

## 1. Для части по SMOOP

Провести импорт таблицы из вашего сервера БД в Hadoop с использованием SMOOP 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 SMOOP 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/xinclude' is not recognized.
    at org.apache.xerces.jaxp.DocumentBuilderFactoryImpl.newDocumentBuilder(Unknown Source)
    at org.apache.hadoop.conf.Configuration.loadResources(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.Sqoop.main(Sqoop.java:252)
Exception in thread "main" java.lang.RuntimeException: javax.xml.parsers.ParserConfigurationException: Feature 'http://apache.org/xml/features/xinclude' 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.Sqoop.main(Sqoop.java:252)
Caused by: javax.xml.parsers.ParserConfigurationException: Feature 'http://apache.org/xml/features/xinclude' 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/xinclude' is not recognized.
at org.apache.hadoop.conf.Configuration.loadResources(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.Sqoop.main(Sqoop.java:252)
Caused by: javax.xml.parsers.ParserConfigurationException: Feature 'http://apache.org/xml/features/xinclude' 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:~$
[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 $ACCUMULO_HOME to the root of your Accumulo installation.
20/06/20 19:57:50 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.16.2
20/06/20 19:57:50 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
20/06/20 19:57:50 INFO manager.SqlManager: Using default fetchSize of 1000
template1
template0
postgres
pg_db
[student4_10@node3 ~]$ 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 $ACCUMULO_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/
drwxr-xr-x - 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
drwxr-xr-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
drwxr-xr-x - student4_10 student4_10 0 2020-05-27 11:00 /user/student4_10/.Trash
drwxr-xr-x - 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
allswell 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 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

The screenshot shows the Hue web interface for Hive. The top navigation bar includes the Hue logo, a search bar, and a user profile for 'student4\_10'. The left sidebar displays a list of tables under the 'student4\_10' schema, including 'border\_crossing', 'citation\_data\_parquet', 'measures', and 'states'. The main area is the Hive query editor, which contains the following SQL query:

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

The query execution time is shown as 1.93s. The bottom status bar indicates the time is 21:26 on 21.06.2020.

The screenshot shows the Hue web interface in a browser window. The address bar indicates the URL is `89.208.221.132:8888/hue/editor?editor=6427`. The interface includes a top navigation bar with the Hue logo, a 'Query' dropdown, and a search bar for saved documents. Below the navigation bar, a SQL query is entered in the editor: `1 SELECT * from student4_10.work limit 10;`. The execution log shows the following messages:

```
INFO : Compiling command(queryId=hive_20200621183030_af888586-9389-4f15-8ff2-c508754c636a): SELECT * from student4_10.work limit 10
INFO : Semantic Analysis Completed
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:work.workid, type:string, comment:null), FieldSchema(name:work.title, type:string, comment:null), FieldSchema(name:work.longtitle, type:string, comment:null), FieldSchema(name:work.year, type:int, comment:null), FieldSchema(name:work.genretype, type:string, comment:null), FieldSchema(name:work.notes, type:string, comment:null), FieldSchema(name:work.source, type:string, comment:null), FieldSchema(name:work.totalwords, type:int, comment:null)], tableName:student4_10.work)
```

Below the log, the 'Results (10)' tab is active, displaying a table with the following data:

	work.workid	work.title	work.longtitle	work.year	work.genretype	work.notes	work.source	work.totalwords
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

Копируем Paragraph Таблицу в формате Parquet, но сначала создадим таблицу так как в Parquet нельзя импортировать таким же способ как авро 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/92a7a9f36ab4d4d42e684ae923c230ef/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/92a7a9f36ab4d4d42e684ae923c230ef/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

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)

