**---Первая часть задания -- форматы данных ---**

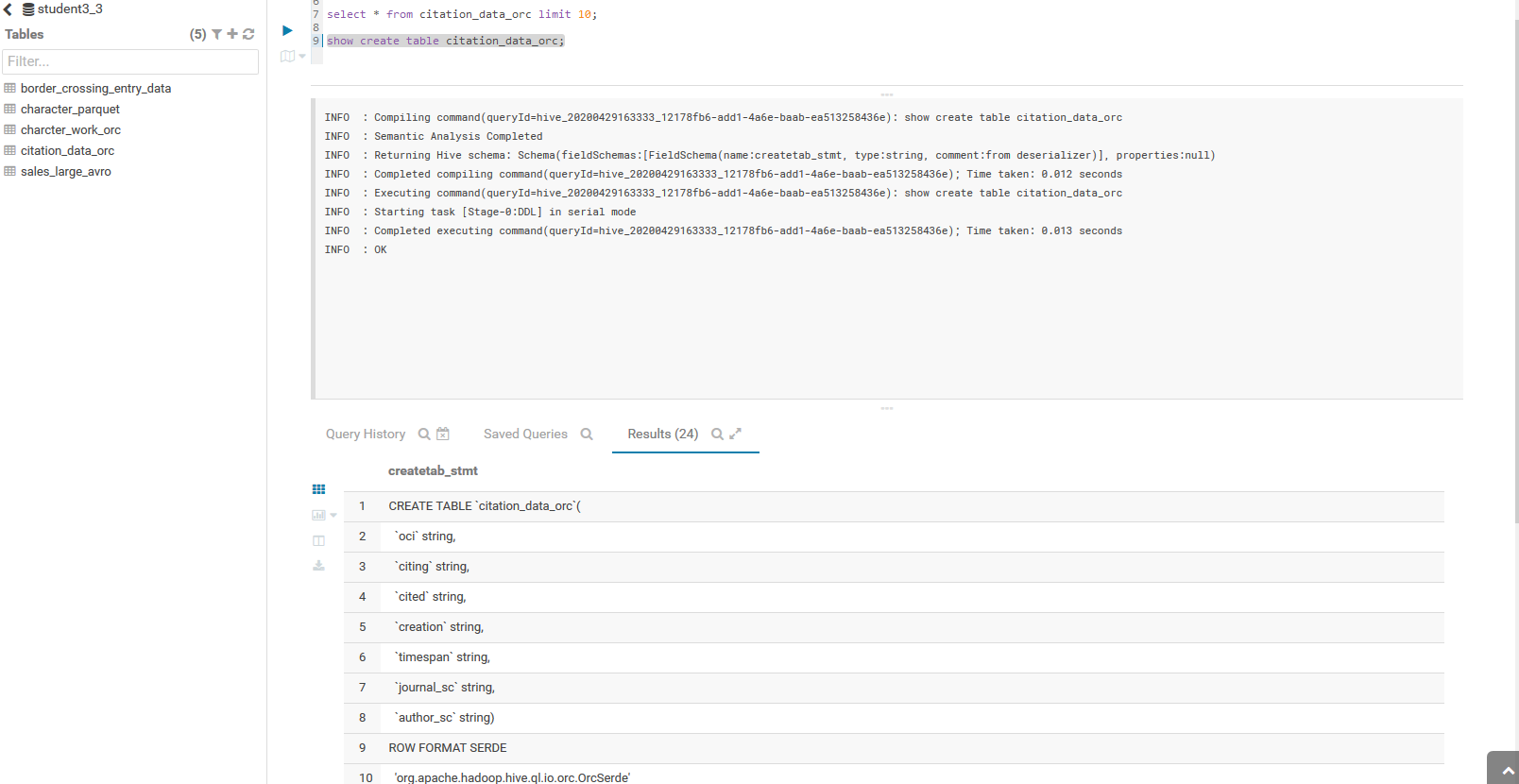
Есть большая таблица по имени

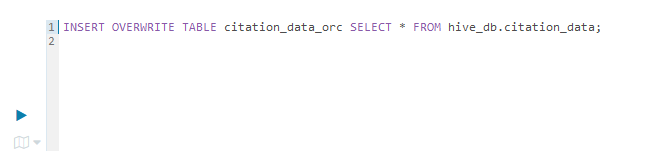
create external table hive\_db.citation\_data  
(  
oci string,  
citing string,  
cited string,  
creation string,  
timespan string,  
journal\_sc string,  
author\_sc string  
)  
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'  
location '/test\_datasets/citation'

