

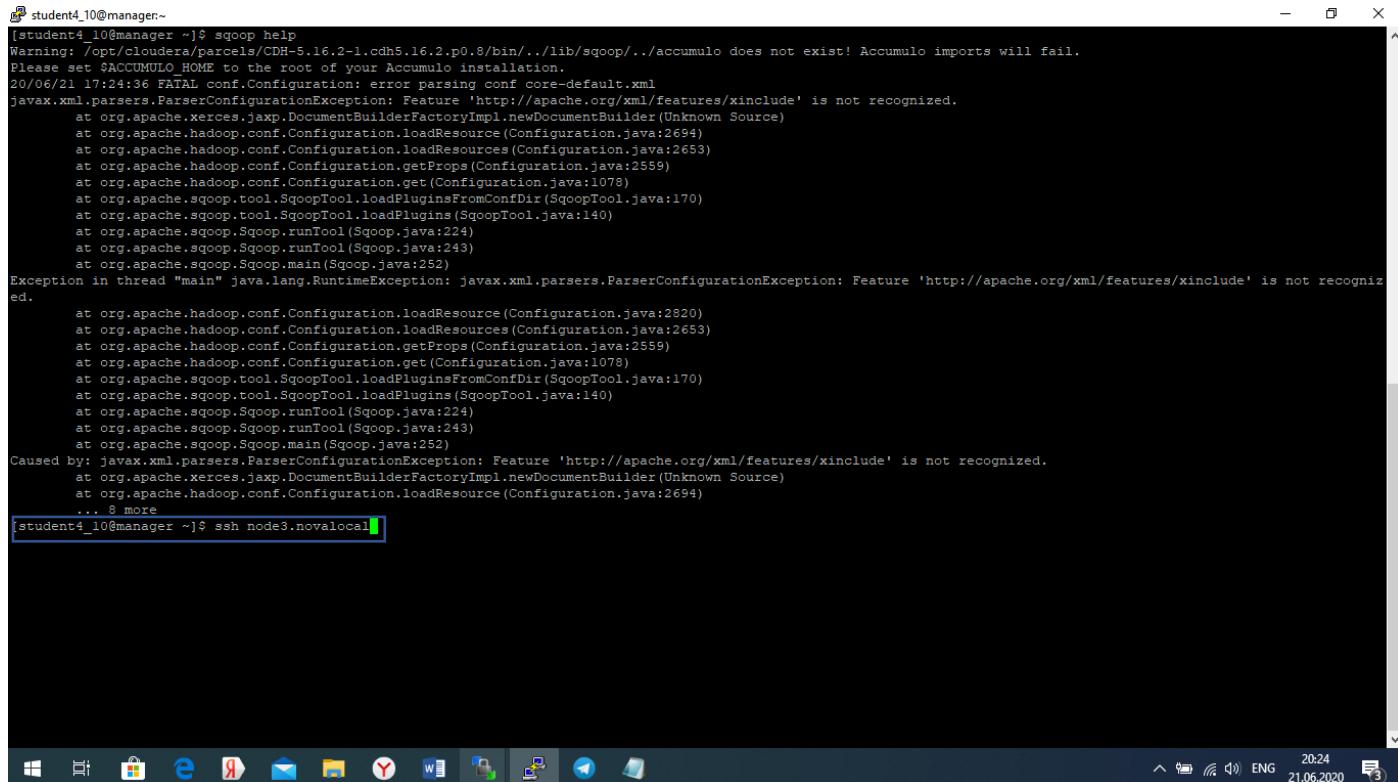
1. Для части по SQOOP

Провести импорт таблицы из вашего сервера БД в Hadoop с использованием SQOOP в любых двух вариантах из перечисленных ниже.

- a. в Hive-таблицу (--hive-import)
- b. в HDFS в формате avro (--as-avrodatafile)
- c. в HDFS в формате sequencefile (--as-sequencefile)

Если у вас нет своего сервера то можно использовать тот Postgres, который я показал на лекции. Пароль expoter_pass

Посмотрим при помощи SQOOP содержимое в PosgreSQL.



```
[student4_10@manager ~]$ sqoop help
Warning: /opt/cloudera/parcels/CDH-5.16.2-1.cdh5.16.2.p0.8/bin/../lib/sqoop ./accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
20/06/21 17:24:36 FATAL conf.Configuration: error parsing conf/core-default.xml
javax.xml.parsers.ParserConfigurationException: Feature 'http://apache.org/xml/features/xinclud
e' is not recognized.
    at org.apache.xerces.jaxp.DocumentBuilderFactoryImpl.newDocumentBuilder(Unknown Source)
    at org.apache.hadoop.conf.Configuration.loadResource(Configuration.java:2694)
    at org.apache.hadoop.conf.Configuration.loadResources(Configuration.java:2653)
    at org.apache.hadoop.conf.Configuration.getProps(Configuration.java:2559)
    at org.apache.hadoop.conf.Configuration.get(Configuration.java:1078)
    at org.apache.sqoop.tool.SqoopTool.loadPluginsFromConfDir(SqoopTool.java:170)
    at org.apache.sqoop.tool.SqoopTool.loadPlugins(SqoopTool.java:140)
    at org.apache.sqoop.Sqoop.runTool(Sqoop.java:224)
    at org.apache.sqoop.Sqoop.runTool(Sqoop.java:243)
    at org.apache.sqoop.main(Sqoop.java:252)
Exception in thread "main" java.lang.RuntimeException: javax.xml.parsers.ParserConfigurationException: Feature 'http://apache.org/xml/features/xinclud
e' is not recognized.
    at org.apache.hadoop.conf.Configuration.loadResource(Configuration.java:2820)
    at org.apache.hadoop.conf.Configuration.loadResources(Configuration.java:2653)
    at org.apache.hadoop.conf.Configuration.getProps(Configuration.java:2559)
    at org.apache.hadoop.conf.Configuration.get(Configuration.java:1078)
    at org.apache.sqoop.tool.SqoopTool.loadPluginsFromConfDir(SqoopTool.java:170)
    at org.apache.sqoop.tool.SqoopTool.loadPlugins(SqoopTool.java:140)
    at org.apache.sqoop.Sqoop.runTool(Sqoop.java:224)
    at org.apache.sqoop.Sqoop.runTool(Sqoop.java:243)
    at org.apache.sqoop.main(Sqoop.java:252)
Caused by: javax.xml.parsers.ParserConfigurationException: Feature 'http://apache.org/xml/features/xinclud
e' is not recognized.
    at org.apache.xerces.jaxp.DocumentBuilderFactoryImpl.newDocumentBuilder(Unknown Source)
    at org.apache.hadoop.conf.Configuration.loadResource(Configuration.java:2694)
... 8 more
[student4_10@manager ~]$ ssh node3.novalocal
```

Sqoop Help не работает, через ssh перехожу на node3.novalocal

```
student4_10@node3:~  
at org.apache.sqoop.Sqoop.runTool(Sqoop.java:224)  
at org.apache.sqoop.Sqoop.runTool(Sqoop.java:243)  
at org.apache.sqoop.Sqoop.main(Sqoop.java:252)  
Exception in thread "main" java.lang.RuntimeException: javax.xml.parsers.ParserConfigurationException: Feature 'http://apache.org/xml/features/xincluder' is not recognized.  
at org.apache.hadoop.conf.Configuration.loadResource(Configuration.java:2820)  
at org.apache.hadoop.conf.Configuration.loadResources(Configuration.java:2653)  
at org.apache.hadoop.conf.Configuration.getProps(Configuration.java:2559)  
at org.apache.hadoop.conf.Configuration.get(Configuration.java:1078)  
at org.apache.sqoop.tool.SqoopTool.loadPluginsFromConfDir(SqoopTool.java:170)  
at org.apache.sqoop.tool.SqoopTool.loadPlugins(SqoopTool.java:140)  
at org.apache.sqoop.Sqoop.runTool(Sqoop.java:224)  
at org.apache.sqoop.Sqoop.main(Sqoop.java:243)  
at org.apache.sqoop.Sqoop.main(Sqoop.java:252)  
Caused by: javax.xml.parsers.ParserConfigurationException: Feature 'http://apache.org/xml/features/xincluder' is not recognized.  
at org.apache.xerces.jaxp.DocumentBuilderFactoryImpl.newDocumentBuilder(Unknown Source)  
at org.apache.hadoop.conf.Configuration.loadResource(Configuration.java:2694)  
... 8 more  
[student4_10@manager ~]$ ssh node3.novalocal  
Last login: Sat Jun 20 18:27:03 2020 from manager.novalocal  
[student4_10@node3 ~]$ sqoop help  
Warning: /opt/cloudera/parcels/CDH-5.16.2-1.cdh5.16.2.p0.8/bin/../lib/sqoop/..../accumulo does not exist! Accumulo imports will fail.  
Please set $ACCUMULO_HOME to the root of your Accumulo installation.  
20/06/21 17:26:38 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.16.2  
usage: sqoop COMMAND [ARGS]  
  
Available commands:  
codegen      Generate code to interact with database records  
create-hive-table Import a table definition into Hive  
eval          Evaluate a SQL statement and display the results  
export        Export an HDFS directory to a database table  
help          List available commands  
import        Import a table from a database to HDFS  
import-all-tables Import tables from a database to HDFS  
import-mainframe Import datasets from a mainframe server to HDFS  
job           Work with saved jobs  
list-databases List available databases on a server  
list-tables   List available tables in a database  
merge         Merge results of incremental imports  
metastore     Run a standalone Sqoop metastore  
version       Display version information  
  
See 'sqoop help COMMAND' for information on a specific command.  
[student4_10@node3 ~]$
```

работает !

