Домашнее задание

Нужно загрузить достаточно большой датасет и провести сравнительные эксперименты. Ниже примеры запросов

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'

create external table citation_data (oci string, citing string, cited string, creation string, timespan string, journal_sc string, author_sc string);

0: jdbc:hive2://bigdataanalytics-worker-0.nov> c**reate external table 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 'user/student1_13/citation';**
INFO : Compiling command(queryId=hive_20210214213641_9a317e37-63c7-4781-af09-22c7762c00b0): create external table 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 'user/student1_13/citation'
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO : Completed compiling command(queryId=hive_20210214213641_9a317e37-63c7-4781-af09-22c7762c00b0); Time taken: 0.028 seconds
INFO : Executing command(queryId=hive_20210214213641_9a317e37-63c7-4781-af09-22c7762c00b0): create external table 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 'user/student1_13/citation'
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20210214213641_9a317e37-63c7-4781-af09-22c7762c00b0); Time taken: 0.053 seconds
INFO : OK
No rows affected (0.119 seconds)
0: jdbc:hive2://bigdataanalytics-worker-0.nov> **show tables;**
INFO : Compiling command(queryId=hive_20210214213650_d3a88405-3a4f-4b8e-81e8-7ca7746c2c75): show tables
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:tab_name, type:string, comment:from deserializer)], properties:null)
INFO : Completed compiling command(queryId=hive_20210214213650_d3a88405-3a4f-4b8e-81e8-7ca7746c2c75); Time taken: 0.005 seconds
INFO : Executing command(queryId=hive_20210214213650_d3a88405-3a4f-4b8e-81e8-7ca7746c2c75): show tables
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20210214213650_d3a88405-3a4f-4b8e-81e8-7ca7746c2c75); Time taken: 0.005 seconds
INFO : OK
+----------------+

```
|   tab_name    |
+---------------+
| churners      |
| citation_data |
| extch         |
+---------------+
3 rows selected (0.022 seconds)
[student1_13@bigdataanalytics-head-0 ~]$ hdfs dfs -du -h -s user/student1_13/citation
0  0  user/student1_13/citation
```

Её размер можно узнать вот так : hdfs dfs -du -h -s /test_datasets/citation

Что вам нужно сделать

1. Создать таблицы в форматах PARQUET/ORC/AVRO с компрессией и без. (Выберите несколько вариантов, например ORC с компрессией)

create external table citation_parquet (clientnum_id string, attrition_flag_id string, customer_age string, gender string, dependent_count string, education_level string, marital_status string) stored as parquet;

create external table citation_orc (clientnum_id string, attrition_flag_id string, customer_age string, gender string, dependent_count string, education_level string, marital_status string) stored as orc;

create external table citation_avro (clientnum_id string, attrition_flag_id string, customer_age string, gender string, dependent_count string, education_level string, marital_status string) stored as avro;

```
0: jdbc:hive2://bigdataanalytics-worker-0.nov> SET parquet.compression =
snappy;
No rows affected (0.004 seconds)
0: jdbc:hive2://bigdataanalytics-worker-0.nov> SET parquet.compression =
SNAPPY;
No rows affected (0.004 seconds)
0: jdbc:hive2://bigdataanalytics-worker-0.nov> SET ORC.compression =
SNAPPY;
No rows affected (0.005 seconds)
0: jdbc:hive2://bigdataanalytics-worker-0.nov> SET parquet.compression =
GZIP;
No rows affected (0.004 seconds)
0: jdbc:hive2://bigdataanalytics-worker-0.nov> SET avro.compression =
GZIP;
No rows affected (0.004 seconds)
```

create external table citation_parquet_comp (clientnum_id string, attrition_flag_id string, customer_age string, gender string, dependent_count string, education_level string, marital_status string) stored as parquet;

create external table citation_orc_comp (clientnum_id string, attrition_flag_id string, customer_age string, gender string, dependent_count string, education_level string, marital_status string) stored as orc;

