BIG DATA

Студент группы МПИ-22-2-1
Королев Никита

Отчет по лабораторной работе № 2

Open

docker-compose.yml
~/bigdata/hive

output.csv

```yaml
version: "3"

services:
  namenode:
    image: bde2020/hadoop-namenode:2.0.0-hadoop2.7.4-java8
    container_name: namenode
    volumes:
      - ./hdfs/namenode:/hadoop/dfs/name
    environment:
      - CLUSTER_NAME=hive
    env_file:
      - ./hadoop-hive.env
    ports:
      - "50070:50070"

  datanode:
    image: bde2020/hadoop-datanode:2.0.0-hadoop2.7.4-java8
    container_name: datanode
    volumes:
      - ./hdfs/datanode:/hadoop/dfs/data
    env_file:
      - ./hadoop-hive.env
    environment:
      SERVICE_PRECONDITION: "namenode:50070"
    depends_on:
      - namenode
    ports:
      - "50075:50075"

  hive-server:
    image: bde2020/hive:2.3.2-postgresql-metastore
    container_name: hive-server
    volumes:
      - ./zoostore:/zoostore
    env_file:
      - ./hadoop-hive.env
    environment:
      HIVE_CORE_CONF_javax_jdo_option_ConnectionURL: "jdbc:postgresql://hive-metastore/metastore"
      SERVICE_PRECONDITION: "hive-metastore:9083"
    depends_on:
      - hive-metastore
    ports:
      - "10000:10000"

  hive-metastore:
    image: bde2020/hive:2.3.2-postgresql-metastore
    container_name: hive-metastore
    env_file:
      - ./hadoop-hive.env
    command: /opt/hive/bin/hive --service metastore
    environment:
      SERVICE_PRECONDITION: "namenode:50070 datanode:50075 hive-metastore-postgresql:5432"
```

