# Урок 4. Заливка данных в Hadoop. Форматы хранения файлов

Выполнил Колеганов Н.Д.

Задание

1. Создать таблицы в форматах PARQUET/ORC/AVRO с компрессией и без оной. (Выберите один вариант, например ORC с компрессией)
2. Заполнить данными из большой таблицы hive_db.citation_data
3. Посмотреть на получившийся размер данных
4. Сделать выводы о эффективности хранения и компресии.

1. Создал таблицу. Выбрал gzip компресиию.

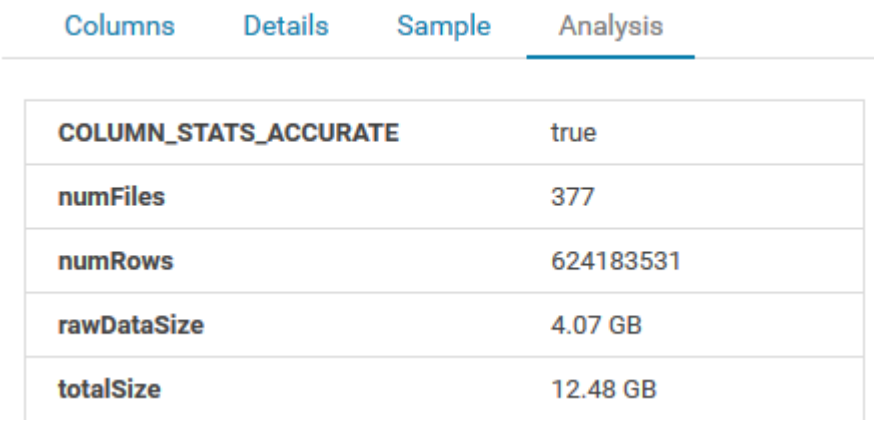
```
SET parquet.compression=gzip;

CREATE TABLE stroganov_db.parquet_test (
                oci string,
                citing string,
                cited string,
                creation string,
                timespan string,
                journal_sc string,
                author_sc string)
STORED AS PARQUET
```

2. Запустил заполнение данными из таблицы hive_db.citation_data

```
INSERT INTO TABLE stroganov_db.parquet_test SELECT * FROM hive_db.citation_data
```

3. Посмотрел размер данных




4. Копирование в новый формат и сжатие длилось долго (около 5 часов) но оно того стоило данные сжались в примерно 24 раза.