Проверим таблицы в базе pg_db

```
student4_10@node3:~$ sqoop list-databases --connect jdbc:postgresql://node3.novalocal/pg_db --username exporter --password exporter_pass  
Warning: /opt/cloudera/parcels/CDH-5.16.2-1.cdh5.16.2.p0.8/bin/../lib/sqoop/..../accumulo does not exist! Accumulo imports will fail.  
Please set $ACCUMULO_HOME to the root of your Accumulo installation.  
20/06/20 19:57:50 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.16.2  
20/06/20 19:57:50 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.  
20/06/20 19:57:50 INFO manager.SqlManager: Using default fetchSize of 1000  
templatel  
template0  
postgres  
pg_db  
[student4_10@node3 ~]$ sqoop list-tables --connect jdbc:postgresql://node3.novalocal/pg_db --username exporter --password exporter_pass  
Warning: /opt/cloudera/parcels/CDH-5.16.2-1.cdh5.16.2.p0.8/bin/../lib/sqoop/..../accumulo does not exist! Accumulo imports will fail.  
Please set $ACCUMULO_HOME to the root of your Accumulo installation.  
20/06/20 19:58:06 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.16.2  
20/06/20 19:58:06 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.  
20/06/20 19:58:06 INFO manager.SqlManager: Using default fetchSize of 1000  
character  
character_work  
paragraph  
sales_large  
wordform  
work  
chapter  
[student4_10@node3 ~]$
```

sqoop list-tables --connect jdbc:postgresql://node3.novalocal/pg_db --username exporter --password exporter_pass

Скопируем Таблицу Work в локальную папку

```
[student4_10@node3 ~]$ sqoop import --connect jdbc:postgresql://node3.novalocal/pg_db --username exporter --password exporter_pass --table work --target-dir /user/student4_10/hw_5/work --as-avrodatafile
Warning: /opt/cloudera/parcels/CDH-5.16.2-1.cdh5.16.2.p0.8/bin/../lib/sqoop/.../accumulo does not exist! Accumulo imports will fail.
Please set $ACUMULO_HOME to the root of your Accumulo installation.
20/06/21 18:00:45 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.16.2
20/06/21 18:00:45 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
20/06/21 18:00:45 INFO manager.SqlManager: Using default fetchSize of 1000
20/06/21 18:00:45 INFO tool.CodeGenTool: Beginning code generation
20/06/21 18:00:45 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM "work" AS t LIMIT 1
20/06/21 18:00:45 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce
Note: /tmp/sqoop-student4_10/compile/618b15e8a75495be294c889b3e6f1766/work.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
20/06/21 18:00:47 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-student4_10/compile/618b15e8a75495be294c889b3e6f1766/work.jar
20/06/21 18:00:47 WARN manager.PostgresqlManager: It looks like you are importing from postgresql.
20/06/21 18:00:47 WARN manager.PostgresqlManager: This transfer can be faster! Use the --direct
20/06/21 18:00:47 WARN manager.PostgresqlManager: option to exercise a postgresql-specific fast path.
20/06/21 18:00:47 INFO mapreduce.ImportJobBase: Beginning import of work
20/06/21 18:00:48 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
20/06/21 18:00:49 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM "work" AS t LIMIT 1
20/06/21 18:00:49 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM "work" AS t LIMIT 1
20/06/21 18:00:49 INFO mapreduce.DataDrivenImportJob: Writing Avro schema file: /tmp/sqoop-student4_10/compile/618b15e8a75495be294c889b3e6f1766/work.avsc
20/06/21 18:00:49 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
```

Проверим папку /user/student4_10

```
[student4_10@node3 ~]$ hdfs dfs -ls /user/student4_10
drwx-----  - hdfs      supergroup          0 2020-06-01 16:11 /user/hdfs
drwxr-xr-x  - mapred    supergroup          0 2019-11-18 19:33 /user/history
drwxr-xr-x  - hive       hive              0 2019-11-18 19:57 /user/hive
drwxrwxr-x  - hue        hue               0 2019-12-08 22:25 /user/hue
drwxr-xr-x  - instructor instructor         0 2020-03-17 19:35 /user/instructor
drwxr-xr-x  - spark      spark             0 2020-01-19 20:20 /user/spark
drwxr-xr-x  - student4_6 student4_6        0 2020-05-23 20:27 /user/student4_6
drwxr-xr-x  - student   student            0 2019-12-02 15:02 /user/student
drwxr-xr-x  - student4_1 student4_1        0 2020-06-10 13:44 /user/student4_1
drwxr-xr-x  - student4_10 student4_10       0 2020-06-21 18:01 /user/student4_10
drwxr-xr-x  - student4_11 student4_11       0 2020-06-07 06:05 /user/student4_11
drwxr-xr-x  - student4_12 student4_12       0 2020-06-21 15:49 /user/student4_12
drwxr-xr-x  - student4_13 student4_13       0 2020-06-16 11:35 /user/student4_13
drwxr-xr-x  - student4_14 student4_14       0 2020-06-16 08:47 /user/student4_14
drwxr-xr-x  - student4_15 student4_15       0 2020-06-15 20:09 /user/student4_15
drwxr-xr-x  - student4_16 student4_16       0 2020-06-07 06:08 /user/student4_16
drwxr-xr-x  - student4_17 student4_17       0 2020-06-07 06:10 /user/student4_17
drwxr-xr-x  - student4_18 student4_18       0 2020-06-07 06:10 /user/student4_18
drwxr-xr-x  - student4_19 student4_19       0 2020-06-07 06:10 /user/student4_19
drwxr-xr-x  - student4_2 student4_2         0 2020-06-06 20:47 /user/student4_2
drwxr-xr-x  - student4_20 student4_20       0 2020-06-07 06:11 /user/student4_20
drwxr-xr-x  - student4_3 student4_3         0 2020-06-21 12:22 /user/student4_3
drwxr-xr-x  - student4_4 student4_4         0 2020-06-14 03:36 /user/student4_4
drwxr-xr-x  - student4_5 student4_5         0 2020-06-08 15:26 /user/student4_5
drwxr-xr-x  - student4_6 student4_6         0 2020-06-07 13:34 /user/student4_6
drwxr-xr-x  - student4_7 student4_7         0 2020-05-31 12:55 /user/student4_7
drwxr-xr-x  - student4_8 student4_8         0 2020-05-23 20:27 /user/student4_8
drwxr-xr-x  - student4_9 student4_9         0 2020-06-05 12:49 /user/student4_9
drwxr-xr-x  - student4_3 student4_3         0 2020-05-19 19:19 /user/student4_3
[student4_10@node3 ~]$ hdfs dfs -ls /user/student4_10/
Found 4 items
drwx-----  - student4_10 student4_10        0 2020-05-27 11:00 /user/student4_10/.Trash
drwx-----  - student4_10 student4_10        0 2020-06-21 18:01 /user/student4_10/.staging
drwxr-xr-x  - student4_10 student4_10        0 2020-06-05 17:02 /user/student4_10/datasets
drwxr-xr-x  - student4_10 student4_10        0 2020-06-21 18:01 /user/student4_10/hw_5
[student4_10@node3 ~]$ hdfs dfs -ls /user/student4_10/hw_5/
Found 1 items
drwxr-xr-x  - student4_10 student4_10        0 2020-06-21 18:01 /user/student4_10/hw_5/work
[student4_10@node3 ~]$
```

Скопируем схему структуры таблицы с локальной директории через команду COPYFROMLOCAL

```

student4_10@node3:~$ hdfs dfs -copyFromLocal work.avsc /user/student4_10/hw_5/
[student4_10@node3 ~]$ hdfs dfs -cat /user/student4_10/hw_5/
cat: `/user/student4_10/hw_5': Is a directory
[student4_10@node3 ~]$ hdfs dfs -ls /user/student4_10/hw_5/
Found 2 items
drwxr-xr-x - student4_10 student4_10 0 2020-06-21 18:01 /user/student4_10/hw_5/work
-rw-r--r-- 3 student4_10 student4_10 1368 2020-06-21 18:21 /user/student4_10/hw_5/work.avsc
[student4_10@node3 ~]$ ll
total 32
-rw-rw-r--. 1 student4_10 student4_10 1368 Jun 21 18:00 work.avsc
-rw-rw-r--. 1 student4_10 student4_10 25203 Jun 21 18:00 work.java
[student4_10@node3 ~]$ ls
work.avsc work.java
[student4_10@node3 ~]$ hive
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=512M; support was removed in 8.0
Java HotSpot(TM) 64-Bit Server VM warning: Using incremental CMS is deprecated and will likely be removed in a future release
Java HotSpot(TM) 64-Bit Server VM warning: ignoring option MaxPermSize=512M; support was removed in 8.0

Logging initialized using configuration in jar:file:/opt/cloudera/parcels/CDH-5.16.2-1.cdh5.16.2.p0.8/jars/hive-common-1.1.0-cdh5.16.2
.jar!/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> select* from student4_10.work limit 10;
OK
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 12night | Twelfth Night | Twelfth Night, Or What You Will | 1599 | c | NULL | Moby | 19837 | 1031 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| alliswell | All's Well That Ends Well | All's Well That Ends Well | 1602 | c | NULL | Moby | 22997 | 1025 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| antonycleo | Antony and Cleopatra | Antony and Cleopatra | 1606 | t | NULL | Moby | 24905 | 1344 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| asyoulikeit | As You Like It | As You Like It | 1599 | c | NULL | Gutenberg | 21690 | 872 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| comedyerrors | Comedy of Errors | The Comedy of Errors | 1589 | c | NULL | Moby | 14692 | 661 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| coriolanus | Coriolanus | Coriolanus | 1607 | t | NULL | Moby | 27577 | 1226 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| cymbeline | Cymbeline | Cymbeline, King of Britain | 1609 | h | NULL | Moby | 27565 | 971 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| hamlet | Hamlet | Tragedy of Hamlet, Prince of Denmark, The | 1600 | t | NULL | Gutenberg | 30558 | 1275 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| henry4p1 | Henry IV, Part I | History of Henry IV, Part I | 1597 | h | NULL | Moby | 24579 | 884 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| henry4p2 | Henry IV, Part II | History of Henry IV, Part II | 1597 | h | NULL | Gutenberg | 25692 | 1013 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+
Time taken: 4.499 seconds, Fetched: 10 row(s)
hive> clear
>

