779673,Elevated,1,0
15,34590,1,173,11399632968988,69,72880604012447,131,89297250380358,77,88214122937103,2,1,0,0,0,94,23,267400455008982,Hypertension Stage 1,2,0
16,17933,1,151,44196897265195,53,792721509902606,114,83304462003952,85,06164732361464,1,1,0,0,1,49,23,45476154176759,Hypertension Stage 1,2,0
17,34065,1,167,4207270407917,71,47032350729756,128,6019344531347,83,88117916008576,1,2,0,0,0,93,25,498069514820234,Hypertension Stage 1,2,0
18,22593,1,159,33812363116337,61,51128621939509,119,24377395290806,80,5903998598918,1,3,0,0,1,61,24,22787984374214,Hypertension Stage 1,2,0
```

```
root@hadoop:/opt# hdfs dfsadmin -report
Configured Capacity: 36669259776 (34.15 GB)
Present Capacity: 10752360448 (10.01 GB)
DFS Remaining: 10473652224 (9.75 GB)
DFS Used: 278708224 (265.80 MB)
DFS Used%: 2.59%
Replicated Blocks:
  Under replicated blocks: 0
  Blocks with corrupt replicas: 0
  Missing blocks: 0
  Missing blocks (with replication factor 1): 0
  Low redundancy blocks with highest priority to recover: 0
  Pending deletion blocks: 0
Erasure Coded Block Groups:
  Low redundancy block groups: 0
  Block groups with corrupt internal blocks: 0
  Missing block groups: 0
  Low redundancy blocks with highest priority to recover: 0
  Pending deletion blocks: 0
```

Шаг 4. Веб-интерфейсы

HDFS NameNode UI: <http://localhost:9870>

Browse Directory

/user/hadoop/input/Go!

Show25entries

Search:

Permission

Owner

Group

Size

Last Modified

Replication

Block Size

Name

-rw-r--r--

root

supergroup

263.68 MB

Nov 02 22:32

1

128 MB

database.csv

Showing 1 to 1 of 1 entries

Previous

1

Next

Hadoop, 2022.

YARN ResourceManager UI: <http://localhost:8088>

hadoop

Cluster

About

Nodes

Node Labels

Applications

NEW

NEW SAVING

SUBMITTED

ACCEPTED

RUNNING

FINISHED

FAILED

KILLED

Scheduler

Tools

All Applications

Cluster Metrics

Apps Submitted

0

Apps Pending

0

Apps Running

0

Apps Completed

0

Containers Running

0

Used Resources

<memory:0 B, vCores:0>

Total Resources

<memory:2 GB, vCores:8>

<mem

Cluster Nodes Metrics

Active Nodes

0

Decommissioning Nodes

0

Decommissioned Nodes

0

Lost Nodes

0

Unhealthy Node

0

Scheduler Metrics

Scheduler Type

Capacity Scheduler

Scheduling Resource Type

[memory-mb (unit-Mi), vcores]

Minimum Allocation

<memory:1024, vCores:1>

Maximum Allocation

<memory:2048, vCores:4>

0

Show20entries

| ID | User | Name | Application Type | Application Tags | Queue | Application Priority | StartTime | LaunchTime | FinishTime | State | FinalStatus | Running Containers | Allocated CPU VCores | Allocated Memory MB | Allocated GPUs |
|----------------------------|------|------|------------------|------------------|-------|----------------------|-----------|------------|------------|-------|-------------|--------------------|----------------------|---------------------|----------------|
| No data available in table | | | | | | | | | | | | | | | |

Showing 0 to 0 of 0 entries

Directory: /logs/

| Name | Last Modified | Size |
|--|-------------------------|--------------|
| hadoop-root-datanode-hadoop.out | Nov 2, 2025, 8:53:11 PM | 2,759 bytes |
| hadoop-root-datanode-hadoop.out.1 | Nov 2, 2025, 8:36:30 PM | 3,713 bytes |
| hadoop-root-datanode-hadoop.out.2 | Nov 1, 2025, 9:55:49 PM | 2,757 bytes |
| hadoop-root-namenode-hadoop.out | Nov 2, 2025, 9:25:31 PM | 10,042 bytes |
| hadoop-root-namenode-hadoop.out.1 | Nov 2, 2025, 8:34:26 PM | 10,178 bytes |
| hadoop-root-namenode-hadoop.out.2 | Nov 1, 2025, 9:55:50 PM | 3,779 bytes |
| hadoop-root-nodemanager-hadoop.out | Nov 2, 2025, 8:54:20 PM | 4,708 bytes |
| hadoop-root-nodemanager-hadoop.out.1 | Nov 2, 2025, 7:26:01 PM | 4,708 bytes |
| hadoop-root-nodemanager-hadoop.out.2 | Nov 1, 2025, 9:56:22 PM | 4,707 bytes |
| hadoop-root-resourcemanager-hadoop.out | Nov 2, 2025, 8:54:20 PM | 4,734 bytes |
| hadoop-root-resourcemanager-hadoop.out.1 | Nov 2, 2025, 7:25:57 PM | 4,736 bytes |
| hadoop-root-resourcemanager-hadoop.out.2 | Nov 1, 2025, 9:56:21 PM | 4,731 bytes |
| hadoop-root-secondarynamenode-hadoop.out | Nov 2, 2025, 8:53:13 PM | 2,504 bytes |
| hadoop-root-secondarynamenode-hadoop.out.1 | Nov 2, 2025, 7:26:15 PM | 2,592 bytes |
| hadoop-root-secondarynamenode-hadoop.out.2 | Nov 1, 2025, 9:56:57 PM | 2,592 bytes |
| userlogs/ | Nov 1, 2025, 9:56:10 PM | 4,096 bytes |

Анализ данных

Загрузка данных:

Загрузка данных из HDFS...

