

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 компрессию.

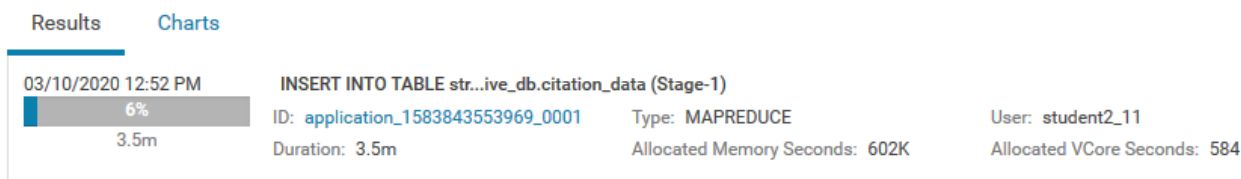
```

1 SET parquet.compression=gzip;
2
3 CREATE TABLE stroganov_db.parquet_test (
4     oci string,
5     citing string,
6     cited string,
7     creation string,
8     timespan string,
9     journal_sc string,
10    author_sc string)
11 STORED AS PARQUET

```

2. Запустил заполнение данными из таблицы hive_db.citation_data

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



3. Посмотрел размер данных

stroganov_db.parquet_test	
Columns	Details
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.