```

Создадим Таблицу с путями для схемы work.avsc и файла таблицы work

```

CREATE EXTERNAL TABLE student4_10.work
STORED AS AVRO
LOCATION '/user/student4_10/hw_5/work'
TBLPROPERTIES ('avro.schema.url'='/user/student4_10/hw_5/work.avsc');

```

The screenshot shows the Hue Editor interface. At the top, there's a browser-like header with 'Hue - Editor' and a URL '89.208.221.132:8888/hue/editor?editor=6427'. Below it is the Hue logo and a search bar. A sidebar on the left has a tree view with one node expanded, showing 'student4_10.work'. The main area contains a query editor with the SQL command 'SELECT * from student4_10.work limit 10;'. Below the editor is a log window showing compilation and semantic analysis messages. The bottom section displays the results of the query in a table format.

work.workid	work.title	work.longtitle	work.year	work.genretype	work.notes	work.source	work.totally	
1	12night	Twelfth Night	Twelfth Night, Or What You Will	1599	c	NULL	Moby	19837
2	allswell	All's Well That Ends Well	All's Well That Ends Well	1602	c	NULL	Moby	22997
3	antonycleo	Antony and Cleopatra	Antony and Cleopatra	1606	t	NULL	Moby	24905
4	asyoulikeit	As You Like It	As You Like It	1599	c	NULL	Gutenberg	21690
5	comedyerrors	Comedy of Errors	The Comedy of Errors	1589	c	NULL	Moby	14692
6								

Копируем Paragraph Таблицу в формате Parquet, но сначала создадим таблицу так как в Parquet нельзя импортировать таким же способом как avro таблицу , без готовой структуры таблицы в базе данных, так как как не импортируется файл со схемой таблицы

Поэтому через Sqoop проверим схему команды

Схема для таблицы work

```
sqoop import --m 1 --connect jdbc:postgresql://node3.novalocal/pg_db --username exporter --password
exporter_pass --query "SELECT column_name, DATA_TYPE FROM INFORMATION_SCHEMA.Columns WHERE
table_name='paragraph' AND \$CONDITIONS" --target-dir '/user/student4_10/hw_5_1/work/'
```

Схема для таблицы paragraph

```
sqoop import --m 1 --connect jdbc:postgresql://node3.novalocal/pg_db --username exporter --password
exporter_pass --query "SELECT column_name, DATA_TYPE FROM INFORMATION_SCHEMA.Columns WHERE
table_name='paragraph' AND \$CONDITIONS" --target-dir '/user/student4_10/hw_5_1/paragraph/'
```

```
[student4_10@node3 ~]$ sqoop import --m 1 --connect jdbc:postgresql://node3.novalocal/pg_db --username exporter --password exporter_password --query "SELECT column_name, DATA_TYPE FROM INFORMATION_SCHEMA.Columns WHERE table_name='paragraph' AND \$CONDITIONS" --target-dir '/user/student4_10/hw_5_1/paragraph'"  
Warning: /opt/cloudera/parcels/CDH-5.16.2-1.cdh5.16.2.p0.8/bin/../lib/sqoop.../accumulo does not exist! Accumulo imports will fail.  
Please set $ACCUMULO_HOME to the root of your Accumulo installation.  
20/06/22 00:09:03 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.16.2  
20/06/22 00:09:03 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.  
20/06/22 00:09:03 INFO manager.SqlManager: Using default fetchSize of 1000  
20/06/22 00:09:03 INFO tool.CodeGenTool: Beginning code generation  
20/06/22 00:09:03 INFO manager.SqlManager: Executing SQL statement: SELECT column_name, DATA_TYPE FROM INFORMATION_SCHEMA.Columns WHERE table_name='paragraph' AND (1 = 0)  
20/06/22 00:09:03 INFO manager.SqlManager: Executing SQL statement: SELECT column_name, DATA_TYPE FROM INFORMATION_SCHEMA.Columns WHERE table_name='paragraph' AND (1 = 0)  
20/06/22 00:09:03 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce  
Note: /tmp/sqoop-student4_10/compile/92a7a9f36ab4d42e684ae923c230ef/QueryResult.java uses or overrides a deprecated API.  
Note: Recompile with -Xlint:deprecation for details.  
20/06/22 00:09:05 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-student4_10/compile/92a7a9f36ab4d42e684ae923c230ef/QueryResult.jar  
20/06/22 00:09:05 INFO mapreduce.ImportJobBase: Beginning query import.  
20/06/22 00:09:05 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar  
20/06/22 00:09:06 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps  
20/06/22 00:09:06 INFO client.RMProxy: Connecting to ResourceManager at manager.novalocal/89.208.221.132:8032  
20/06/22 00:09:12 INFO db.DBInputFormat: Using read committed transaction isolation  
20/06/22 00:09:12 INFO mapreduce.JobSubmitter: number of splits:1  
20/06/22 00:09:12 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1592246524221_0074  
20/06/22 00:09:13 INFO impl.YarnClientImpl: Submitted application application_1592246524221_0074  
20/06/22 00:09:13 INFO mapreduce.Job: The url to track the job: http://manager.novalocal:8088/proxy/application_1592246524221_0074/  
20/06/22 00:09:13 INFO mapreduce.Job: Running job: job_1592246524221_0074
```

```
[student4_10@node3 ~]$  
20/06/23 11:01:41 INFO mapreduce.Job: Running job: job_1592839005008_0003  
20/06/23 11:01:50 INFO mapreduce.Job: Job job_1592839005008_0003 running in uber mode : false  
20/06/23 11:01:50 INFO mapreduce.Job: map 0% reduce 0%  
20/06/23 11:02:06 INFO mapreduce.Job: map 100% reduce 0%  
20/06/23 11:02:10 INFO mapreduce.Job: Job job_1592839005008_0003 completed successfully  
20/06/23 11:02:10 INFO mapreduce.Job: Counters: 30  
File System Counters  
FILE: Number of bytes read=0  
FILE: Number of bytes written=175434  
FILE: Number of read operations=0  
FILE: Number of large read operations=0  
FILE: Number of write operations=0  
HDFS: Number of bytes read=87  
HDFS: Number of bytes written=238  
HDFS: Number of read operations=4  
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)=44804  
Total time spent by all reduces in occupied slots (ms)=0  
Total time spent by all map tasks (ms)=11201  
Total vcore-milliseconds taken by all map tasks=11201  
Total megabyte-milliseconds taken by all map tasks=11469824  
Map-Reduce Framework  
Map input records=12  
Map output records=12  
Input split bytes=87  
Spilled Records=0  
Failed Shuffles=0  
Merged Map outputs=0  
GC time elapsed (ms)=101  
CPU time spent (ms)=1040  
Physical memory (bytes) snapshot=225079296  
Virtual memory (bytes) snapshot=2800836608  
Total committed heap usage (bytes)=190316544  
File Input Format Counters  
Bytes Read=0  
File Output Format Counters  
Bytes Written=238  
20/06/23 11:02:10 INFO mapreduce.ImportJobBase: Transferred 238 bytes in 36.659 seconds (6.4923 bytes/sec)  
20/06/23 11:02:10 INFO mapreduce.ImportJobBase: Retrieved 12 records.  
[student4_10@node3 ~]$
```

Добавил snapshot полного лога

Проверим что импортировалось с базы PG_DATABASE (таблица параграф)

```
student4_10@node3:~$ hdfs dfs -ls /user/student4_10/
Found 5 items
drwx-----  - student4_10 student4_10          0 2020-06-22 00:02 /user/student4_10/.Trash
drwx-----  - student4_10 student4_10          0 2020-06-22 00:20 /user/student4_10/.staging
drwxr-xr-x  - student4_10 student4_10          0 2020-06-05 17:02 /user/student4_10/Datasets
drwxr-xr-x  - student4_10 student4_10          0 2020-06-21 18:21 /user/student4_10/hw_5
drwxr-xr-x  - student4_10 student4_10          0 2020-06-22 00:20 /user/student4_10/hw_5_1
[student4_10@node3 ~]$ hdfs dfs -ls /user/student4_10/hw_5_1/
Found 1 items
drwxr-xr-x  - student4_10 student4_10          0 2020-06-22 00:20 /user/student4_10/hw_5_1/paragraph
[student4_10@node3 ~]$ hdfs dfs -ls -r /user/student4_10/hw_5_1/paragraph/
Found 2 items
-rw-r--r--  3 student4_10 student4_10      238 2020-06-22 00:20 /user/student4_10/hw_5_1/paragraph/part-m-00000
-rw-r--r--  3 student4_10 student4_10          0 2020-06-22 00:20 /user/student4_10/hw_5_1/paragraph/_SUCCESS
[student4_10@node3 ~]$
```

```
student4_10@node3:~$ hdfs dfs -ls /user/student4_10/
Found 5 items
drwx-----  - student4_10 student4_10          0 2020-06-22 00:02 /user/student4_10/.Trash
drwx-----  - student4_10 student4_10          0 2020-06-22 00:20 /user/student4_10/.staging
drwxr-xr-x  - student4_10 student4_10          0 2020-06-05 17:02 /user/student4_10/Datasets
drwxr-xr-x  - student4_10 student4_10          0 2020-06-21 18:21 /user/student4_10/hw_5
drwxr-xr-x  - student4_10 student4_10          0 2020-06-22 00:20 /user/student4_10/hw_5_1
[student4_10@node3 ~]$ hdfs dfs -ls /user/student4_10/hw_5_1/
Found 1 items
drwxr-xr-x  - student4_10 student4_10          0 2020-06-22 00:20 /user/student4_10/hw_5_1/paragraph
[student4_10@node3 ~]$ hdfs dfs -ls -r /user/student4_10/hw_5_1/paragraph/
Found 2 items
-rw-r--r--  3 student4_10 student4_10      238 2020-06-22 00:20 /user/student4_10/hw_5_1/paragraph/part-m-00000
-rw-r--r--  3 student4_10 student4_10          0 2020-06-22 00:20 /user/student4_10/hw_5_1/paragraph/_SUCCESS
[student4_10@node3 ~]$ hdfs dfs -cat /user/student4_10/hw_5_1/paragraph/part-m-00000
workid,character varying
paragraphid,integer
paragraphnum,integer
charid,character varying
plaintext,text
phonetictext,text
stemtext,text
paragraphtype,character varying
section,integer
chapter,integer
charcount,integer
wordcount,integer
[student4_10@node3 ~]$
```