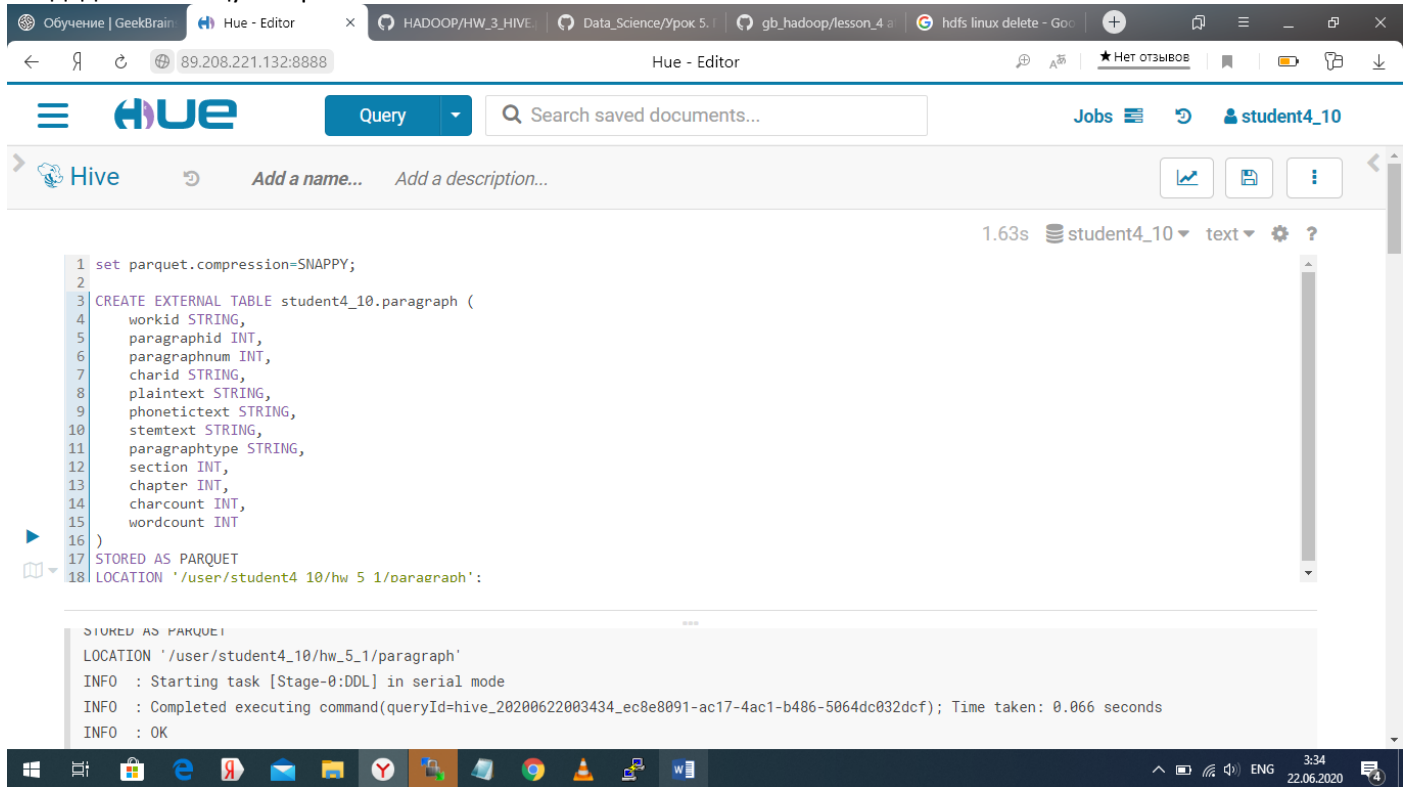


```
student4_10@node3:~$ hdfs dfs -ls /user/student4_10/
Found 5 items
drwx----- - student4_10 student4_10      0 2020-06-22 00:02 /user/student4_10/.Trash
drwx----- - student4_10 student4_10      0 2020-06-22 00:20 /user/student4_10/.staging
drwxr-xr-x - student4_10 student4_10      0 2020-06-05 17:02 /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

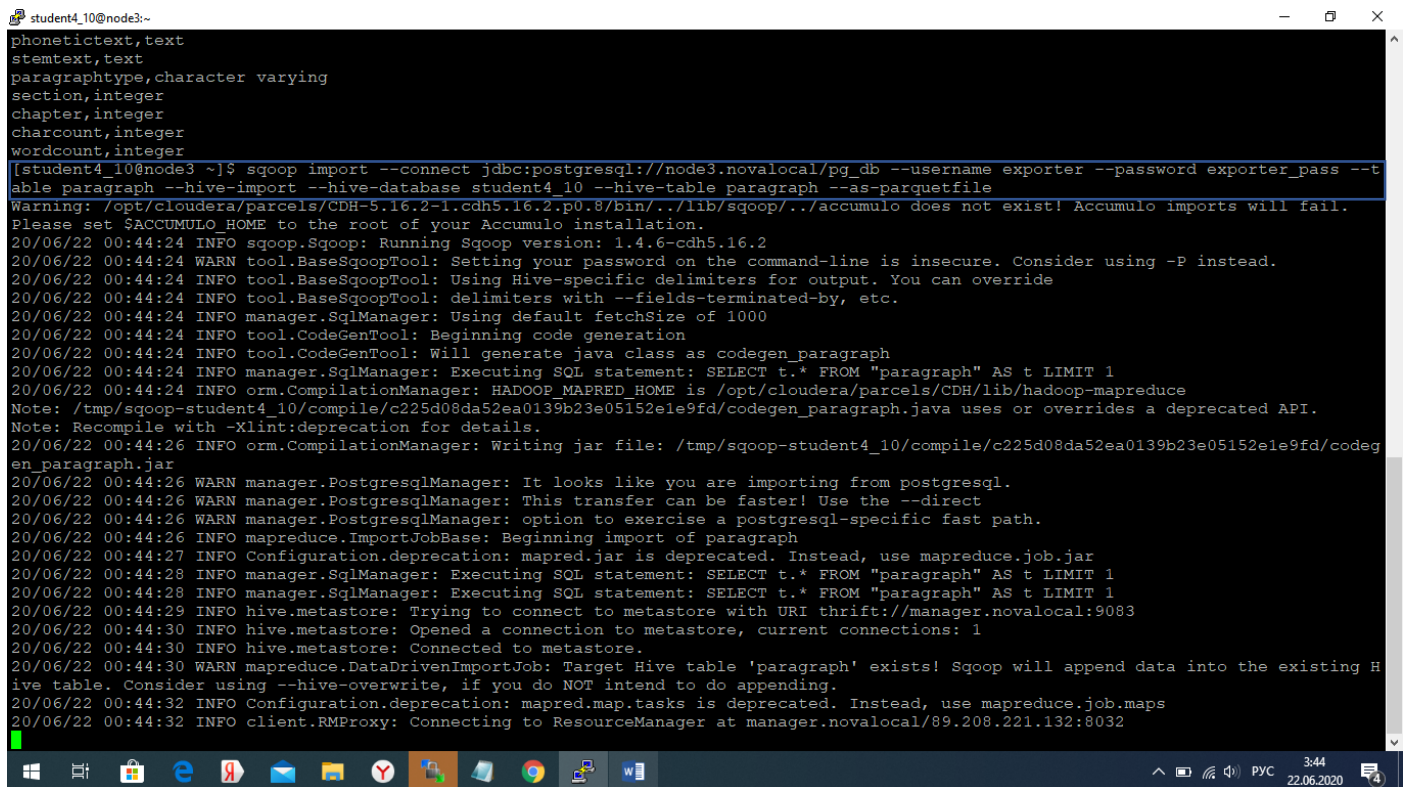