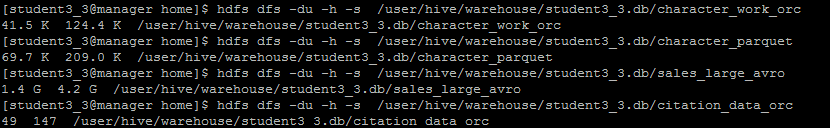
Её размер вот такой:

hdfs dfs -du -h -s /test\_datasets/citation  
97.2 G 291.5 G /test\_datasets/citation

Что вам нужно сделать  
1. Создать таблицы в форматах PARQUET/ORC/AVRO c 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.

Вывод сделать не удалось в связи с тем, что данные либо грузятся чудовищно долго или застревают за 73%. Потом цифра MAP становится 0.

INFO : Compiling command(queryId=hive\_20200430063333\_499ad081-9e0a-4559-ab07-0599e87f54a1): INSERT OVERWRITE TABLE student3\_3.citation\_data\_parquet SELECT \* FROM hive\_db.citation\_data

INFO : Semantic Analysis Completed

INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:citation\_data.oci, type:string, comment:null), FieldSchema(name:citation\_data.citing, type:string, comment:null), FieldSchema(name:citation\_data.cited, type:string, comment:null), FieldSchema(name:citation\_data.creation, type:string, comment:null), FieldSchema(name:citation\_data.timespan, type:string, comment:null), FieldSchema(name:citation\_data.journal\_sc, type:string, comment:null), FieldSchema(name:citation\_data.author\_sc, type:string, comment:null)], properties:null)

INFO : Completed compiling command(queryId=hive\_20200430063333\_499ad081-9e0a-4559-ab07-0599e87f54a1); Time taken: 0.096 seconds

INFO : Executing command(queryId=hive\_20200430063333\_499ad081-9e0a-4559-ab07-0599e87f54a1): INSERT OVERWRITE TABLE student3\_3.citation\_data\_parquet SELECT \* FROM hive\_db.citation\_data

INFO : Query ID = hive\_20200430063333\_499ad081-9e0a-4559-ab07-0599e87f54a1

INFO : Total jobs = 3

INFO : Launching Job 1 out of 3

INFO : Starting task [Stage-1:MAPRED] in serial mode

INFO : Number of reduce tasks is set to 0 since there's no reduce operator

INFO : number of splits:377

INFO : Submitting tokens for job: job\_1583843553969\_0566

INFO : The url to track the job: http://manager.novalocal:8088/proxy/application\_1583843553969\_0566/

INFO : Starting Job = job\_1583843553969\_0566, Tracking URL = http://manager.novalocal:8088/proxy/application\_1583843553969\_0566/

INFO : Kill Command = /opt/cloudera/parcels/CDH-5.16.2-1.cdh5.16.2.p0.8/lib/hadoop/bin/hadoop job -kill job\_1583843553969\_0566

INFO : Hadoop job information for Stage-1: number of mappers: 377; number of reducers: 0

INFO : 2020-04-30 06:33:58,781 Stage-1 map = 0%, reduce = 0%

INFO : 2020-04-30 06:34:39,253 Stage-1 map = 1%, reduce = 0%, Cumulative CPU 63.38 sec

INFO : 2020-04-30 06:35:32,834 Stage-1 map = 2%, reduce = 0%, Cumulative CPU 172.38 sec

INFO : 2020-04-30 06:36:22,030 Stage-1 map = 3%, reduce = 0%, Cumulative CPU 267.44 sec

INFO : 2020-04-30 06:37:11,867 Stage-1 map = 4%, reduce = 0%, Cumulative CPU 366.76 sec

INFO : 2020-04-30 06:38:00,644 Stage-1 map = 5%, reduce = 0%, Cumulative CPU 456.31 sec

INFO : 2020-04-30 06:38:58,145 Stage-1 map = 6%, reduce = 0%, Cumulative CPU 559.85 sec

INFO : 2020-04-30 06:39:48,883 Stage-1 map = 7%, reduce = 0%, Cumulative CPU 655.28 sec

INFO : 2020-04-30 06:40:33,908 Stage-1 map = 8%, reduce = 0%, Cumulative CPU 736.97 sec

INFO : 2020-04-30 06:41:34,643 Stage-1 map = 8%, reduce = 0%, Cumulative CPU 842.27 sec