Рис. 1 – конфигурация файла docker-compose.yml

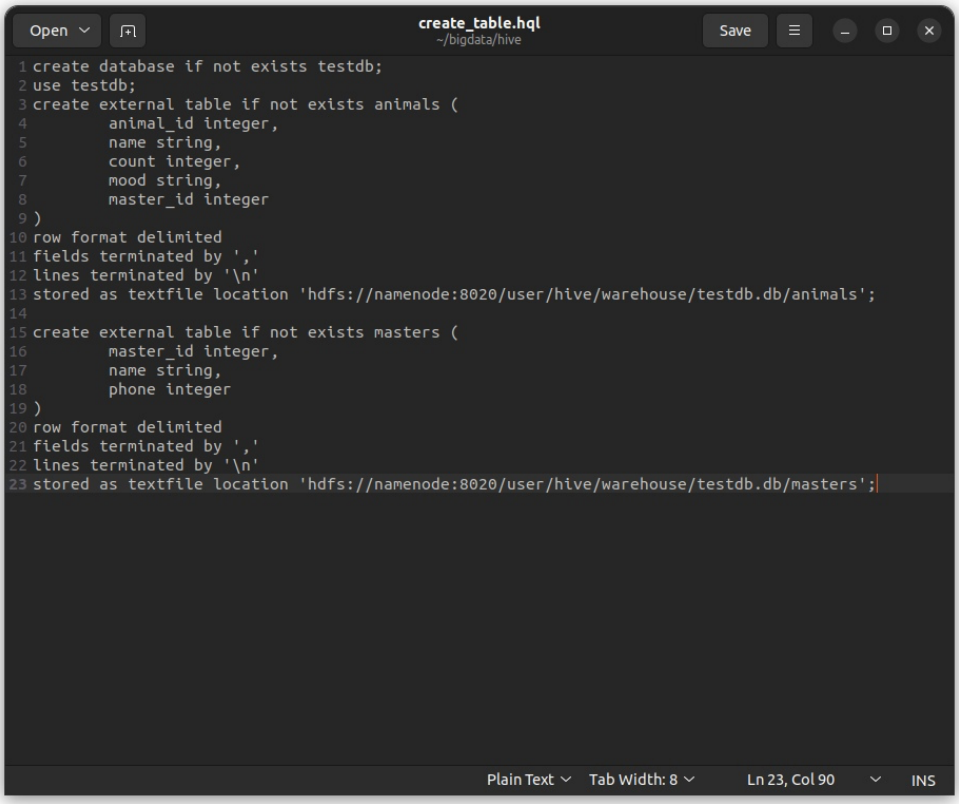

Рис. 2 – скрипт hql для создания базы данных и hive

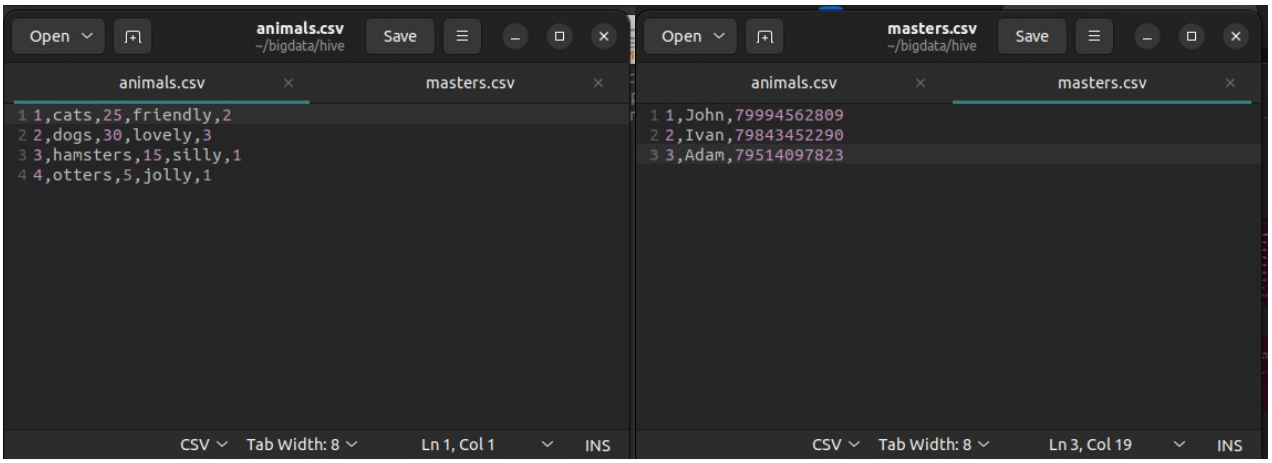

Рис. 3, 4 – содержание файлов animals.csv и masters.csv

```
iratewarrior@iratewarrior-DP156:~/bigdata/hive$ ls
animals.csv  create_table.hql  docker-compose.yml  hadoop-hive.env  hdfs  masters.csv  metastore-postgresql  zoostore
iratewarrior@iratewarrior-DP156:~/bigdata/hive$ sudo docker ps -a
[sudo] password for iratewarrior:
CONTAINER ID   IMAGE                                          COMMAND                  CREATED          STATUS                    PORTS                                                                                  NAME
S
620f2004dd8b   bde2020/hive-metastore-postgresql:2.3.0        "/docker-entrypoint.…"   41 minutes ago   Up 41 minutes             5432/tcp                                                                               hive
-metastore-postgresql
060517e9b20d   bde2020/hadoop-datanode:2.0.0-hadoop2.7.4-java8   "/entrypoint.sh /run…"   41 minutes ago   Up 41 minutes (healthy)   0.0.0.0:50075->50075/tcp, :::50075->50075/tcp                                          data
node
6b464f28d1cd   bde2020/hive:2.3.2-postgresql-metastore         "entrypoint.sh /bin/…"   41 minutes ago   Up 41 minutes             0.0.0.0:10000->10000/tcp, :::10000->10000/tcp, 10002/tcp                               hive
-server
