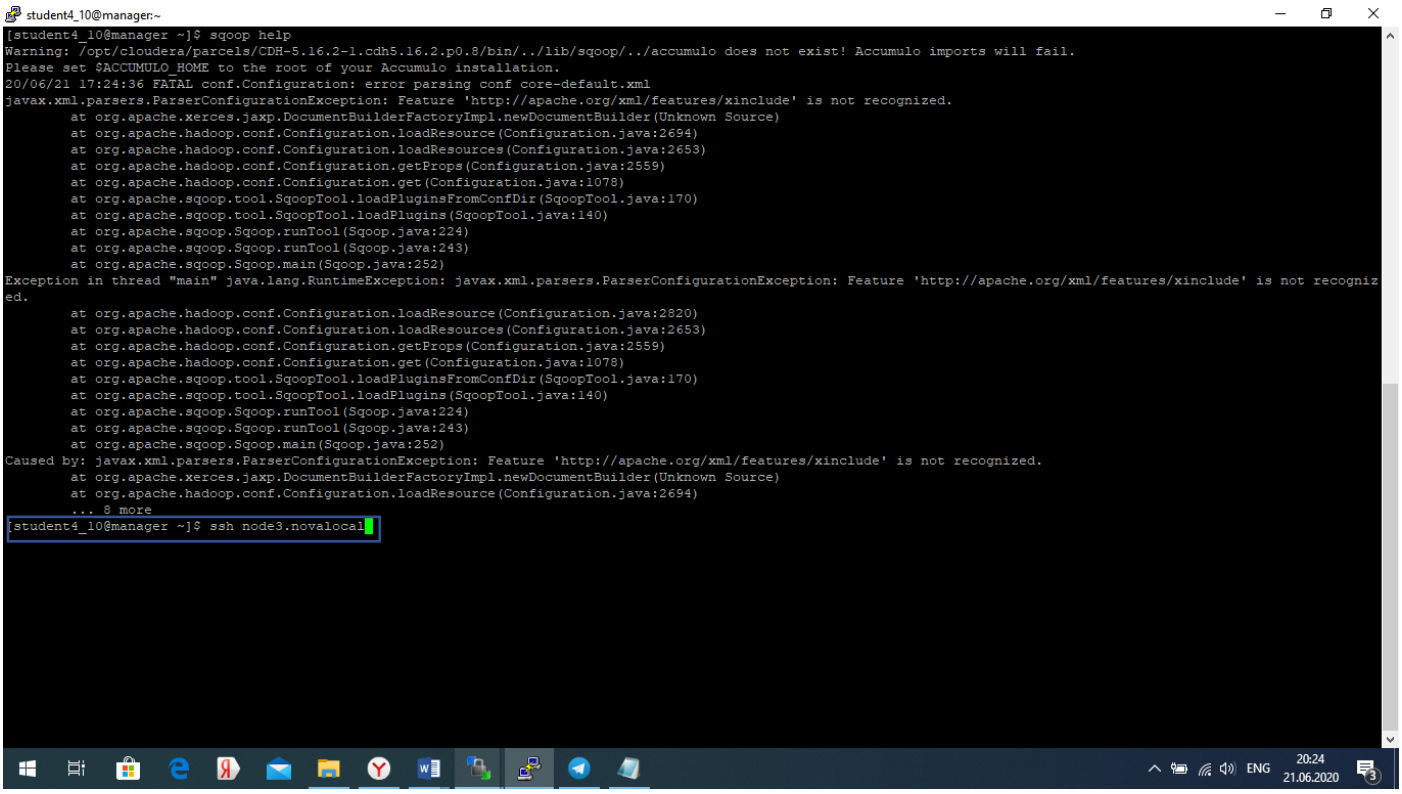
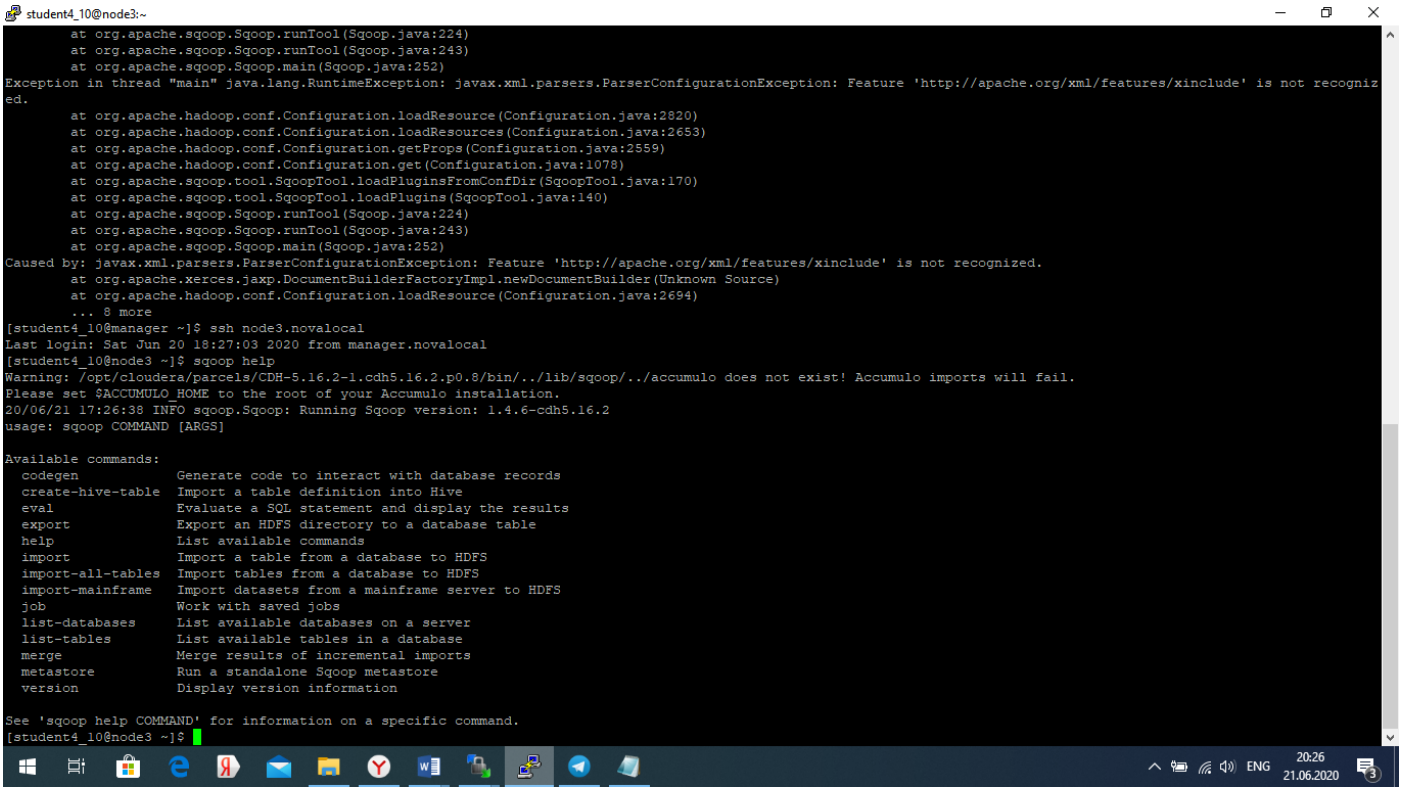
1. Для части по SQOOP
Провести импорт таблицы из вашего сервера БД в Hadoop с использованием SQOOP в любых двух вариантах из перечисленных ниже.
a. в Hive-таблицу (--hive-import)
b. в HDFS в формате avro (--as-avrodatafile)
c. в HDFS в формате sequencefile (--as-sequencefile)
Если у вас нет своего сервера то можно использовать тот Postgres, который я показал на лекции. Пароль expoter_pass

Посмотрим при помощи SQOOP содержимое в PosgreSQL.

