Федеральное агентство связи
Ордена трудового Красного Знамени
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский технический университет связи и информатики»

Практическая работа № 3

По дисциплине
Введение в большие данные

Группа: МБД2431
ФИО: Киреев Артём Александрович

Москва, 2025

Цель работы: получить навыки работы с Apache Hive

1. Запуск оболочки Hive

```
hive
```

```
SHOW DATABASES;
```

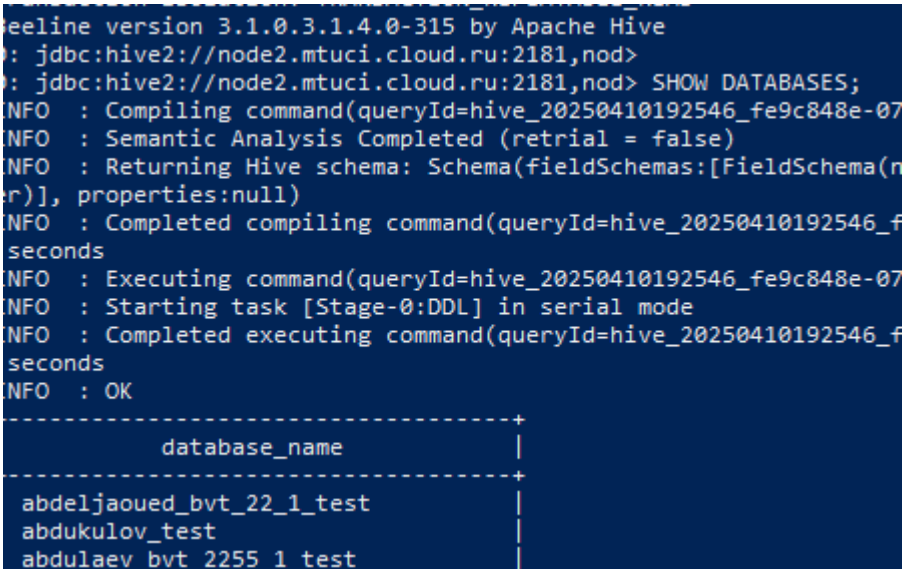


Рисунок 1 - Запуск оболочки Hive

2. Создание базы данных

```
CREATE DATABASE IF NOT EXISTS kireev_mbd_2431_test LOCATION
'/data/kireev_mbd_2431_test';
```

```
USE kireev_mbd_2431_test;
```

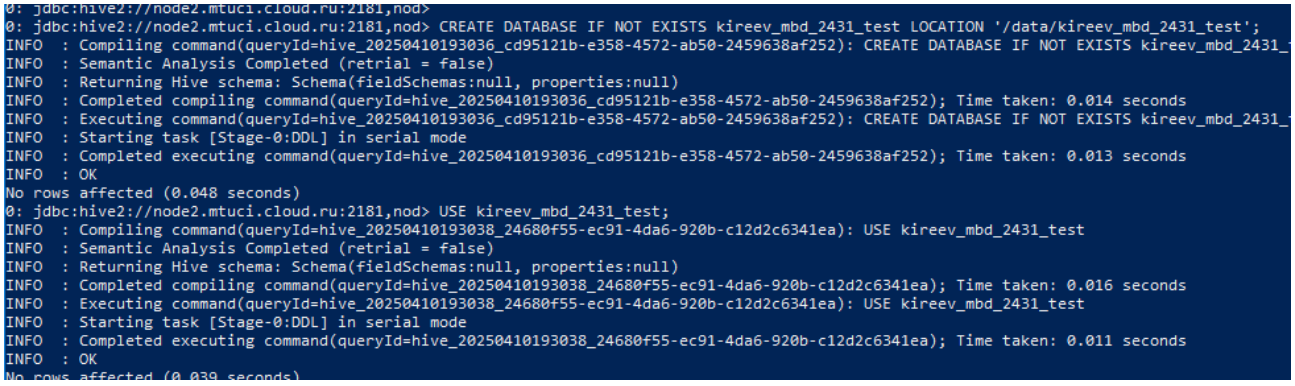


Рисунок 2 - Создание тестовой БД

```
DROP DATABASE IF EXISTS kireev_mbd_2431_test CASCADE;
```