В схеме указаны колонки с типами данных

Создадим Таблицу в паркет

The screenshot shows the Hue Editor interface. At the top, there are tabs for 'Обучение | GeekBrain', 'Hue - Editor', 'HADOOP/HW_3_HIVE', 'Data_Science/Ypok 5.', 'gb_hadoop/lesson_4', and 'hdfs linux delete - Goo'. Below the tabs, the URL is 89.208.221.132:8888 and the title is 'Hue - Editor'. On the right, there are icons for search, refresh, and user 'student4_10'. The main area is titled 'Hive' with fields 'Add a name...' and 'Add a description...'. A query editor shows the following SQL code:

```
1 set parquet.compression=SNAPPY;
2
3 CREATE EXTERNAL TABLE student4_10.paragraph (
4   workid STRING,
5   paragraphid INT,
6   paragraphnum INT,
7   charid STRING,
8   plaintext STRING,
9   phonetictext STRING,
10  stemtext STRING,
11  paragraphtype STRING,
12  section INT,
13  chapter INT,
14  charcount INT,
15  wordcount INT
16 )
17 STORED AS PARQUET
18 LOCATION '/user/student4_10/hw_5_1/paragraph';
```

Below the code, the output of the command is shown:

```
STORED AS PARQUET
LOCATION '/user/student4_10/hw_5_1/paragraph'
INFO : Starting task [Stage-0:DDL] in serial mode
INFO : Completed executing command(queryId=hive_20200622003434_ec8e8091-ac17-4ac1-b486-5064dc032dcf); Time taken: 0.066 seconds
INFO : OK
```

The system tray at the bottom shows various icons and the date/time: 3:34 22.06.2020.

Импортируем данные в таблицу

```
sqoop import --connect jdbc:postgresql://node3.novalocal/pg_db --username exporter --password exporter_pass --table paragraph --hive-import --hive-database student4_10 --hive-table paragraph_1 --as-parquetfile
```

The screenshot shows a terminal window with the title 'student4_10@node3:~'. The command 'sqoop import --connect jdbc:postgresql://node3.novalocal/pg_db --username exporter --password exporter_pass --table paragraph --hive-import --hive-database student4_10 --hive-table paragraph --as-parquetfile' is being run. The terminal output shows several informational messages and warnings, including:

```
[student4_10@node3 ~]$ sqoop import --connect jdbc:postgresql://node3.novalocal/pg_db --username exporter --password exporter_pass --table paragraph --hive-import --hive-database student4_10 --hive-table paragraph --as-parquetfile
Warning: /opt/cloudera/parcels/CDH-5.16.2-1.cdh5.16.2.p0.8/bin/../lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
20/06/22 00:44:24 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.16.2
20/06/22 00:44:24 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
20/06/22 00:44:24 INFO tool.BaseSqoopTool: Using Hive-specific delimiters for output. You can override
20/06/22 00:44:24 INFO tool.BaseSqoopTool: delimiters with --fields-terminated-by, etc.
20/06/22 00:44:24 INFO manager.SqlManager: Using default fetchSize of 1000
20/06/22 00:44:24 INFO tool.CodeGenTool: Beginning code generation
20/06/22 00:44:24 INFO tool.CodeGenTool: Will generate java class as codegen_paragraph
20/06/22 00:44:24 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM "paragraph" AS t LIMIT 1
20/06/22 00:44:26 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce
Note: /tmp/sqoop-student4_10/compile/c225d08da52ea0139b23e05152e1e9fd/codegen_paragraph.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
20/06/22 00:44:26 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-student4_10/compile/c225d08da52ea0139b23e05152e1e9fd/codegen_paragraph.jar
20/06/22 00:44:26 WARN manager.PostgresqlManager: It looks like you are importing from postgresql.
20/06/22 00:44:26 WARN manager.PostgresqlManager: This transfer can be faster! Use the --direct
20/06/22 00:44:26 WARN manager.PostgresqlManager: option to exercise a postgresql-specific fast path.
20/06/22 00:44:26 INFO mapreduce.ImportJobBase: Beginning import of paragraph
20/06/22 00:44:27 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
20/06/22 00:44:28 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM "paragraph" AS t LIMIT 1
20/06/22 00:44:28 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM "paragraph" AS t LIMIT 1
20/06/22 00:44:29 INFO hive.metastore: Trying to connect to metastore with URI thrift://manager.novalocal:9083
20/06/22 00:44:30 INFO hive.metastore: Opened a connection to metastore, current connections: 1
20/06/22 00:44:30 INFO hive.metastore: Connected to metastore.
20/06/22 00:44:30 WARN mapreduce.DataDrivenImportJob: Target Hive table 'paragraph' exists! Sqoop will append data into the existing Hive table. Consider using --hive-overwrite, if you do NOT intend to do appending.
20/06/22 00:44:32 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
20/06/22 00:44:32 INFO client.RMProxy: Connecting to ResourceManager at manager.novalocal/89.208.221.132:8032
```

The system tray at the bottom shows various icons and the date/time: 3:44 22.06.2020.

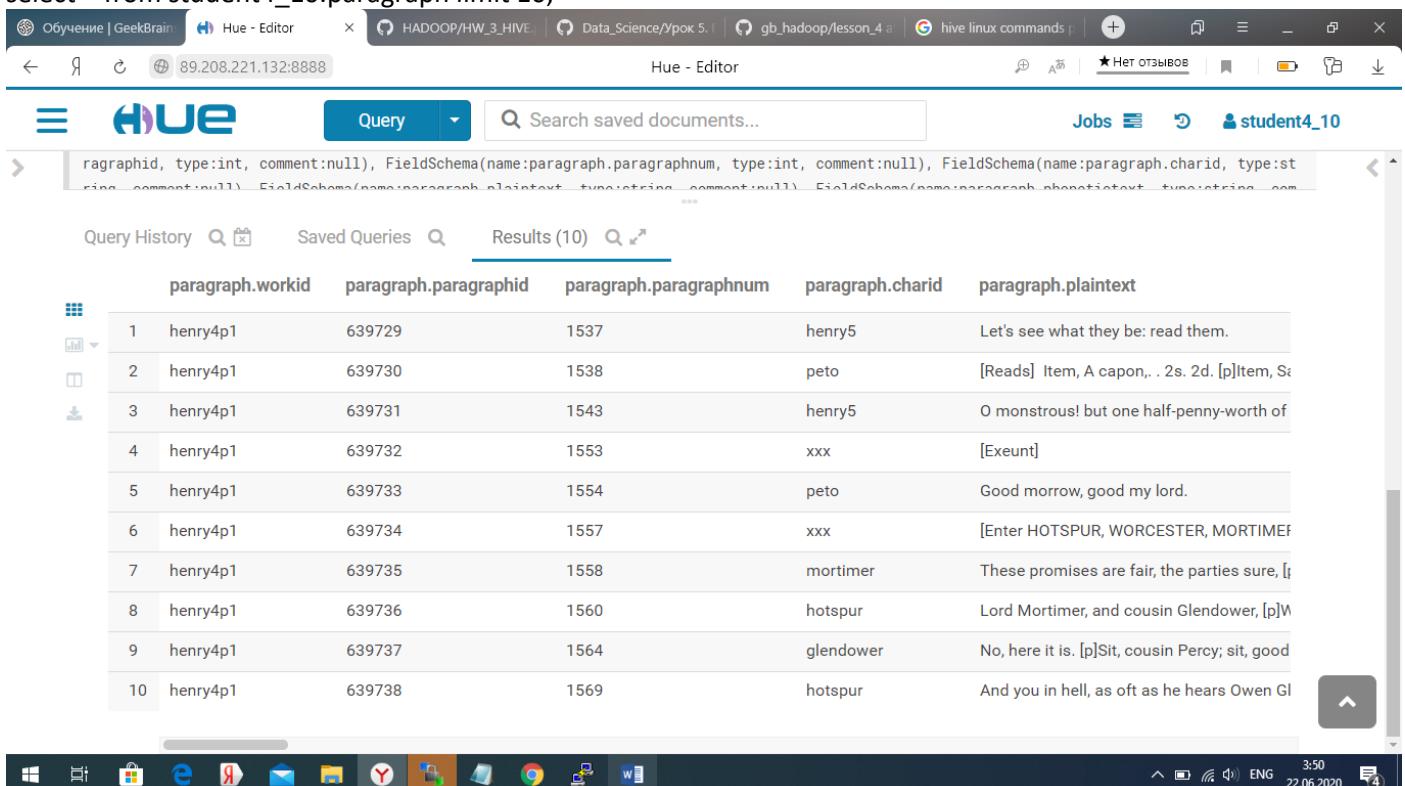
```

student4_10@node3:~ 
20/06/23 10:44:21 INFO mapreduce.Job: map 0% reduce 0%
20/06/23 10:44:38 INFO mapreduce.Job: map 50% reduce 0%
20/06/23 10:44:39 INFO mapreduce.Job: map 75% reduce 0%
20/06/23 10:44:40 INFO mapreduce.Job: map 100% reduce 0%
20/06/23 10:44:41 INFO mapreduce.Job: Job job_1592839005008_0002 completed successfully
20/06/23 10:44:41 INFO mapreduce.Job: Counters: 30
  File System Counters
    FILE: Number of bytes read=0
    FILE: Number of bytes written=990172
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=62345
    HDFS: Number of bytes written=8868427
    HDFS: Number of read operations=192
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=40
  Job Counters
    Launched map tasks=4
    Other local map tasks=4
    Total time spent by all maps in occupied slots (ms)=198908
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=49727
    Total vcore-milliseconds taken by all map tasks=49727
    Total megabyte-milliseconds taken by all map tasks=50920448
  Map-Reduce Framework
    Map input records=35465
    Map output records=35465
    Input split bytes=505
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=1208
    CPU time spent (ms)=29420
    Physical memory (bytes) snapshot=1709895680
    Virtual memory (bytes) snapshot=11352760320
    Total committed heap usage (bytes)=1388838912
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=0
20/06/23 10:44:41 INFO mapreduce.ImportJobBase: Transferred 8.4576 MB in 64.9173 seconds (133.4093 KB/sec)
20/06/23 10:44:41 INFO mapreduce.ImportJobBase: Retrieved 35465 records.
[student4_10@node3 ~]$ 