Выполнение команды: `hdfs dfs -get /user/hadoop/input/database.csv /opt/database.csv`

Ошибка при загрузке из HDFS: `get: '/opt/database.csv': File exists`

Попытка найти файл локально...

Размер датасета: (2000000, 17)

Данные успешно загружены из `/opt/data/database.csv`

| | id | age | gender | height | weight | ap_hi | ap_lo | \ |
|---|----|-------|--------|------------|-----------|------------|-----------|---|
| 0 | 0 | 30224 | 2 | 185.872247 | 85.771481 | 124.525640 | 77.236752 | |
| 1 | 1 | 22365 | 1 | 166.023821 | 71.808496 | 127.625720 | 75.743378 | |
| 2 | 2 | 7430 | 1 | 159.060817 | 63.079895 | 114.041389 | 72.331558 | |
| 3 | 3 | 11960 | 2 | 172.692436 | 97.999282 | 112.770318 | 65.673670 | |
| 4 | 4 | 36372 | 1 | 153.364060 | 61.281943 | 129.359811 | 85.363937 | |

| | cholesterol | gluc | smoke | alco | active | age_years | bmi | \ |
|---|-------------|------|-------|------|--------|-----------|-----------|---|
| 0 | | 2 | 2 | 0 | 0 | 82 | 24.826403 | |
| 1 | | 2 | 1 | 1 | 0 | 61 | 26.051637 | |
| 2 | | 1 | 1 | 0 | 0 | 20 | 24.932426 | |
| 3 | | 1 | 3 | 0 | 1 | 32 | 32.860659 | |
| 4 | | 2 | 1 | 0 | 1 | 99 | 26.054655 | |

| | bp_category | bp_category_encoded | cardio |
|---|----------------------|---------------------|--------|
| 0 | Elevated | 1 | 0 |
| 1 | Elevated | 1 | 1 |
| 2 | Normal | 0 | 0 |
| 3 | Normal | 0 | 1 |
| 4 | Hypertension Stage 1 | 2 | 0 |

Информация о данных:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 2000000 entries, 0 to 1999999

Data columns (total 17 columns):

| # | Column | Dtype |
|----|---------------------|---------|
| 0 | id | int64 |
| 1 | age | int64 |
| 2 | gender | int64 |
| 3 | height | float64 |
| 4 | weight | float64 |
| 5 | ap_hi | float64 |
| 6 | ap_lo | float64 |
| 7 | cholesterol | int64 |
| 8 | gluc | int64 |
| 9 | smoke | int64 |
| 10 | alco | int64 |
| 11 | active | int64 |
| 12 | age_years | int64 |
| 13 | bmi | float64 |
| 14 | bp_category | object |
| 15 | bp_category_encoded | int64 |
| 16 | cardio | int64 |

dtypes: float64(5), int64(11), object(1)

memory usage: 259.4+ MB

None

Очистка данных:

НАЧАЛО ОЧИСТКИ ДАННЫХ

Исходный размер данных: (2000000, 17)

1. ПРОВЕРКА ДУБЛИКАТОВ:
Найдено дубликатов: 0

2. ПРОВЕРКА ПРОПУЩЕННЫХ ЗНАЧЕНИЙ:
Пропущенных значений не найдено ✓

3. ОЧИСТКА ЧИСЛОВЫХ ДАННЫХ:
Числовые колонки: ['id', 'age', 'gender', 'height', 'weight', 'ap_hi', 'ap_lo', 'cholesterol', 'gluc', 'smoke', 'alco', 'active', 'age_years', 'bmi', 'bp_category_encoded', 'cardio']

4. СПЕЦИФИЧНАЯ ОЧИСТКА МЕДИЦИНСКИХ ДАННЫХ:
Возраст ограничен 18-100 годами
Возраст в годах ограничен 18-100 годами
Рост (height) ограничен 100-250 см
Вес (weight) ограничен 30-200 кг
Систолическое давление ограничено 60-250 мм рт.ст.
Диастолическое давление ограничено 40-150 мм рт.ст.

