Министерство науки и высшего образования Российской Федерации
Федеральное государсвенное бюджетное образовательное учреждение
высшего образования
«Курганский Государственный Университет» (КГУ)
Кафедра «Безопасность информационных и автоматизированных систем»

Отчет
По лабораторной работе
«Clickhouse»

Дисциплина: Методы и инструменты анализа больших данных
Выполнили студенты: ПТ-40917 группы
/Молоков И. А./
/Мухортиков Д. Д./
/Раевский М. С./
/Смирнов А. А./
Преподаватель: Мирвода С. Г./

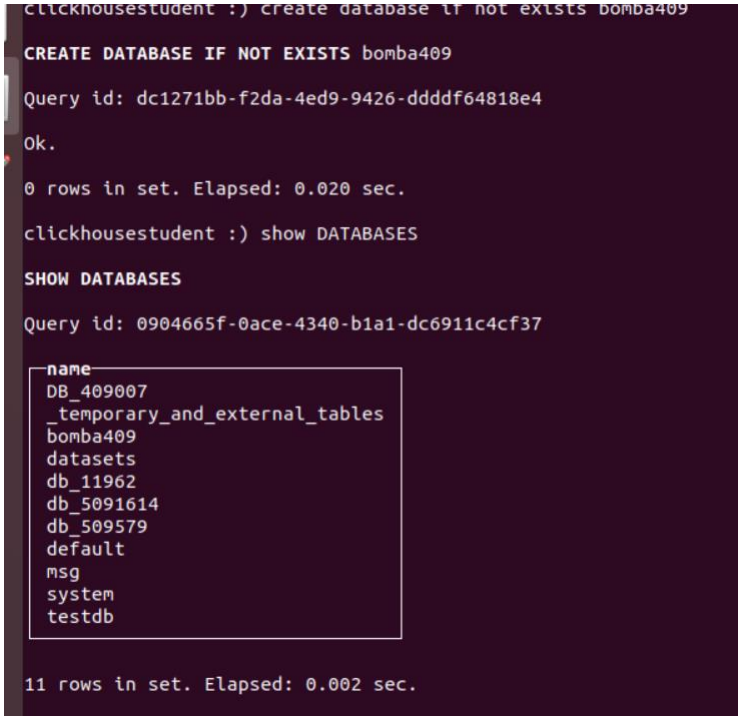
Курган, 2020

Задачи:

1. Объединиться в группы, придумать своей группе код
2. Подключиться к лабораторной ВМ по SSH
3. Создать БД с названием содержащим код своей группы
4. На диске находится файл hits_v1.tsv требуется загрузить данные из него в таблицу
5. Пользуясь примерами команд из раздела "Получение таблиц из сжатых tsv-файлов" https://clickhouse.tech/docs/ru/getting-started/example-datasets/metrica/
   4.1 Создайте в своей БД таблицу hits_v1 со структурой указанной в разделе 4
   4.2 При помощи команды `head` загрузите первые 10000 строк из файла `hits_v1.tsv` в свою таблицу (в примере нужно заменить команду `cat` на команду `head --lines=число_строк`)
   4.3 Подключитесь к БД с помощью `clickhouse-client --user default --password qwerty12345` и проверьте число строк
   4.4 Повторите п 4., но только с командой `tail`, в отчёте укажите чем они отличаются
6. При помощи команды `describe` https://clickhouse.tech/docs/ru/sql-reference/statements/describe-table/ посмотрите структуру созданной вами таблицы
   5.1 Соответствует ли она тому SQL, который вы запускали?
   5.2 Руководствуясь полученными данными о структуре подумайте какие секции (partition) можно создать, ответ обоснуйте
7. Объём данных
   6.1 Узнайте размеры вашей таблицы на диске при помощи следующего запроса: `SELECT formatReadableSize(sum(bytes)) AS size, sum(rows) AS rows FROM system.parts WHERE active and table ='имя_вашей таблицы'`
   6.2 Соответствует ли число строк количеству загруженных?
   6.3 Запомните размер данных. Загрузите ещё 10000 строк пропустив первые 10000.
   6.4 Проверьте размер ещё раз. На сколько вырос объём данных?
   6.5 Сохраните эти же 10000 строк на диск (примерно так: `head --lines=10 hits_v1.tsv > temp.txt`)
   6.6 Узнайте полученного файла и сравните его с размером этих же данных в КХ
   6.7 Сделайте вывод
8. В итоге в вашей таблице должно получиться 30000 строк