create external table citation_avro_comp (clientnum_id string, attrition_flag_id string, customer_age string, gender string, dependent_count string, education_level string, marital_status string) stored as avro;

```
0: jdbc:hive2://bigdataanalytics-worker-0.nov> create external table
citation_avro_comp (clientnum_id string, attrition_flag_id string,
customer_age string, gender string, dependent_count string,
education_level string, marital_status string) stored as avro;
INFO  : Compiling
command(queryId=hive_20210214221827_41b58883-5e05-4ba2-b807-
f88a42ab68d7): create external table citation_avro_comp (clientnum_id
string, attrition_flag_id string, customer_age string, gender string,
dependent_count string, education_level string, marital_status string)
stored as avro
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null,
properties:null)
INFO  : Completed compiling
command(queryId=hive_20210214221827_41b58883-5e05-4ba2-b807-
f88a42ab68d7); Time taken: 0.024 seconds
INFO  : Executing
command(queryId=hive_20210214221827_41b58883-5e05-4ba2-b807-
f88a42ab68d7): create external table citation_avro_comp (clientnum_id
string, attrition_flag_id string, customer_age string, gender string,
dependent_count string, education_level string, marital_status string)
stored as avro
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing
command(queryId=hive_20210214221827_41b58883-5e05-4ba2-b807-
f88a42ab68d7); Time taken: 0.038 seconds
INFO  : OK
0: jdbc:hive2://bigdataanalytics-worker-0.nov> create external table
citation_parquet_comp (clientnum_id string, attrition_flag_id string,
customer_age string, gender string, dependent_count string,
education_level string, marital_status string) stored as parquet;
INFO  : Compiling command(queryId=hive_20210214222325_cbe5d0c6-
ab7e-40c0-9bc1-e9c2872128e3): create external table
citation_parquet_comp (clientnum_id string, attrition_flag_id string,
customer_age string, gender string, dependent_count string,
education_level string, marital_status string) stored as parquet
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null,
properties:null)
INFO  : Completed compiling
command(queryId=hive_20210214222325_cbe5d0c6-ab7e-40c0-9bc1-
e9c2872128e3); Time taken: 0.022 seconds
INFO  : Executing command(queryId=hive_20210214222325_cbe5d0c6-
ab7e-40c0-9bc1-e9c2872128e3): create external table
citation_parquet_comp (clientnum_id string, attrition_flag_id string,
```

```
customer_age string, gender string, dependent_count string,
education_level string, marital_status string) stored as parquet
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing
command(queryId=hive_20210214222325_cbe5d0c6-ab7e-40c0-9bc1-
e9c2872128e3); Time taken: 0.044 seconds
INFO  : OK
No rows affected (0.1 seconds)
0: jdbc:hive2://bigdataanalytics-worker-0.nov> create external table
citation_orc_comp (clientnum_id string, attrition_flag_id string,
customer_age string, gender string, dependent_count string,
education_level string, marital_status string) stored as orc;
INFO  : Compiling command(queryId=hive_20210214222026_2dd89bd1-
ab53-40f5-a36d-97c70dc57fa5): create external table citation_orc_comp
(clientnum_id string, attrition_flag_id string, customer_age string,
gender string, dependent_count string, education_level string,
marital_status string) stored as orc
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null,
properties:null)
INFO  : Completed compiling
command(queryId=hive_20210214222026_2dd89bd1-ab53-40f5-
a36d-97c70dc57fa5); Time taken: 0.022 seconds
INFO  : Executing command(queryId=hive_20210214222026_2dd89bd1-
ab53-40f5-a36d-97c70dc57fa5): create external table citation_orc_comp
(clientnum_id string, attrition_flag_id string, customer_age string,
gender string, dependent_count string, education_level string,
marital_status string) stored as orc
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing
command(queryId=hive_20210214222026_2dd89bd1-ab53-40f5-
a36d-97c70dc57fa5); Time taken: 0.058 seconds
INFO  : OK
No rows affected (0.116 seconds)



0: jdbc:hive2://bigdataanalytics-worker-0.nov> use bank;
INFO  : Compiling
command(queryId=hive_20210214220413_b4e1fedb-4e87-4907-88f5-98b155cdd015
): use bank
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null,
properties:null)
INFO  : Completed compiling
command(queryId=hive_20210214220413_b4e1fedb-4e87-4907-88f5-98b155cdd015
); Time taken: 0.004 seconds
INFO  : Executing
command(queryId=hive_20210214220413_b4e1fedb-4e87-4907-88f5-98b155cdd015
): use bank
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing
command(queryId=hive_20210214220413_b4e1fedb-4e87-4907-88f5-98b155cdd015
); Time taken: 0.004 seconds
INFO  : OK
```

```
No rows affected (0.017 seconds)
0: jdbc:hive2://bigdataanalytics-worker-0.nov> create external table