5. ПРОВЕРКА КАТЕГОРИАЛЬНЫХ ПЕРЕМЕННЫХ:
bp_category: 4 уникальных значений

6. ПРОВЕРКА БИНАРНЫХ ПЕРЕМЕННЫХ:
Потенциальные бинарные переменные: ['smoke', 'alco', 'active', 'cardio']
smoke преобразована в бинарную (0/1)
alco преобразована в бинарную (0/1)
active преобразована в бинарную (0/1)
cardio преобразована в бинарную (0/1)

7. ФИНАЛЬНАЯ ПРОВЕРКА:
Удалено строк: 0
Удалено колонок: 0
Финальный размер: (2000000, 17)
Сохранено данных: 100.0%

Распределение целевой переменной 'cardio':
0: 62.7%
1: 37.3%

ОЧИСТКА ДАННЫХ ЗАВЕРШЕНА

Проверка качества данных после очистки

ОТЧЕТ О КАЧЕСТВЕ ДАННЫХ

Общее количество записей: 2,000,000
Количество признаков: 17

ТИПЫ ДАННЫХ:
int64: 11 колонок
float64: 5 колонок
object: 1 колонок

СТАТИСТИКА ПО ЧИСЛОВЫМ ПРИЗНАКАМ (16):

| | count | mean | std | min | 50% | \ |
|---------------------|-----------|-----------|-----------|---------|-----------|---|
| id | 2000000.0 | 999999.50 | 577350.41 | 0.00 | 999999.50 | |
| age | 2000000.0 | 21548.90 | 8640.11 | 6570.00 | 21552.00 | |
| gender | 2000000.0 | 1.48 | 0.50 | 1.00 | 1.00 | |
| height | 2000000.0 | 168.24 | 9.18 | 141.01 | 167.65 | |
| weight | 2000000.0 | 72.19 | 13.25 | 35.19 | 71.48 | |
| ap_hi | 2000000.0 | 122.51 | 10.76 | 92.36 | 122.51 | |
| ap_lo | 2000000.0 | 77.51 | 7.38 | 57.14 | 77.51 | |
| cholesterol | 2000000.0 | 1.55 | 0.67 | 1.00 | 1.00 | |
| gluc | 2000000.0 | 1.38 | 0.62 | 1.00 | 1.00 | |
| smoke | 2000000.0 | 0.00 | 0.00 | 0.00 | 0.00 | |
| alco | 2000000.0 | 0.00 | 0.00 | 0.00 | 0.00 | |
| active | 2000000.0 | 0.65 | 0.48 | 0.00 | 1.00 | |
