**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;

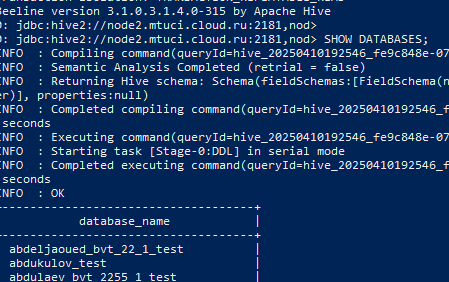


Рисунок 1 - Запуск оболочки Hive

1. Создание базы данных

CREATE DATABASE IF NOT EXISTS kireev\_mbd\_2431\_test LOCATION '/data/kireev\_mbd\_2431\_test';

USE kireev\_mbd\_2431\_test;

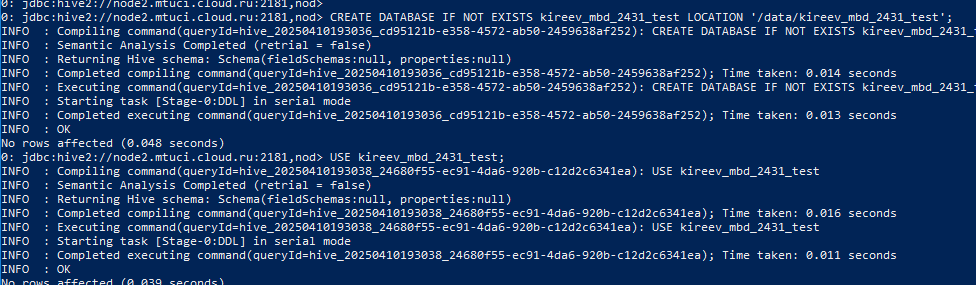


Рисунок 2 - Создание тестовой БД

DROP DATABASE IF EXISTS kireev\_mbd\_2431\_test CASCADE;

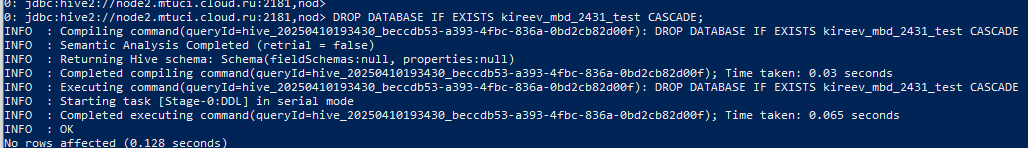


Рисунок 3 - Удаление БД

CREATE DATABASE kireev\_mbd\_2431\_test;

USE kireev\_mbd\_2431\_test;

DESCRIBE DATABASE kireev\_mbd\_2431\_test;

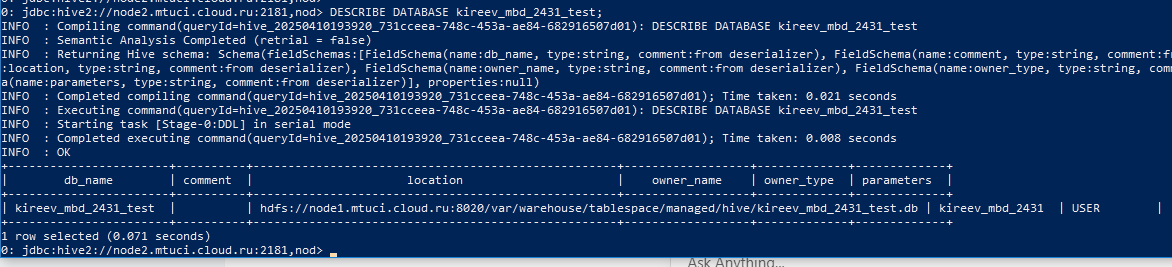


Рисунок 4 - Вывод информации о БД

1. 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';

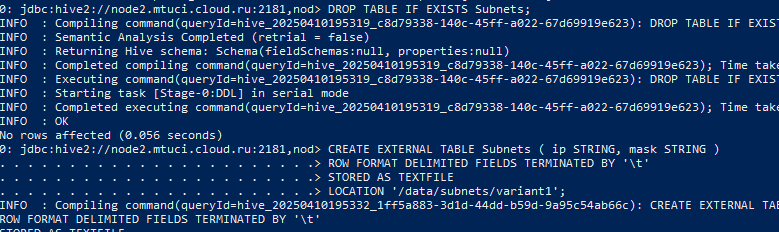


Рисунок 5 - Создание таблицы

Проверка таблицы

SELECT \* FROM Subnets LIMIT 10;

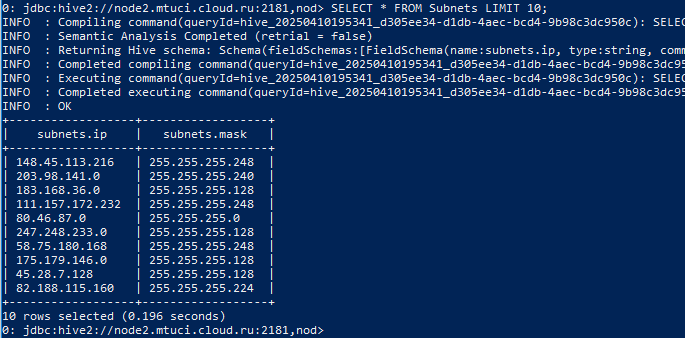


Рисунок 6 - Проверка таблицы

SHOW TABLES;