citation_parquet (clientnum_id string, attrition_flag_id string,
customer_age string, gender string, dependent_count string,
education_level string, marital_status string) stored as parquet;
INFO  : Compiling
command(queryId=hive_20210214220418_ca39bcde-79e5-4e3a-9689-
bf4e529f524a): create external table citation_parquet (clientnum_id
string, attrition_flag_id string, customer_age string, gender string,
dependent_count string, education_level string, marital_status string)
stored as parquet
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null,
properties:null)
INFO  : Completed compiling
command(queryId=hive_20210214220418_ca39bcde-79e5-4e3a-9689-
bf4e529f524a); Time taken: 0.017 seconds
INFO  : Executing
command(queryId=hive_20210214220418_ca39bcde-79e5-4e3a-9689-
bf4e529f524a): create external table citation_parquet (clientnum_id
string, attrition_flag_id string, customer_age string, gender string,
dependent_count string, education_level string, marital_status string)
stored as parquet
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing
command(queryId=hive_20210214220418_ca39bcde-79e5-4e3a-9689-
bf4e529f524a); Time taken: 0.085 seconds
INFO  : OK
No rows affected (0.251 seconds)
0: jdbc:hive2://bigdataanalytics-worker-0.nov> create external table
citation_orc (clientnum_id string, attrition_flag_id string,
customer_age string, gender string, dependent_count string,
education_level string, marital_status string) stored as orc;
INFO  : Compiling command(queryId=hive_20210214220435_0fc7884a-
a03f-47a9-b15d-c6a6a1838d01): create external table citation_orc
(clientnum_id string, attrition_flag_id string, customer_age string,
gender string, dependent_count string, education_level string,
marital_status string) stored as orc
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null,
properties:null)
INFO  : Completed compiling
command(queryId=hive_20210214220435_0fc7884a-a03f-47a9-b15d-
c6a6a1838d01); Time taken: 0.016 seconds
INFO  : Executing command(queryId=hive_20210214220435_0fc7884a-
a03f-47a9-b15d-c6a6a1838d01): create external table citation_orc
(clientnum_id string, attrition_flag_id string, customer_age string,
gender string, dependent_count string, education_level string,
marital_status string) stored as orc
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing
command(queryId=hive_20210214220435_0fc7884a-a03f-47a9-b15d-
c6a6a1838d01); Time taken: 0.035 seconds
INFO  : OK
No rows affected (0.089 seconds)
```

```
0: jdbc:hive2://bigdataanalytics-worker-0.nov> create external table
citation_avro (clientnum_id string, attrition_flag_id string,
customer_age string, gender string, dependent_count string,
education_level string, marital_status string) stored as avro;
INFO  : Compiling
command(queryId=hive_20210214220445_fb205ff4-31aa-48dd-
b6c1-7046886b5fa5): create external table citation_avro (clientnum_id
string, attrition_flag_id string, customer_age string, gender string,
dependent_count string, education_level string, marital_status string)
stored as avro
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null,
properties:null)
INFO  : Completed compiling
command(queryId=hive_20210214220445_fb205ff4-31aa-48dd-
b6c1-7046886b5fa5); Time taken: 0.016 seconds
INFO  : Executing
command(queryId=hive_20210214220445_fb205ff4-31aa-48dd-
b6c1-7046886b5fa5): create external table citation_avro (clientnum_id
string, attrition_flag_id string, customer_age string, gender string,
dependent_count string, education_level string, marital_status string)
stored as avro
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing
command(queryId=hive_20210214220445_fb205ff4-31aa-48dd-
b6c1-7046886b5fa5); Time taken: 0.038 seconds
INFO  : OK
No rows affected (0.09 seconds)
0: jdbc:hive2://bigdataanalytics-worker-0.nov> show tables;
INFO  : Compiling
command(queryId=hive_20210214220451_91054174-3492-41ed-a53b-
e44f3b296a61): show tables
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:
[FieldSchema(name:tab_name, type:string, comment:from deserializer)],
properties:null)
INFO  : Completed compiling
command(queryId=hive_20210214220451_91054174-3492-41ed-a53b-
e44f3b296a61); Time taken: 0.005 seconds
INFO  : Executing
command(queryId=hive_20210214220451_91054174-3492-41ed-a53b-
e44f3b296a61): show tables
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing
command(queryId=hive_20210214220451_91054174-3492-41ed-a53b-
e44f3b296a61); Time taken: 0.005 seconds
INFO  : OK
+--------------------+
|     tab_name       |
+--------------------+
| churners           |
| citation_avro      |
| citation_data      |
| citation_orc       |
| citation_parquet   |
| extch              |
```

```
+-------------------+
6 rows selected (0.023 seconds)
```