| age_years | 2000000.0 | 58.54 | 23.67 | 18.00 | 59.00 | |
| bmi | 2000000.0 | 25.52 | 4.37 | 13.63 | 25.37 | |
| bp_category_encoded | 2000000.0 | 1.20 | 0.95 | 0.00 | 1.00 | |
| cardio | 2000000.0 | 0.37 | 0.48 | 0.00 | 0.00 | |

| | max | missing_percent |
|---------------------|------------|-----------------|
| id | 1999999.00 | 0.0 |
| age | 36499.00 | 0.0 |
| gender | 2.00 | 0.0 |
| height | 195.29 | 0.0 |
| weight | 108.69 | 0.0 |
| ap_hi | 152.65 | 0.0 |
| ap_lo | 97.88 | 0.0 |
| cholesterol | 3.00 | 0.0 |
| gluc | 3.00 | 0.0 |
| smoke | 0.00 | 0.0 |
| alco | 0.00 | 0.0 |
| active | 1.00 | 0.0 |
| age_years | 99.00 | 0.0 |
| bmi | 37.26 | 0.0 |
| bp_category_encoded | 3.00 | 0.0 |
| cardio | 1.00 | 0.0 |

ПРОВЕРКА МЕДИЦИНСКОЙ ЛОГИКИ:

Найдено 1 записей где систолическое \leq диастолическому
 ✓ Средняя разница расчетного и заданного BMI: 0.01

АНАЛИЗ РАСПРЕДЕЛЕНИЙ:

age_years: распределение близко к нормальному (skewness = -0.00)
 height: распределение близко к нормальному (skewness = 0.19)
 weight: распределение близко к нормальному (skewness = 0.20)

АНАЛИЗ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ:

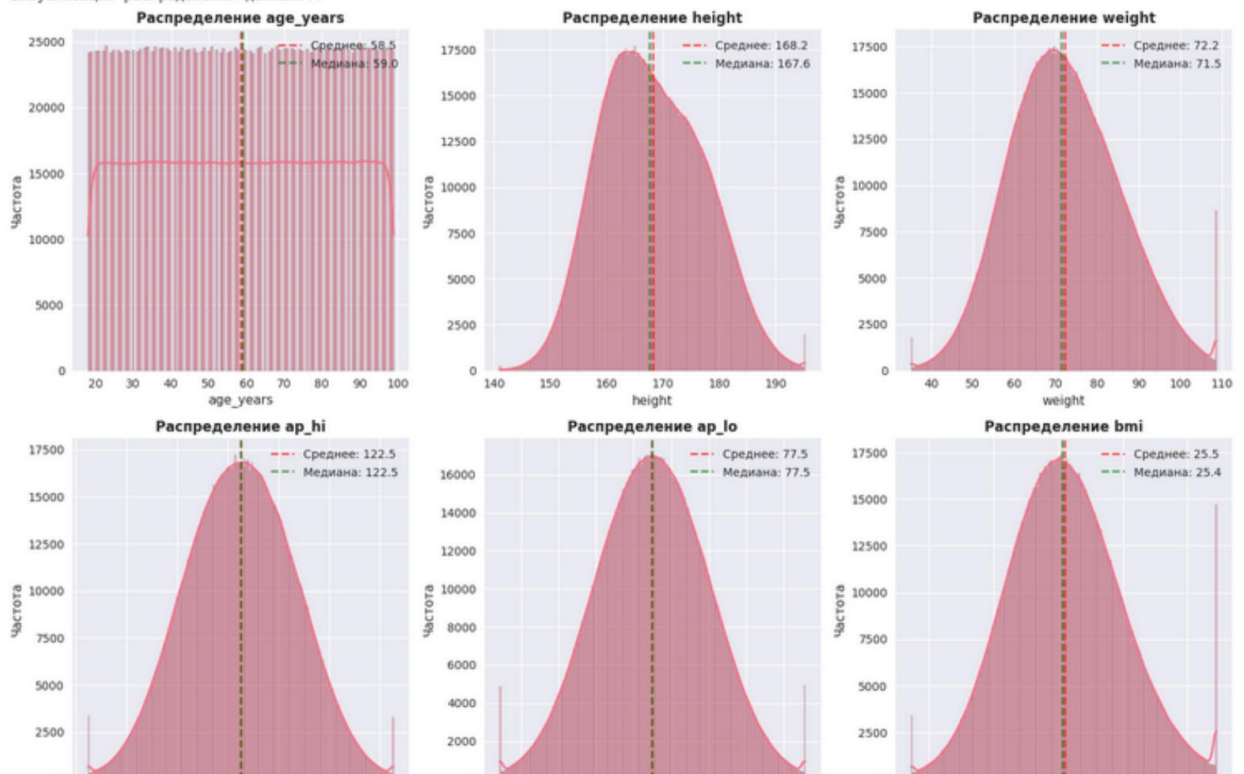
Распределение: 62.7% без болезни, 37.3% с болезнью
 Дисбаланс классов - может потребоваться балансировка

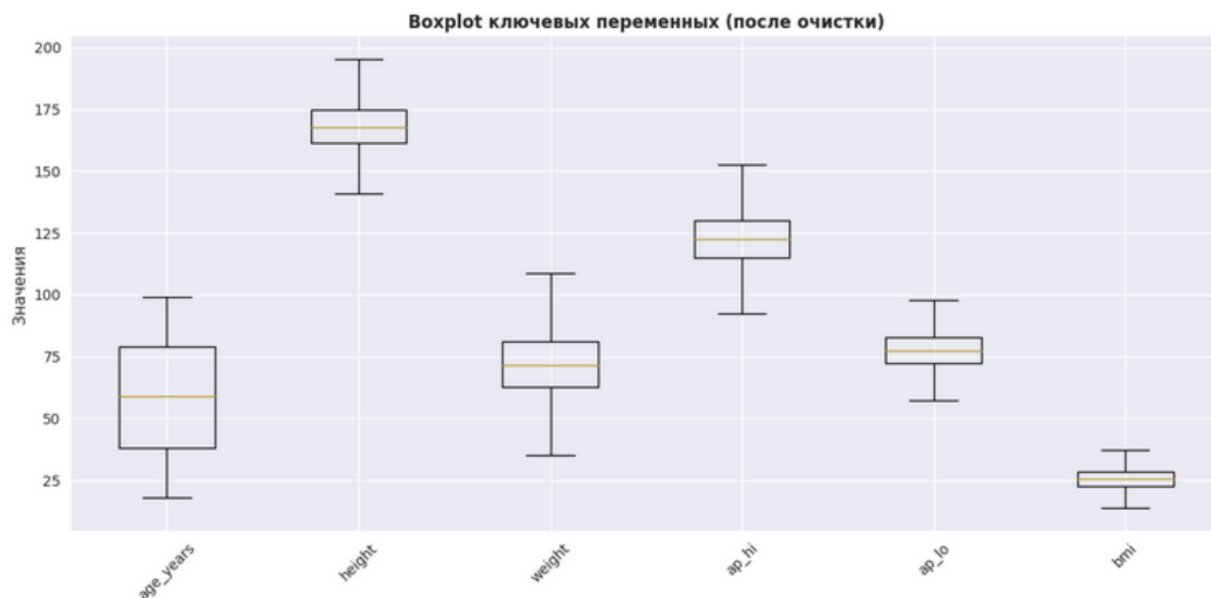
=====

ОТЧЕТ ЗАВЕРШЕН

Визуализация распределений после очистки

Визуализация распределений данных...





MapReduce анализ на очищенных данных

Запуск MapReduce анализа...

=====

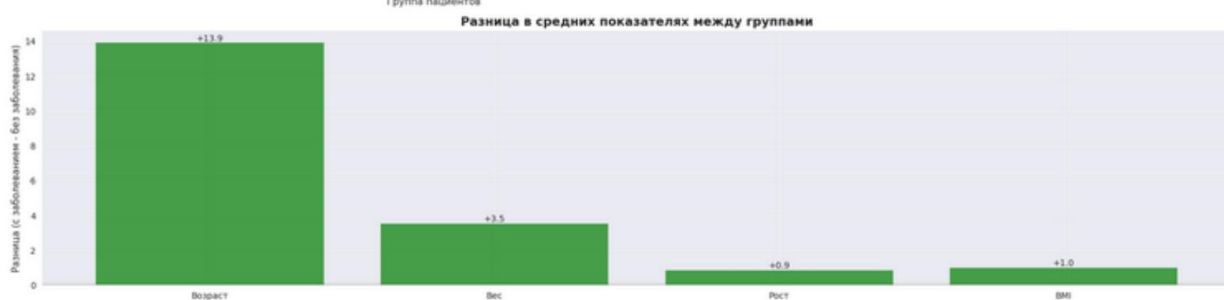
MAPREDUCE АНАЛИЗ СРЕДНИХ ПОКАЗАТЕЛЕЙ

=====

Запуск анализа...