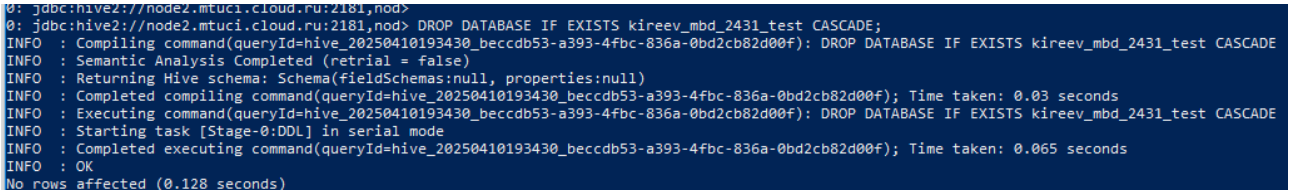


Рисунок 3 - Удаление БД

```
CREATE DATABASE kireev_mbd_2431_test;

USE kireev_mbd_2431_test;

DESCRIBE DATABASE kireev_mbd_2431_test;
```

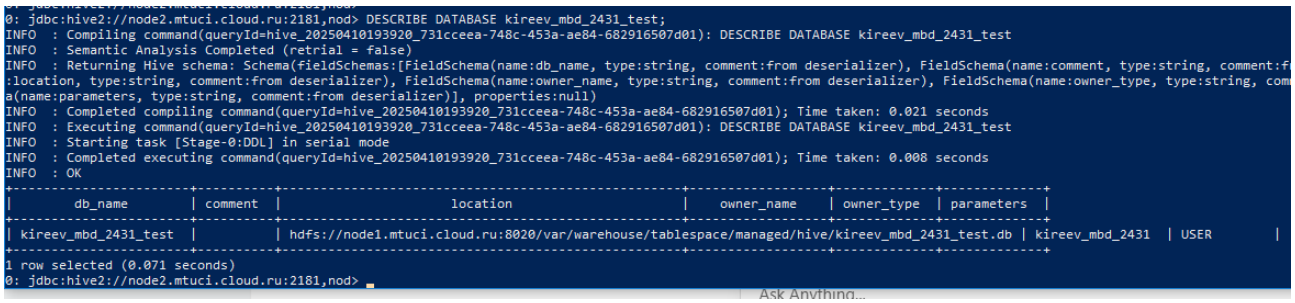


Рисунок 4 - Вывод информации о БД

3. Создание таблиц

Создадим таблицу в тестовой базе

```
DROP TABLE IF EXISTS Subnets;

CREATE EXTERNAL TABLE Subnets ( ip STRING, mask STRING )

ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'

STORED AS TEXTFILE

LOCATION '/data/subnets/variant1';
```

