

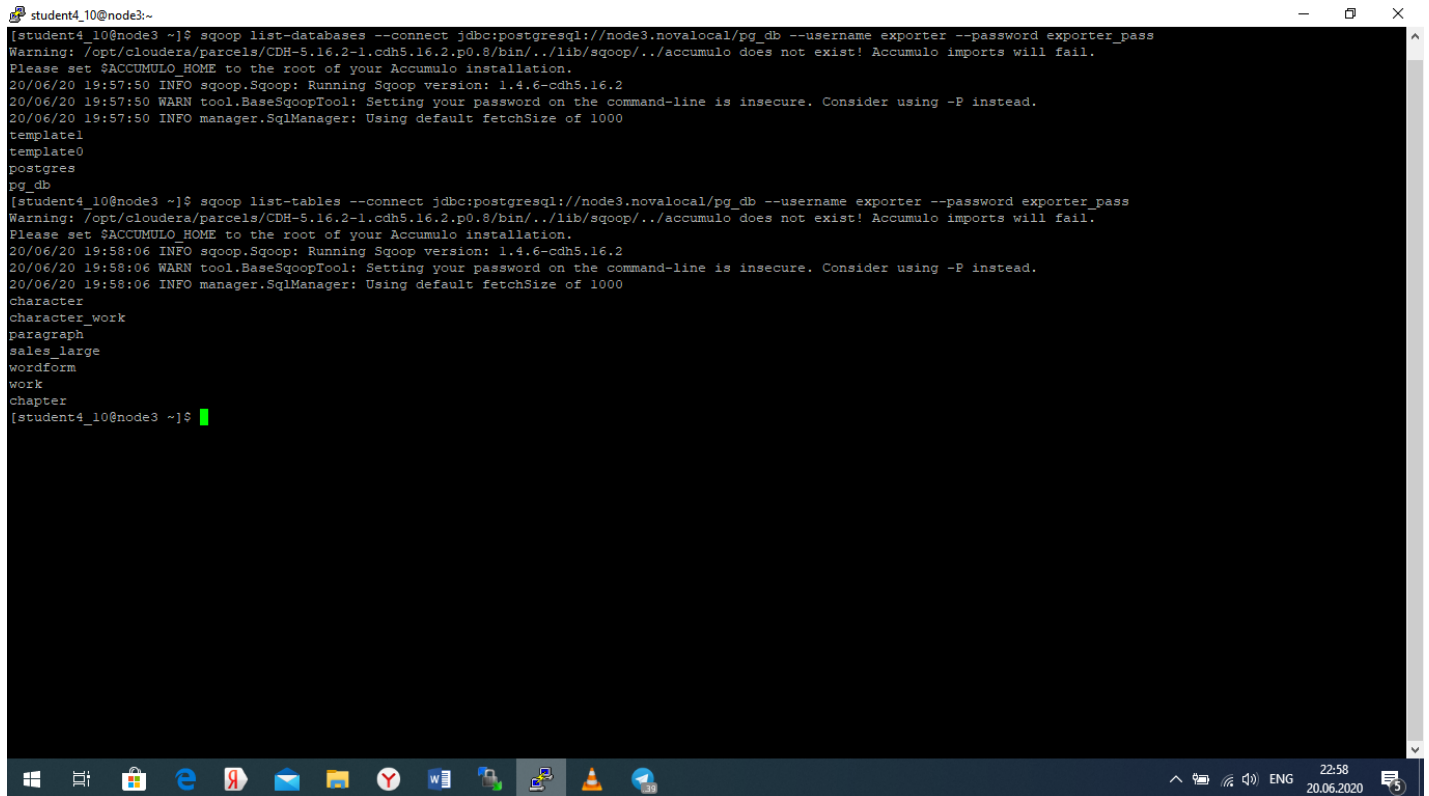
## 1. Для части по SQOOP

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

- в Hive-таблицу (--hive-import)
- в HDFS в формате avro (--as-avrodatafile)
- в HDFS в формате sequencefile (--as-sequencefile)

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

Посмотреть при помощи SQOOP содержимое в PostgreSQL.