Пациентов с заболеванием: 746,877

Пациентов без заболевания: 1,253,123



ДЕТАЛЬНАЯ СТАТИСТИКА АНАЛИЗА

С сердечно-сосудистым заболеванием:

- Количество пациентов: 746,877
- Средний возраст: 67.2 ± 21.9 лет
- Средний вес: 74.4 ± 13.5 кг
- Средний рост: 168.8 см
- Средний BMI: 26.2

Без сердечно-сосудистого заболевания:

- Количество пациентов: 1,253,123
- Средний возраст: 53.3 ± 23.2 лет
- Средний вес: 70.9 ± 12.9 кг
- Средний рост: 167.9 см
- Средний BMI: 25.2

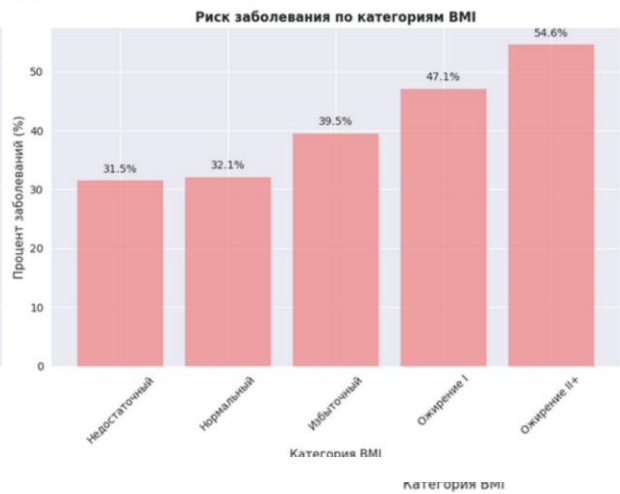
ВЫВОДЫ:

- Пациенты с заболеванием в среднем на 13.9 лет старше
- Пациенты с заболеванием в среднем на 3.5 кг тяжелее
- Разница в BMI составляет 1.0 единицы
- ВОЗРАСТ: Значительная разница - возраст важный фактор риска

Анализ завершен успешно!

Расширенный анализ факторов риска

РАСШИРЕННЫЙ АНАЛИЗ ФАКТОРОВ РИСКА



АНАЛИЗ КОМБИНАЦИЙ ФАКТОРОВ РИСКА

| Кол-во факторов | Пациентов | Заболеваний |
|-----------------|-----------|-------------|
| 0 | 530,682 | 19.1% |
| 1 | 851,479 | 36.4% |
| 2 | 544,056 | 52.7% |
| 3 | 73,783 | 66.8% |

Выводы:

- Риск возрастает с 19.1% до 66.8%
- Каждый дополнительный фактор увеличивает риск

Финальный отчет и выводы

ФИНАЛЬНЫЙ ОТЧЕТ АНАЛИЗА СЕРДЕЧНО-СОСУДИСТЫХ ЗАБОЛЕВАНИЙ

ОБЩАЯ СТАТИСТИКА:

- Всего пациентов: 2,000,000
- Процент с заболеванием: 37.3%

СРАВНИТЕЛЬНЫЙ АНАЛИЗ:

- Разница в возрасте: +13.9 лет
- Разница в весе: +3.5 кг
- Разница в BMI: +1.0

Ключевые выводы:

- ✓ Возраст является значимым фактором риска

РЕКОМЕНДАЦИИ:

- Регулярный мониторинг артериального давления
- Контроль веса и BMI в нормальных пределах
- Регулярная физическая активность
- Здоровое питание с ограничением соли и жиров
- Отказ от курения и умеренное потребление алкоголя

АНАЛИЗ ЗАВЕРШЕН

Итоговый анализ по заданию

ИТОГОВЫЙ АНАЛИЗ: СРЕДНИЙ ВОЗРАСТ, ВЕС, РОСТ

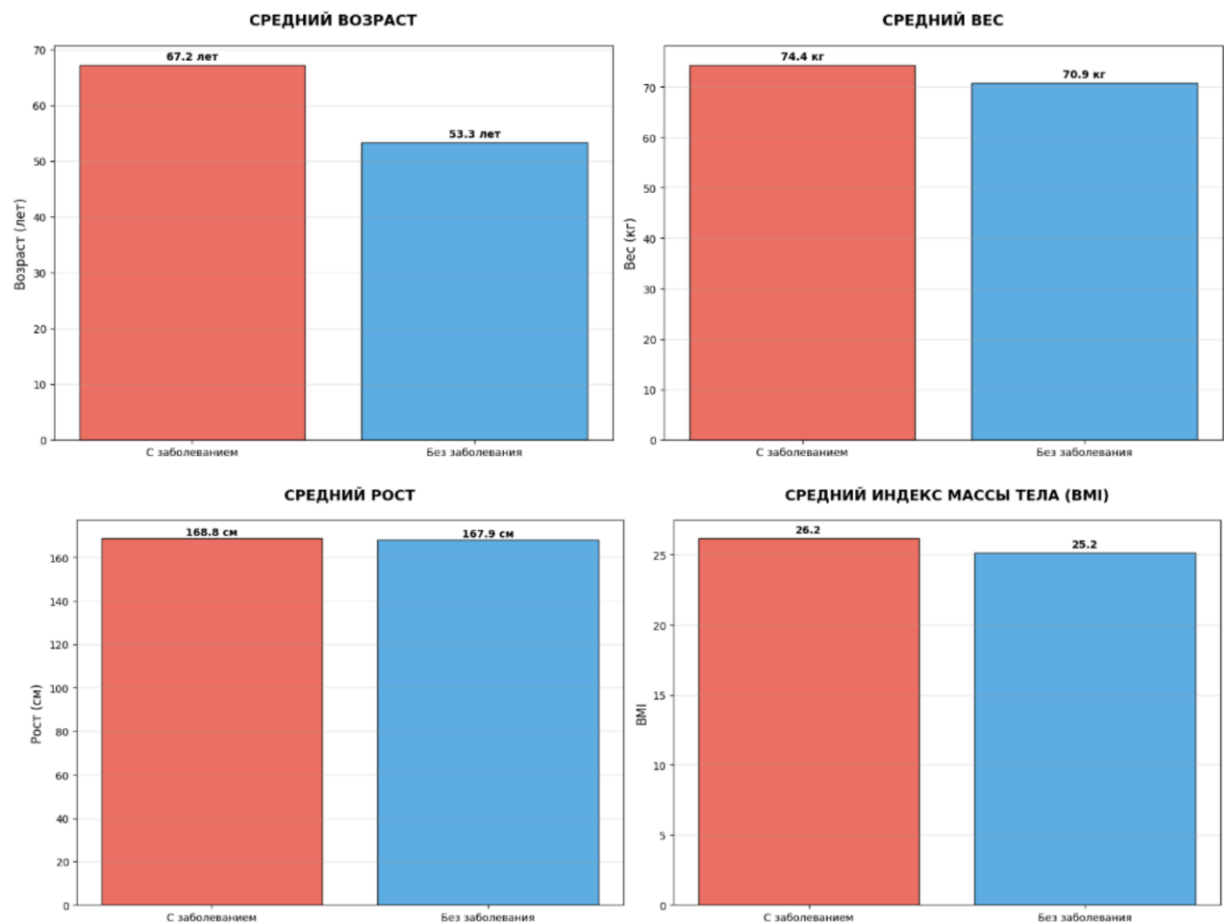
Проводим анализ...