```
1 set parquet.compression=SNAPPY;
2
3 CREATE EXTERNAL TABLE student4_10.paragraph (
4     workid STRING,
5     paragraphid INT,
6     paragraphnum INT,
7     charid STRING,
8     plaintext STRING,
9     phonetictext STRING,
10    stemtext STRING,
11    paragraphtype STRING,
12    section INT,
13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

```
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
```

## 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
```



```
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_1 --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 00:44:24 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 H
ive 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.RMPProxy: Connecting to ResourceManager at manager.novalocal/89.208.221.132:8032
```



```
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;

Обучение | GeekBrain | Hue - Editor | HADOOP/HW\_3\_HIVE | Data\_Science/Урок 5. | gb\_hadoop/lesson\_4 | hive linux commands |

← | ↻ | 🔍 89.208.221.132:8888 | Hue - Editor | ★ Нет отзывов | 📄 | 📁 | 📥 | 📦 |

HUE

Query

🔍 Search saved documents...

Jobs | 🔄 | 👤 student4\_10

🔍

paragraphid, type:int, comment:null), FieldSchema(name:paragraph.paragraphnum, type:int, comment:null), FieldSchema(name:paragraph.charid, type:st

Query History | 🔍 Saved Queries | Results (10) | 🔍

	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, [
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

🏠 | 📁 | 📄 | 📥 | 📦 |

📶 | 🔊 | 🌐 ENG | 3:50 22.06.2020 | 🗨️ 4

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-a113-03325e3b
397e.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-00a0d5c8
8979.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-3490ae59
a669.parquet
-rw-r--r--  3 student4_10 supergroup 2389502 2020-06-22 00:45 /user/student4_10/hw_5_1/paragraph/11592d45-8a3f-402b-9312-2ef781b1
e859.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 – exes
  - 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 сливов 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)

Home - Cloudera Manager

GeekBrains Clu... (CDH 5.15.2, Paros)

4 Hosts

Add Service

Add Hosts

Start

Stop

Restart

Rolling Restart

Deploy Client Configuration

Deploy Kerberos Client Configuration

Upgrade Cluster

Refresh Cluster

Refresh Dynamic Resource Pools

Inspect Hosts in Cluster

Enable Kerberos

Delete Kerberos Credentials

Charts

HDFS IO

Cluster Network IO

Cluster Disk IO

Cluster CPU

cloudera MANAGER

Support student4

### 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 HBase. Impala requires the Hive service and shares the Hive Metastore with Hive.