```
[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`

Finish it later

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)

The screenshot shows the Cloudera Manager web interface. The top navigation bar includes 'Home', 'Status', 'All Health Issues', 'Configuration', 'Audits', 'Charts', and 'Administration'. The main content area is titled 'Add Service to GeekBrains Cluster'. It displays a list of service types with their descriptions. The 'Flume' service is selected. Below this, there is a table to select dependencies for the new Flume service. The table has columns for HBase, HDFS, Kafka, and ZooKeeper. The 'Flume' service is selected, and the 'HDFS' dependency is chosen. The 'Continue' button is visible at the bottom right of the page.

**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 1.8 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](#)

**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](#)

**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.

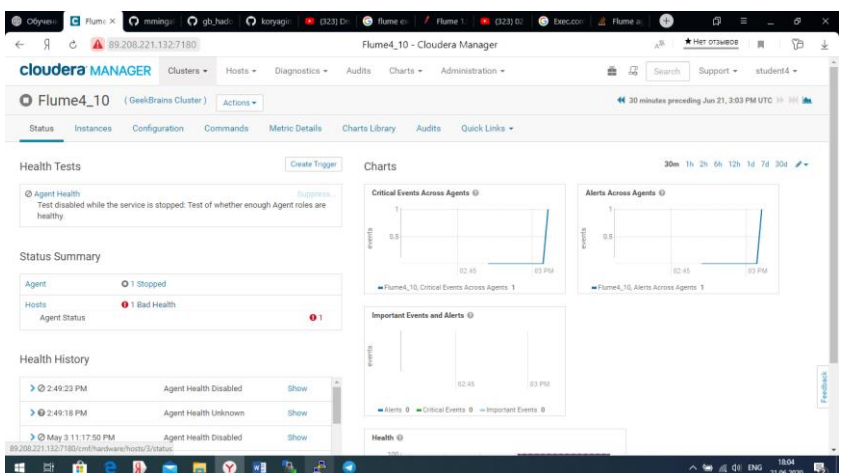
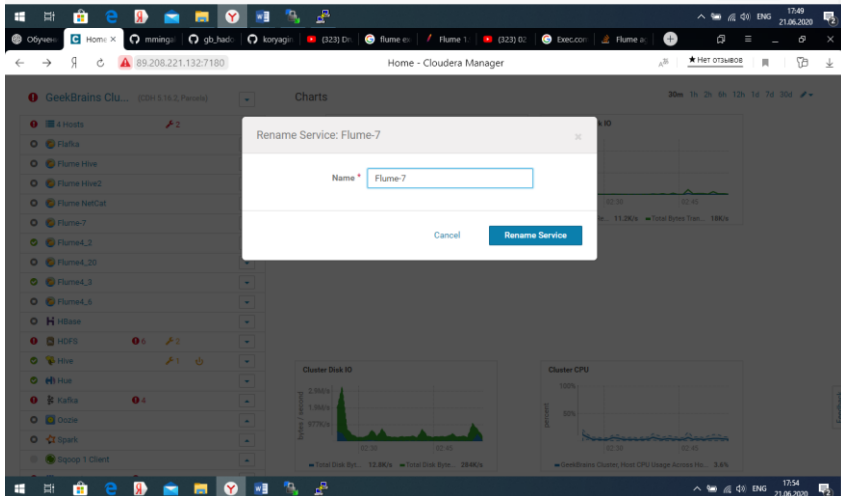
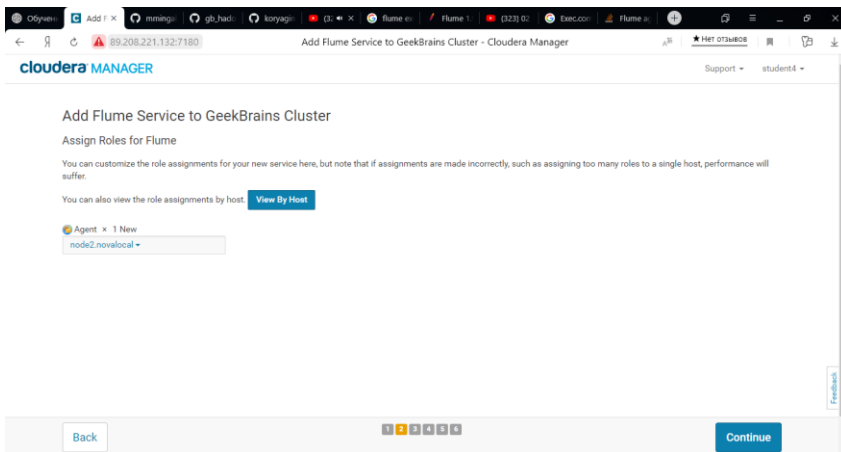
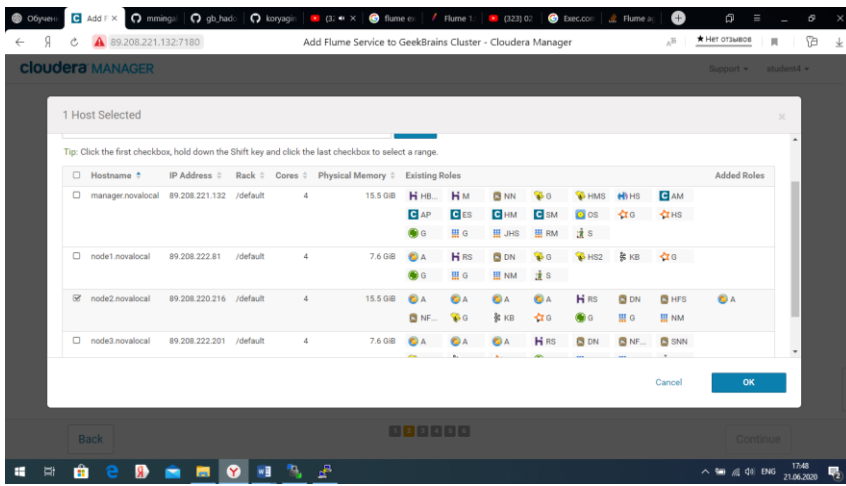
You can also view the role assignments by host. [View By Host](#)

**Agent**

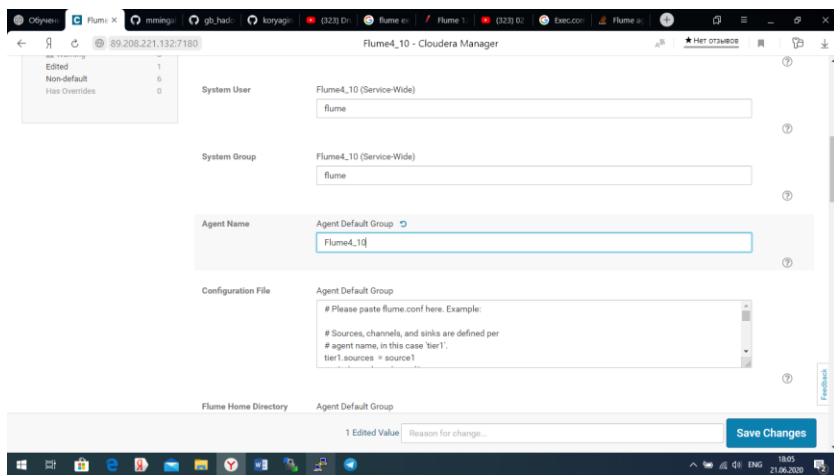
Select hosts

Too few hosts assigned, minimum is 1.

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



BAD HEALTH!