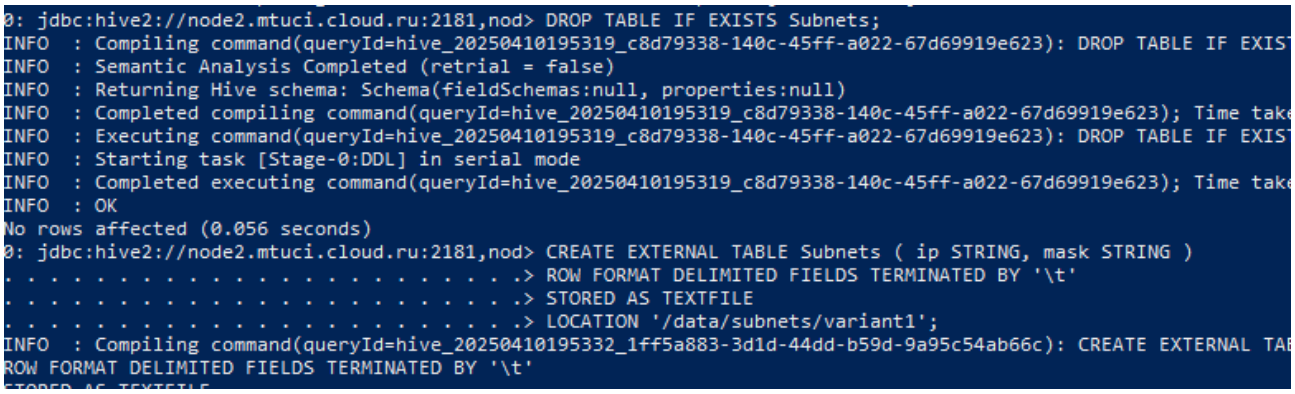


Рисунок 5 - Создание таблицы

Проверка таблицы

```
SELECT * FROM Subnets LIMIT 10;
```

```
INFO  : Completed executing command(queryId=hive_20250410195341_d
INFO  : OK
+------------------+------------------+
|    subnets.ip    |   subnets.mask   |
+------------------+------------------+
| 148.45.113.216   | 255.255.255.248  |
| 203.98.141.0     | 255.255.255.240  |
| 183.168.36.0     | 255.255.255.128  |
| 111.157.172.232  | 255.255.255.248  |
| 80.46.87.0       | 255.255.255.0    |
| 247.248.233.0    | 255.255.255.128  |
| 58.75.180.168    | 255.255.255.248  |
| 175.179.146.0    | 255.255.255.128  |
| 45.28.7.128      | 255.255.255.128  |
| 82.188.115.160   | 255.255.255.224  |
+------------------+------------------+
10 rows selected (0.196 seconds)
0: jdbc:hive2://node2.mtuci.cloud.ru:2181,nod>
```

Рисунок 6 - Проверка таблицы

```
SHOW TABLES;
```

```
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 3.1.0.3.1.4.0-315 by Apache Hive
0: jdbc:hive2://node2.mtuci.cloud.ru:2181,nod> SHOW TABLES;
INFO  : Compiling command(queryId=hive_20250410195744_4f07676
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSche
INFO  : Completed compiling command(queryId=hive_202504101957
INFO  : Executing command(queryId=hive_20250410195744_4f07676
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_202504101957
INFO  : OK
+-------------------------+
|        tab_name         |
+-------------------------+
| applestore              |
| students                |
| subneets                |
| subnetd                 |
```

Рисунок 7 - Проверка таблицы

4. Партиционирование

Создадим партиционированную таблицу

```
SET hive.exec.dynamic.partition.mode=nonstrict;

CREATE EXTERNAL TABLE IF NOT EXISTS SubnetsPart_new ( ip STRING )

PARTITIONED BY (mask STRING)

STORED AS TEXTFILE;
```

```
0: jdbc:hive2://node2.mtuci.cloud.ru:2181,nod> CREATE EXTERNAL TABLE IF NOT EXISTS SubnetsPart_new ( ip STRING )
. . . . . .> PARTITIONED BY (mask STRING)
. . . . . .> STORED AS TEXTFILE;
INFO  : Compiling command(queryId=hive_20250410200643_ae60cb8e-1110-42de-9ddf-586f21f1e7d1): CREATE EXTERNAL TABLE IF
PARTITIONED BY (mask STRING)
STORED AS TEXTFILE
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
```