```

Скриншоты лога

Делаем проверку
select * from student4_10.paragraph limit 10;



The screenshot shows the Hue web interface with the following details:

- Header:** Shows tabs for "Обучение | GeekBrain", "Hue - Editor", "HADOOP/HW_3_HIVE", "Data_Science/Урок 5.", "gb_hadoop/lesson_4", and "hive linux commands".
- Toolbar:** Includes icons for back, forward, search, and other navigation.
- Query Bar:** Displays the query: `paragraphid, type:int, comment:null), FieldSchema(name:paragraph.paragraphnum, type:int, comment:null), FieldSchema(name:paragraph.charid, type:st`.
- Results:** The "Results (10)" tab is selected, showing a table with 10 rows of data from the "paragraph" table. The columns are: paragraph.workid, paragraph.paragraphid, paragraph.paragraphnum, paragraph.charid, and paragraph.plaintext.
- Data:** The table contains the following data:

	paragraph.workid	paragraph.paragraphid	paragraph.paragraphnum	paragraph.charid	paragraph.plaintext
1	henry4p1	639729	1537	henry5	Let's see what they be: read them.
2	henry4p1	639730	1538	peto	[Reads] Item, A capon,. . 2s. 2d. [p]Item, S
3	henry4p1	639731	1543	henry5	O monstrous! but one half-penny-worth of
4	henry4p1	639732	1553	xxx	[Exeunt]
5	henry4p1	639733	1554	peto	Good morrow, good my lord.
6	henry4p1	639734	1557	xxx	[Enter HOTSPUR, WORCESTER, MORTIMER
7	henry4p1	639735	1558	mortimer	These promises are fair, the parties sure, [i
8	henry4p1	639736	1560	hotspur	Lord Mortimer, and cousin Glendower, [p]W
9	henry4p1	639737	1564	glendower	No, here it is. [p]Sit, cousin Percy; sit, good
10	henry4p1	639738	1569	hotspur	And you in hell, as oft as he hears Owen Gl

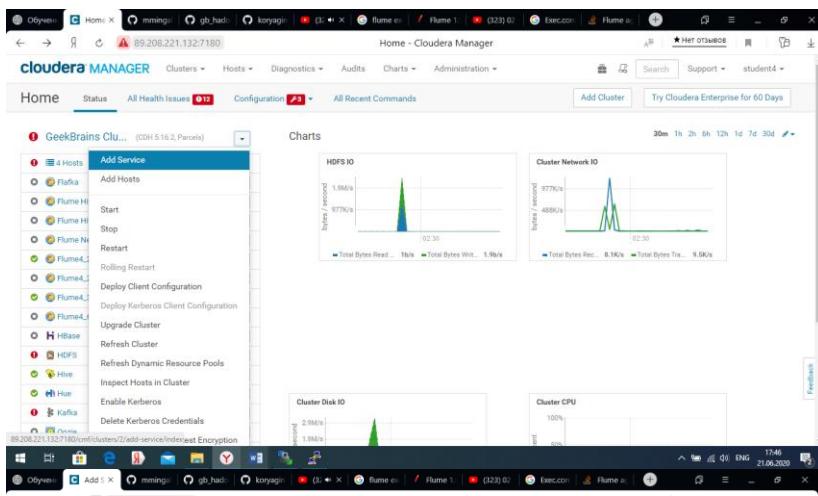
Проверим папку paragraph

```
student4_10@node3:~$ hdfs dfs -ls -r /user/student4_10/hw_5_1/paragraph/
Found 7 items
-rw-r--r-- 3 student4_10 student4_10          238 2020-06-22 00:20 /user/student4_10/hw_5_1/paragraph/part-m-00000
-rw-r--r-- 3 student4_10 supergroup  2025889 2020-06-22 00:45 /user/student4_10/hw_5_1/paragraph/afd67b6e-f858-46e1-a113-03325e3b397e.parquet
-rw-r--r-- 3 student4_10 student4_10          0 2020-06-22 00:20 /user/student4_10/hw_5_1/paragraph/_SUCCESS
-rw-r--r-- 3 student4_10 supergroup  2155202 2020-06-22 00:45 /user/student4_10/hw_5_1/paragraph/8892996f-cb9c-427f-ad5a-00a0d5c8979.parquet
-rw-r--r-- 3 student4_10 supergroup  2281342 2020-06-22 00:45 /user/student4_10/hw_5_1/paragraph/7536a816-71ea-4102-a22e-3490ae59a669.parquet
-rw-r--r-- 3 student4_10 supergroup  2389502 2020-06-22 00:45 /user/student4_10/hw_5_1/paragraph/11592d45-8a3f-402b-9312-2ef781b1e859.parquet
drwxr-xr-x - student4_10 student4_10          0 2020-06-22 00:45 /user/student4_10/hw_5_1/paragraph/.signals
[student4_10@node3 ~]$
```

Точно таким способом можно выполнить работу и с Avro файлом

1. Создать Flume-агент с именем, соответствующим имени своего пользователя (например Flume4_20)
2. Создать любой Flume поток используя Flume сервис соответствующего номера.
 - Тип источника источник – exec
 - Тип канала – memory
 - Тип слива – hdfs
3. Убедиться что данные поступают в слив.
4. Создать поверх данных в hdfs таблицу через которую можно просмотреть полученные данные.
5. [Продвинутый вариант] Сделать то-же самое используя несколько слипов в разные места, например в HDFS и в Hive одновременно
6. [Продвинутый вариант] Повторить стандартный пример с выборкой сообщений из Twitter.

1. Создать Flume-агент с именем, соответствующим имени своего пользователя (например Flume4_20)



Add Service to GeekBrains Cluster

Select the type of service you want to add.

Service Type	Description
<input type="radio"/> ADLS Connector	The ADLS Connector service provides key management for accessing Azure Data Lake Stores from CDH services.
<input type="radio"/> Accumulo	The Apache Accumulo sorted, distributed key/value store is a robust, scalable, high performance data storage and retrieval system. This service only works with releases based on Apache Accumulo 1.6 or later.
<input checked="" type="radio"/> Flume	Flume collects and aggregates data from almost any source into a persistent store such as HDFS.
<input type="radio"/> HBase	Apache HBase provides random, real-time, read/write access to large data sets (requires HDFS and ZooKeeper).
<input type="radio"/> HDFS	Apache Hadoop Distributed File System (HDFS) is the primary storage system used by Hadoop applications. HDFS creates multiple replicas of data blocks and distributes them on compute hosts throughout a cluster to enable reliable, extremely rapid computations.
<input type="radio"/> Hive	Hive is a data warehouse system that offers a SQL-like language called HiveQL.
<input type="radio"/> Hue	Hue is a graphical user interface to work with the Cloudera Distribution Including Apache Hadoop (requires HDFS, MapReduce, and Hive).
<input type="radio"/> Impala	Impala provides a real-time SQL query interface for data stored in HDFS and Hive. Impala requires the Hive service and shares the Hive Metastore with Hive.

[Back](#) [Continue](#)

Add Flume Service to GeekBrains Cluster

Select the set of dependencies for your new Flume

HBase	HDFS	Kafka	ZooKeeper
<input type="radio"/>	Kafka	ZooKeeper	
<input type="radio"/>	HDFS	Kafka	ZooKeeper
<input checked="" type="radio"/> HBase	HDFS	Kafka	ZooKeeper
<input type="radio"/> HBase	HDFS		ZooKeeper
<input type="radio"/>	HDFS		ZooKeeper

[Back](#) [Continue](#)

Add Flume Service to GeekBrains Cluster

Assign Roles for Flume

You can customize the role assignments for your new service here, but note that if assignments are made incorrectly, such as assigning too many roles to a single host, performance will suffer.

You can also view the role assignments by host: [View By Host](#)

Agent	Select hosts
	Too few hosts assigned, minimum is 1.

[Back](#) [Continue](#)



Add Flume Service to GeekBrains Cluster - Cloudera Manager

1 Host Selected

Tip: Click the first checkbox, hold down the Shift key and click the last checkbox to select a range.

Hostname	IP Address	Rack	Cores	Physical Memory	Existing Roles	Added Roles
manager.novalocal	89.208.221.132	/default	4	15.5 GB	H B... H M... NN G... H MS HS AM	C AP CS CM RM OS G... S
node1.novalocal	89.208.222.81	/default	4	7.6 GB	A H RS DN G... H S2 KB G...	G B... NM S
node2.novalocal	89.208.220.216	/default	4	15.5 GB	A A A A A H RS DN H FS A	NF... G... KB G... G... NM
node3.novalocal	89.208.222.201	/default	4	7.6 GB	A A A A H RS DN NF... SNN	

Cancel OK

Add Flume Service to GeekBrains Cluster

Assign Roles for Flume

You can customize the role assignments for your new service here, but note that if assignments are made incorrectly, such as assigning too many roles to a single host, performance will suffer.

You can also view the role assignments by host [View By Host](#)

Agent x 1 New node2.novalocal

Home - Cloudera Manager

GeekBrains Clu... (GSH 5.16.2, Panels)

Charts

Rename Service: Flume-7

Name * Flume-7

Cancel Rename Service

Cluster Disk IO

Cluster CPU

30m 1h 2h 6h 12h 1d 7d 30d

Flume4_10 - Cloudera Manager

cloudera MANAGER Clusters Hosts Diagnostics Audits Charts Administration

Flume4_10 (GeekBrains Cluster) Actions

Status Instances Configuration Commands Metric Details Charts Library Audits Quick Links

Health Tests

Agent Health Test disabled while the service is stopped. Test of whether enough Agent roles are healthy.