05fb15c3ad74   bde2020/hive:2.3.2-postgresql-metastore         "entrypoint.sh /opt/…"   41 minutes ago   Up 41 minutes             10000/tcp, 0.0.0.0:9083->9083/tcp, :::9083->9083/tcp, 10002/tcp                        hive
-metastore
59ca84fae401   bde2020/hadoop-namenode:2.0.0-hadoop2.7.4-java8   "/entrypoint.sh /run…"   41 minutes ago   Up 41 minutes (healthy)   0.0.0.0:50070->50070/tcp, :::50070->50070/tcp                                          name
node
60b9284fac42   hello-world                                    "/hello"                 17 hours ago     Exited (0) 17 hours ago                                                                                          wiza
rdly_bose
iratewarrior@iratewarrior-DP156:~/bigdata/hive$ sudo docker cp create_table.hql hive-server
must specify at least one container source
iratewarrior@iratewarrior-DP156:~/bigdata/hive$ sudo docker cp create_table.hql hive-server:/tmp
Successfully copied 2.56kB to hive-server:/tmp
iratewarrior@iratewarrior-DP156:~/bigdata/hive$ sudo docker exec -it hive-server /bin/bash
```

```
root@6b464f28d1cd:/opt# cd ..
root@6b464f28d1cd:/# cd tmp
root@6b464f28d1cd:/tmp# hive -f create_table.hql
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hadoop-2.7.4/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in file:/opt/hive/conf/hive-log4j2.properties Async: true
OK
Time taken: 1.098 seconds
OK
Time taken: 0.015 seconds
OK
Time taken: 0.246 seconds
OK
Time taken: 0.054 seconds
root@6b464f28d1cd:/tmp#
```

Рис. 5, 6 – копирование скрипта hql для создания базы данных в hive с хоста в docker контейнер и создание этой базы данных