Рисунок 8 - Создание партиционированой таблицы

Вставим данные из исходной таблицы Subnets в новую партиционированную таблицу SubnetsPart_new

```
INSERT OVERWRITE TABLE SubnetsPart_new PARTITION (mask)

SELECT ip, mask FROM Subnets;
```

```
0: jdbc:hive2://node2.mtuci.cloud.ru:2181,nod>
0: jdbc:hive2://node2.mtuci.cloud.ru:2181,nod> INSERT OVERWRITE TABLE SubnetsPart_new PARTITION (mask)
. . . . . .> SELECT ip, mask FROM Subnets;
INFO  : Compiling command(queryId=hive_20250410200658_61835fdf-5bf4-4057-9801-1c8ffbcf1578): INSERT OVER
SELECT ip, mask FROM Subnets
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:ip, type:string, comment:null), Fie
INFO  : Completed compiling command(queryId=hive_20250410200658_61835fdf-5bf4-4057-9801-1c8ffbcf1578); T
INFO  : Executing command(queryId=hive_20250410200658_61835fdf-5bf4-4057-9801-1c8ffbcf1578): INSERT OVER
SELECT ip, mask FROM Subnets
INFO  : Query ID = hive_20250410200658_61835fdf-5bf4-4057-9801-1c8ffbcf1578
INFO  : Total jobs = 1
INFO  : Launching Job 1 out of 1
INFO  : Starting task [Stage-1:MAPRED] in serial mode
INFO  : Subscribed to counters: [] for queryId: hive_20250410200658_61835fdf-5bf4-4057-9801-1c8ffbcf1578
INFO  : Tez session hasn't been created yet. Opening session
INFO  : Dag name: INSERT OVERWRITE TABLE SubnetsPart...Subnets (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1743703651672_0374)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      2          2        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 4.62 s
----------------------------------------------------------------------------------------------
```

Рисунок 9 - Вставка данных

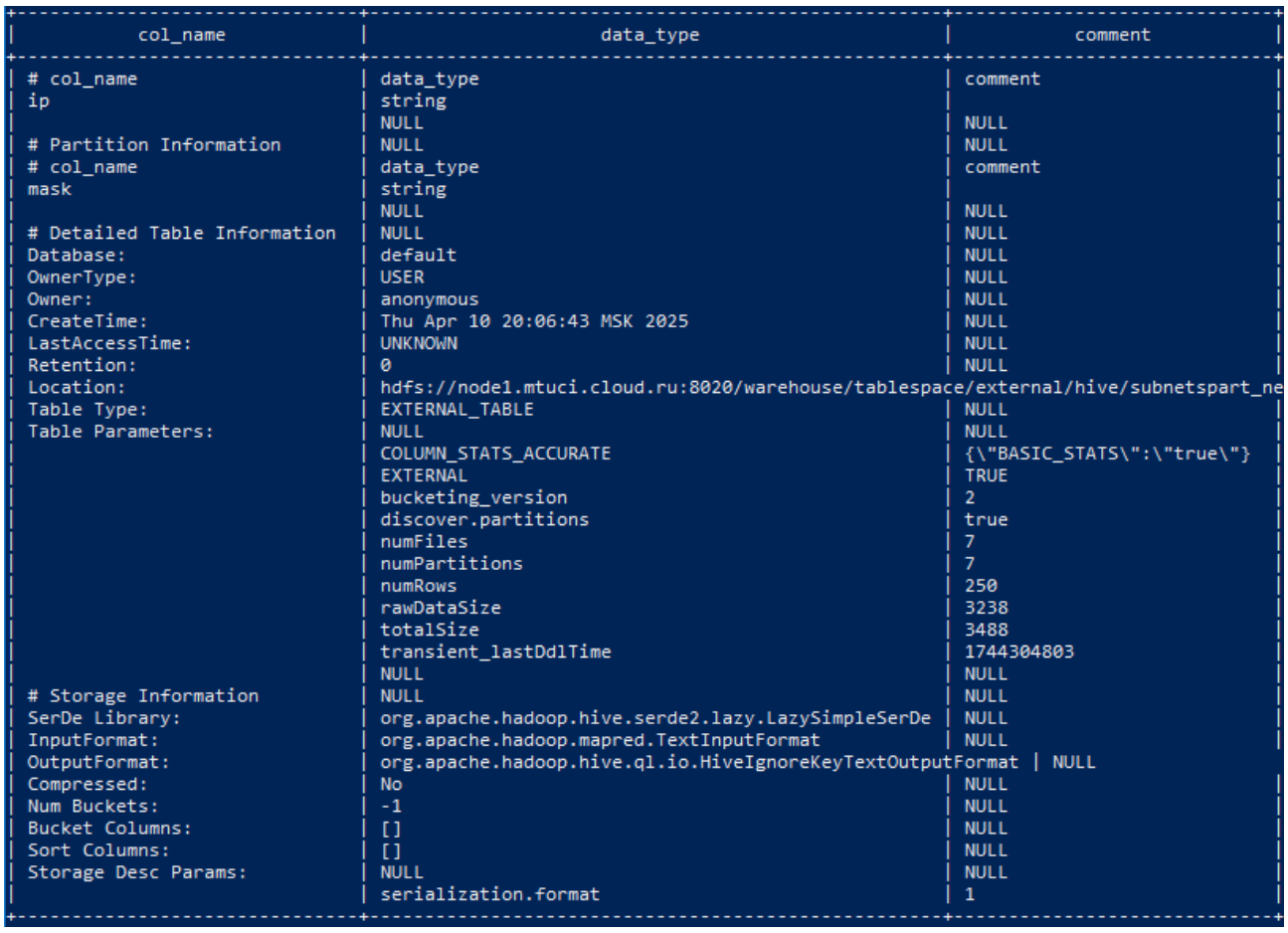
```
SHOW PARTITIONS SubnetsPart_new;
```

```
0: jdbc:hive2://node2.mtuci.cloud.ru:2181,nod> SHOW PARTITIONS SubnetsPart_new;
INFO  : Compiling command(queryId=hive_20250410200719_94f1922e-9f21-47f8-9526-bfb6ba6
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:partition, type:
INFO  : Completed compiling command(queryId=hive_20250410200719_94f1922e-9f21-47f8-95
INFO  : Executing command(queryId=hive_20250410200719_94f1922e-9f21-47f8-9526-bfb6ba6
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20250410200719_94f1922e-9f21-47f8-95
INFO  : OK
+-----------------------+
|       partition       |
+-----------------------+
| mask=255.255.255.0    |
| mask=255.255.255.128  |
| mask=255.255.255.192  |
| mask=255.255.255.224  |
| mask=255.255.255.240  |
| mask=255.255.255.248  |
| mask=255.255.255.252  |
+-----------------------+
```