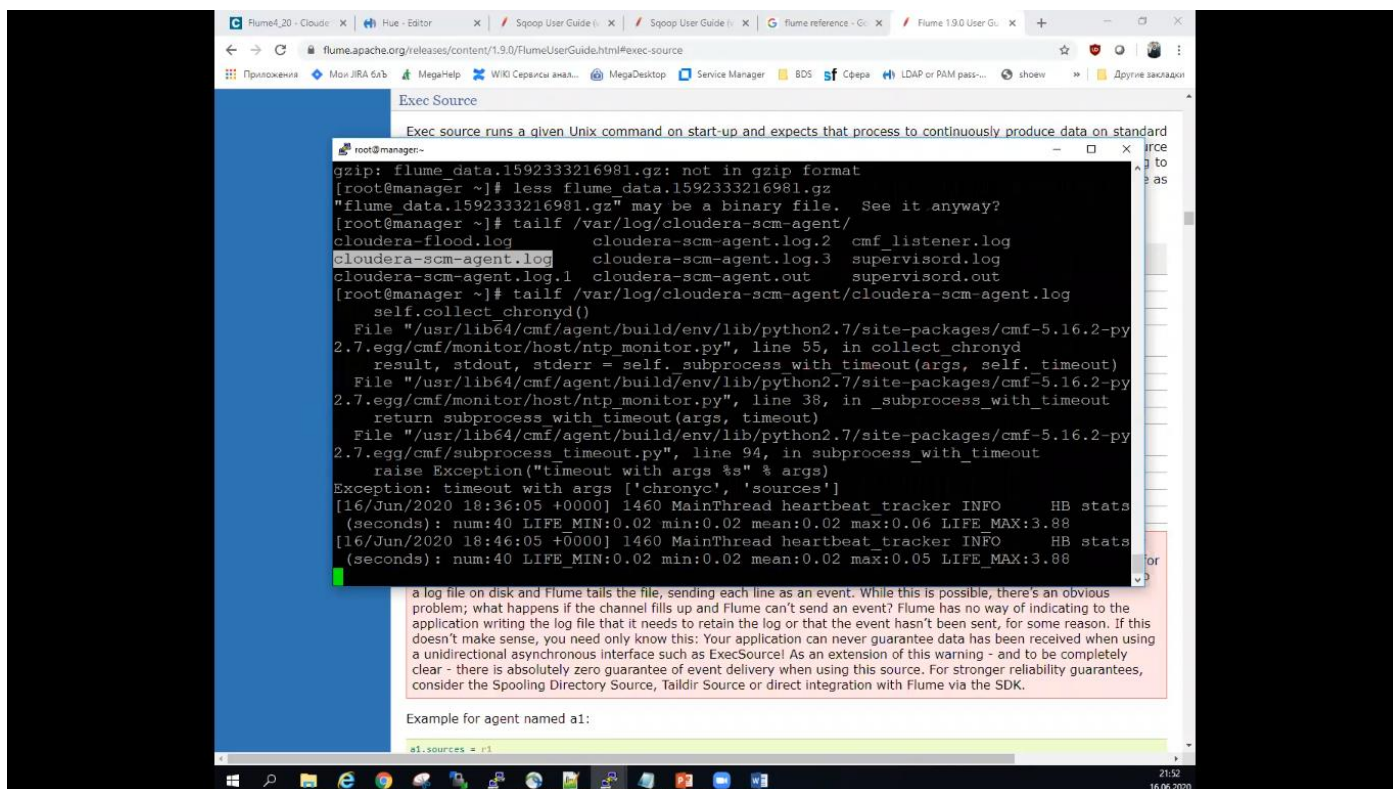


## 2. Создать любой Flume поток используя Flume сервис соответствующего номера.

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

tailf /var/log/cloudera-scm-agent/cloudera-scm-agent.log

работает с этим лог файлом



## 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-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 ~]$
```

```
*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.HdfsSink.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

Обучени Flume x mmingal gb\_had koryagin (32) X flume ex Flume 1. (323) 02 Exec.com Flume ac

89.208.221.132:7180 Flume4\_10 - Cloudera Manager

Configuration File Agent Default Group

```
# 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
```

Configuration changes have been saved successfully.