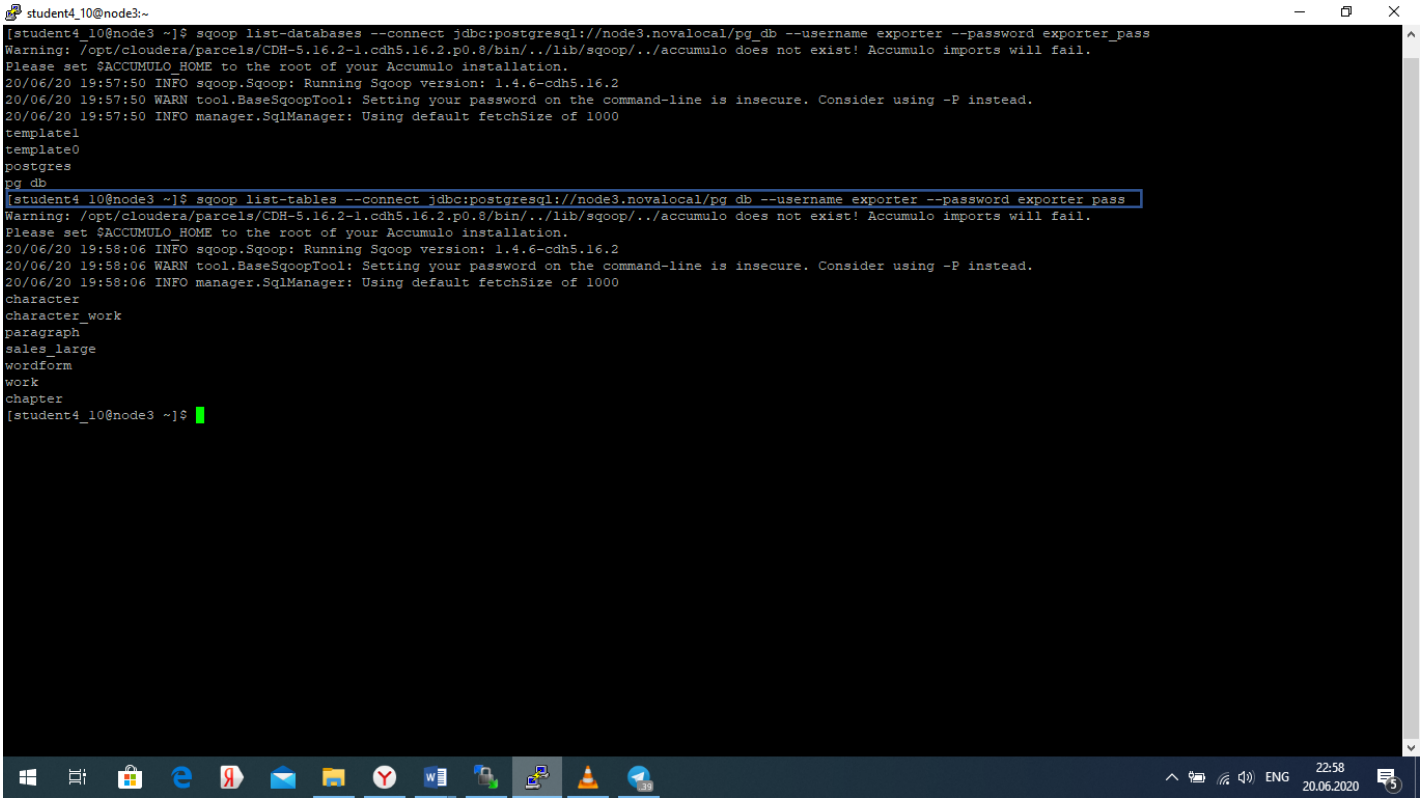


Sqoop Help не работает, через ssh перехожу на node3.novalocal