Back Continue

cloudera MANAGER

Support student4

### Add Flume Service to GeekBrains Cluster

Select the set of dependencies for your new Flume

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

Back Continue

cloudera MANAGER

Support student4

### 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.

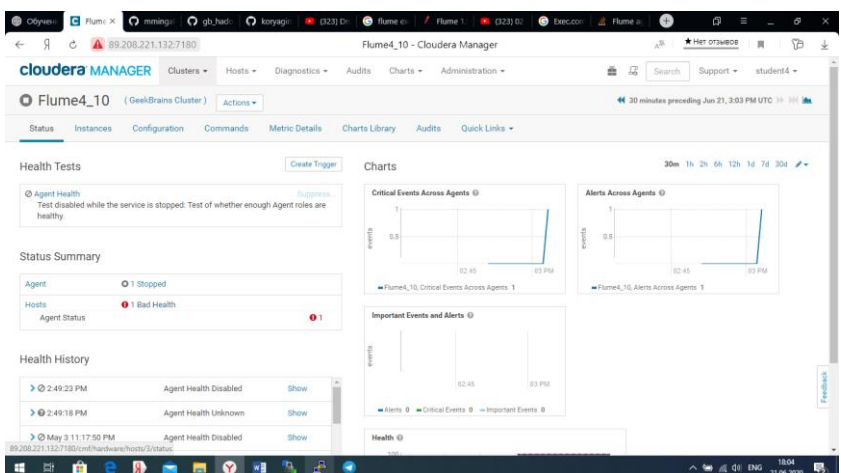
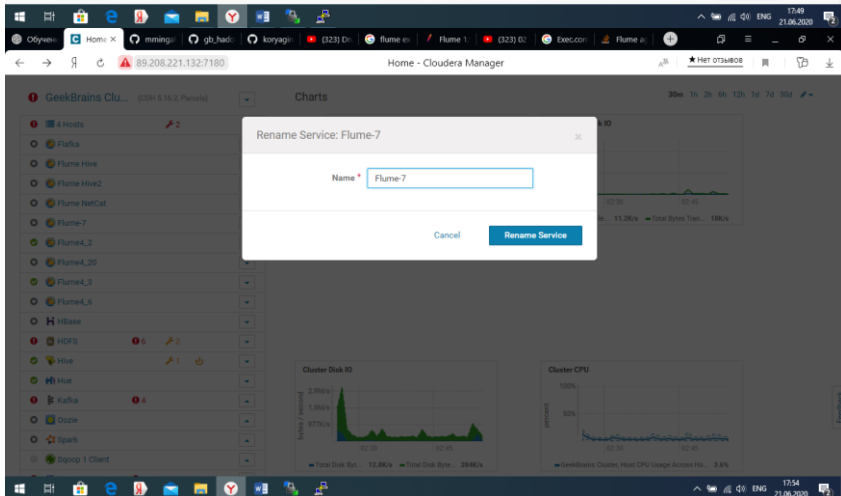
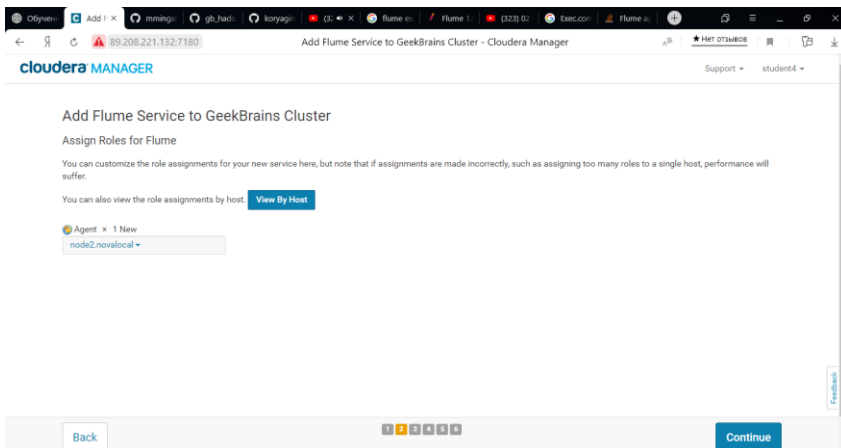
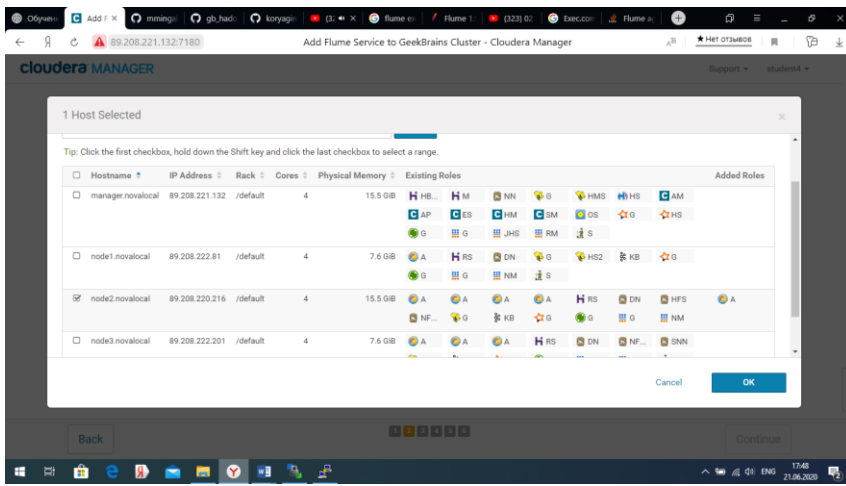
You can also view the role assignments by host. [View By Host](#)