Рисунок 10 - Вывод информации о партиции

Выведем полную информацию о таблице

```
DESCRIBE FORMATTED SubnetsPart_new;
```



col_name	data_type	comment
# col_name	data_type	comment
ip	string	
	NULL	NULL
# Partition Information	NULL	NULL
# col_name	data_type	comment
mask	string	
	NULL	NULL
# Detailed Table Information	NULL	NULL
Database:	default	NULL
OwnerType:	USER	NULL
Owner:	anonymous	NULL
CreateTime:	Thu Apr 10 20:06:43 MSK 2025	NULL
LastAccessTime:	UNKNOWN	NULL
Retention:	0	NULL
Location:	hdfs://node1.mtuci.cloud.ru:8020/warehouse/tablespace/external/hive/subnetspart_ne	
Table Type:	EXTERNAL_TABLE	NULL
Table Parameters:	NULL	NULL
	COLUMN_STATS_ACCURATE	{\"BASIC_STATS\":\"true\"}
	EXTERNAL	TRUE
	bucketing_version	2
	discover.partitions	true
	numFiles	7
	numPartitions	7
	numRows	250
	rawDataSize	3238
	totalSize	3488
	transient_lastDdlTime	1744304803
	NULL	NULL
# Storage Information	NULL	NULL
SerDe Library:	org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe	NULL
InputFormat:	org.apache.hadoop.mapred.TextInputFormat	NULL
OutputFormat:	org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat	NULL
Compressed:	No	NULL
Num Buckets:	-1	NULL
Bucket Columns:	[]	NULL
Sort Columns:	[]	NULL
Storage Desc Params:	NULL	NULL
	serialization.format	1

Рисунок 11 - Вывод информации о таблице

5. Парсинг входных данных с помощью регулярных выражений

```
DROP TABLE IF EXISTS SerDeExample;

CREATE EXTERNAL TABLE SerDeExample (
    ip STRING,
    log_date STRING,
    request STRING,
    responsecode STRING
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'
WITH SERDEPROPERTIES (
    "input.regex" = '^(\\S*)\\t(\\S*)\\t(\\S*)\\t(\\S*)\\t.*$'
)
STORED AS TEXTFILE
LOCATION '/data/user_logs/user_logs_M';
```

```
0: jdbc:hive2://node2.mtuci.cloud.ru:2181,nod> CREATE EXTERNAL TABLE SerDeExample (
. . . . . .>     ip STRING,
. . . . . .>     log_date STRING,
. . . . . .>     request STRING,
. . . . . .>     responsecode STRING
. . . . . .> )
. . . . . .> ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'
. . . . . .> WITH SERDEPROPERTIES (
. . . . . .>     "input.regex" = '^(\\S*)\\t(\\S*)\\t(\\S*)\\t(\\S*)\\t.*$'
. . . . . .> )
. . . . . .> STORED AS TEXTFILE
. . . . . .> LOCATION '/data/user_logs/user_logs_M';
INFO  : Compiling command(queryId=hive_20250410202920_09a65aab-60c2-408e-b313-60934e26806c): CREATE EXTERNAL TABL
ip STRING,
log_date STRING,
request STRING,
responsecode STRING
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'
```

Рисунок 12 - Создание таблицы SerDeExample

```
SELECT * FROM SerDeExample LIMIT 10;

DESCRIBE SerDeExample;
```

```
0: jdbc:hive2://node2.mtuci.cloud.ru:2181,nod> SELECT * FROM SerDeExample LIMIT 10;
INFO  : Compiling command(queryId=hive_20250410202929_b59255cb-d245-40b7-87c9-d89a062e3210): SELE
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:serdeexample.ip, type:string
deexample.request, type:string, comment:null), FieldSchema(name:serdeexample.responsecode, type:s
INFO  : Completed compiling command(queryId=hive_20250410202929_b59255cb-d245-40b7-87c9-d89a062e3
INFO  : Executing command(queryId=hive_20250410202929_b59255cb-d245-40b7-87c9-d89a062e3210): SELE
INFO  : Completed executing command(queryId=hive_20250410202929_b59255cb-d245-40b7-87c9-d89a062e3
INFO  : OK
+------------------+-------------------------+-----------------------+-----------------------------
| serdeexample.ip  | serdeexample.log_date   | serdeexample.request  | serdeexample.responsecode
+------------------+-------------------------+-----------------------+-----------------------------
| 33.49.147.163    |                         |                       | 20140101014611
| 197.72.248.141   |                         |                       | 20140101020306
| 33.49.147.163    |                         |                       | 20140101023103
| 75.208.40.166    |                         |                       | 20140101032909
| 197.72.248.141   |                         |                       | 20140101033626
| 222.131.187.37   |                         |                       | 20140101033837
| 197.72.248.141   |                         |                       | 20140101034726
| 33.49.147.163    |                         |                       | 20140101041149
| 197.72.248.141   |                         |                       | 20140101050543
| 181.217.177.35   |                         |                       | 20140101052930
+------------------+-------------------------+-----------------------+-----------------------------
```

```
0: jdbc:hive2://node2.mtuci.cloud.ru:2181,nod> DESCRIBE SerDeExample
INFO  : Compiling command(queryId=hive_20250410202934_e778da9f-bb2f-
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name
ame:comment, type:string, comment:from deserializer)], properties:nu
INFO  : Completed compiling command(queryId=hive_20250410202934_e778
INFO  : Executing command(queryId=hive_20250410202934_e778da9f-bb2f-
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20250410202934_e778
INFO  : OK
+---------------+------------+----------+
|   col_name    | data_type  | comment  |
+---------------+------------+----------+
| ip            | string     |          |
| log_date      | string     |          |
| request       | string     |          |
| responsecode  | string     |          |
+---------------+------------+----------+
4 rows selected (0.066 seconds)
```

Рисунок 13 - Проверка результатов

6. Практика

Задача 0. Посчитать количество различных масок подсети

```
SELECT COUNT(DISTINCT mask) FROM Subnets;
```