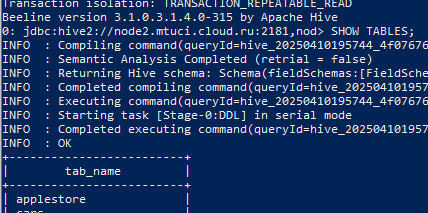




Рисунок 7 - Проверка таблицы

1. 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;

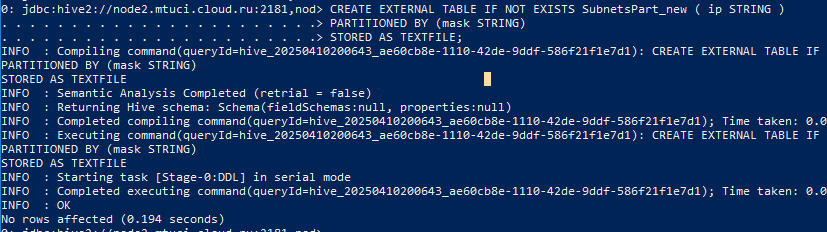


Рисунок 8 - Создание партиционированой таблицы

Вставим данные из исходной таблицы Subnets в новую партиционированную таблицу SubnetsPart\_new

INSERT OVERWRITE TABLE SubnetsPart\_new PARTITION (mask)

SELECT ip, mask FROM Subnets;

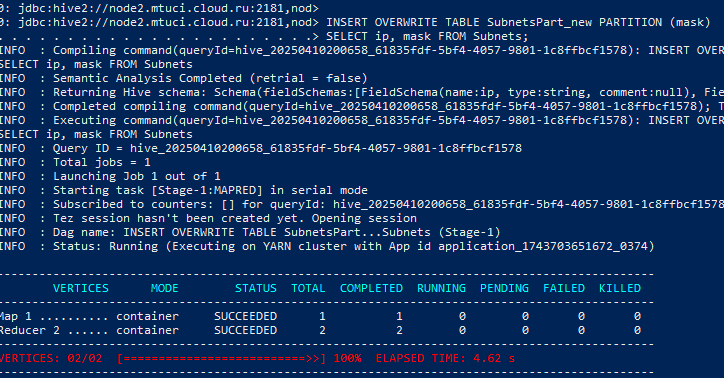


Рисунок 9 - Вставка данных

SHOW PARTITIONS SubnetsPart\_new;

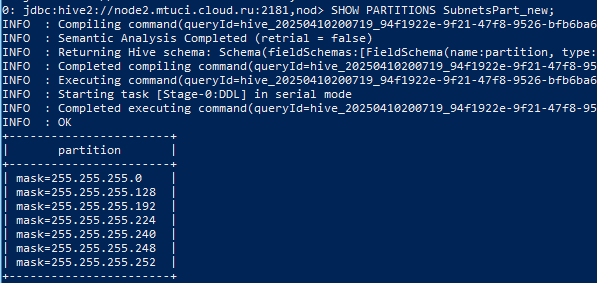


Рисунок 10 - Вывод информации о партиции

Выведем полную информацию о таблице

DESCRIBE FORMATTED SubnetsPart\_new;

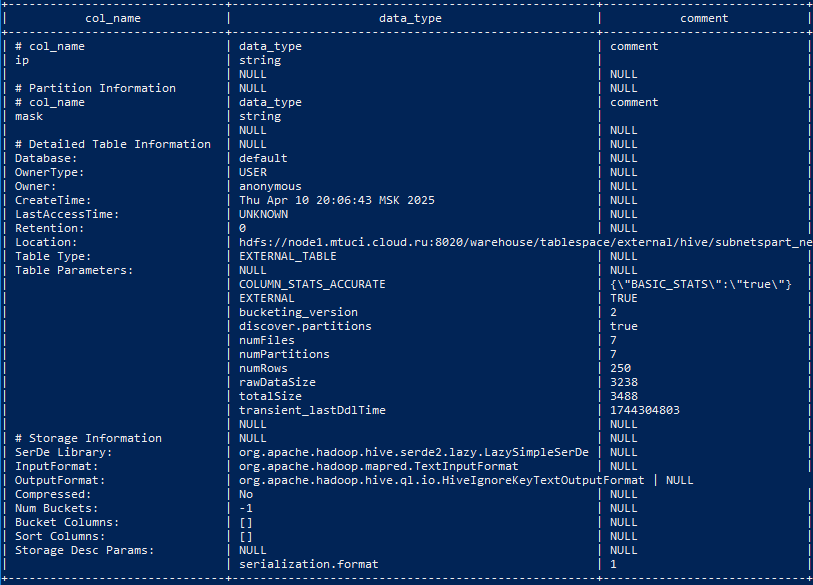


Рисунок 11 - Вывод информации о таблице

1. 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';