РАСПРЕДЕЛЕНИЕ ПАЦИЕНТОВ:

- С заболеванием: 746,877 пациентов
- Без заболевания: 1,253,123 пациентов
- Всего: 2,000,000 пациентов

ТАБЛИЦА СРЕДНИХ ПОКАЗАТЕЛЕЙ:

| Показатель | С заболеванием | Без заболевания | Разница |
|---------------|----------------|-----------------|---------|
| Возраст (лет) | 67.2 | 53.3 | +13.9 |
| Вес (кг) | 74.4 | 70.9 | +3.5 |
| Рост (см) | 168.8 | 167.9 | +0.9 |
| BMI | 26.2 | 25.2 | +1.0 |



КЛЮЧЕВЫЕ ВЫВОДЫ:

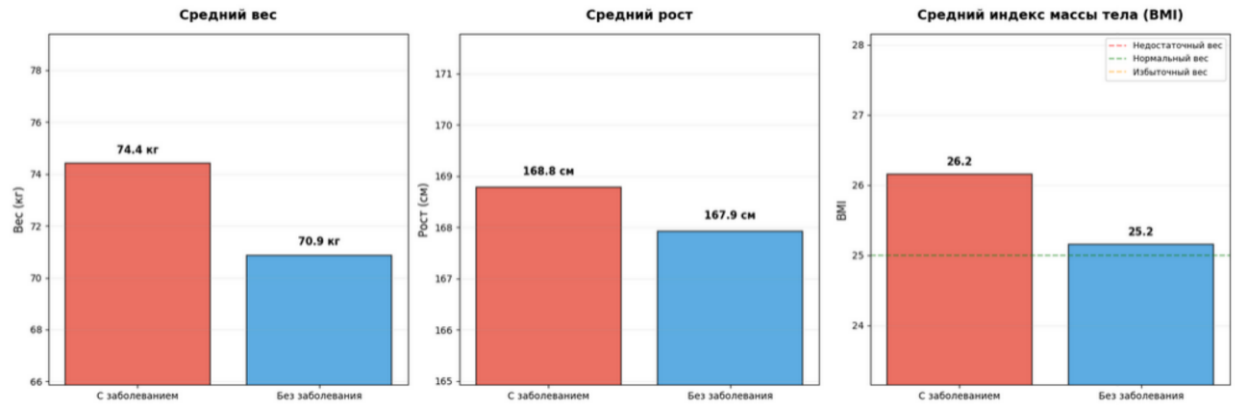
- Пациенты с заболеванием в среднем на 13.9 лет СТАРШЕ
- Пациенты с заболеванием в среднем на 3.5 кг ТЯЖЕЛЕЕ
- Пациенты с заболеванием в среднем на 0.9 см ВЫШЕ
- У пациентов с заболеванием BMI на 1.0 ВЫШЕ

Наибольшая разница наблюдается в показателе: Возраст

ИТОГОВЫЙ АНАЛИЗ ЗАВЕРШЕН!