insert into citation_data (select clientnum_id as oci, attrition_flag_id as citing, customer_age as cited, gender as creation, dependent_count as timespan, education_level as journal_sc, marital_status as author_sc from churners);

churners.clientnum_id | churners.attrition_flag_id | churners.customer_age | churners.gender | churners.dependent_count | churners.education_level | churners.marital_status | churners.income_category | churners.card_category | churners.months_on_book | churners.total_relationship_count | churners.months_inactive_12_mon | churners.contacts_count_12_mon | churners.credit_limit | churners.total_revolving_bal | churners.avg_open_to_buy | churners.total_amt_chng_q4_q1 | churners.total_trans_amt | churners.total_trans_ct | churners.total_ct_chng_q4_q1 | churners.avg_utilization_ratio | churners.naive_bayes | churners.naive_bayes_2 |

0: jdbc:hive2://bigdataanalytics-worker-0.nov> insert into citation_data (select clientnum_id as oci, attrition_flag_id as citing, customer_age as cited, gender as creation, dependent_count as timespan, education_level as journal_sc, marital_status as author_sc from churners);
Interrupting... Please be patient this may take some time.
INFO : Compiling command(queryId=hive_20210214215410_b64c6a01-52b9-4e55-826c-a1512eb8cad2): insert into citation_data (select clientnum_id as oci, attrition_flag_id as citing, customer_age as cited, gender as creation, dependent_count as timespan, education_level as journal_sc, marital_status as author_sc from churners)
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:oci, type:string, comment:null), FieldSchema(name:citing, type:string, comment:null), FieldSchema(name:cited, type:string, comment:null), FieldSchema(name:creation, type:string, comment:null), FieldSchema(name:timespan, type:string, comment:null), FieldSchema(name:journal_sc, type:string, comment:null), FieldSchema(name:author_sc, type:string, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20210214215410_b64c6a01-52b9-4e55-826c-a1512eb8cad2); Time taken: 0.179 seconds
INFO : Executing command(queryId=hive_20210214215410_b64c6a01-52b9-4e55-826c-a1512eb8cad2): insert into citation_data (select clientnum_id as oci, attrition_flag_id as citing, customer_age as cited, gender as creation, dependent_count as timespan, education_level as journal_sc, marital_status as author_sc from churners)
INFO : Query ID = hive_20210214215410_b64c6a01-52b9-4e55-826c-a1512eb8cad2
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Subscribed to counters: [] for queryId: hive_20210214215410_b64c6a01-52b9-4e55-826c-a1512eb8cad2
INFO : Tez session hasn't been created yet. Opening session
ERROR : Failed to execute tez graph.
java.io.IOException: Previous writer likely failed to write hdfs://bigdataanalytics-head-0.novalocal:8020/user/student1_13/.hiveJars/hive-exec-3.1.0.3.1.4.0-315-effe339ef81326ce093bc0a1516e86d9d189e11126de97212254c005859b949e.jar. Failing because I am unlikely to write too.
at org.apache.hadoop.hive.ql.exec.tez.DagUtils.localizeResource(DagUtils.java:1296) ~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
at org.apache.hadoop.hive.ql.exec.tez.TezSessionState.createJarLocalResource(TezSessionState.java:917) ~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]