Save Changes

19:14 21.06.2020

Обучени Flume x mmingal gb\_had koryagin (32) X flume ex Flume 1. (323) 02 Exec.com Flume ac

89.208.221.132:7180 Flume4\_10 - Cloudera Manager

cloudera MANAGER Clusters Hosts Diagnostics Audits Charts Administration

Flume4\_10 (GeekBrains Cluster) Actions

Status Instances Configuration

Start Stop Restart Rolling Restart Add Role Instances Rename Enter Maintenance Mode Update Configuration

Health Tests

Agent Health

Status Summary

Agent 1 Stopped

Hosts 1 Bad Health

Health History

Time	Event	Action
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
May 3 11:17:28 PM	Agent Health Unknown	Show
Apr 29 10:16 PM	Agent Health Disabled	Show

Charts

Critical Events Across Agents

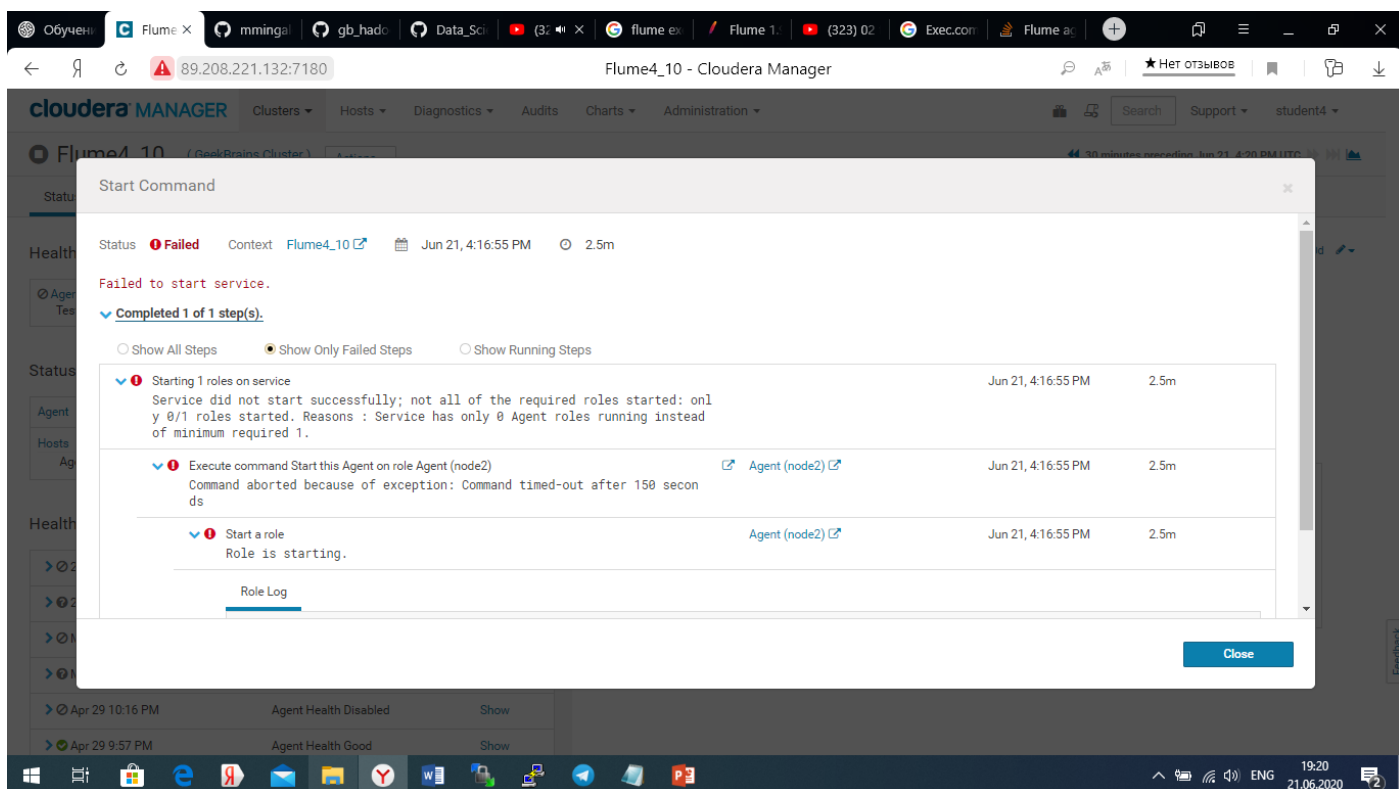
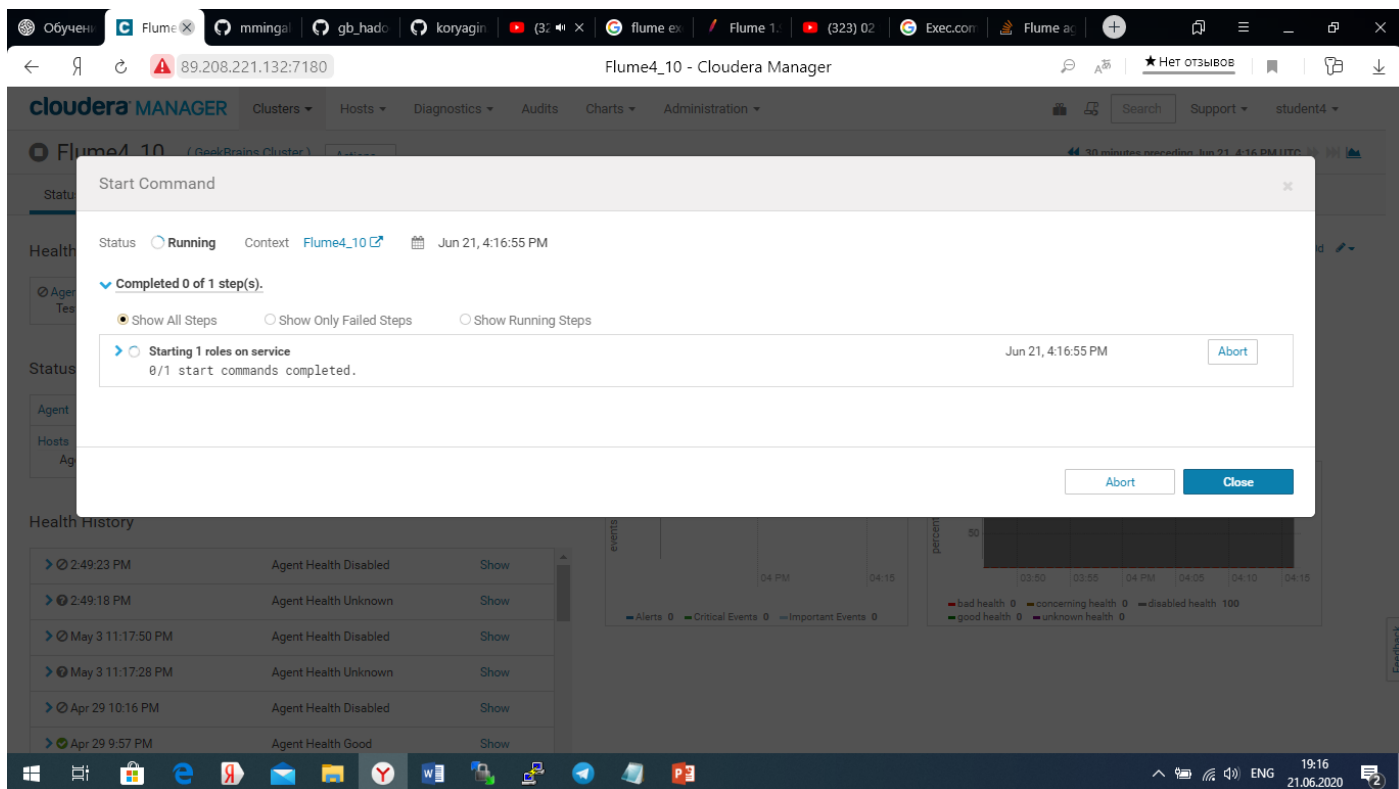
Alerts Across Agents

Important Events and Alerts

Health

19:16 21.06.2020

**BAD HEALTH !!!!**



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



Обучени Flume X mmingal gb\_had Data\_Sci (32 4 X flume ex Flume 1. (323) 02 Exec.con Flume a

89.208.221.132:7180 Flume4\_2 - Cloudera Manager

Agent Name Agent Default Group Flume4\_2

Configuration File

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.HdfsSink.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

19:27 21.06.2020