☒ Agent

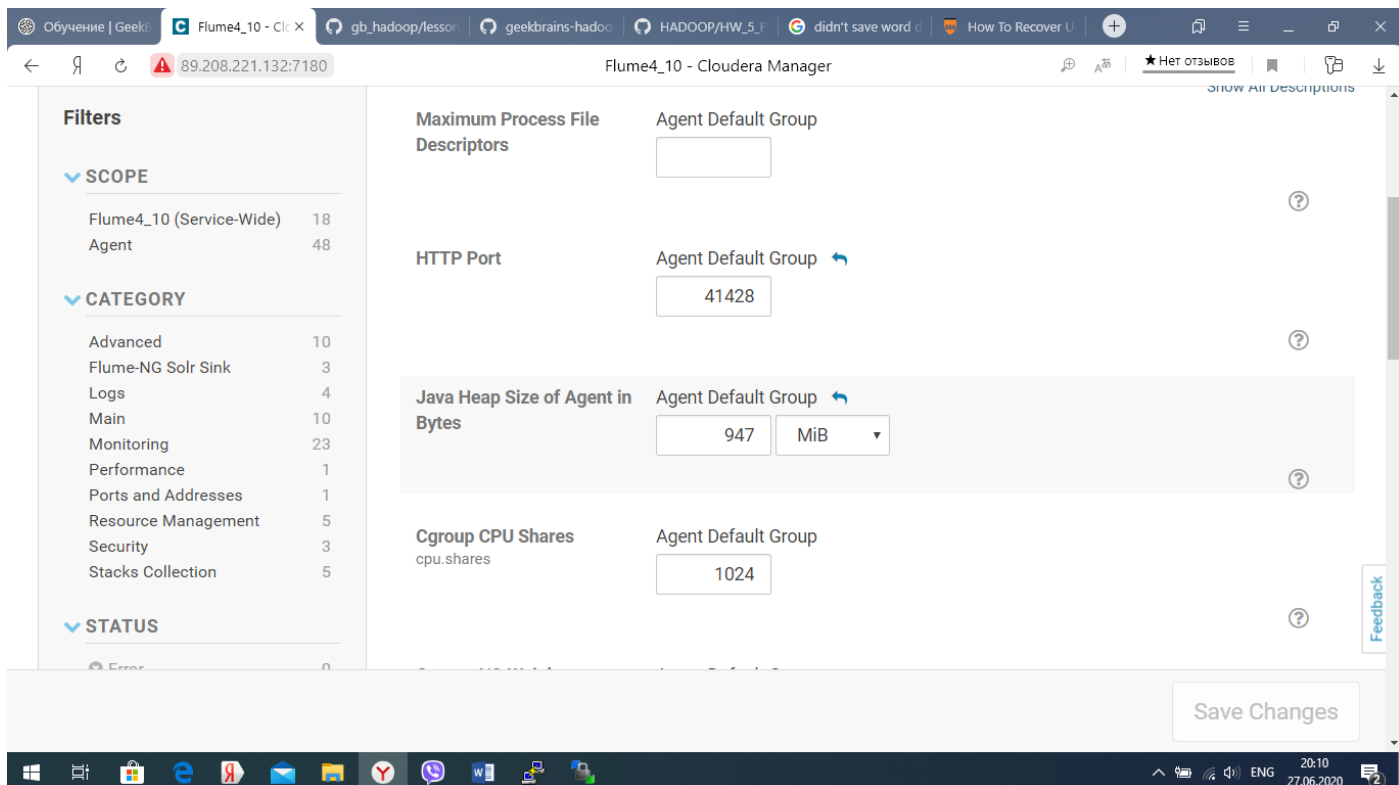
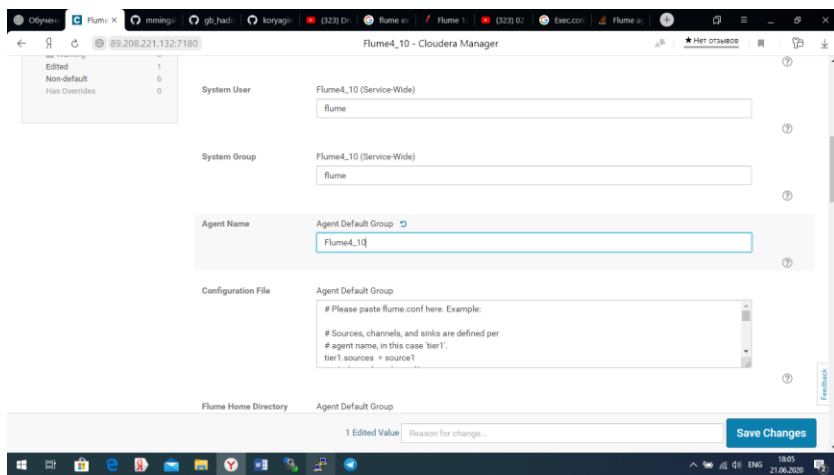
Select hosts

Too few hosts assigned, minimum is 1.

Back Continue



BAD HEALTH!



Поменяем лог на 41428

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

- Тип источника источник – **exes**
- Тип канала – **memory**
- Тип слива – **hdfs**

```
student4_10@manager:~  
student4_9 test  
systemd-private-59c34e5ab2ae47c89143caa30f443a67-chronyd.service-2Izgpm  
tmp_9fwltp  
tmpFNjPEg  
tmphhCooV  
tmpQag35  
[student4_10@manager ~]$ ls -r /var/log/cron  
/var/log/cron  
[student4_10@manager ~]$ less /var/log/cron  
Jun 22 03:06:02 manager run-parts(/etc/cron.daily)[8899]: finished logrotate  
Jun 22 03:06:02 manager run-parts(/etc/cron.daily)[8887]: starting man-db.cron  
Jun 22 03:06:14 manager run-parts(/etc/cron.daily)[9024]: finished man-db.cron  
Jun 22 03:06:14 manager run-parts(/etc/cron.daily)[8887]: starting mlocate  
Jun 22 03:07:40 manager run-parts(/etc/cron.daily)[9183]: finished mlocate  
Jun 22 03:07:40 manager anacron[8167]: Job `cron.daily' terminated  
Jun 22 03:26:01 manager anacron[8167]: Job `cron.weekly' started  
Jun 22 03:26:01 manager anacron[8167]: Job `cron.weekly' terminated  
Jun 22 03:26:01 manager anacron[8167]: Normal exit (2 jobs run)  
Jun 22 04:01:01 manager CROND[17396]: (root) CMD (run-parts /etc/cron.hourly)  
Jun 22 04:01:01 manager run-parts(/etc/cron.hourly)[17396]: starting 0anacron  
Jun 22 04:01:01 manager run-parts(/etc/cron.hourly)[17405]: finished 0anacron  
Jun 22 05:01:02 manager CROND[27032]: (root) CMD (run-parts /etc/cron.hourly)  
Jun 22 05:01:02 manager run-parts(/etc/cron.hourly)[27032]: starting 0anacron  
Jun 22 05:01:02 manager run-parts(/etc/cron.hourly)[27041]: finished 0anacron  
Jun 22 06:01:01 manager CROND[4784]: (root) CMD (run-parts /etc/cron.hourly)  
Jun 22 06:01:01 manager run-parts(/etc/cron.hourly)[4784]: starting 0anacron  
Jun 22 06:01:01 manager run-parts(/etc/cron.hourly)[4793]: finished 0anacron  
Jun 22 07:01:01 manager CROND[14107]: (root) CMD (run-parts /etc/cron.hourly)  
Jun 22 07:01:01 manager run-parts(/etc/cron.hourly)[14107]: starting 0anacron  
Jun 22 07:01:01 manager run-parts(/etc/cron.hourly)[14116]: finished 0anacron  
Jun 22 08:01:01 manager CROND[23604]: (root) CMD (run-parts /etc/cron.hourly)  
Jun 22 08:01:01 manager run-parts(/etc/cron.hourly)[23604]: starting 0anacron  
Jun 22 08:01:01 manager run-parts(/etc/cron.hourly)[23613]: finished 0anacron  
Jun 22 09:01:01 manager CROND[3052]: (root) CMD (run-parts /etc/cron.hourly)  
Jun 22 09:01:01 manager run-parts(/etc/cron.hourly)[3052]: starting 0anacron  
Jun 22 09:01:01 manager run-parts(/etc/cron.hourly)[3061]: finished 0anacron  
Jun 22 10:01:01 manager CROND[13502]: (root) CMD (run-parts /etc/cron.hourly)  
Jun 22 10:01:01 manager run-parts(/etc/cron.hourly)[13502]: starting 0anacron  
Jun 22 10:01:01 manager run-parts(/etc/cron.hourly)[13511]: finished 0anacron  
Jun 22 11:01:02 manager CROND[22980]: (root) CMD (run-parts /etc/cron.hourly)  
Jun 22 11:01:02 manager run-parts(/etc/cron.hourly)[22980]: starting 0anacron  
Jun 22 11:01:02 manager run-parts(/etc/cron.hourly)[22989]: finished 0anacron  
Jun 22 12:01:01 manager CROND[500]: (root) CMD (run-parts /etc/cron.hourly)  
Jun 22 12:01:01 manager run-parts(/etc/cron.hourly)[500]: starting 0anacron
```