INFO : 2020-04-30 06:41:37,772 Stage-1 map = 9%, reduce = 0%, Cumulative CPU 857.5 sec

INFO : 2020-04-30 06:42:32,151 Stage-1 map = 10%, reduce = 0%, Cumulative CPU 960.67 sec

INFO : 2020-04-30 06:43:25,390 Stage-1 map = 11%, reduce = 0%, Cumulative CPU 1064.71 sec

INFO : 2020-04-30 06:44:22,018 Stage-1 map = 12%, reduce = 0%, Cumulative CPU 1171.61 sec

INFO : 2020-04-30 06:45:22,661 Stage-1 map = 13%, reduce = 0%, Cumulative CPU 1284.67 sec

INFO : 2020-04-30 06:46:13,861 Stage-1 map = 14%, reduce = 0%, Cumulative CPU 1382.66 sec

INFO : 2020-04-30 06:47:07,278 Stage-1 map = 15%, reduce = 0%, Cumulative CPU 1482.56 sec

INFO : 2020-04-30 06:47:59,567 Stage-1 map = 16%, reduce = 0%, Cumulative CPU 1570.67 sec

INFO : 2020-04-30 06:48:54,939 Stage-1 map = 17%, reduce = 0%, Cumulative CPU 1688.46 sec

INFO : 2020-04-30 06:49:33,688 Stage-1 map = 18%, reduce = 0%, Cumulative CPU 1756.38 sec

INFO : 2020-04-30 06:50:27,124 Stage-1 map = 19%, reduce = 0%, Cumulative CPU 1856.61 sec

INFO : 2020-04-30 06:51:24,710 Stage-1 map = 20%, reduce = 0%, Cumulative CPU 1958.06 sec

INFO : 2020-04-30 06:52:03,573 Stage-1 map = 21%, reduce = 0%, Cumulative CPU 2037.76 sec

INFO : 2020-04-30 06:52:51,523 Stage-1 map = 22%, reduce = 0%, Cumulative CPU 2130.33 sec

INFO : 2020-04-30 06:53:31,167 Stage-1 map = 23%, reduce = 0%, Cumulative CPU 2206.7 sec

INFO : 2020-04-30 06:54:19,080 Stage-1 map = 24%, reduce = 0%, Cumulative CPU 2298.53 sec

INFO : 2020-04-30 06:55:00,645 Stage-1 map = 25%, reduce = 0%, Cumulative CPU 2382.03 sec

INFO : 2020-04-30 06:55:52,712 Stage-1 map = 26%, reduce = 0%, Cumulative CPU 2478.73 sec

INFO : 2020-04-30 06:56:38,573 Stage-1 map = 27%, reduce = 0%, Cumulative CPU 2564.34 sec

INFO : 2020-04-30 06:57:24,495 Stage-1 map = 28%, reduce = 0%, Cumulative CPU 2656.41 sec

INFO : 2020-04-30 06:58:10,445 Stage-1 map = 29%, reduce = 0%, Cumulative CPU 2732.3 sec

INFO : 2020-04-30 06:59:01,552 Stage-1 map = 30%, reduce = 0%, Cumulative CPU 2827.65 sec

INFO : 2020-04-30 06:59:39,106 Stage-1 map = 31%, reduce = 0%, Cumulative CPU 2909.35 sec

INFO : 2020-04-30 07:00:32,433 Stage-1 map = 32%, reduce = 0%, Cumulative CPU 3007.02 sec

INFO : 2020-04-30 07:01:23,531 Stage-1 map = 33%, reduce = 0%, Cumulative CPU 3100.36 sec

INFO : 2020-04-30 07:02:06,382 Stage-1 map = 34%, reduce = 0%, Cumulative CPU 3186.16 sec

INFO : 2020-04-30 07:02:58,856 Stage-1 map = 35%, reduce = 0%, Cumulative CPU 3286.38 sec

INFO : 2020-04-30 07:03:43,969 Stage-1 map = 36%, reduce = 0%, Cumulative CPU 3363.69 sec

INFO : 2020-04-30 07:04:33,032 Stage-1 map = 37%, reduce = 0%, Cumulative CPU 3456.96 sec

INFO : 2020-04-30 07:05:14,691 Stage-1 map = 38%, reduce = 0%, Cumulative CPU 3546.49 sec

INFO : 2020-04-30 07:06:08,041 Stage-1 map = 39%, reduce = 0%, Cumulative CPU 3642.99 sec