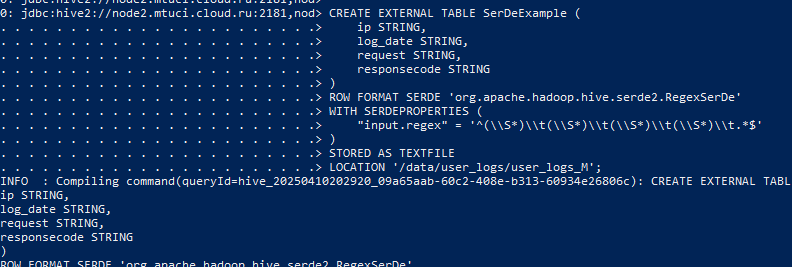
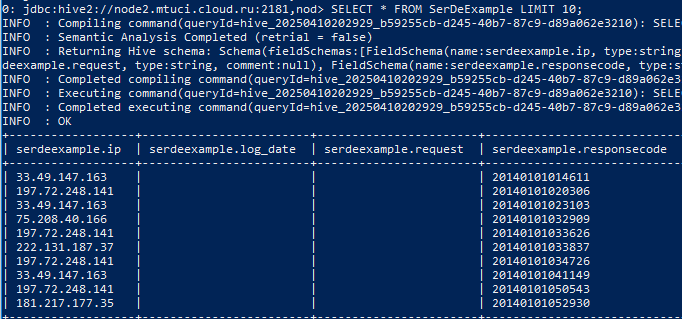


Рисунок 12 - Создание таблицы SerDeExample

SELECT \* FROM SerDeExample LIMIT 10;

DESCRIBE SerDeExample;



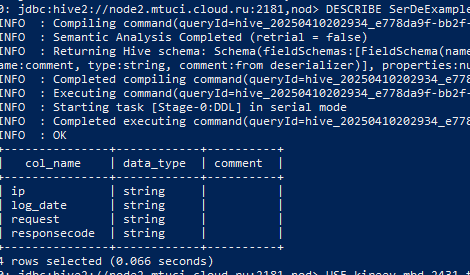


Рисунок 13 - Проверка результатов

1. Практика

Задача 0. Посчитать количество различных масок подсети

SELECT COUNT(DISTINCT mask) FROM Subnets;

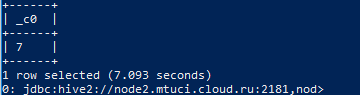


Рисунок 14 - Количество масок подсети

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';

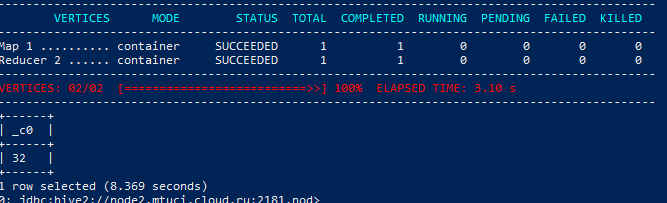


Рисунок 15 - Количество адресов

На партиционированной таблице

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

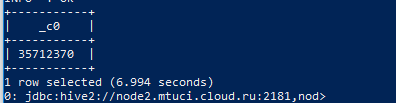


Рисунок 16 - Количество адресов

Задача 2. Посчитать среднее количество адресов по подсетям

SELECT AVG(count) FROM (

SELECT mask, COUNT(\*) as count FROM Subnets GROUP BY mask

) t;

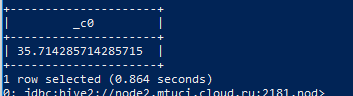


Рисунок 17 - Среднее количество адресов

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';

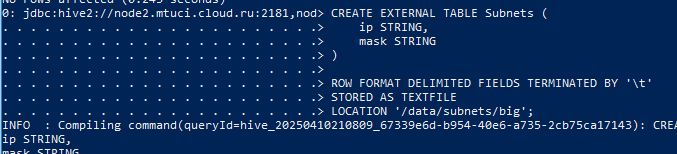


Рисунок 19 - Создание на большом датасете

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;

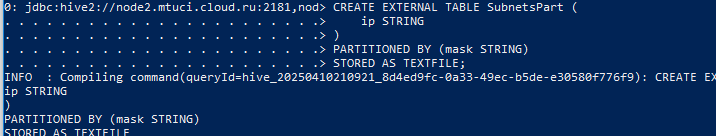


Рисунок 20 - Создание партиционированной датасета

На таблице без партиций

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

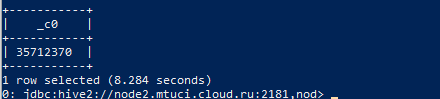


Рисунок 21 - Реультат без партиций

На партиционированной таблице

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

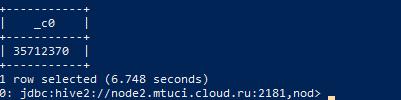


Рисунок 22 - Реультат с партициями