Status Summary

Agent 1 Stopped

Hosts 1 Bad Health

Health History

2:49:23 PM Agent Health Disabled Show

2:49:18 PM Agent Health Unknown Show

May 3 11:17:50 PM Agent Health Disabled Show

99.208.221.132:cmf/hardware/host?/status

BAD HEALTH!

Flume4_10 - Cloudera Manager

System User: Flume4_10 (Service-Wide)

System Group: flume

Agent Name: Agent Default Group

Configuration File:

```
# Please paste flume.conf here. Example:
# Sources, channels, and sinks are defined per
# agent name, in this case 'tier1'.
tier1.sources = source1
```

Flume Home Directory: Agent Default Group

Save Changes

2. Создать любой Flume поток используя Flume сервис соответствующего номера.

- Тип источника источник – exec
- Тип канала – memory
- Тип слива – hdfs

`tailif /var/log/cloudera-scm-agent/cloudera-scm-agent.log`

работает с этим лог файлом

```
Exec Source
Exec source runs a given Unix command on start-up and expects that process to continuously produce data on standard output. This source is useful for reading log files or monitoring system metrics.

root@manager:~#
gz: flume_data.1592333216981.gz: not in gzip format
[root@manager ~]# less flume_data.1592333216981.gz
"flume_data.1592333216981.gz" may be a binary file. See it anyway?
[root@manager ~]# tailf /var/log/cloudera-scm-agent/
cloudera-flood.log    cloudera-scm-agent.log.2  cmf_listener.log
cloudera-scm-agent.log.1 cloudera-scm-agent.out   supervisord.log
cloudera-scm-agent.log.1 cloudera-scm-agent.out   supervisord.out
[root@manager ~]# tailf /var/log/cloudera-scm-agent/cloudera-scm-agent.log
self.collect_chronyd()
File "/usr/lib64/cmfl/agent/build/env/lib/python2.7/site-packages/cmfl-5.16.2-py2.7.egg/cmfl/monitor/host/ntp_monitor.py", line 55, in collect_chronyd
result, stdout, stderr = self._subprocess_with_timeout(args, self, timeout)
File "/usr/lib64/cmfl/agent/build/env/lib/python2.7/site-packages/cmfl-5.16.2-py2.7.egg/cmfl/monitor/host/ntp_monitor.py", line 38, in _subprocess_with_timeout
return subprocess_with_timeout(args, timeout)
File "/usr/lib64/cmfl/agent/build/env/lib/python2.7/site-packages/cmfl-5.16.2-py2.7.egg/cmfl/subprocess_timeout.py", line 94, in subprocess_with_timeout
raise Exception("timeout with args '%s' % args")
Exception: timeout with args ['chronyd', 'sources']
[16/Jun/2020 18:36:05 +0000] 1460 MainThread heartbeat tracker INFO      HB stats
(seconds): num:40 LIFE_MIN:0.02 min:0.02 mean:0.02 max:0.06 LIFE_MAX:3.88
[16/Jun/2020 18:46:05 +0000] 1460 MainThread heartbeat tracker INFO      HB stats
(seconds): num:40 LIFE_MIN:0.02 min:0.02 mean:0.02 max:0.05 LIFE_MAX:3.88
```

a log file on disk and Flume tails the file, sending each line as an event. While this is possible, there's an obvious problem; what happens if the channel fills up and Flume can't send an event? Flume has no way of indicating to the application writing the log file that it needs to retain the log or that the event hasn't been sent, for some reason. If this doesn't make sense, you need only know this: Your application can never guarantee data has been received when using a unidirectional asynchronous interface such as ExecSource! As an extension of this warning - and to be completely clear - there is absolutely zero guarantee of event delivery when using this source. For stronger reliability guarantees, consider the Spooling Directory Source, Taildir Source or direct integration with Flume via the SDK.

Example for agent named a1:

a1.sources = r1

Писать буду в Flume папку !

```
student4_10@manager:~$ hdfs dfs -ls /flume/
Found 10 items
drwxr-xr-x  - flume flume      0 2020-04-20 22:59 /flume/flume-7
drwxr-xr-x  - flume flume      0 2020-04-19 19:46 /flume/flume10
drwxr-xr-x  - flume flume      0 2020-03-12 14:59 /flume/flume11
drwxr-xr-x  - flume flume      0 2020-04-21 11:27 /flume/flume3_2
drwxr-xr-x  - flume flume      0 2020-04-20 09:36 /flume/student3_10
drwxr-xr-x  - flume flume      0 2020-05-06 22:06 /flume/student3_14
drwxr-xr-x  - flume flume      0 2020-05-24 00:01 /flume/student3_14_1
drwxr-xr-x  - flume flume      0 2020-05-06 22:04 /flume/student3_14_2
drwxr-xr-x  - flume flume      0 2020-04-30 09:18 /flume/student3_3
drwxr-xr-x  - flume flume      0 2020-04-23 02:17 /flume/student3_5
[student4_10@manager ~]$ hdfs dfs -ls /flume/flume-7/
Found 10 items
drwxr-xr-x  - flume flume      0 2020-04-19 18:58 /flume/flume-7/exec-file-hdfs
drwxr-xr-x  - flume flume      0 2020-04-20 22:59 /flume/flume-7/exec-file-hdfs-v1
drwxr-xr-x  - flume flume      0 2020-04-19 19:02 /flume/flume-7/exec-file-hdfs-v2
drwxr-xr-x  - flume flume      0 2020-04-19 19:05 /flume/flume-7/exec-file-hdfs-v3
drwxr-xr-x  - flume flume      0 2020-04-19 19:21 /flume/flume-7/exec-file-hdfs-v4
drwxr-xr-x  - flume flume      0 2020-04-20 18:47 /flume/flume-7/exec-file-hdfs-v5
drwxr-xr-x  - flume flume      0 2020-04-20 22:06 /flume/flume-7/exec-file-hdfs-v6
drwxr-xr-x  - flume flume      0 2020-04-20 22:20 /flume/flume-7/exec-file-hdfs-v7
drwxr-xr-x  - flume flume      0 2020-04-20 22:30 /flume/flume-7/exec-file-hdfs-v8
drwxr-xr-x  - flume flume      0 2020-04-20 22:33 /flume/flume-7/exec-file-hdfs-v9
[student4_10@manager ~]$
```

```
*Безымянный – Блокнот
Файл Правка Формат Вид Справка
# Naming the components on the current agent
Flume4_10.sources = ExecSource
Flume4_10.channels = MemChannel
Flume4_10.sinks = HdfsSink

# Describing/Configuring the source
Flume4_10.sources.ExecSource.type = exec
Flume4_10.sources.ExecSource.command = tailf /var/log/cloudera-scm-agent/cloudera-scm-agent.log
Flume4_10.sources.ExecSource.interceptors = TimestampInterceptor
Flume4_10.sources.ExecSource.interceptors.TimestampInterceptor.type = timestamp

# Describing/Configuring the HDFS sink
Flume4_10.sinks.HdfsSink.type = hdfs
Flume4_10.sinks.LoggerSink.hdfs.path= /flume/flume4_10/%y-%m-%d/
Flume4_10.sinks.HdfsSink.hdfs.filePrefix = events

# Describing/Configuring the channel
Flume4_10.channels.MemChannel.type = memory
Flume4_10.channels.MemChannel.capacity = 10000
Flume4_10.channels.MemChannel.transactionCapacity = 10

# Bind the source and sink to the channel
Flume4_10.sources.ExecSource.channels = MemChannel
Flume4_10.sinks.HdfsSink.channel = MemChannel
```

Код для configuration file

```
# Naming the components on the current agent
Flume4_10.sources = ExecSource
Flume4_10.channels = MemChannel
Flume4_10.sinks = HdfsSink

# Describing/Configuring the source
Flume4_10.sources.ExecSource.type = exec
Flume4_10.sources.ExecSource.command = tailf /var/log/cloudera-scm-agent/cloudera-scm-agent.log
Flume4_10.sources.ExecSource.interceptors = TimestampInterceptor
Flume4_10.sources.ExecSource.interceptors.Timestampinterceptor.type = timestamp

# Describing/Configuring the HDFS sink
Flume4_10.sinks.HdfsSink.type = hdfs
Flume4_10.sinks.LoggerSink.hdfs.path= /flume/flume4_10/%y-%m-%d/
Flume4_10.sinks.HdfsSink.hdfs.filePrefix = events

# Describing/Configuring the channel
Flume4_10.channels.MemChannel.type = memory
Flume4_10.channels.MemChannel.capacity = 10000
Flume4_10.channels.MemChannel.transactionCapacity = 10

# Bind the source and sink to the channel
Flume4_10.sources.ExecSource.channels = MemChannel
Flume4_10.sinks.HdfsSink.channel = MemChannel
```

Flume4_10 (GeekBrains Cluster)

Status Instances Configuration Co... Actions ▾

Start Start a service and its associated Roles

Stop

Restart Create Trigger

Rolling Restart Suppress...

Add Role Instances healthy...

Rename

Enter Maintenance Mode

Update Configuration 1

Health Tests

Agent Health Test disabled while the service is stopped: Test of v...

Status Summary

Agent 1 Stopped

Hosts 1 Bad Health

Agent Status

Health History

249:23 PM Agent Health Disabled Show

249:18 PM Agent Health Unknown Show

May 3 11:17:50 PM Agent Health Disabled Show

May 3 11:17:28 PM Agent Health Unknown Show

Apr 29 10:16 PM Agent Health Disabled Show

89.208.221.132:7180/cmf/services/83/do?command=Start Good Show

30 minutes preceding Jun 21, 4:15 PM UTC

30m 1h 2h 6h 12h 1d 7d 30d

Critical Events Across Agents

Alerts Across Agents

Important Events and Alerts

Health

Feedback

BAD HEALTH !!!!

