Результат работы.

1. Копируем информацию из репозитория с данными в заранее созданную папку с помощью Git Bash:

**$ git clone** [**git@github.com:sultanmurad/csv\_files.git**](mailto:git@github.com:sultanmurad/csv_files.git)

1. Предобработка файлов выполнена средствами python + pandas(4\_5.ipynb).

Предобработанные файлы размещены в папке csv\_files\_with\_groups

1. Развертывание контейнера по аналогии с заданием 4\_3:

Копируем информацию из репозитория с образом Hadoop в заранее созданную папку с помощью Git Bash:

**$ git clone** [**git@github.com:tech4242/docker-hadoop-hive-parquet.git**](mailto:git@github.com:tech4242/docker-hadoop-hive-parquet.git)

Для развертывания контейнера необходимо в командной строке войти в скачанную папку docker-hadoop-hive-parquet и выполнить команду:

**docker-compose up**

1. После подняния контейнера попадаем в HUE по адресу <http://localhost:8888/hue>

Для авторизации использовался логин *shiva*

Личная папка в файловой системе Hadoop(hdfs) /user/shiva

1. Подключаемся к контейнеру “datanode-1”, создаем внутри папку “data” и переносим в нее скачанные файлы

C:\Projects\4\_3\docker-hadoop-hive-parquet>docker ps

CONTAINER ID IMAGE COMMAND CREATED STATUS PORTS NAMES

4d4d447ee441 gethue/hue:4.6.0 "./startup.sh" 12 minutes ago Up 12 minutes 0.0.0.0:8888->8888/tcp docker-hadoop-hive-parquet-hue-1

38b960bff8e4 bde2020/hadoop-namenode:2.0.0-hadoop2.7.4-java8 "/entrypoint.sh /run…" 12 minutes ago Up 12 minutes (healthy) 0.0.0.0:50070->50070/tcp docker-hadoop-hive-parquet-namenode-1

**07fb66487bb2** bde2020/hadoop-datanode:2.0.0-hadoop2.7.4-java8 "/entrypoint.sh /run…" 12 minutes ago Up 12 minutes (healthy) 0.0.0.0:50075->50075/tcp docker-hadoop-hive-parquet-datanode-1

c356ac0e000c bde2020/hive-metastore-postgresql:2.3.0 "/docker-entrypoint.…" 12 minutes ago Up 12 minutes 0.0.0.0:5432->5432/tcp docker-hadoop-hive-parquet-hive-metastore-postgresql-1

7420819cf352 postgres:12.1-alpine "docker-entrypoint.s…" 12 minutes ago Up 12 minutes 0.0.0.0:56579->5432/tcp docker-hadoop-hive-parquet-huedb-1

cb71a1c56259 bde2020/hive:2.3.2-postgresql-metastore "entrypoint.sh /bin/…" 12 minutes ago Up 12 minutes 0.0.0.0:10000->10000/tcp, 10002/tcp docker-hadoop-hive-parquet-hive-server-1

b97f5897e8af bde2020/hive:2.3.2-postgresql-metastore "entrypoint.sh /opt/…" 12 minutes ago Up 12 minutes 10000/tcp, 0.0.0.0:9083->9083/tcp, 10002/tcp docker-hadoop-hive-parquet-hive-metastore-1

7afc6d4d3a18 bde2020/hadoop-resourcemanager:2.0.0-hadoop2.7.4-java8 "/entrypoint.sh /run…" 12 minutes ago Up 12 minutes (healthy) 8088/tcp docker-hadoop-hive-parquet-resourcemanager-1

C:\Projects\4\_3\docker-hadoop-hive-parquet>docker exec -it 07fb66487bb2 bash

root@07fb66487bb2:/# ls

bin dev etc hadoop-data lib media opt root run.sh srv tmp var

boot entrypoint.sh hadoop home lib64 mnt proc run sbin sys usr

root@07fb66487bb2:/# cd usr

root@07fb66487bb2:/usr# mkdir data

root@07fb66487bb2:/usr# ls

bin **data** games include lib local sbin share src

C:\Projects\4\_3\docker-hadoop-hive-parquet>docker cp C:\Projects\4\_5\csv\_files\_with\_groups\customers\_with\_groups.csv 07fb66487bb2:/usr/data/

Successfully copied 17.9MB to 07fb66487bb2:/usr/data/

C:\Projects\4\_3\docker-hadoop-hive-parquet>docker cp C:\Projects\4\_5\csv\_files\_with\_groups\organizations\_with\_groups.csv 07fb66487bb2:/usr/data/

Successfully copied 14.3MB to 07fb66487bb2:/usr/data/

C:\Projects\4\_3\docker-hadoop-hive-parquet>docker cp C:\Projects\4\_5\csv\_files\_with\_groups\people\_with\_groups.csv 07fb66487bb2:/usr/data/