at org.apache.hadoop.hive.ql.exec.tez.TezSessionState.makeCombinedJarMap(TezSessionState.java:349) ~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
at org.apache.hadoop.hive.ql.exec.tez.TezSessionState.openInternal(TezSessionState.java:418) ~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
at org.apache.hadoop.hive.ql.exec.tez.TezSessionPoolSession.openInternal(TezSessionPoolSession.java:124) ~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
at org.apache.hadoop.hive.ql.exec.tez.TezSessionState.open(TezSessionState.java:373) ~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
at org.apache.hadoop.hive.ql.exec.tez.TezTask.ensureSessionHasResources(TezTask.java:373) ~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
at org.apache.hadoop.hive.ql.exec.tez.TezTask.execute(TezTask.java:200) ~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
at org.apache.hadoop.hive.ql.exec.Task.executeTask(Task.java:212) ~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
at org.apache.hadoop.hive.ql.exec.TaskRunner.runSequential(TaskRunner.java:103) ~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
at org.apache.hadoop.hive.ql.Driver.launchTask(Driver.java:2712) ~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
at org.apache.hadoop.hive.ql.Driver.execute(Driver.java:2383) ~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
at org.apache.hadoop.hive.ql.Driver.runInternal(Driver.java:2055) ~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1753) ~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1747) ~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
at org.apache.hadoop.hive.ql.reexec.ReExecDriver.run(ReExecDriver.java:157) ~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
at org.apache.hive.service.cli.operation.SQLOperation.runQuery(SQLOperation.java:226) ~[hive-service-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
at org.apache.hive.service.cli.operation.SQLOperation.access$700(SQLOperation.java:87) ~[hive-service-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
at org.apache.hive.service.cli.operation.SQLOperation$BackgroundWork$1.run(SQLOperation.java:324) ~[hive-service-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
at java.security.AccessController.doPrivileged(Native Method) ~[?:1.8.0_191]
at javax.security.auth.Subject.doAs(Subject.java:422) ~[?:1.8.0_191]
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1730) ~[hadoop-common-3.1.1.3.1.4.0-315.jar:?]
at org.apache.hive.service.cli.operation.SQLOperation$BackgroundWork.run(SQLOperation.java:342) ~[hive-service-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511) ~[?:1.8.0_191]
at java.util.concurrent.FutureTask.run(FutureTask.java:266) ~[?:1.8.0_191]
at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511) ~[?:1.8.0_191]
at java.util.concurrent.FutureTask.run(FutureTask.java:266) ~[?:1.8.0_191]
at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149) ~[?:1.8.0_191]
at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624) ~[?:1.8.0_191]

at java.lang.Thread.run(Thread.java:748) [?:1.8.0_191]
ERROR : FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.tez.TezTask
INFO : Completed executing command(queryId=hive_20210214215410_b64c6a01-52b9-4e55-826c-a1512eb8cad2); Time taken: 25.094 seconds
Error: Error while processing statement: FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.tez.TezTask (state=08S01,code=1)
0: jdbc:hive2://bigdataanalytics-worker-0.nov> insert into citation_data (select clientnum_id as oci, attrition_flag_id as citing, customer_age as cited, gender as creation, dependent_count as timespan, education_level as journal_sc, marital_status as author_sc from churners);
INFO : Compiling command(queryId=hive_20210214215442_d9ef7191-04ce-4b07-9274-bd6fa80c6a5d): insert into citation_data (select clientnum_id as oci, attrition_flag_id as citing, customer_age as cited, gender as creation, dependent_count as timespan, education_level as journal_sc, marital_status as author_sc from churners)
INFO : Semantic Analysis Completed (retrial = false)
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:oci, type:string, comment:null), FieldSchema(name:citing, type:string, comment:null), FieldSchema(name:cited, type:string, comment:null), FieldSchema(name:creation, type:string, comment:null), FieldSchema(name:timespan, type:string, comment:null), FieldSchema(name:journal_sc, type:string, comment:null), FieldSchema(name:author_sc, type:string, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20210214215442_d9ef7191-04ce-4b07-9274-bd6fa80c6a5d); Time taken: 0.143 seconds
INFO : Executing command(queryId=hive_20210214215442_d9ef7191-04ce-4b07-9274-bd6fa80c6a5d): insert into citation_data (select clientnum_id as oci, attrition_flag_id as citing, customer_age as cited, gender as creation, dependent_count as timespan, education_level as journal_sc, marital_status as author_sc from churners)
INFO : Query ID = hive_20210214215442_d9ef7191-04ce-4b07-9274-bd6fa80c6a5d
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
WARN : The session: sessionId=5988fe25-6875-4972-873b-84663a70b447, queueName=null, user=student1_13, doAs=true, isOpen=false, isDefault=false has not been opened
INFO : Subscribed to counters: [] for queryId: hive_20210214215442_d9ef7191-04ce-4b07-9274-bd6fa80c6a5d
INFO : Tez session hasn't been created yet. Opening session
ERROR : Failed to execute tez graph.
java.io.IOException: Previous writer likely failed to write hdfs://bigdataanalytics-head-0.novalocal:8020/user/student1_13/.hiveJars/hive-exec-3.1.0.3.1.4.0-315-effe339ef81326ce093bc0a1516e86d9d189e11126de97212254c005859b949e.jar. Failing because I am unlikely to write too.
at org.apache.hadoop.hive.ql.exec.tez.DagUtils.localizeResource(DagUtils.java:1296) ~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
at org.apache.hadoop.hive.ql.exec.tez.TezSessionState.createJarLocalResource(TezSessionState.java:917) ~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
at org.apache.hadoop.hive.ql.exec.tez.TezSessionState.makeCombinedJarMap(TezSessionState.java:349) ~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
at org.apache.hadoop.hive.ql.exec.tez.TezSessionState.openInternal(TezSessionState.java:418) ~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
at org.apache.hadoop.hive.ql.exec.tez.TezSessionPoolSession.openInternal(TezSessionPoolSession.java:124) ~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]

```
	at
org.apache.hadoop.hive.ql.exec.tez.TezSessionState.open(TezSessionState.java:373)
~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
	at
org.apache.hadoop.hive.ql.exec.tez.TezTask.ensureSessionHasResources(TezTask.java:373
) ~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
	at org.apache.hadoop.hive.ql.exec.tez.TezTask.execute(TezTask.java:200) ~[hive-
exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
	at org.apache.hadoop.hive.ql.exec.Task.executeTask(Task.java:212) ~[hive-
exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
	at org.apache.hadoop.hive.ql.exec.TaskRunner.runSequential(TaskRunner.java:103)
~[hive-exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
	at org.apache.hadoop.hive.ql.Driver.launchTask(Driver.java:2712) ~[hive-
exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
	at org.apache.hadoop.hive.ql.Driver.execute(Driver.java:2383) ~[hive-
exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
	at org.apache.hadoop.hive.ql.Driver.runInternal(Driver.java:2055) ~[hive-
exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
	at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1753) ~[hive-
exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
	at org.apache.hadoop.hive.ql.Driver.run(Driver.java:1747) ~[hive-
exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
	at org.apache.hadoop.hive.ql.reexec.ReExecDriver.run(ReExecDriver.java:157) ~[hive-
exec-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
	at
org.apache.hive.service.cli.operation.SQLOperation.runQuery(SQLOperation.java:226)
~[hive-service-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
	at
org.apache.hive.service.cli.operation.SQLOperation.access$700(SQLOperation.java:87)
~[hive-service-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
	at
org.apache.hive.service.cli.operation.SQLOperation$BackgroundWork$1.run(SQLOperation
.java:324) ~[hive-service-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
	at java.security.AccessController.doPrivileged(Native Method) ~[?:1.8.0_191]
	at javax.security.auth.Subject.doAs(Subject.java:422) ~[?:1.8.0_191]
	at
org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1730)
~[hadoop-common-3.1.1.3.1.4.0-315.jar:?]
	at
org.apache.hive.service.cli.operation.SQLOperation$BackgroundWork.run(SQLOperation.ja
va:342) ~[hive-service-3.1.0.3.1.4.0-315.jar:3.1.0.3.1.4.0-315]
	at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511)
~[?:1.8.0_191]
	at java.util.concurrent.FutureTask.run(FutureTask.java:266) ~[?:1.8.0_191]
	at java.util.concurrent.Executors$RunnableAdapter.call(Executors.java:511)
~[?:1.8.0_191]
	at java.util.concurrent.FutureTask.run(FutureTask.java:266) ~[?:1.8.0_191]
	at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
~[?:1.8.0_191]
	at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
~[?:1.8.0_191]
	at java.lang.Thread.run(Thread.java:748) [?:1.8.0_191]
ERROR : FAILED: Execution Error, return code 1 from
org.apache.hadoop.hive.ql.exec.tez.TezTask
INFO  : Completed executing
command(queryId=hive_20210214215442_d9ef7191-04ce-4b07-9274-bd6fa80c6a5d); Time
taken: 25.106 seconds
Error: Error while processing statement: FAILED: Execution Error, return code 1 from
org.apache.hadoop.hive.ql.exec.tez.TezTask (state=08S01,code=1)
```