работает с этим лог файлом

Var/log/cron

Создадим новую папку в разделе FLUME под именем Flume4\_10

```
student4_10@manager:~  
[student4_10@manager ~]$ hdfs dfs -ls /flume/  
Found 14 items  
drwxr-xr-x - flume flume 0 2020-06-23 17:54 /flume/Flume4_10_1  
drwxr-xr-x - flume flume 0 2020-06-25 16:09 /flume/Flume4_9_1  
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 2020-06-26 13:50 /flume/flume4_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  
drwxr-xr-x - flume flume 0 2020-06-25 11:58 /flume/student4_12  
[student4_10@manager ~]$ export HADOOP_USER_NAME=flume  
[student4_10@manager ~]$ hdfs dfs -mkdir /flume/Flume4_10  
[student4_10@manager ~]$ hdfs dfs -ls /flume/  
Found 15 items  
drwxr-xr-x - flume flume 0 2020-06-27 17:30 /flume/Flume4_10  
drwxr-xr-x - flume flume 0 2020-06-23 17:54 /flume/Flume4_10_1  
drwxr-xr-x - flume flume 0 2020-06-25 16:09 /flume/Flume4_9_1  
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 2020-06-26 13:50 /flume/flume4_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  
drwxr-xr-x - flume flume 0 2020-06-25 11:58 /flume/student4_12  
[student4_10@manager ~]$
```

Код для FLUME



```
*Новый текстовый документ - 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 = /bin/tailf /var/log/cron|
#Flume4_10.sources.ExecSource.command = /bin/tailf /tmp/myfile
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.HdfsSink.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
```

Обучение | GeekBrains | Flume4\_10 - Cloudera X | gb\_hadoop/lesson\_5 | geekbrains-hadoop-in | HADOOP/HW\_5\_FLUM | koryagin2006 (Andrey)

89.208.221.132:7180 Flume4\_10 - Cloudera Manager

cloudera MANAGER

Clusters Hosts Diagnostics Audits Charts Administration

✓ Flume4\_10 (GeekBrains Cluster) Actions Jun 27, 5:37 PM UTC

Status Instances Configuration Commands Metric Details Charts Library Audits Quick Links

