

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 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
```

03/10/2020 12:04 PM	INSERT INTO TABLE stroganov_db.citation_data (stage=1)	
6%	ID: application_1583843553969_0001	Type: MAPREDUCE
3.5m	Duration: 3.5m	Allocated Memory Seconds: 602K

3. Посмотрел размер данных

Columns	Details	Sample	Analysis
COLUMN_STATS_ACCURATE		true	
numFiles		377	
numRows		624183531	
rawDataSize		4.07 GB	
totalSize		12.48 GB	

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