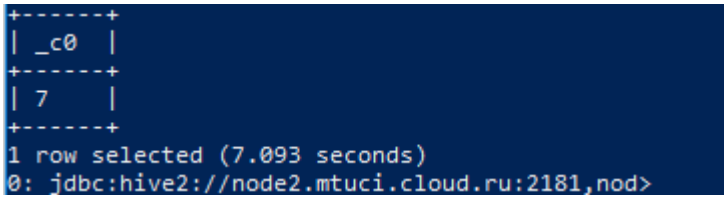


Рисунок 14 - Количество масок подсети

Задача 1. Посчитать количество адресов, имеющих маску 255.255.255.128

На таблице без партиций

```
SELECT COUNT(*) FROM Subnets WHERE mask = '255.255.255.128';
```

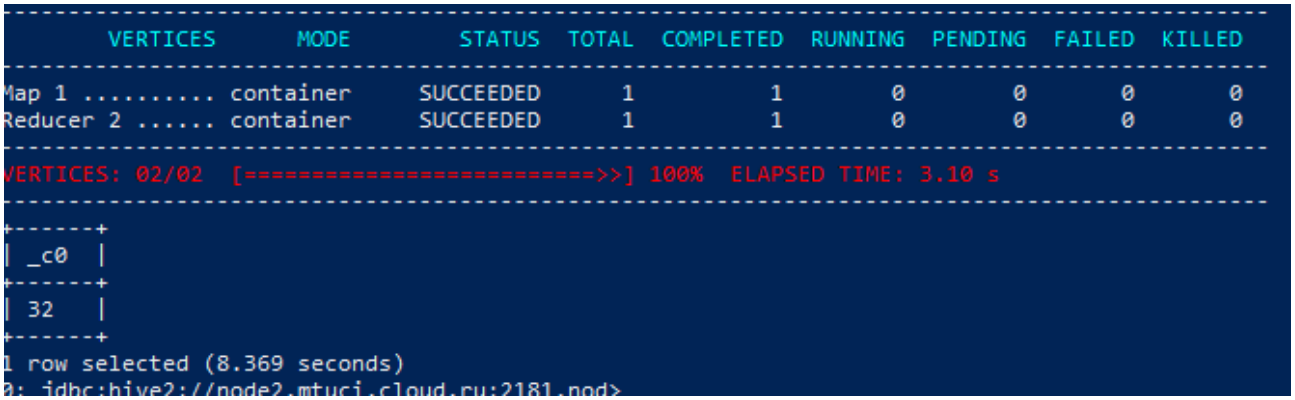


Рисунок 15 - Количество адресов

На партиционированной таблице

```
SELECT COUNT(*) FROM SubnetsPart WHERE mask = '255.255.255.128';
```

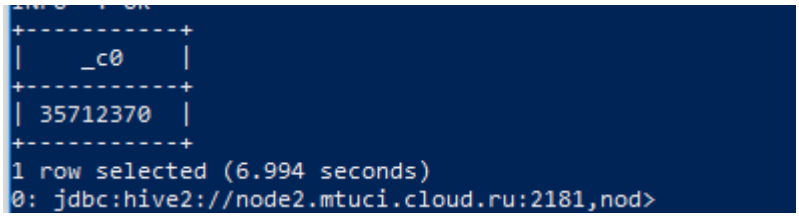


Рисунок 16 - Количество адресов

Задача 2. Посчитать среднее количество адресов по подсетям

```
SELECT AVG(count) FROM (
    SELECT mask, COUNT(*) as count FROM Subnets GROUP BY mask
) t;
```