Search Flume4\_10 on GeekBrains Cluster Switch to the classic layout Role Groups

Filters

SCOPE

Flume4_10 (Service-Wide)	18
Agent	48

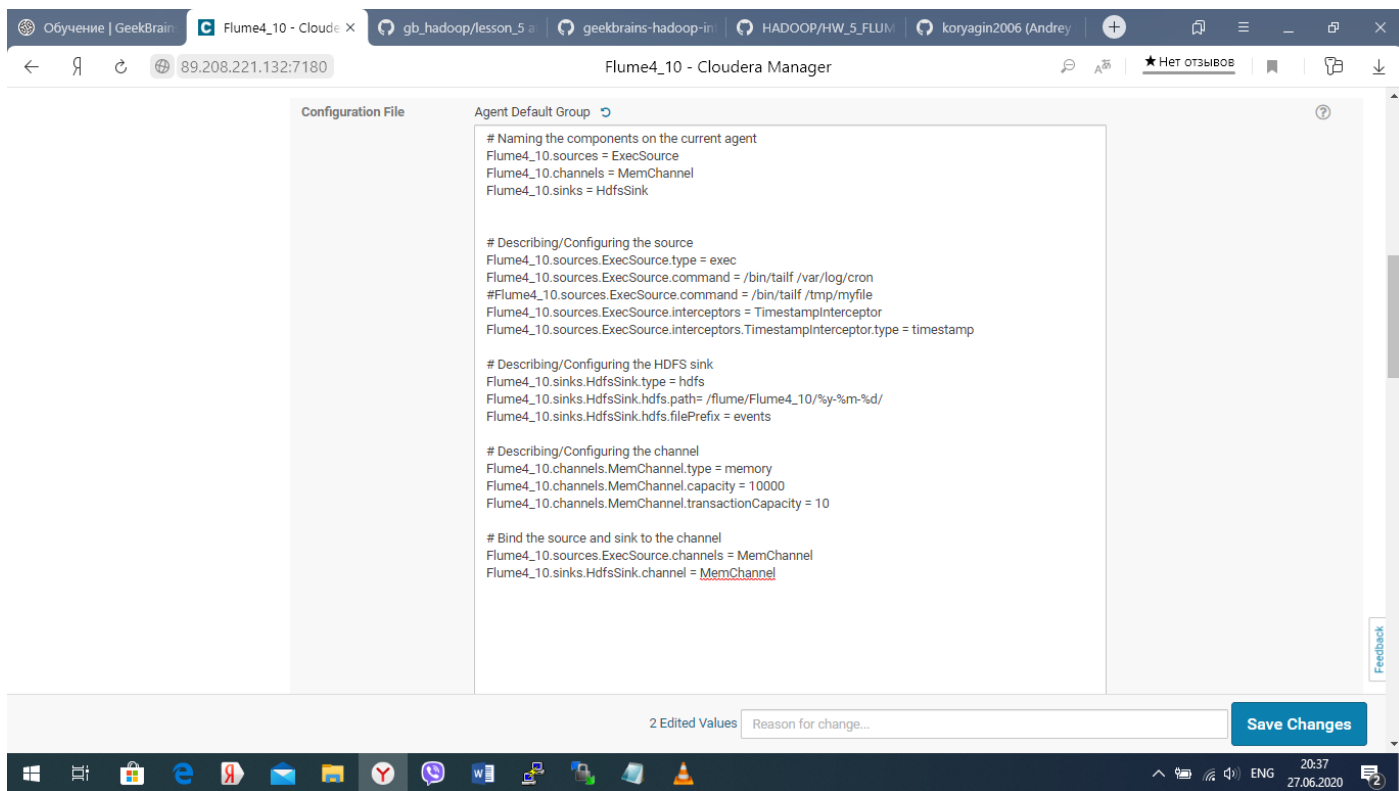
HDFS Service

Flume4\_10 (Service-Wide)