INFO : 2020-04-30 07:06:52,400 Stage-1 map = 40%, reduce = 0%, Cumulative CPU 3723.08 sec

INFO : 2020-04-30 07:07:37,698 Stage-1 map = 41%, reduce = 0%, Cumulative CPU 3814.12 sec

INFO : 2020-04-30 07:08:18,462 Stage-1 map = 42%, reduce = 0%, Cumulative CPU 3896.84 sec

INFO : 2020-04-30 07:09:06,565 Stage-1 map = 43%, reduce = 0%, Cumulative CPU 3987.97 sec

INFO : 2020-04-30 07:09:51,393 Stage-1 map = 44%, reduce = 0%, Cumulative CPU 4070.22 sec

INFO : 2020-04-30 07:10:39,591 Stage-1 map = 45%, reduce = 0%, Cumulative CPU 4161.7 sec

INFO : 2020-04-30 07:11:29,795 Stage-1 map = 46%, reduce = 0%, Cumulative CPU 4254.53 sec

INFO : 2020-04-30 07:12:09,351 Stage-1 map = 47%, reduce = 0%, Cumulative CPU 4327.24 sec

INFO : 2020-04-30 07:12:57,237 Stage-1 map = 48%, reduce = 0%, Cumulative CPU 4417.47 sec

INFO : 2020-04-30 07:13:30,645 Stage-1 map = 49%, reduce = 0%, Cumulative CPU 4489.92 sec

INFO : 2020-04-30 07:14:22,876 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 4583.92 sec

INFO : 2020-04-30 07:15:03,679 Stage-1 map = 51%, reduce = 0%, Cumulative CPU 4669.07 sec

INFO : 2020-04-30 07:15:56,094 Stage-1 map = 52%, reduce = 0%, Cumulative CPU 4767.23 sec

INFO : 2020-04-30 07:16:38,889 Stage-1 map = 53%, reduce = 0%, Cumulative CPU 4842.12 sec

INFO : 2020-04-30 07:17:29,996 Stage-1 map = 54%, reduce = 0%, Cumulative CPU 4947.2 sec

INFO : 2020-04-30 07:18:08,566 Stage-1 map = 55%, reduce = 0%, Cumulative CPU 5017.21 sec

INFO : 2020-04-30 07:18:52,412 Stage-1 map = 56%, reduce = 0%, Cumulative CPU 5108.16 sec

INFO : 2020-04-30 07:19:40,410 Stage-1 map = 57%, reduce = 0%, Cumulative CPU 5198.0 sec

INFO : 2020-04-30 07:20:22,382 Stage-1 map = 58%, reduce = 0%, Cumulative CPU 5274.35 sec

INFO : 2020-04-30 07:21:10,735 Stage-1 map = 59%, reduce = 0%, Cumulative CPU 5367.95 sec

INFO : 2020-04-30 07:21:52,585 Stage-1 map = 60%, reduce = 0%, Cumulative CPU 5450.26 sec

INFO : 2020-04-30 07:22:38,602 Stage-1 map = 61%, reduce = 0%, Cumulative CPU 5538.18 sec

INFO : 2020-04-30 07:23:20,443 Stage-1 map = 62%, reduce = 0%, Cumulative CPU 5620.1 sec

INFO : 2020-04-30 07:24:08,437 Stage-1 map = 63%, reduce = 0%, Cumulative CPU 5709.04 sec

INFO : 2020-04-30 07:24:53,301 Stage-1 map = 64%, reduce = 0%, Cumulative CPU 5781.79 sec

INFO : 2020-04-30 07:25:40,470 Stage-1 map = 65%, reduce = 0%, Cumulative CPU 5873.69 sec

INFO : 2020-04-30 07:26:19,139 Stage-1 map = 66%, reduce = 0%, Cumulative CPU 5948.86 sec

INFO : 2020-04-30 07:27:12,498 Stage-1 map = 67%, reduce = 0%, Cumulative CPU 6044.44 sec

INFO : 2020-04-30 07:27:54,494 Stage-1 map = 68%, reduce = 0%, Cumulative CPU 6128.77 sec

INFO : 2020-04-30 07:28:38,558 Stage-1 map = 69%, reduce = 0%, Cumulative CPU 6218.86 sec

INFO : 2020-04-30 07:29:28,074 Stage-1 map = 70%, reduce = 0%, Cumulative CPU 6311.06 sec

INFO : 2020-04-30 07:30:13,066 Stage-1 map = 71%, reduce = 0%, Cumulative CPU 6384.37 sec