## Save changes and restart

Обучени Flume X mmingal gb\_had Data\_Sci (32 4 X flume ex Flume 1. (323) 02 Exec.con Flume a

89.208.221.132:7180 Flume4\_2 - Cloudera Manager

cloudera MANAGER Clusters Hosts Diagnostics Audits Charts Administration

Flume4\_2 (GeekBrains Cluster)

Status

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	<a href="#">Flume4_2</a>	Jun 21, 4:29:58 PM	1.55s
Execute command Start on service Flume4_2	<a href="#">Flume4_2</a>	Jun 21, 4:30:00 PM	22.3s

Close

Save Changes

19:30 21.06.2020



```
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 18:02 /flume/flume-7/exec-file-hdfs-v2
drwxr-xr-x - flume flume 0 2020-04-19 18:05 /flume/flume-7/exec-file-hdfs-v3
drwxr-xr-x - flume flume 0 2020-04-19 18:21 /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/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/
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 ~]$
```

нету ничего

Попробую flume-7 в configuration file

Обучение | Flume4\_2 | mmingalov | Data\_Science | gb\_hadoop | (324) | Flume 1.9 | (324) 02 fl | Exec.com | Flume ag |

89.208.221.132:7180 Flume4\_2 - Cloudera Manager

cloudra MANAGER Clusters Hosts Diagnostics Audits Charts Administration

Flume4\_2 (Spark/BigData Cluster)

Status

Start Command

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	Jun 21, 4:51:27 PM	22.35s
