Создала таблицу **student4_1.citation_data_parquet_st4_1** в формате PARQUET с компрессией и заполнила ее данными из таблицы hive_db.citation_data:

```
1 set parquet.compression=SNAPPY;
2
3 create  table student4_1.citation_data_parquet_st4_1
4 STORED AS PARQUET
5 as select * from hive_db.citation_data;
6
7 show create table student4_1.citation_data_parquet_st4_1
```

CREATE TABLE `student4_1.citation_data_parquet_st4_1`(	
2	`oci` string,
3	`citing` string,
4	`cited` string,
5	`creation` string,
6	`timespan` string,
7	`journal_sc` string,
8	`author_sc` string)
9	ROW FORMAT SERDE
10	org.apache.hadoop.hive.ql.io.parquet.serde.ParquetHiveSerDe'
11	STORED AS INPUTFORMAT
12	org.apache.hadoop.hive.ql.io.parquet.MapredParquetInputFormat'
13	OUTPUTFORMAT
14	org.apache.hadoop.hive.ql.io.parquet.MapredParquetOutputFormat'
15	LOCATION
16	hdfs://manager.novalocal:8020/user/hive/warehouse/student4_1.db/citation_data_parquet_st4_1'
17	TBLPROPERTIES (
18	COLUMN_STATS_ACCURATE'='true',
19	numFiles'='377',
20	numRows'='624183594',
21	rawDataSize'='4369285158',
22	totalSize'='24412470352',
23	transient_lastDdlTime'='1591992244')

В результате сжатия размер данных сократился в 4 раза: