1. SQOOP

Apache Sqoop — это инструмент, предназначенный для передачи данных между Hadoop и реляционными базами данных или мэйнфреймами.

Запрашиваю список БД в PosrgeSQL, чтобы в дальнейшем выполнить импорт БД PSQL в hive

```
[student3_16@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ sqoop list-databases --connect jdbc:postgresql://10.0.0.7/lesson5 --username exporter --password exporter_pass
Warning: /usr/hdp/3.1.4.0-315/accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/hdp/3.1.4.0-315/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/hdp/3.1.4.0-315/phoenix/phoenix-5.0.0.3.1.4.0-315-server.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/hdp/3.1.4.0-315/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
21/04/25 16:37:59 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7.3.1.4.0-315
21/04/25 16:37:59 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
21/04/25 16:37:59 INFO manager.SqlManager: Using default fetchSize of 1000
postgres
ambari
template1
template0
hive
oozie
lesson5
student21
[student3_16@bigdataanalytics2-head-shdpt-v31-1-0 ~]$
```

Запрашиваю список таблиц

```
[student3_16@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ sqoop list-tables --connect jdbc:postgresql://10.0.0.7/lesson5 --username exporter --password exporter_pass
Warning: /usr/hdp/3.1.4.0-315/accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/hdp/3.1.4.0-315/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/hdp/3.1.4.0-315/phoenix/phoenix-5.0.0.3.1.4.0-315-server.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/hdp/3.1.4.0-315/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
21/04/25 16:44:14 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7.3.1.4.0-315
21/04/25 16:44:14 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
21/04/25 16:44:14 INFO manager.SqlManager: Using default fetchSize of 1000
test
[student3_16@bigdataanalytics2-head-shdpt-v31-1-0 ~]$
```

Выполняю команду импорта с помощью sqoop из БД PSQL в Hive использую параметр -m 1

```
[student3_16@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ sqoop import --connect jdbc:postgresql://10.0.0.7/lesson5 --username exporter --password exporter_pass --table test --hive-import --hive-database student3_16 --hive-table test -m 1
Warning: /usr/hdp/3.1.4.0-315/accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/hdp/3.1.4.0-315/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/hdp/3.1.4.0-315/phoenix/phoenix-5.0.0.3.1.4.0-315-server.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/hdp/3.1.4.0-315/hive/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
21/04/25 16:46:34 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7.3.1.4.0-315
21/04/25 16:46:34 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
21/04/25 16:46:34 INFO tool.BaseSqoopTool: Using Hive-specific delimiters for output. You can override
21/04/25 16:46:34 INFO tool.BaseSqoopTool: delimiters with --fields-terminated-by, etc.
21/04/25 16:46:35 INFO manager.SqlManager: Using default fetchSize of 1000
21/04/25 16:46:35 INFO tool.CodeGenTool: Beginning code generation
21/04/25 16:46:35 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM "test" AS t LIMIT 1
21/04/25 16:46:35 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/hdp/3.1.4.0-315/hadoop-mapreduce
21/04/25 16:46:37 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-student3_16/compile/992170afc5a1cf6d4b758091ef4150c4/test.jar
21/04/25 16:46:37 WARN manager.PostgresqlManager: It looks like you are importing from postgresql.
21/04/25 16:46:37 WARN manager.PostgresqlManager: This transfer can be faster! Use the --direct
21/04/25 16:46:37 WARN manager.PostgresqlManager: option to exercise a postgresql-specific fast path.
21/04/25 16:46:37 INFO mapreduce.ImportJobBase: Beginning import of test
21/04/25 16:46:39 INFO client.RMProxy: Connecting to ResourceManager at bigdataanalytics2-head-shdpt-v31-1-0.novalocal/10.0.0.7:8050
21/04/25 16:46:39 INFO client.AHSProxy: Connecting to Application History server at bigdataanalytics2-head-shdpt-v31-1-0.novalocal/10.0.0.7:10200
21/04/25 16:46:40 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /user/student3_16/.staging/job_1618415779628_0964
21/04/25 16:46:42 INFO db.DBInputFormat: Using read commited transaction isolation
21/04/25 16:46:42 INFO mapreduce.JobSubmitter: number of splits:1
21/04/25 16:46:43 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1618415779628_0964
21/04/25 16:46:43 INFO mapreduce.JobSubmitter: Executing with tokens: []
21/04/25 16:46:43 INFO conf.Configuration: found resource resource-types.xml at file:/etc/hadoop/3.1.4.0-315/0/resource-types.xml
21/04/25 16:46:43 INFO impl.YarnClientImpl: Submitted application application_1618415779628_0964
21/04/25 16:46:43 INFO mapreduce.Job: The url to track the job: http://bigdataanalytics2-head-shdpt-v31-1-0.novalocal:8888/proxy/application_1618415779628_0964/
21/04/25 16:46:43 INFO mapreduce.Job: Running job: job_1618415779628_0964
21/04/25 16:46:50 INFO mapreduce.Job: Job job_1618415779628_0964 running in uber mode : false
21/04/25 16:46:50 INFO mapreduce.Job:  map 0% reduce 0%
21/04/25 16:46:56 INFO mapreduce.Job:  map 100% reduce 0%
21/04/25 16:51:52 INFO mapreduce.Job: Job job_1618415779628_0964 completed successfully
21/04/25 16:51:52 INFO mapreduce.Job: Counters: 32
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=243957
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=85
                HDFS: Number of bytes written=19
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=14852
```

Захожу Hive в БД student3_16 и проверяю содержимое таблицы Test

Чтобы удостовериться что в импортируемой базе 2 записи, захожу в PSQL и проверяю содержимое таблицы test

```
[student3_16@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ psql -d lesson5 -U exporter
psql (9.2.24, server 9.6.21)
WARNING: psql version 9.2, server version 9.6.
         Some psql features might not work.
Type "help" for help.

lesson5=> select * from test limit 10;
 id |   str
----+---------
  1 | myText
  2 | myText2
(2 rows)

lesson5=>
```

Захожу в hdfs dfs и проверяю содержимое папки test. Наличие файла _SUCCESS говорит об успешном импорте БД из PSQL в Hive

```
[student3_16@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ hdfs dfs -ls test
Found 2 items
-rw-r--r--   2 student3_16 student3_16          0 2021-04-25 16:46 test/_SUCCESS
-rw-r--r--   2 student3_16 student3_16         19 2021-04-25 16:46 test/part-m-00000
[student3_16@bigdataanalytics2-head-shdpt-v31-1-0 ~]$
```

Удаляю папку test в hdfs

```
[student3_16@bigdataanalytics2-head-shdpt-v31-1-0 ~]$ hdfs dfs -rm -r -skipTrash test
Deleted test
[student3_16@bigdataanalytics2-head-shdpt-v31-1-0 ~]$
```

Удаляю таблицу test в Hive

```
Beeline version 3.1.0.3.1.4.0-315 by Apache Hive
0: jdbc:hive2://bigdataanalytics2-worker-shdp> use student3_16;
INFO  : Compiling command(queryId=hive_20210425170846_89bc9d38-9748-4321-9723-2f811397ccdd): use student3_16
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20210425170846_89bc9d38-9748-4321-9723-2f811397ccdd); Time taken: 0.004 seconds
INFO  : Executing command(queryId=hive_20210425170846_89bc9d38-9748-4321-9723-2f811397ccdd): use student3_16
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20210425170846_89bc9d38-9748-4321-9723-2f811397ccdd); Time taken: 0.012 seconds
INFO  : OK
No rows affected (0.124 seconds)
0: jdbc:hive2://bigdataanalytics2-worker-shdp> drop table test;
INFO  : Compiling command(queryId=hive_20210425170856_84fdbefa-4935-4820-b84a-618c62115551): drop table test
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:null, properties:null)
INFO  : Completed compiling command(queryId=hive_20210425170856_84fdbefa-4935-4820-b84a-618c62115551); Time taken: 0.041 seconds
INFO  : Executing command(queryId=hive_20210425170856_84fdbefa-4935-4820-b84a-618c62115551): drop table test
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hive_20210425170856_84fdbefa-4935-4820-b84a-618c62115551); Time taken: 0.115 seconds
INFO  : OK
No rows affected (0.191 seconds)
0: jdbc:hive2://bigdataanalytics2-worker-shdp>
```

Выполняю команду импорта с помощью sqoop из БД PSQL в Hive использую параметр –split-by id

Захожу в hdfs dfs и проверяю содержимое папки test. Наличие файла _SUCCESS говорит об успешном импорте БД из PSQL в Hive. Т.к. в таблице БД PSQL было 2 записи (2 id), то соответственно и 2 файла part-m-

Захожу Hive в БД student3_16 и проверяю содержимое таблицы Test

2. FLUME

Apache Flume – распределенная и высоконадежная система для эффективного сбора, агрегации и сохранения больших объемов логов из множества различных источников в централизованное хранилище данных

Подготавливаю скрипт heartbeat.sh в директории flm и файл конфигурации flm.conf

Выполняю команду импорта с помощью flume с использованием скрипта и конфигурационного файла, созданных на предыдущем этапе

Проверяю наличие созданных логов

Создаю таблицу в hive

```
create external table flm_logs (first string, second string, third string,
fourth string, fifth string)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ';'
STORED AS INPUTFORMAT 'org.apache.hadoop.mapred.SequenceFileInputFormat'
OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io.HiveSequenceFileOutputFormat'
location '/user/student3_16/flm'
;
```


Проверяю содержимое таблицы

```
select * from flm_logs;
```

```
0: jdbc:hive2://bigdataanalytics2-worker-shdp> select * from flm_logs;
INFO  : Compiling command(queryId=hive_20210426151921_4c7c0d67-9f8f-4559-8379-d794e408eb44): select * from flm_logs
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:flm_logs.first, type:string, comment:null), FieldSchema(name:flm_logs
, type:string, comment:null), FieldSchema(name:flm_logs.fourth, type:string, comment:null), FieldSchema(name:flm_logs.fifth, type:string,
INFO  : Completed compiling command(queryId=hive_20210426151921_4c7c0d67-9f8f-4559-8379-d794e408eb44); Time taken: 0.152 seconds
INFO  : Executing command(queryId=hive_20210426151921_4c7c0d67-9f8f-4559-8379-d794e408eb44): select * from flm_logs
INFO  : Completed executing command(queryId=hive_20210426151921_4c7c0d67-9f8f-4559-8379-d794e408eb44); Time taken: 0.001 seconds
INFO  : OK
```

flm_logs.first	flm_logs.second	flm_logs.third	flm_logs.fourth	flm_logs.fifth
I live for 0 seconds	2021-04-25:17.59.56	2021-04-25:17.59.56	I did it 1	NULL
0m0.002s 0m0.007s	NULL	NULL	NULL	NULL
0m0.011s 0m0.004s	NULL	NULL	NULL	NULL
I live for 10 seconds	2021-04-25:17.59.56	2021-04-25:17.59.56	I did it 2	NULL
0m0.005s 0m0.007s	NULL	NULL	NULL	NULL
0m0.015s 0m0.014s	NULL	NULL	NULL	NULL
I live for 20 seconds	2021-04-25:17.59.56	2021-04-25:17.59.56	I did it 3	NULL
0m0.006s 0m0.008s	NULL	NULL	NULL	NULL
0m0.024s 0m0.019s	NULL	NULL	NULL	NULL
I live for 30 seconds	2021-04-25:17.59.56	2021-04-25:17.59.56	I did it 4	NULL
0m0.006s 0m0.010s	NULL	NULL	NULL	NULL
0m0.028s 0m0.024s	NULL	NULL	NULL	NULL
I live for 40 seconds	2021-04-25:17.59.56	2021-04-25:17.59.56	I did it 5	NULL
0m0.007s 0m0.011s	NULL	NULL	NULL	NULL
0m0.033s 0m0.033s	NULL	NULL	NULL	NULL
I live for 50 seconds	2021-04-25:17.59.56	2021-04-25:17.59.56	I did it 6	NULL
0m0.007s 0m0.013s	NULL	NULL	NULL	NULL
0m0.037s 0m0.041s	NULL	NULL	NULL	NULL
I live for 60 seconds	2021-04-25:17.59.56	2021-04-25:17.59.56	I did it 7	NULL
0m0.009s 0m0.014s	NULL	NULL	NULL	NULL
0m0.043s 0m0.049s	NULL	NULL	NULL	NULL
I live for 70 seconds	2021-04-25:17.59.56	2021-04-25:17.59.56	I did it 8	NULL
0m0.009s 0m0.016s	NULL	NULL	NULL	NULL
0m0.050s 0m0.054s	NULL	NULL	NULL	NULL
I live for 80 seconds	2021-04-25:17.59.56	2021-04-25:17.59.56	I did it 9	NULL
0m0.010s 0m0.017s	NULL	NULL	NULL	NULL
0m0.055s 0m0.063s	NULL	NULL	NULL	NULL
I live for 90 seconds	2021-04-25:17.59.56	2021-04-25:17.59.56	I did it 10	NULL
0m0.010s 0m0.018s	NULL	NULL	NULL	NULL
0m0.060s 0m0.068s	NULL	NULL	NULL	NULL
I live for 100 seconds	2021-04-25:17.59.56	2021-04-25:17.59.56	I did it 11	NULL
0m0.012s 0m0.018s	NULL	NULL	NULL	NULL
0m0.065s 0m0.075s	NULL	NULL	NULL	NULL
I live for 110 seconds	2021-04-25:17.59.56	2021-04-25:17.59.56	I did it 12	NULL
0m0.013s 0m0.020s	NULL	NULL	NULL	NULL
0m0.070s 0m0.082s	NULL	NULL	NULL	NULL
I live for 120 seconds	2021-04-25:17.59.56	2021-04-25:17.59.56	I did it 13	NULL
0m0.013s 0m0.022s	NULL	NULL	NULL	NULL
0m0.079s 0m0.086s	NULL	NULL	NULL	NULL
I live for 130 seconds	2021-04-25:17.59.56	2021-04-25:17.59.56	I did it 14	NULL
0m0.013s 0m0.024s	NULL	NULL	NULL	NULL
0m0.089s 0m0.088s	NULL	NULL	NULL	NULL
I live for 140 seconds	2021-04-25:17.59.56	2021-04-25:17.59.56	I did it 15	NULL
0m0.013s 0m0.027s	NULL	NULL	NULL	NULL
0m0.093s 0m0.099s	NULL	NULL	NULL	NULL

```
45 rows selected (0.318 seconds)
```