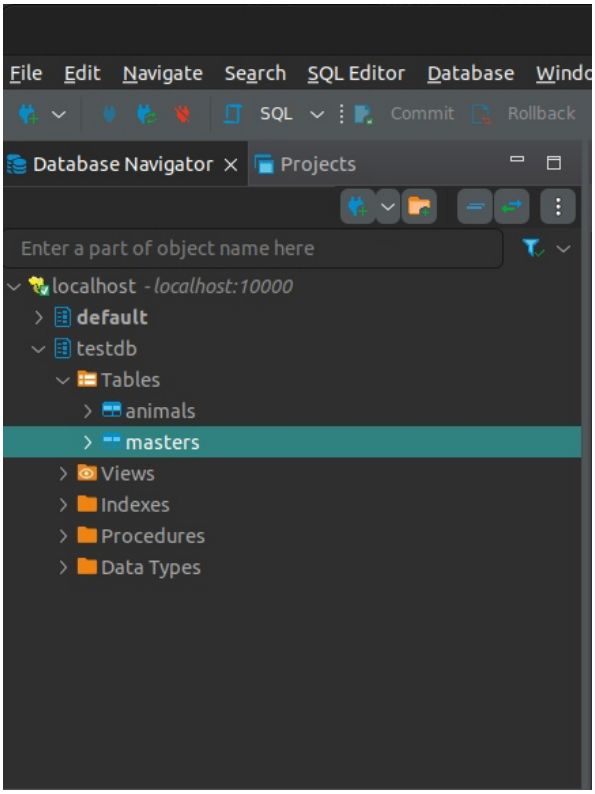

Рис. 7 – созданная база данных с двумя таблицами в DBeaver

```
iratewarrior@iratewarrior-DP156:~/bigdata/hive$ sudo docker cp animals.csv datanode:/tmp
Preparing to copy...
Copying to container - 2.048kB
Successfully copied 2.048kB to datanode:/tmp
iratewarrior@iratewarrior-DP156:~/bigdata/hive$ sudo docker cp masters.csv datanode:/tmp
Preparing to copy...
Copying to container - 2.048kB
Successfully copied 2.048kB to datanode:/tmp
iratewarrior@iratewarrior-DP156:~/bigdata/hive$ sudo docker exec -it datanode /bin/bash
root@060517e9b20d:/# cd tmp/
root@060517e9b20d:/tmp# ls
Jetty_localhost_38487_datanode____.6vsinf  animals.csv  hsperfdata_root  masters.csv
root@060517e9b20d:/tmp# hdfs dfs -copyFromLocal animals.csv /user/hive/warehouse/testdb.db/animals
root@060517e9b20d:/tmp# hdfs dfs -copyFromLocal masters.csv /user/hive/warehouse/testdb.db/masters
root@060517e9b20d:/tmp#
```