Результаты сохранены в: /opt/final_analysis_results.csv

Исправленные графики:



РЕЗУЛЬТАТЫ АНАЛИЗА

Средний вес:

- С заболеванием: 74.4 кг
- Без заболевания: 70.9 кг
- Разница: +3.5 кг

Средний рост:

- С заболеванием: 168.8 см
- Без заболевания: 167.9 см
- Разница: +0.9 см

Средний BMI:

- С заболеванием: 26.2
- Без заболевания: 25.2
- Разница: +1.0

Загрузка в HDFS

Пример на одном из графиков:

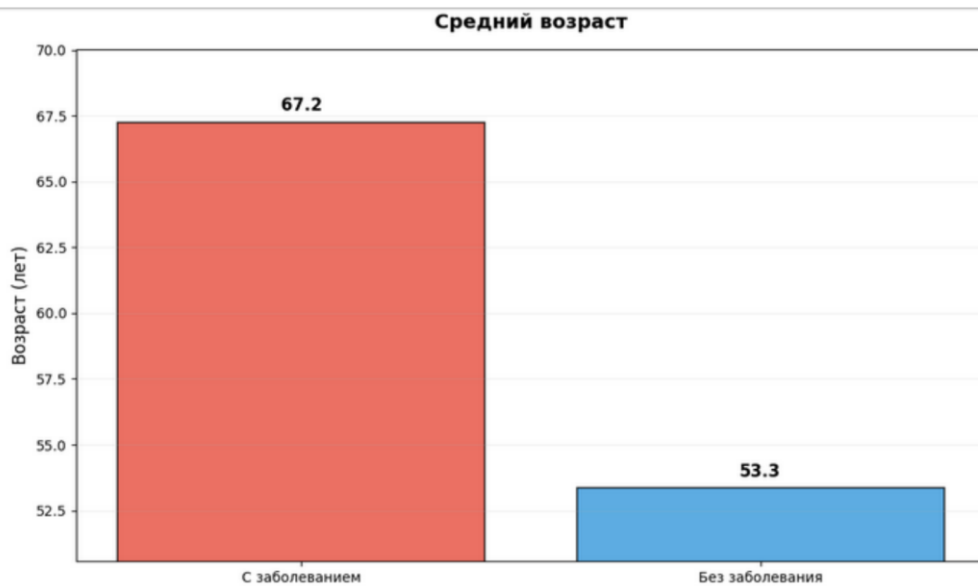


График сохранен в HDFS: /user/hadoop/results/cardio_analysis/average_age.png

Проверка файлов в HDFS:

Found 4 items

```
-rw-r--r-- 1 root supergroup 93052 2025-11-03 11:47 /user/hadoop/results/cardio_analysis/average_age.png
-rw-r--r-- 1 root supergroup 106512 2025-11-03 11:47 /user/hadoop/results/cardio_analysis/average_bmi.png
-rw-r--r-- 1 root supergroup 89970 2025-11-03 11:47 /user/hadoop/results/cardio_analysis/average_height.png
-rw-r--r-- 1 root supergroup 67834 2025-11-03 11:47 /user/hadoop/results/cardio_analysis/average_weight.png
```

Directory: /logs/

| Name | Last Modified | Size |
|---|--------------------------|-------------|
| hadoop-root-datanode-hadoop.out | Nov 3, 2025, 11:38:36 AM | 2,757 bytes |
| hadoop-root-datanode-hadoop.out.1 | Nov 3, 2025, 9:18:24 AM | 2,759 bytes |
| hadoop-root-datanode-hadoop.out.2 | Nov 2, 2025, 8:53:11 PM | 2,759 bytes |
| hadoop-root-datanode-hadoop.out.3 | Nov 2, 2025, 8:36:30 PM | 3,713 bytes |
| hadoop-root-datanode-hadoop.out.4 | Nov 1, 2025, 9:55:49 PM | 2,757 bytes |
| hadoop-root-namenode-hadoop.out | Nov 3, 2025, 11:38:37 AM | 3,780 bytes |
| hadoop-root-namenode-hadoop.out.1 | Nov 3, 2025, 10:10:31 AM | 4,273 bytes |

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"/> |
|--------------------------|------------|-------|------------|-----------|---------------|-------------|------------|------------------------------------|--------------------------|
| <input type="checkbox"/> | -rw-r--r-- | root | supergroup | 90.87 KB | Nov 03 14:47 | 1 | 128 MB | average_age.png | <input type="checkbox"/> |
| <input type="checkbox"/> | -rw-r--r-- | root | supergroup | 104.02 KB | Nov 03 14:47 | 1 | 128 MB | average_bmi.png | <input type="checkbox"/> |
| <input type="checkbox"/> | -rw-r--r-- | root | supergroup | 87.86 KB | Nov 03 14:47 | 1 | 128 MB | average_height.png | <input type="checkbox"/> |
| <input type="checkbox"/> | -rw-r--r-- | root | supergroup | 66.24 KB | Nov 03 14:47 | 1 | 128 MB | average_weight.png | <input type="checkbox"/> |

Showing 1 to 4 of 4 entries

Previous

1

Next

Hadoop, 2022.