INFO : 2020-04-30 07:31:06,330 Stage-1 map = 72%, reduce = 0%, Cumulative CPU 6483.89 sec

INFO : 2020-04-30 07:31:45,944 Stage-1 map = 73%, reduce = 0%, Cumulative CPU 6568.11 sec

INFO : 2020-04-30 07:32:31,089 Stage-1 map = 0%, reduce = 0%

INFO : 2020-04-30 07:33:28,113 Stage-1 map = 1%, reduce = 0%, Cumulative CPU 48.54 sec

INFO : 2020-04-30 07:34:28,962 Stage-1 map = 1%, reduce = 0%, Cumulative CPU 94.22 sec

INFO : 2020-04-30 07:35:02,916 Stage-1 map = 2%, reduce = 0%, Cumulative CPU 137.75 sec

INFO : 2020-04-30 07:36:03,540 Stage-1 map = 2%, reduce = 0%, Cumulative CPU 138.39 sec

INFO : 2020-04-30 07:37:04,193 Stage-1 map = 2%, reduce = 0%, Cumulative CPU 138.39 sec

INFO : 2020-04-30 07:38:04,702 Stage-1 map = 2%, reduce = 0%, Cumulative CPU 138.39 sec

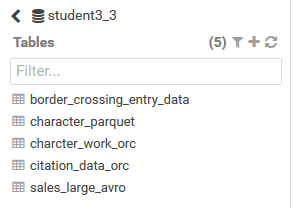
INFO : 2020-04-30 07:39:05,328 Stage-1 map = 2%, reduce = 0%, Cumulative CPU 138.39 sec

INFO : 2020-04-30 07:40:06,248 Stage-1 map = 2%, reduce = 0%, Cumulative CPU 138.39 sec

INFO : 2020-04-30 07:41:07,017 Stage-1 map = 2%, reduce = 0%, Cumulative CPU 138.39 sec

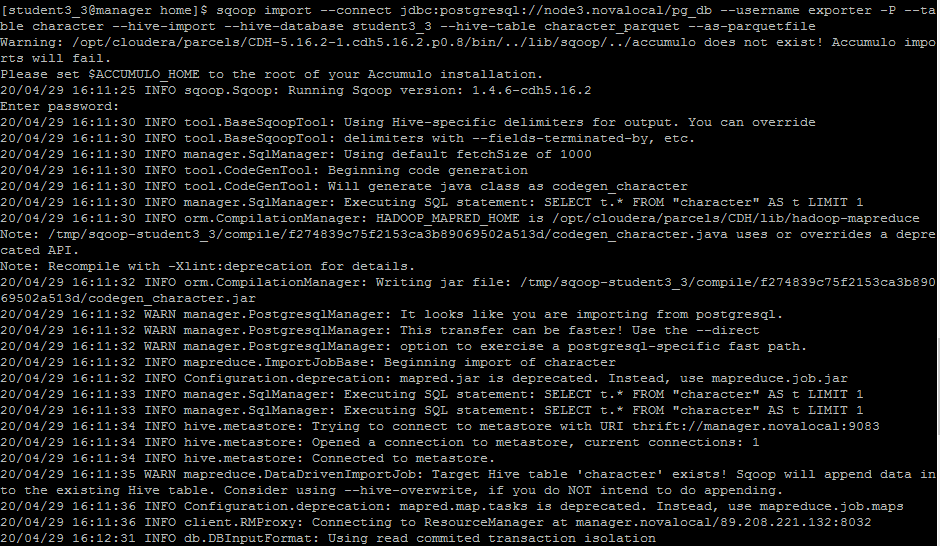
**---Вторая часть задания -- SQOOP ---**  
Простое задание довольно :)  
1. Создать отдельную БД в HIve