Рис. 8 – копирование csv файлов на datanode

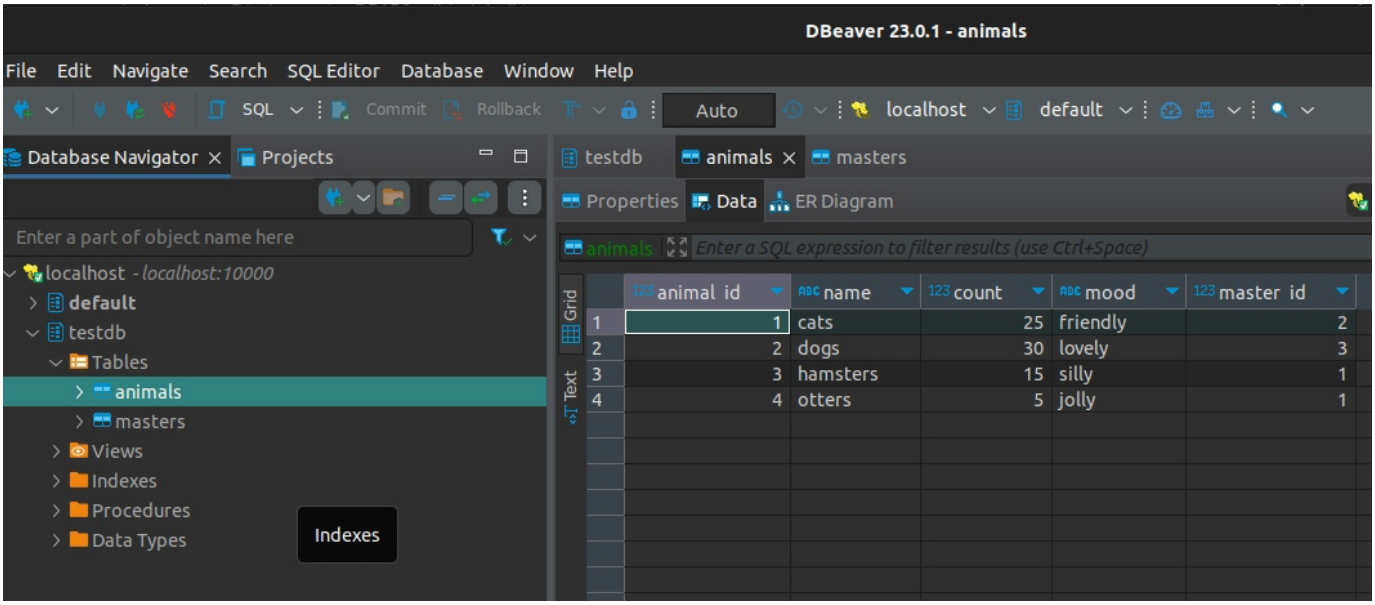

Рис. 9 – таблицы в базе данных заполнены данными из csv файлов

```
iratewarrior@iratewarrior-DP156:~/bigdata/hive$ sudo docker exec -it hive-server /bin/bash
root@6b464f28d1cd:/opt# hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hadoop-2.7.4/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in file:/opt/hive/conf/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive> use testdb
    > ;
OK
Time taken: 0.672 seconds
```

```
hive> select a.*, m.* from animals as a inner join masters as m on a.master_id = m.master_id;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = root_20230330133839_a2dcfad0-8c67-44b8-82e2-1f222bbf119b
Total jobs = 1
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/opt/hive/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/opt/hadoop-2.7.4/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Execution log at: /tmp/root/root_20230330133839_a2dcfad0-8c67-44b8-82e2-1f222bbf119b.log
2023-03-30 13:38:46     Starting to launch local task to process map join;      maximum memory = 477626368
2023-03-30 13:38:46     Dump the side-table for tag: 1 with group count: 3 into file: file:/tmp/root/ac3236f3-c1f9-46db-ad09-b527d95ede6e/hive_2023-03-30_13-38-39_952_1308499187162719818-1/-local-10004/HashTable-Stage-3/MapJoin-mapfile01--.hashtable
2023-03-30 13:38:46     Uploaded 1 File to: file:/tmp/root/ac3236f3-c1f9-46db-ad09-b527d95ede6e/hive_2023-03-30_13-38-39_952_1308499187162719818-1/-local-10004/HashTable-Stage-3/MapJoin-mapfile01--.hashtable (332 bytes)
2023-03-30 13:38:46     End of local task; Time Taken: 0.855 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Job running in-process (local Hadoop)
2023-03-30 13:38:49,327 Stage-3 map = 100%,  reduce = 0%
Ended Job = job_local1023704723_0001
MapReduce Jobs Launched:
Stage-Stage-3:  HDFS Read: 81 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
1       cats    25      friendly        2       2       Ivan    NULL
2       dogs    30      lovely  3       3       Adam    NULL
3       hamsters        15      silly   1       1       John    NULL
4       otters  5       jolly   1       1       John    NULL
Time taken: 9.386 seconds, Fetched: 4 row(s)
hive>
```