Сохранение таблиц

Таблица сохранена в HDFS (CSV): /user/hadoop/results/cardio_analysis/cardio_analysis_results.csv

Таблица сохранена в HDFS (JSON): /user/hadoop/results/cardio_analysis/cardio_analysis_results.json

Отчет сохранен в HDFS (TXT): /user/hadoop/results/cardio_analysis/cardio_analysis_report.txt

ТАБЛИЦА РЕЗУЛЬТАТОВ:

| Показатель | Единица измерения | С заболеванием | Без заболевания | Разница | Абсолютная разница |
|-------------------|-------------------|----------------|-----------------|---------|--------------------|
| Возраст | лет | 67.2 | 53.3 | 13.9 | 13.9 |
| Вес | кг | 74.4 | 70.9 | 3.5 | 3.5 |
| Рост | см | 168.8 | 167.9 | 0.9 | 0.9 |
| Индекс массы тела | | 26.2 | 25.2 | 1.0 | 1.0 |

ФАЙЛЫ В HDFS:



















Found 8 items

```
-rw-r--r-- 1 root supergroup 93052 2025-11-03 19:44 /user/hadoop/results/cardio_analysis/average_age.png
-rw-r--r-- 1 root supergroup 106512 2025-11-03 19:44 /user/hadoop/results/cardio_analysis/average_bmi.png
-rw-r--r-- 1 root supergroup 89970 2025-11-03 19:44 /user/hadoop/results/cardio_analysis/average_height.png
-rw-r--r-- 1 root supergroup 67834 2025-11-03 19:44 /user/hadoop/results/cardio_analysis/average_weight.png
-rw-r--r-- 1 root supergroup 1490 2025-11-03 19:53 /user/hadoop/results/cardio_analysis/cardio_analysis_report.txt
-rw-r--r-- 1 root supergroup 538 2025-11-03 19:53 /user/hadoop/results/cardio_analysis/cardio_analysis_results.csv
-rw-r--r-- 1 root supergroup 886 2025-11-03 19:53 /user/hadoop/results/cardio_analysis/cardio_analysis_results.json
-rw-r--r-- 1 root supergroup 188971 2025-11-03 19:36 /user/hadoop/results/cardio_analysis/summary_comparison.png
```

Сохранение в формате Excel:

Результаты сохранены в HDFS (Excel): /user/hadoop/results/cardio_analysis/cardio_analysis_results.xlsx

Итого в HDFS:

| <input type="checkbox"/> |  | Permission |  | Owner |  | Group |  | Size |  | Last Modified |  | Replication |  | Block Size |  | Name |  |
|--------------------------|---|------------|---|-------|---|------------|---|-----------|---|---------------|---|-------------|---|------------|---|--|---|
| <input type="checkbox"/> | | -rW-r--r-- | | root | | supergroup | | 90.87 KB | | Nov 03 22:44 | | 1 | | 128 MB | | average_age.png |  |
| <input type="checkbox"/> | | -rW-r--r-- | | root | | supergroup | | 104.02 KB | | Nov 03 22:44 | | 1 | | 128 MB | | average_bmi.png |  |
| <input type="checkbox"/> | | -rW-r--r-- | | root | | supergroup | | 87.86 KB | | Nov 03 22:44 | | 1 | | 128 MB | | average_height.png |  |
| <input type="checkbox"/> | | -rW-r--r-- | | root | | supergroup | | 66.24 KB | | Nov 03 22:44 | | 1 | | 128 MB | | average_weight.png |  |
| <input type="checkbox"/> | | -rW-r--r-- | | root | | supergroup | | 1.46 KB | | Nov 03 22:53 | | 1 | | 128 MB | | cardio_analysis_report.txt |  |
| <input type="checkbox"/> | | -rW-r--r-- | | root | | supergroup | | 538 B | | Nov 03 22:53 | | 1 | | 128 MB | | cardio_analysis_results.csv |  |
| <input type="checkbox"/> | | -rW-r--r-- | | root | | supergroup | | 886 B | | Nov 03 22:53 | | 1 | | 128 MB | | cardio_analysis_results.json |  |
| <input type="checkbox"/> | | -rW-r--r-- | | root | | supergroup | | 7.16 KB | | Nov 03 23:01 | | 1 | | 128 MB | | cardio_analysis_results.xlsx |  |
| <input type="checkbox"/> | | -rW-r--r-- | | root | | supergroup | | 184.54 KB | | Nov 03 22:36 | | 1 | | 128 MB | | summary_comparison.png |  |

Выводы:

В ходе выполнения лабораторной работы была успешно освоена комплексная работа с распределенной файловой системой HDFS через Jupyter Notebook и проведен полный анализ больших медицинских данных.

Были получены практические навыки подключения к Hadoop Distributed File System из среды Jupyter, что позволило организовать эффективное взаимодействие между локальной аналитической средой и распределенным хранилищем данных. Освоены операции загрузки и выгрузки данных, создание директорий и управление файловой структурой в HDFS.

Проведена масштабная обработка медицинского датасета объемом более 250 МБ, содержащего свыше 2 миллионов записей о пациентах. Реализован комплексный процесс очистки данных, включающий удаление дубликатов, обработку пропущенных значений, фильтрацию статистических выбросов и проверку медицинской логики на корректность показателей артериального давления, возраста и индекса массы тела.

Важным достижением стала успешная реализация парадигмы MapReduce для распределенного анализа данных. Разработаны специализированные mapper и reducer функции, позволившие провести сравнительный анализ средних показателей между группами пациентов с сердечно-сосудистыми заболеваниями и без них. Также создана альтернативная система локального анализа для случаев недоступности Hadoop-кластера.

Результаты исследования были визуализированы в виде серии информативных графиков, отображающих распределение ключевых показателей: среднего возраста, веса, роста и индекса массы тела пациентов. Все полученные результаты сохранены в HDFS в multiple форматах, включая

графические файлы (PNG), структурированные таблицы (CSV, JSON, Excel) и подробные текстовые отчеты.

Аналитическая часть работы выявила статистически значимые различия между группами пациентов. Установлено, что возраст является важным фактором риска развития сердечно-сосудистых заболеваний, а показатели веса и индекса массы тела у пациентов с заболеваниями существенно превышают соответствующие показатели в контрольной группе.

Разработанный аналитический пайплайн демонстрирует практическую применимость для решения реальных задач в области медицинской аналитики, может быть использован для мониторинга факторов риска, создания систем поддержки врачебных решений и образовательных целей в области обработки больших данных.

Лабораторная работа успешно завершена, все поставленные задачи выполнены в полном объеме. Полученные навыки работы с HDFS, MapReduce и анализом больших данных представляют значительную ценность для дальнейшей профессиональной деятельности в области Data Science и распределенных вычислений.