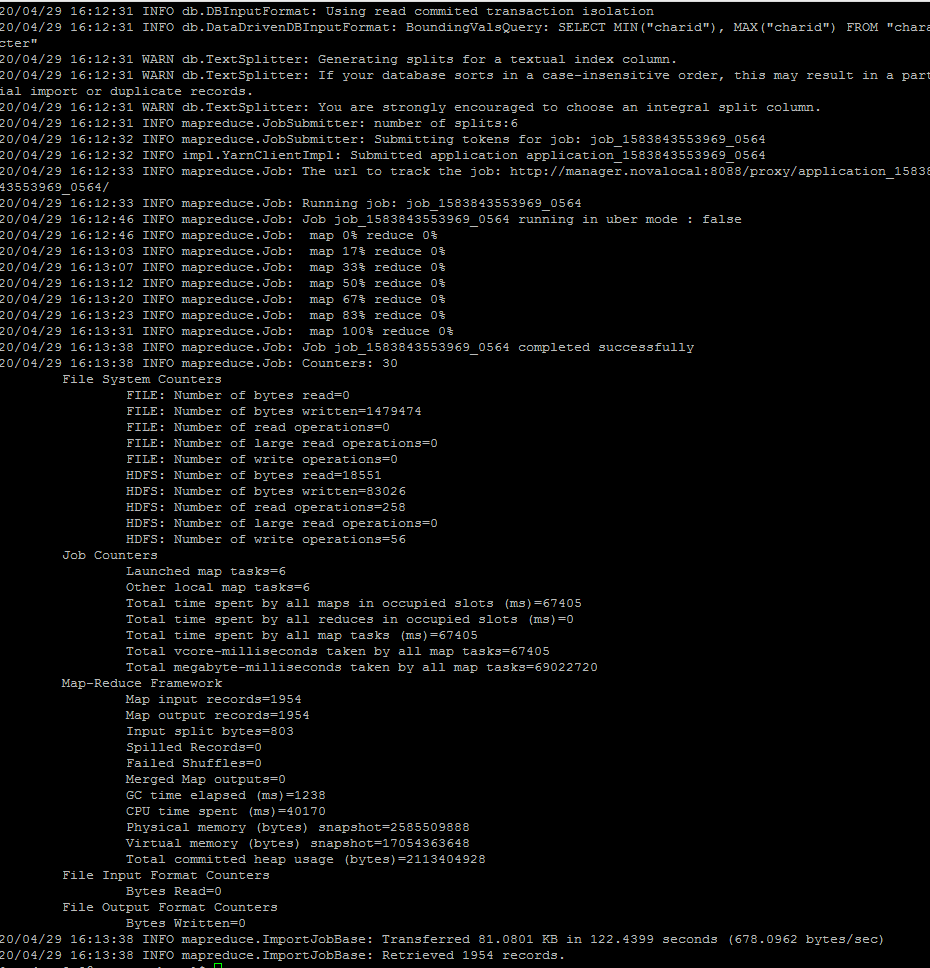
Использовал БД с прошлого урока

  
2. Посмотреть при помощи SQOOP содержимое БД в POsgresql

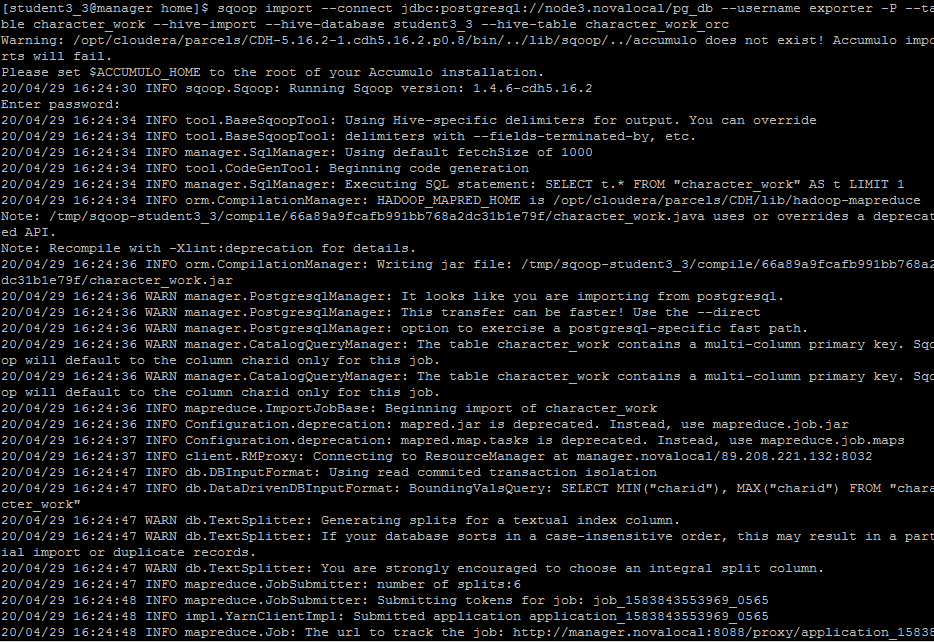
  
3. Импортировать в нее три любые таблицы из базы pg\_db в Postgresql используя SQOOP. Для каждой таблице используйте отдельный формат хранения -- ORC/Parquet/AVRO Рекомендую захватить таблицу sales\_large -- там порядка 10 миллионов записей, она будет достаточно репрезентативна для проверки компрессии.