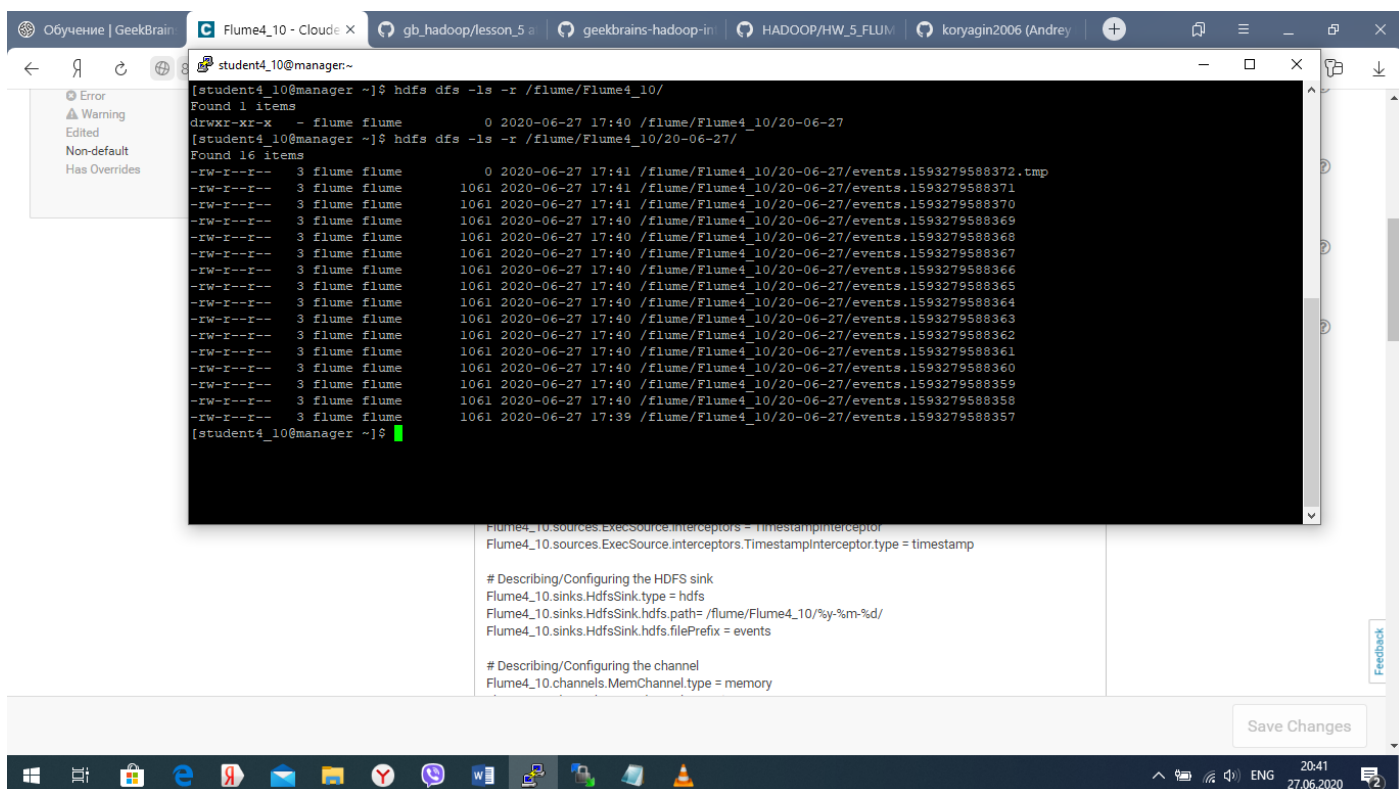
☒ HDFS

☐ none

2 Edited Values Reason for change... Save Changes



Заполняем кодом , потом сохраняем и запускаем Flume4\_10



Проверяем содержимое в папке flume

Обучение | GeekBrain | Flume4\_10 - Cloude X | gb\_hadoop/lesson\_5 | geekbrains-hadoop-in | HADOOP/HW\_5\_FLUM | koryagin2006 (Andrey) | 89.208.221.132:7180 | Flume4\_10 - Cloudera Manager | Нет отзывов

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 = /bin/tailf /var/log/cron
#Flume4_10.sources.ExecSource.command = /bin/tailf /tmp/myfile
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.HdfsSink.hdfs.path = /flume/Flume4_10/%y-%m-%d/
#Flume4_10.sinks.HdfsSink.hdfs.path = /home/student4_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
```

1 Edited Value Reason for change...

Save Changes

Feedback

Windows taskbar: 20:49 27.06.2020

student4\_10@manager~

```
Found 1 items
drwxrwxrwx - flume supergroup 0 2020-06-27 17:56 /home/student4_10/20-06-27
[student4_10@manager ~]$ hdfs dfs -ls -r /home/student4_10/20-06-27
Found 113 items
-rw-r--r-- 3 flume supergroup 0 2020-06-27 17:57 /home/student4_10/20-06-27/events.1593280041210.tmp
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:57 /home/student4_10/20-06-27/events.1593280041209
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:57 /home/student4_10/20-06-27/events.1593280041208
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:56 /home/student4_10/20-06-27/events.1593280041207
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:56 /home/student4_10/20-06-27/events.1593280041206
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:56 /home/student4_10/20-06-27/events.1593280041205
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:56 /home/student4_10/20-06-27/events.1593280041204
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:56 /home/student4_10/20-06-27/events.1593280041203
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:56 /home/student4_10/20-06-27/events.1593280041202
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:56 /home/student4_10/20-06-27/events.1593280041201
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:56 /home/student4_10/20-06-27/events.1593280041200
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:56 /home/student4_10/20-06-27/events.1593280041199
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:56 /home/student4_10/20-06-27/events.1593280041198
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:56 /home/student4_10/20-06-27/events.1593280041197
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:56 /home/student4_10/20-06-27/events.1593280041196
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:55 /home/student4_10/20-06-27/events.1593280041195
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:55 /home/student4_10/20-06-27/events.1593280041194
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:55 /home/student4_10/20-06-27/events.1593280041193
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:55 /home/student4_10/20-06-27/events.1593280041192
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:55 /home/student4_10/20-06-27/events.1593280041191
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:55 /home/student4_10/20-06-27/events.1593280041190
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:55 /home/student4_10/20-06-27/events.1593280041189
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:55 /home/student4_10/20-06-27/events.1593280041188
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:55 /home/student4_10/20-06-27/events.1593280041187
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:55 /home/student4_10/20-06-27/events.1593280041186
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:55 /home/student4_10/20-06-27/events.1593280041185
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:54 /home/student4_10/20-06-27/events.1593280041184
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:54 /home/student4_10/20-06-27/events.1593280041183
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:54 /home/student4_10/20-06-27/events.1593280041182
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:54 /home/student4_10/20-06-27/events.1593280041181
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:54 /home/student4_10/20-06-27/events.1593280041180
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:54 /home/student4_10/20-06-27/events.1593280041179
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:54 /home/student4_10/20-06-27/events.1593280041178
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:54 /home/student4_10/20-06-27/events.1593280041177
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:54 /home/student4_10/20-06-27/events.1593280041176
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:54 /home/student4_10/20-06-27/events.1593280041175
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:54 /home/student4_10/20-06-27/events.1593280041174
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:54 /home/student4_10/20-06-27/events.1593280041173
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:53 /home/student4_10/20-06-27/events.1593280041172
-rw-r--r-- 3 flume supergroup 1061 2020-06-27 17:53 /home/student4_10/20-06-27/events.1593280041171
```

Windows taskbar: 21:04 27.06.2020

## Вопрос

По поводу прав для того чтобы делать SINK и SOURCE с файлами FLUME я должен иметь права на папки в директории как student4\_10,

другими словами я не совсем понимаю какие права должны быть у FLUME agent а какие у STUDENT4\_10

А вот в каких случаях Flume должен иметь права, так как HADOOP\_USER\_NAME не нужен если ты работаешь за пределами папки FLUME

```
[student4_10@manager ~]$ export HADOOP_USER_NAME=flume
```

```
[student4_10@manager ~]$ hdfs dfs -mkdir /home/student4_10/new
```

```
[student4_10@manager ~]$ ls /students/student4_10/
```

```
mapper.py reducer.py
```