Successfully started 1 roles on service.		
Execute command Start this Agent on role Agent (node3)	Agent (node3)	Jun 21, 4:51:27 PM 22.34s

Close

Save Changes

19:52 21.06.2020

Обучение | Flume4\_2 | mmingalov | Data\_Science | gb\_hadoop | (324) | Flume 1.9.0 | (324) 02 | Exec.com | Flume agent | + | \* Нет отзывов

89.208.221.132:7180 Flume4\_2 - Cloudera Manager

Agent Name: Flume4\_2

Configuration File

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
```

Save Changes

## Flume-7

```
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 18:58 /flume/flume-7/exec-file-hdfs
s
drwxr-xr-x - flume flume 0 2020-04-20 22:59 /flume/flume-7/exec-file-hdfs
s-v10
drwxr-xr-x - flume flume 0 2020-04-19 19:02 /flume/flume-7/exec-file-hdfs
s-v2
drwxr-xr-x - flume flume 0 2020-04-19 19:05 /flume/flume-7/exec-file-hdfs
s-v3
drwxr-xr-x - flume flume 0 2020-04-19 19:21 /flume/flume-7/exec-file-hdfs
s-v4
drwxr-xr-x - flume flume 0 2020-04-20 18:47 /flume/flume-7/exec-file-hdfs
s-v5
drwxr-xr-x - flume flume 0 2020-04-20 22:06 /flume/flume-7/exec-file-hdfs
s-v6
drwxr-xr-x - flume flume 0 2020-04-20 22:20 /flume/flume-7/exec-file-hdfs
s-v7
drwxr-xr-x - flume flume 0 2020-04-20 22:30 /flume/flume-7/exec-file-hdfs
s-v8
drwxr-xr-x - flume flume 0 2020-04-20 22:33 /flume/flume-7/exec-file-hdfs
s-v9
[student4_10@manager ~]$
```

Не работает