The screenshot shows the Cloudera Manager interface for a 'GeekBrains Cluster'. A modal dialog box titled 'Start Command' is open, indicating a failure to start the 'Flume4_10' service. The status bar at the top shows 'Status Failed' and the time 'Jun 21, 4:16:55 PM'. The modal contains a summary of the failed steps:

- Starting 1 roles on service**: Service did not start successfully; not all of the required roles started: only 0/1 roles started. Reasons: Service has only 0 Agent roles running instead of minimum required 1.
- Execute command Start this Agent on role Agent (node2)**: Command aborted because of exception: Command timed-out after 150 seconds
- Start a role**: Role is starting.

At the bottom of the modal, there is a 'Role Log' section with a progress bar and a 'Close' button.

Я принял решение запустить с Flume 4_2 так как у него GOOD HEALTH

Agent Name: Flume4_2

```

Agent Default Group
# Naming the components on the current agent
Flume4_2.sources = ExecSource
Flume4_2.channels = MemChannel
Flume4_2.sinks = HdfsSink

# Describing/Configuring the source
Flume4_2.sources.ExecSource.type = exec
Flume4_2.sources.ExecSource.command = tailf /var/log/cloudera-scm-agent/cloudera-scm-agent.log
Flume4_2.sources.ExecSource.interceptors = TimestampInterceptor
Flume4_2.sources.ExecSource.interceptors.TimestampInterceptor.type = timestamp

# Describing/Configuring the HDFS sink
Flume4_2.sinks.HdfsSink.type = hdfs
Flume4_2.sinks.LoggerSink.hdfs.path= /flume/Flume4_2/%y-%m-%d/
Flume4_2.sinks.HdfsSink.hdfs.filePrefix = events

# Describing/Configuring the channel
Flume4_2.channels.MemChannel.type = memory
Flume4_2.channels.MemChannel.capacity = 10000
Flume4_2.channels.MemChannel.transactionCapacity = 10

# Bind the source and sink to the channel
Flume4_2.sources.ExecSource.channels = MemChannel
Flume4_2.sinks.HdfsSink.channel = MemChannel
|
```

1 Edited Value Reason for change... Save Changes

Save changes and restart

Flume4_2 - Cloudera Manager

Restart Command

Status: Finished Context: Flume4_2 Jun 21, 4:29:58 PM 23.9s

Successfully restarted service.

Completed 2 of 2 step(s).

Show All Steps Show Only Failed Steps Show Running Steps

> ✓ Execute command Stop on service Flume4_2	Flume4_2	Jun 21, 4:29:58 PM	1.55s
> ✓ Execute command Start on service Flume4_2	Flume4_2	Jun 21, 4:30:00 PM	22.3s

Close

Performance: none
 Ports and Addresses: 1
 Resource Management: 5
 Security: 3
 Stacks Collection: 5

Solr Service: Flume4_2 (Service-Wide) none

STATUS

Save Changes

```

student4_10@manager:~$ hdfs dfs -ls /flume/flume-7/
Found 10 items
drwxr-xr-x - flume flume 0 2020-04-19 18:58 /flume/flume-7/exec-file-hdfs
drwxr-xr-x - flume flume 0 2020-04-20 22:59 /flume/flume-7/exec-file-hdfs-v10
drwxr-xr-x - flume flume 0 2020-04-19 19:02 /flume/flume-7/exec-file-hdfs-v2
drwxr-xr-x - flume flume 0 2020-04-19 19:05 /flume/flume-7/exec-file-hdfs-v3
drwxr-xr-x - flume flume 0 2020-04-19 19:21 /flume/flume-7/exec-file-hdfs-v4
drwxr-xr-x - flume flume 0 2020-04-20 18:47 /flume/flume-7/exec-file-hdfs-v5
drwxr-xr-x - flume flume 0 2020-04-20 22:06 /flume/flume-7/exec-file-hdfs-v6
drwxr-xr-x - flume flume 0 2020-04-20 22:20 /flume/flume-7/exec-file-hdfs-v7
drwxr-xr-x - flume flume 0 2020-04-20 22:30 /flume/flume-7/exec-file-hdfs-v8
drwxr-xr-x - flume flume 0 2020-04-20 22:33 /flume/flume-7/exec-file-hdfs-v9
[student4_10@manager ~]$ hdfs dfs -ls /var/log/flume-ng
ls: '/var/log/flume-ng': No such file or directory
[student4_10@manager ~]$ hdfs dfs -ls /var/log/
ls: '/var/log/': No such file or directory
[student4_10@manager ~]$ ls /var/log/flume-ng
flume-cmf-flume10-AGENT-manager.novalocal.log flume-cmf-flume4-AGENT-manager.novalocal.log stacks
flume-cmf-flume19-AGENT-manager.novalocal.log flume-cmf-flume6-AGENT-manager.novalocal.log
[student4_10@manager ~]$ hdfs dfs -ls /flume/
Found 10 items
drwxr-xr-x - flume flume 0 2020-04-20 22:59 /flume/flume-7
drwxr-xr-x - flume flume 0 2020-04-19 19:46 /flume/flume10
drwxr-xr-x - flume flume 0 2020-03-12 14:59 /flume/flumell
drwxr-xr-x - flume flume 0 2020-04-21 11:27 /flume/flume3_2
drwxr-xr-x - flume flume 0 2020-04-20 09:36 /flume/student3_10
drwxr-xr-x - flume flume 0 2020-05-06 22:06 /flume/student3_14
drwxr-xr-x - flume flume 0 2020-05-24 00:01 /flume/student3_14_1
drwxr-xr-x - flume flume 0 2020-05-06 22:04 /flume/student3_14_2
drwxr-xr-x - flume flume 0 2020-04-30 09:18 /flume/student3_3
drwxr-xr-x - flume flume 0 2020-04-23 02:17 /flume/student3_5
[student4_10@manager ~]$ hdfs dfs -ls /flume/
Found 10 items
drwxr-xr-x - flume flume 0 2020-04-20 22:59 /flume/flume-7
drwxr-xr-x - flume flume 0 2020-04-19 19:46 /flume/flume10
drwxr-xr-x - flume flume 0 2020-03-12 14:59 /flume/flumell
drwxr-xr-x - flume flume 0 2020-04-21 11:27 /flume/flume3_2
drwxr-xr-x - flume flume 0 2020-04-20 09:36 /flume/student3_10
drwxr-xr-x - flume flume 0 2020-05-06 22:06 /flume/student3_14
drwxr-xr-x - flume flume 0 2020-05-24 00:01 /flume/student3_14_1
drwxr-xr-x - flume flume 0 2020-05-06 22:04 /flume/student3_14_2
drwxr-xr-x - flume flume 0 2020-04-30 09:18 /flume/student3_3
drwxr-xr-x - flume flume 0 2020-04-23 02:17 /flume/student3_5
[student4_10@manager ~]$

```

Нету ничего

Попробую flume-7 в configuration file

Flume4_2 - Cloudera Manager

cloudera MANAGER Clusters ▾ Hosts ▾ Diagnostics ▾ Audits Charts ▾ Administration ▾

Status: **Finished** Context: Flume4_2 ▾ Jun 21, 4:51:27 PM 22.35s

Successfully started service.

Completed 1 of 1 step(s).

Show All Steps Show Only Failed Steps Show Running Steps

✓ Starting 1 roles on service Successfully started 1 roles on service.	Jun 21, 4:51:27 PM	22.35s
➤ Execute command Start this Agent on role Agent (node3)	Agent (node3) ▾	Jun 21, 4:51:27 PM