1 Подключимся по SSH к ВМ

2 Создадим БД группы с названием bomba409

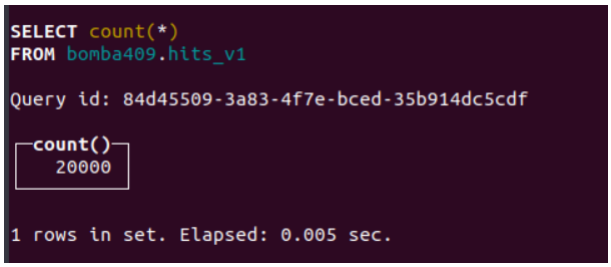


3 Далее создадим таблицу в нашей БД

```
CREATE TABLE bomba409.hits_v1
(
    `WatchID` UInt64,
    `JavaEnable` UInt8,
    `Title` String,
    `GoodEvent` Int16,
    `EventTime` DateTime,
    `EventDate` Date,
    `CounterID` UInt32,
    `ClientIP` UInt32,
    `ClientIP6` FixedString(16),
    `RegionID` UInt32,
    `UserID` UInt64,
    `CounterClass` Int8,
    `OS` UInt8,
    `UserAgent` UInt8,
    `URL` String,
    `Referer` String,
    `URLDomain` String,
    `RefererDomain` String,
    `Refresh` UInt8,
    `IsRobot` UInt8,
    `RefererCategories` Array(UInt16),
    `URLCategories` Array(UInt16),
    `URLRegions` Array(UInt32),
    `RefererRegions` Array(UInt32),
    `ResolutionWidth` UInt16,
    `ResolutionHeight` UInt16,
    `ResolutionDepth` UInt8,
    `FlashMajor` UInt8,
    `FlashMinor` UInt8,
    `FlashMinor2` String,
    `NetMajor` UInt8,
    `NetMinor` UInt8,
    `UserAgentMajor` UInt16,
    `UserAgentMinor` FixedString(2),
    `CookieEnable` UInt8,
    `JavascriptEnable` UInt8,
    `IsMobile` UInt8,
    `MobilePhone` UInt8,
    `MobilePhoneModel` String,
    `Params` String,
    `IPNetworkID` UInt32,
    `TraficSourceID` Int8,
    `SearchEngineID` UInt16,
    `SearchPhrase` String,
    `AdvEngineID` UInt8,
    `IsArtifical` UInt8,
    `WindowClientWidth` UInt16,
    `WindowClientHeight` UInt16,
    `ClientTimeZone` Int16,
    `ClientEventTime` DateTime,
    `SilverlightVersion1` UInt8,
    `SilverlightVersion2` UInt8,
    `SilverlightVersion3` UInt32,
    `SilverlightVersion4` UInt16,
    `PageCharset` String,
    `CodeVersion` UInt32,
    `IsLink` UInt8,
    `IsDownload` UInt8,
    `IsNotBounce` UInt8,
    `FUniqID` UInt64,
    `HID` UInt32,
    `IsOldCounter` UInt8,
    `IsEvent` UInt8,
    `IsParameter` UInt8,
    `DontCountHits` UInt8
```

4 Добавим 10000 строк из файла в нашу таблицу (head)

Сделаем то же самое с командой tail



Различие команд в том, каким образом происходит выборка из файла
head – из начала
tail – из конца

5 Посмотрим структуру таблицы

```
clickhousestudent :) desc bomba409.hits_v1

DESCRIBE TABLE bomba409.hits_v1

Query id: e7c86e1f-9f4c-4cb1-ae7e-02903b594ad0

┌─name──────────────────────┬─type──────────┬─default_type─┬─default_expression
pression─┬─ttl_expression─┐
│ WatchID                   │ UInt64        │              │
                  │                │
│ JavaEnable                │ UInt8         │              │
                  │                │
│ Title                     │ String        │              │
                  │                │
│ GoodEvent                 │ Int16         │              │
                  │                │
│ EventTime                 │ DateTime      │              │
                  │                │
│ EventDate                 │ Date          │              │
                  │                │
│ CounterID                 │ UInt32        │              │
                  │                │
│ ClientIP                  │ UInt32        │              │
                  │                │
│ ClientIP6                 │ FixedString(16) │            │
                  │                │
│ RegionID                  │ UInt32        │              │
                  │                │
│ UserID                    │ UInt64        │              │
                  │                │
│ CounterClass              │ Int8          │              │
                  │                │
│ OS                        │ UInt8         │              │
                  │                │
│ UserAgent                 │ UInt8         │              │
                  │                │
│ URL                       │ String        │              │
```

Вероятно, партиционировать стоит по дате события или региона. Зависит от того, что нужнее знать, когда или откуда.

6 Узнаем объем таблицы на диске

```
SELECT
    formatReadableSize(sum(bytes)) AS size,
    sum(rows) AS rows
FROM system.parts
WHERE active AND (database = 'bomba409') AND (table = 'hits_v1')

Query id: 2dd6619d-e27b-4756-838b-e7358cfd26bc

┌─size─────┬─rows─┐
│ 3.06 MiB │ 20000 │
└──────────┴──────┘
```

Загрузим еще 10000 строк, пропустив первые 10000

```
administrator@clickhousestudent:~$ head -n20000  hits_v1.tsv | tail -n10000 |
clickhouse-client --user='default' --password='qwerty12345' --query "insert in
to bomba409.hits_v1 format TSV" --max_insert_block_size=10000
```

```
SELECT
    formatReadableSize(sum(bytes)) AS size,
    sum(rows) AS rows
FROM system.parts
WHERE active AND (database = 'bomba409') AND (table = 'hits_v1')

Query id: aace7be9-95e5-442d-bd86-2750474f597e

┌─size─────┬─rows─┐
│ 4.24 MiB │ 30000 │
└──────────┴──────┘

1 rows in set. Elapsed: 0.004 sec.
```

Объем вырос на 1.18 MiB

Размер txt-файла из 10000 строк значительно больше тех же данных в clickhouse

```
administrator@clickhousestudent:~$ ls -l --block-size=M
total 7440M
-rw-rw-r-- 1 administrator administrator    8M Nov 12 10:21 btext.txt
-rw-rw-r-- 1 administrator administrator 7424M Nov 11 07:06 hits_v1.tsv
-rw-rw-r-- 1 administrator administrator    1M Oct 28 12:35 index.html
-rw-rw-r-- 1 administrator administrator    8M Nov 11 14:47 test.txt
```

Вероятно, потому что числовые значения в БД занимают от 4 до 8 байт в то время как в виде текста они могут занимать до 20 байт. Также в txt-файле между полями есть знаки табуляции.