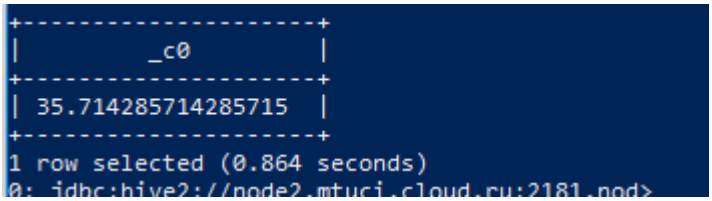


Рисунок 17 - Среднее количество адресов

Проверка размера данных

```
hdfs dfs -du -h /data/subnets/variant1
```

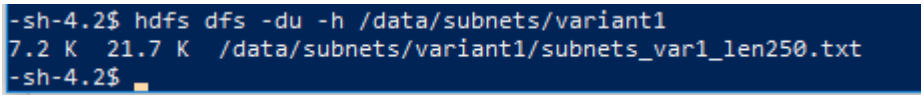


Рисунок 18 - Вывод размера

Пересоздание таблиц на большом датасете

Создание таблицы Subnets на большом датасете

```
DROP TABLE IF EXISTS Subnets;

CREATE EXTERNAL TABLE Subnets (

    ip STRING,

    mask STRING

)

ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'

STORED AS TEXTFILE

LOCATION '/data/subnets/big';
```

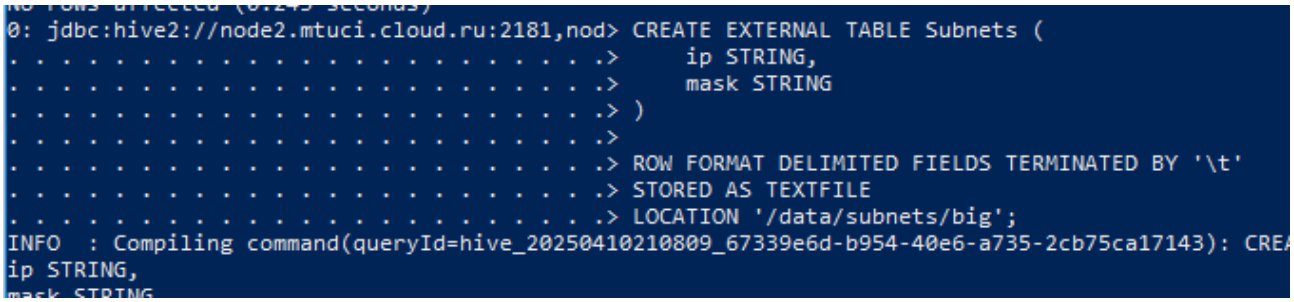


Рисунок 19 - Создание на большом датасете

Создание партиционированной таблицы SubnetsPart на большом датасете

```
DROP TABLE IF EXISTS SubnetsPart;

CREATE EXTERNAL TABLE SubnetsPart (

    ip STRING

)

PARTITIONED BY (mask STRING)

STORED AS TEXTFILE;
```

```
INSERT OVERWRITE TABLE SubnetsPart PARTITION (mask)

SELECT ip, mask FROM Subnets;
```

```
0: jdbc:hive2://node2.mtuci.cloud.ru:2181,nod> CREATE EXTERNAL TABLE SubnetsPart (
. . . . . .> ip STRING
. . . . . .> )
. . . . . .> PARTITIONED BY (mask STRING)
. . . . . .> STORED AS TEXTFILE;
INFO  : Compiling command(queryId=hive_20250410210921_8d4ed9fc-0a33-49ec-b5de-e30580f776f9): CREATE EX
ip STRING
)
PARTITIONED BY (mask STRING)
STORED AS TEXTFILE
```

Рисунок 20 - Создание партиционированной датасета

На таблице без партиций

```
SELECT COUNT(*) FROM Subnets WHERE mask = '255.255.255.128';
```

```
+-----------+
|    _c0    |
+-----------+
| 35712370  |
+-----------+
1 row selected (8.284 seconds)
0: jdbc:hive2://node2.mtuci.cloud.ru:2181,nod>
```

Рисунок 21 - Реультат без партиций

На партиционированной таблице

```
SELECT COUNT(*) FROM SubnetsPart WHERE mask = '255.255.255.128';
```

```
+-----------+
|    _c0    |
+-----------+
| 35712370  |
+-----------+
1 row selected (6.748 seconds)
0: jdbc:hive2://node2.mtuci.cloud.ru:2181,nod>
```

Рисунок 22 - Реультат с партициями