Successfully copied 11.8MB to 07fb66487bb2:/usr/data/

C:\Projects\4\_3\docker-hadoop-hive-parquet>docker exec -it 07fb66487bb2 bash

root@07fb66487bb2:/# ls

bin dev etc hadoop-data lib media opt root run.sh srv tmp var

boot entrypoint.sh hadoop home lib64 mnt proc run sbin sys usr

root@07fb66487bb2:/# cd usr

root@07fb66487bb2:/usr# cd data

root@07fb66487bb2:/usr/data# ls

customers\_with\_groups.csv organizations\_with\_groups.csv people\_with\_groups.csv

1. Загружаем файлы на hdfs в личную папку

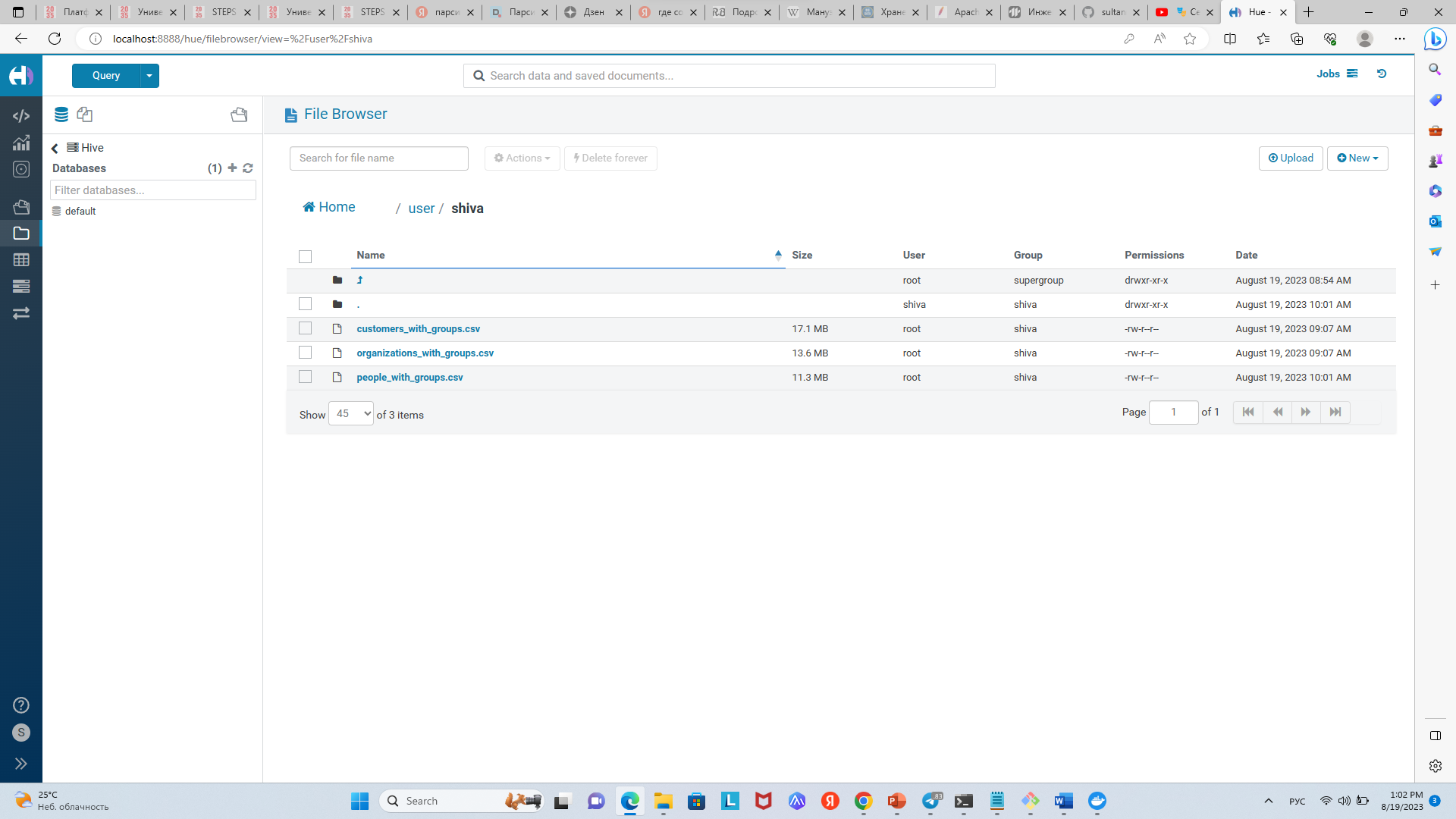
root@07fb66487bb2:/usr/data# hdfs dfs -copyFromLocal customers\_with\_groups.csv /user/shiva

root@07fb66487bb2:/usr/data# hdfs dfs -copyFromLocal organizations\_with\_groups.csv /user/shiva

root@07fb66487bb2:/usr/data# hdfs dfs -copyFromLocal people\_with\_groups.csv /user/shiva

root@07fb66487bb2:/usr/data#

Отображение файлов в HUE



1. Команды для создания отдельной таблицы под каждый csv -файл, а также сводная статистика на их основе выполнялись в HUE в разделе “Query” и представлены в файле “Запросы в Hive.txt”