2. Заполнить данными из большой таблицы hive_db.citation_data
3. Посмотреть на получившийся размер данных
4. Посчитать count некоторых колонок в разных форматах хранения.
5. Посчитать агрегаты по одной и нескольким колонкам в разных форматах.
6. Сделать выводы о эффективности хранения и компрессии.

```
ekaterina@MacBook-Pro-Ekaterina ~ % cd .ssh
ekaterina@MacBook-Pro-Ekaterina .ssh % ssh -i ../.ssh/id_rsa_mac
student1_13@185.241.193.174
Last login: Wed Feb 10 13:36:33 2021 from
89-64-64-122.dynamic.chello.pl
-bash: warning: setlocale: LC_CTYPE: cannot change locale (UTF-8):
No such file or directory
[student1_13@bigdataanalytics-head-0 ~]$ hdfs dfs -cat hw4/
BankChurners.csv | head
"CLIENTNUM","Attrition_Flag","Customer_Age","Gender","Dependent_co
unt","Education_Level","Marital_Status","Income_Category","Card_Ca
tegory","Months_on_book","Total_Relationship_Count","Months_Inacti
ve_12_mon","Contacts_Count_12_mon","Credit_Limit","Total_Revolving
_Bal","Avg_Open_To_Buy","Total_Amt_Chng_Q4_Q1","Total_Trans_Amt","
Total_Trans_Ct","Total_Ct_Chng_Q4_Q1","Avg_Utilization_Ratio","Nai
ve_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12
_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_1","Na
ive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_1
2_mon_Dependent_count_Education_Level_Months_Inactive_12_mon_2"
768805383,"Existing Customer",45,"M",3,"High
School","Married","$60K -
$80K","Blue",39,5,1,3,12691,777,11914,1.335,1144,42,1.625,0.061,9.
3448e-05,0.99991
818770008,"Existing Customer",49,"F",5,"Graduate","Single","Less
than
$40K","Blue",44,6,1,2,8256,864,7392,1.541,1291,33,3.714,0.105,5.68
61e-05,0.99994
713982108,"Existing Customer",51,"M",3,"Graduate","Married","$80K
-
$120K","Blue",36,4,1,0,3418,0,3418,2.594,1887,20,2.333,0,2.1081e-0
5,0.99998
769911858,"Existing Customer",40,"F",4,"High
School","Unknown","Less than
$40K","Blue",34,3,4,1,3313,2517,796,1.405,1171,20,2.333,0.76,0.000
13366,0.99987
709106358,"Existing
Customer",40,"M",3,"Uneducated","Married","$60K -
$80K","Blue",21,5,1,0,4716,0,4716,2.175,816,28,2.5,0,2.1676e-05,0.
99998
713061558,"Existing Customer",44,"M",2,"Graduate","Married","$40K
-
```

$60K","Blue",36,3,1,2,4010,1247,2763,1.376,1088,24,0.846,0.311,5.5
077e-05,0.99994
810347208,"Existing Customer",51,"M",4,"Unknown","Married","$120K
+","Gold",46,6,1,3,34516,2264,32252,1.975,1330,31,0.722,0.066,0.00
012303,0.99988
818906208,"Existing Customer",32,"M",0,"High
School","Unknown","$60K -
$80K","Silver",27,2,2,2,29081,1396,27685,2.204,1538,36,0.714,0.048
,8.5795e-05,0.99991
710930508,"Existing Customer",37,"M",3,"Uneducated","Single","$60K
-
$80K","Blue",36,5,2,0,22352,2517,19835,3.355,1350,24,1.182,0.113,4
.4796e-05,0.99996
cat: Unable to write to output stream.