Рис. 10, 11 – объединение двух таблиц внутри hive-server

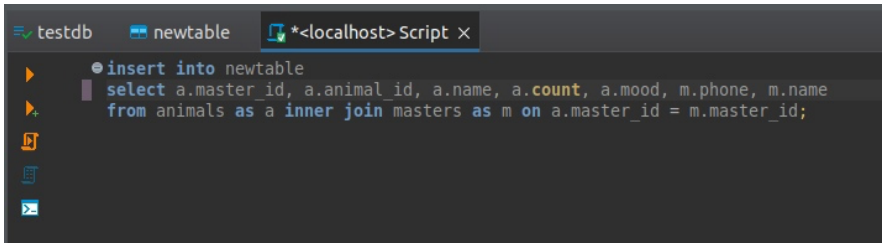

Рис. 12 – создание третьей объединяющей две предыдущие таблицы в DBeaver

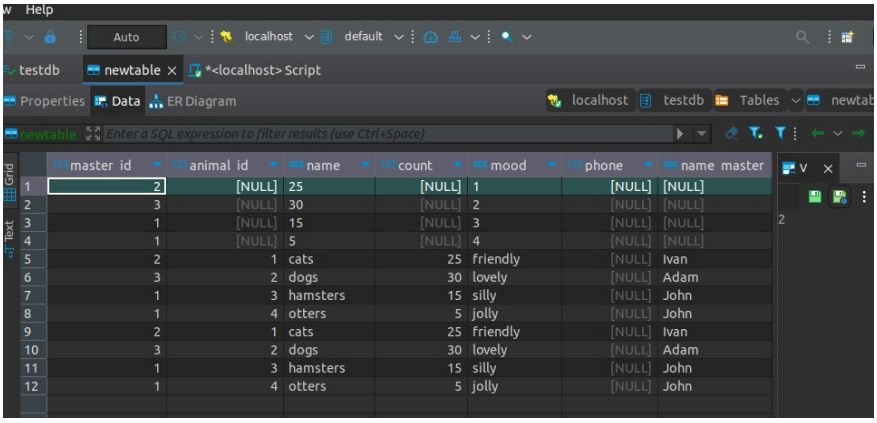

Рис. 13 – таблицы объединены

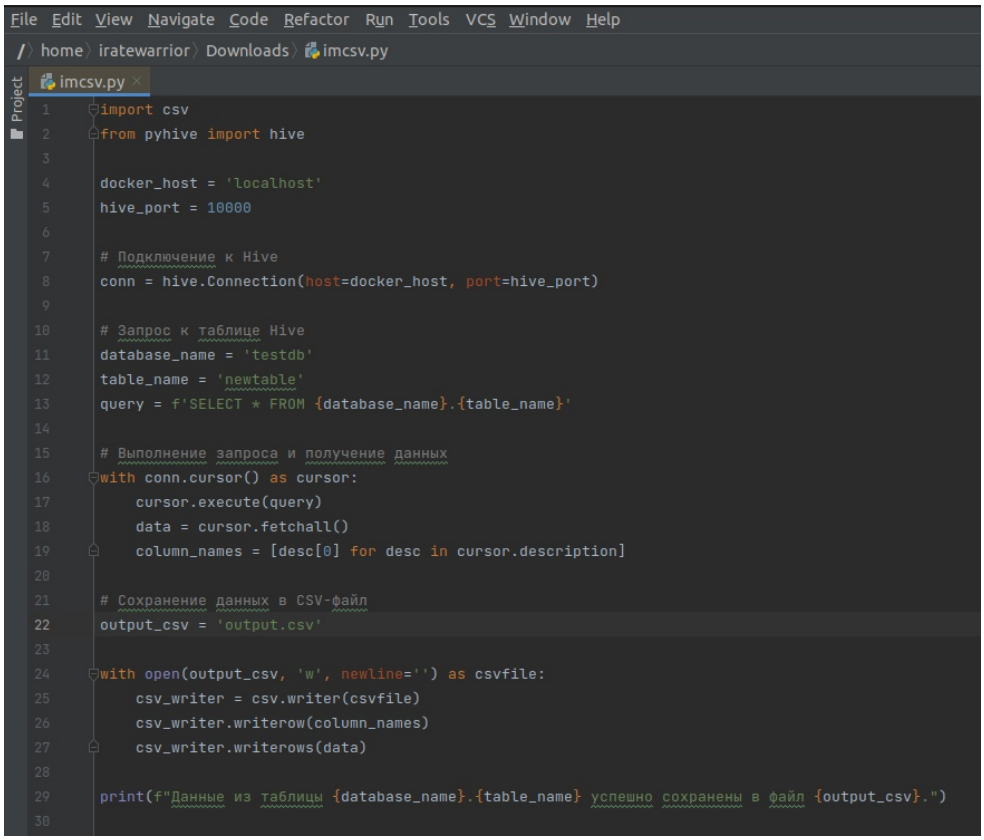

Рис. 14 – python скрипт для извлечения данных из newtable в csv файл (используются библиотеки csv и pyhive)

```
iratewarrior@iratewarrior-DP156:~/Downloads$ sudo python3 imcsv.py
Данные из таблицы testdb.newtable успешно сохранены в файл output.csv.
iratewarrior@iratewarrior-DP156:~/Downloads$
```

Рис. 15 – выполнение скрипта успешно

Open | output.csv [Read-Only] ~/Downloads | Save

```
1 newtable.master_id,newtable.animal_id,newtable.name,newtable.count,newtable.mood,newtable.phone,newtable.name_master
2 2,,25,,1,,
3 3,,30,,2,,
4 1,,15,,3,,
5 1,,5,,4,,
6 2,1,cats,25,friendly,,Ivan
7 3,2,dogs,30,lovely,,Adam
8 1,3,hamsters,15,silly,,John
9 1,4,otters,5,jolly,,John
10 2,1,cats,25,friendly,,Ivan
11 3,2,dogs,30,lovely,,Adam
12 1,3,hamsters,15,silly,,John
13 1,4,otters,5,jolly,,John
```

Рис. 16 – данные из таблицы перенесены в output.csv