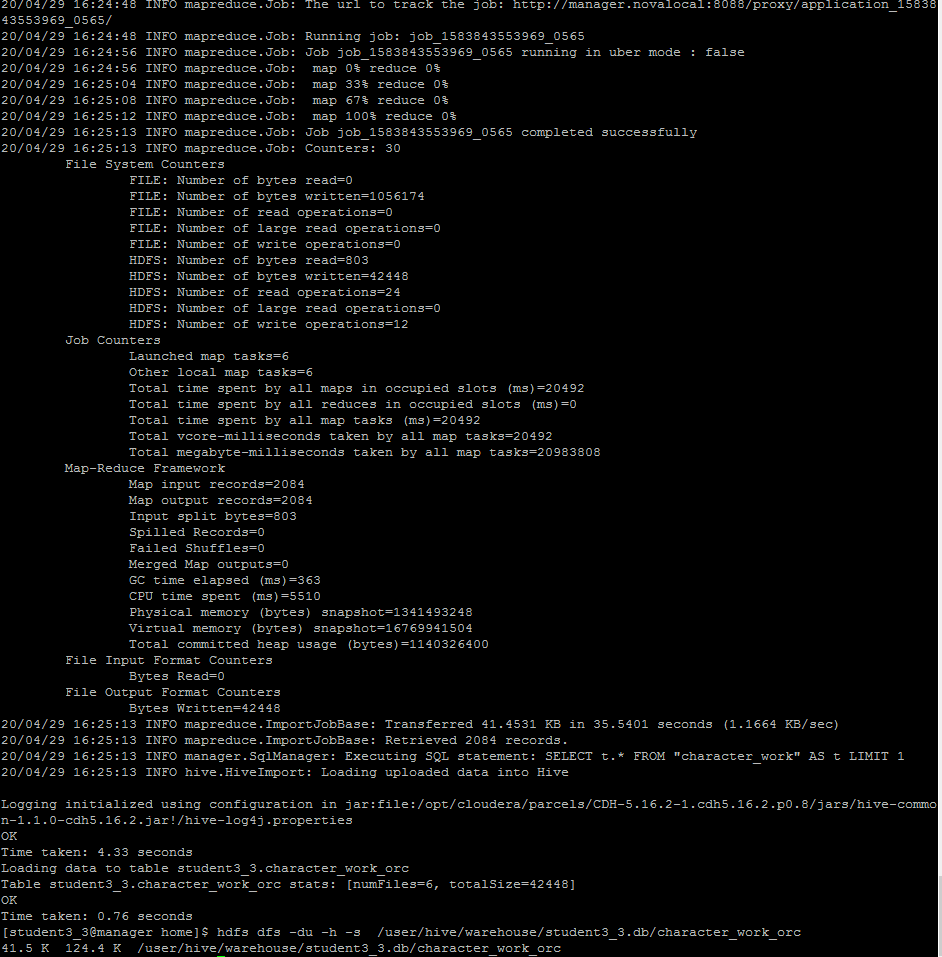
**PARQUET**



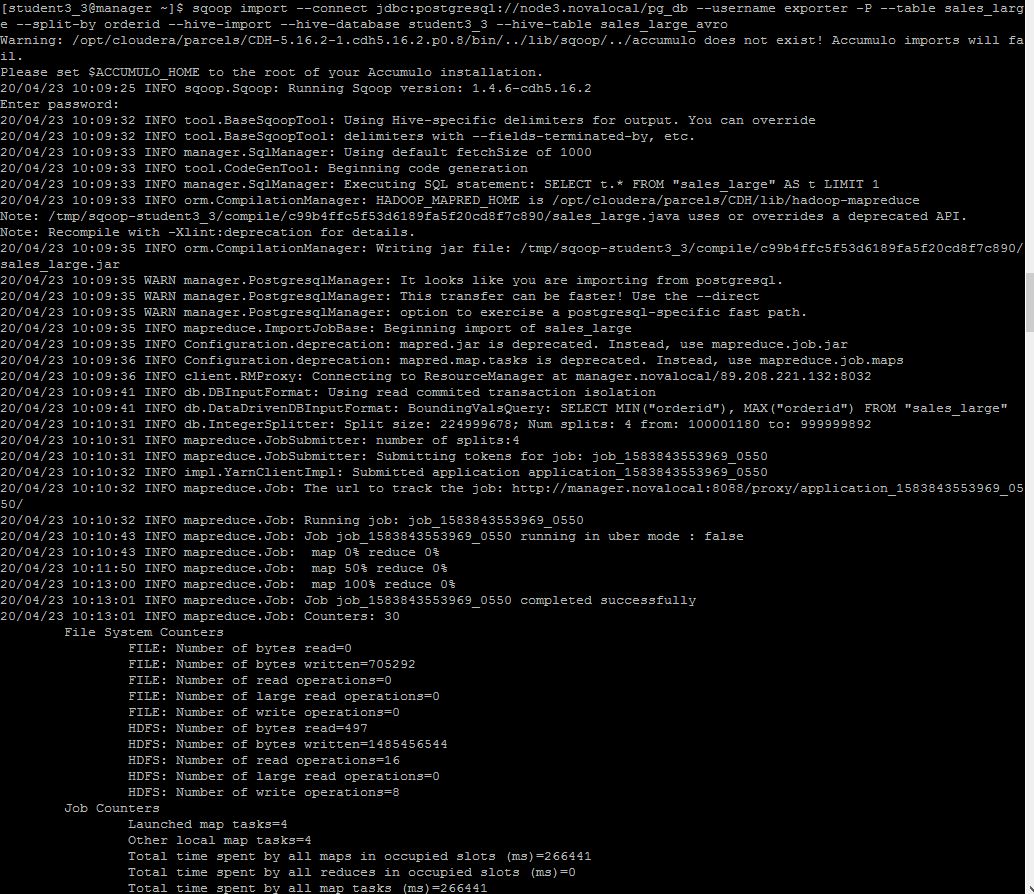


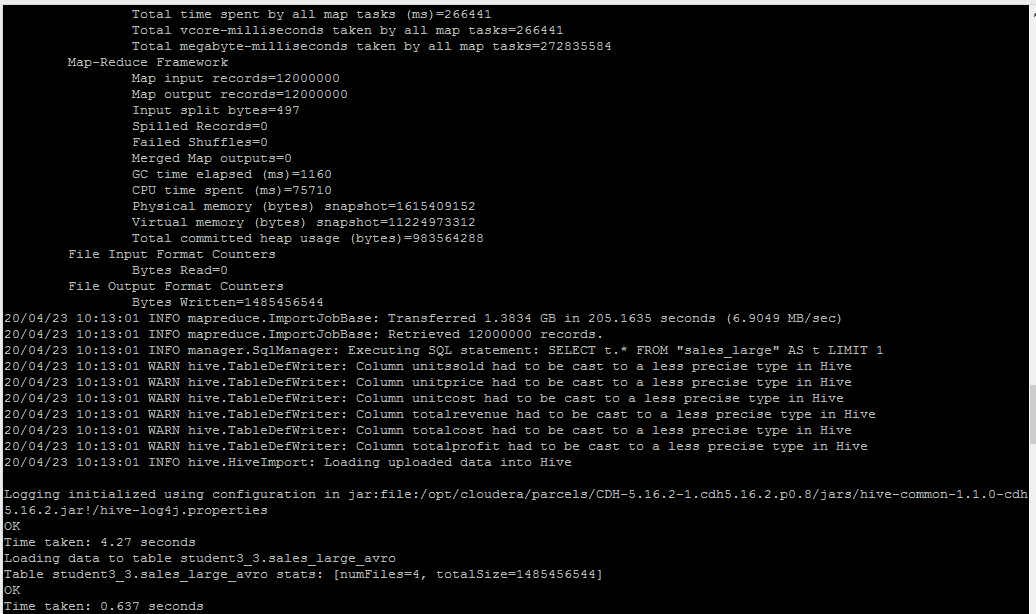
**ORC**



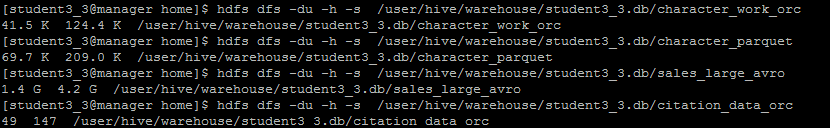


**AVRO**



  
4. Найдите папки на файловой системе куда были сохранены данные. Посмотрите их

размер.

  
5. Сделайте несколько произвольных запросов к этим таблицам.