Ports and Addresses 1
Resource Management 5
Security 3
Stacks Collection 5

Solr Service Flume4_2 (Service-Wide)
none

STATUS

Save Changes

Jun 21, 4:51:27 PM UTC

19:52 ENG 21.06.2020

Flume4_2 - Cloudera Manager

Agent Default Group

```
# Naming the components on the current agent
Flume4_2.sources = ExecSource
Flume4_2.channels = MemChannel
Flume4_2.sinks = HdfsSink

# Describing/Configuring the source
Flume4_2.sources.ExecSource.type = exec
Flume4_2.sources.ExecSource.command = tailf /var/log/cloudera-scm-agent/cloudera-scm-agent.log
Flume4_2.sources.ExecSource.interceptors = TimestampInterceptor
Flume4_2.sources.ExecSource.interceptors.TimestampInterceptor.type = timestamp

# Describing/Configuring the HDFS sink
Flume4_2.sinks.HdfsSink.type = hdfs
Flume4_2.sinks.LoggerSink.hdfs.path= flume/flume-7/log/%y-%m-%d/
Flume4_2.sinks.HdfsSink.hdfs.filePrefix = events

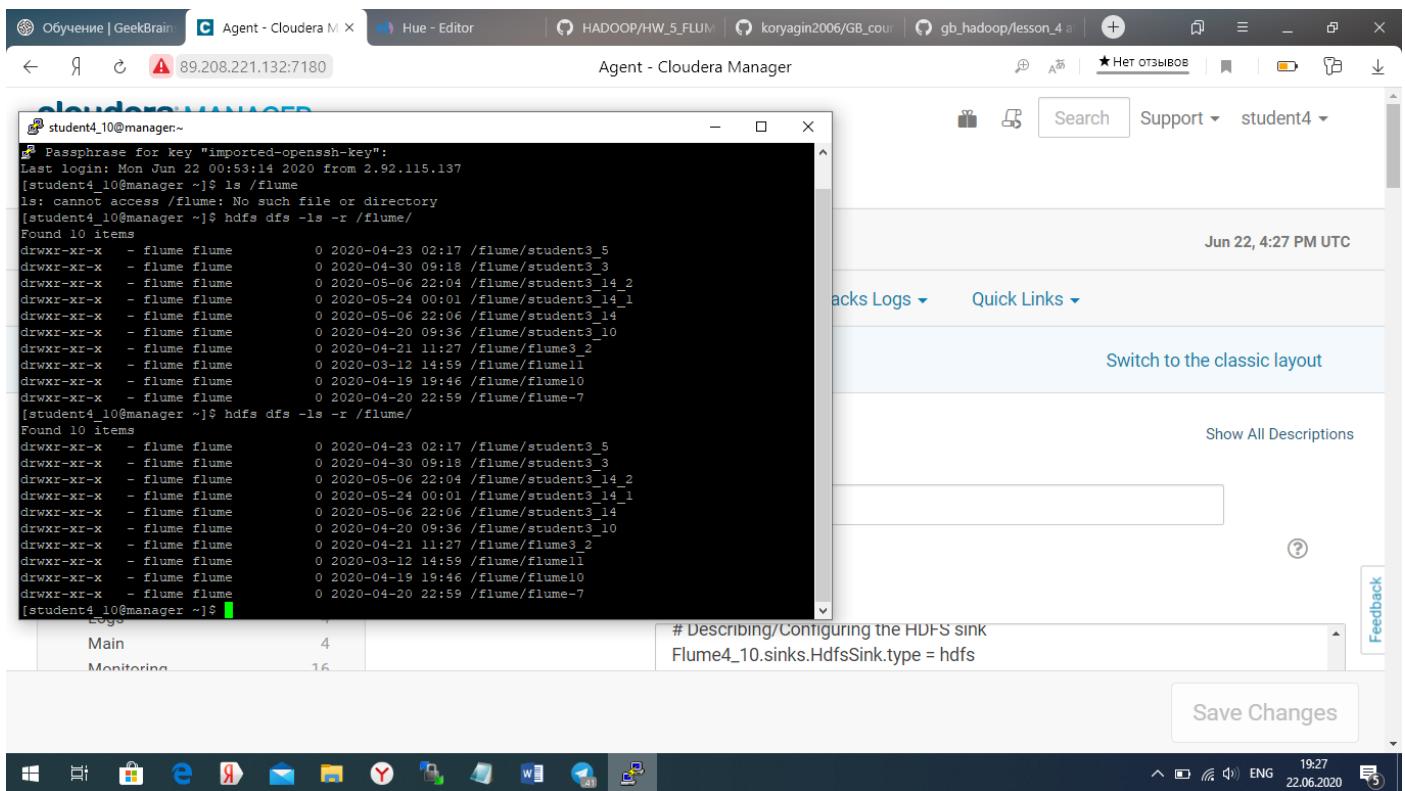
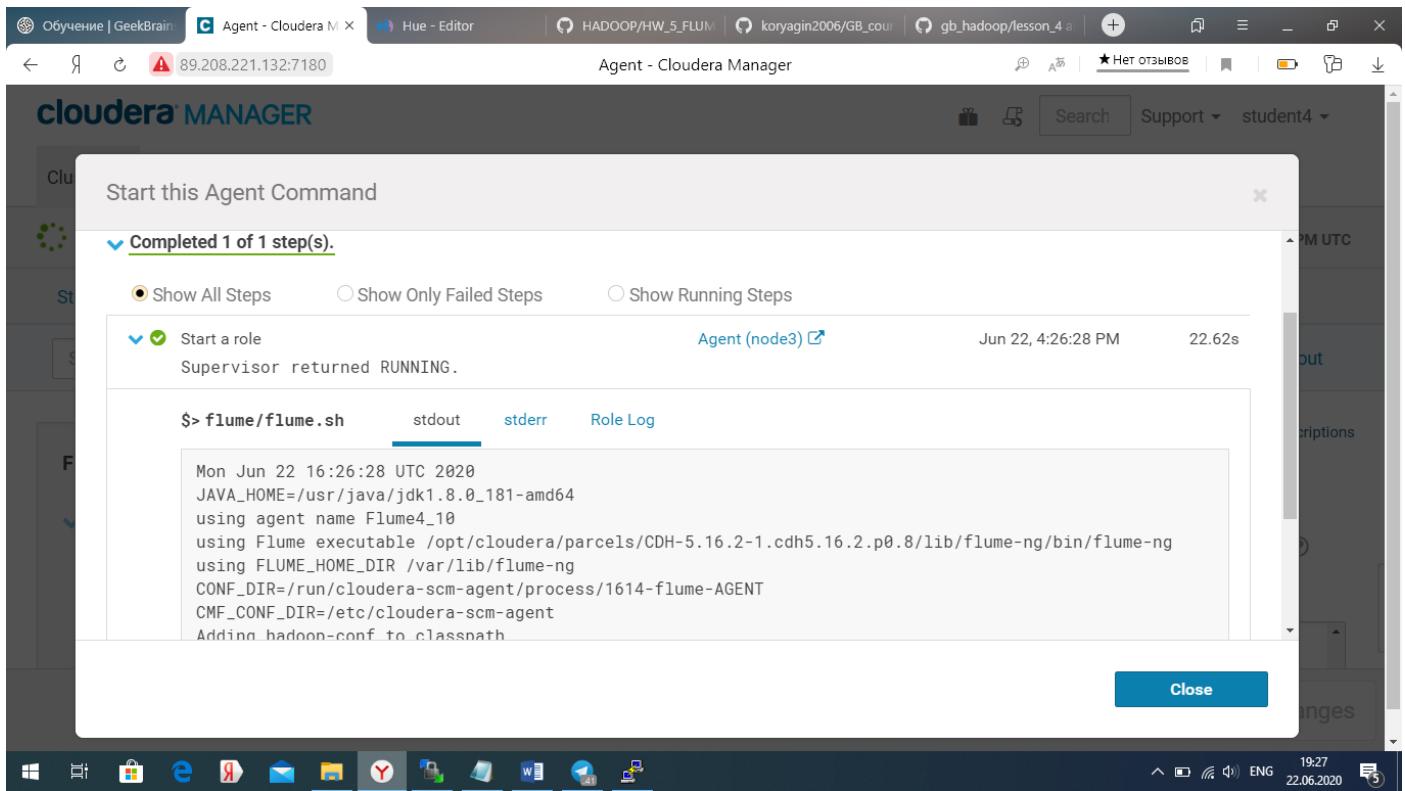
# Describing/Configuring the channel
Flume4_2.channels.MemChannel.type = memory
Flume4_2.channels.MemChannel.capacity = 10000
Flume4_2.channels.MemChannel.transactionCapacity = 10

# Bind the source and sink to the channel
Flume4_2.sources.ExecSource.channels = MemChannel
Flume4_2.sinks.HdfsSink.channel = MemChannel
```

Flume-7

```
student4_10@manager:~$ hdfs dfs -ls /user/student4_10/
drwxr-xr-x - student4_6 student4_6 0 2020-06-07 13:34 /user/student4_6
drwxr-xr-x - student4_7 student4_7 0 2020-05-31 12:55 /user/student4_7
drwxr-xr-x - student4_8 student4_8 0 2020-05-23 20:27 /user/student4_8
drwxr-xr-x - student4_9 student4_9 0 2020-06-05 12:49 /user/student4_9
drwxr-xr-x - student4_3 student4_3 0 2020-05-19 19:19 /user/student4_3
[student4_10@manager ~]$ hdfs dfs -ls /user/student4_10/
Found 3 items
drwx----- - student4_10 student4_10 0 2020-05-27 11:00 /user/student4_10/.Trash
drwx----- - student4_10 student4_10 0 2020-06-10 21:13 /user/student4_10/.staging
drwxr-xr-x - student4_10 student4_10 0 2020-06-05 17:02 /user/student4_10/Datasets
[student4_10@manager ~]$ hdfs dfs -ls /user/student4_2/
Found 4 items
drwx----- - student4_2 student4_2 0 2020-06-07 20:00 /user/student4_2/.Trash
drwx----- - student4_2 student4_2 0 2020-06-21 14:25 /user/student4_2/.staging
-rw-r--r-- 3 student4_2 student4_2 37054236 2020-06-06 20:40 /user/student4_2/Border_Crossing_Entry_Data.csv
-rw-r--r-- 3 student4_2 student4_2 260933 2020-06-06 20:39 /user/student4_2/amazon.csv
[student4_10@manager ~]$ hdfs dfs -ls /user/student4_10/
Found 3 items
drwx----- - student4_10 student4_10 0 2020-05-27 11:00 /user/student4_10/.Trash
drwx----- - student4_10 student4_10 0 2020-06-10 21:13 /user/student4_10/.staging
drwxr-xr-x - student4_10 student4_10 0 2020-06-05 17:02 /user/student4_10/Datasets
[student4_10@manager ~]$ hdfs dfs -ls /flume/flume-7/
Found 10 items
drwxr-xr-x - flume flume 0 2020-04-19 19:58 /flume/flume-7/exec-file-hdfs
drwxr-xr-x - flume flume 0 2020-04-20 22:59 /flume/flume-7/exec-file-hdfs-y10
drwxr-xr-x - flume flume 0 2020-04-19 19:02 /flume/flume-7/exec-file-hdfs-y2
drwxr-xr-x - flume flume 0 2020-04-19 19:05 /flume/flume-7/exec-file-hdfs-y3
drwxr-xr-x - flume flume 0 2020-04-19 19:21 /flume/flume-7/exec-file-hdfs-y4
drwxr-xr-x - flume flume 0 2020-04-20 18:47 /flume/flume-7/exec-file-hdfs-y5
drwxr-xr-x - flume flume 0 2020-04-20 22:06 /flume/flume-7/exec-file-hdfs-y6
drwxr-xr-x - flume flume 0 2020-04-20 22:20 /flume/flume-7/exec-file-hdfs-y7
drwxr-xr-x - flume flume 0 2020-04-20 22:30 /flume/flume-7/exec-file-hdfs-y8
drwxr-xr-x - flume flume 0 2020-04-20 22:33 /flume/flume-7/exec-file-hdfs-y9
[student4_10@manager ~]$
```

Не работает



Запустил Flume 4_10 с node3.novalocal но в FLUME папке ничего не появилось

Вопрос

- Я написал правильный код

- Результаты работы надо искать в папке Flume/student4_10, если так то почему туда ничего не поступает

