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
```

Results Charts

03/10/2020 12:52 PM
6%
3.5m

INSERT INTO TABLE str...ive_db.citation_data (Stage-1)
ID: application_1583843553969_0001 Type: MAPREDUCE User: student2_11
Duration: 3.5m Allocated Memory Seconds: 602K Allocated VCore Seconds: 584

3. Посмотрел размер данных

stroganov_db.parquet_test

Columns Details Sample Analysis

COLUMN_STATS_ACCURATE	true
numFiles	377
numRows	624183531
rawDataSize	4.07 GB
totalSize	12.48 GB
transient_lastDdlTime	10.03.2020 16:41

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