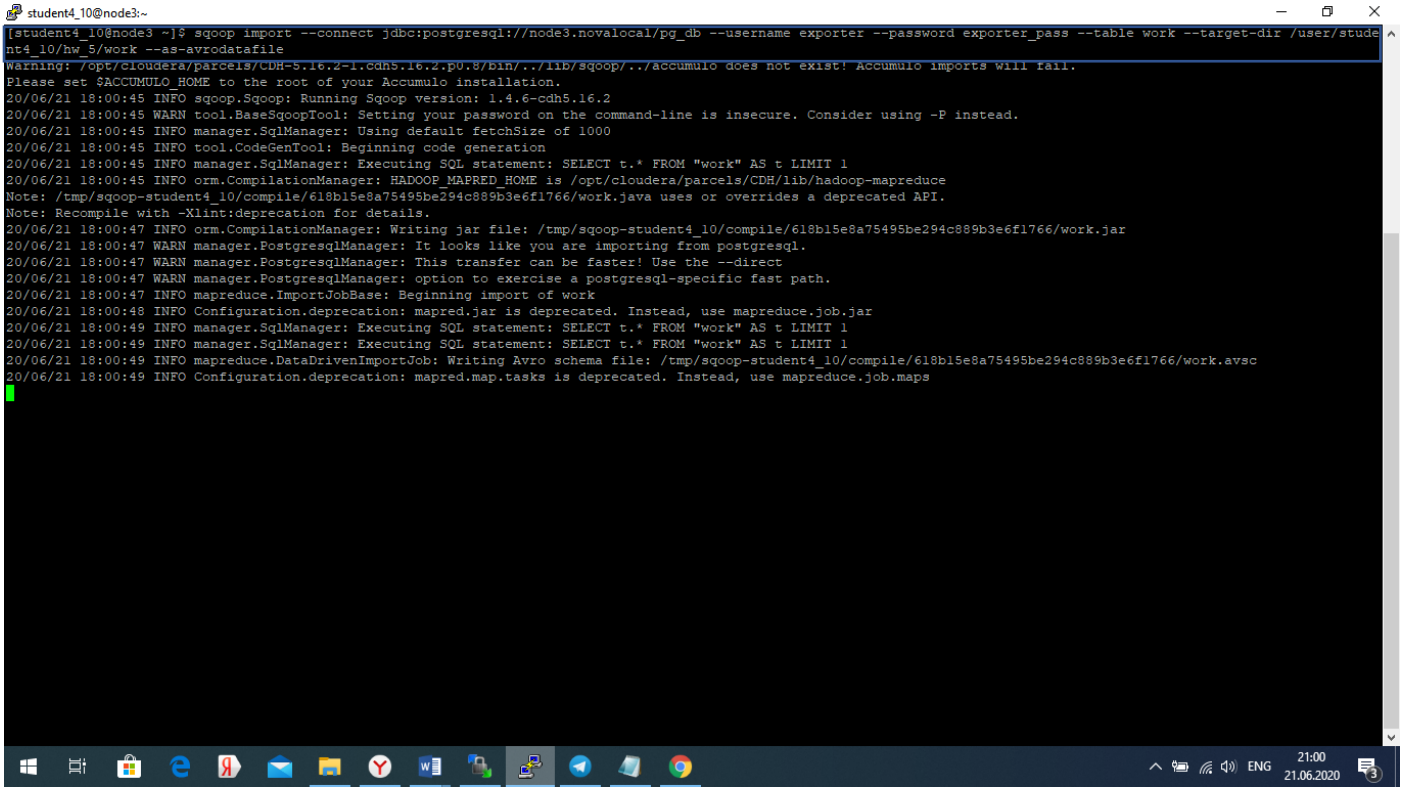
работает !

Проверим таблицы в базе pg_db

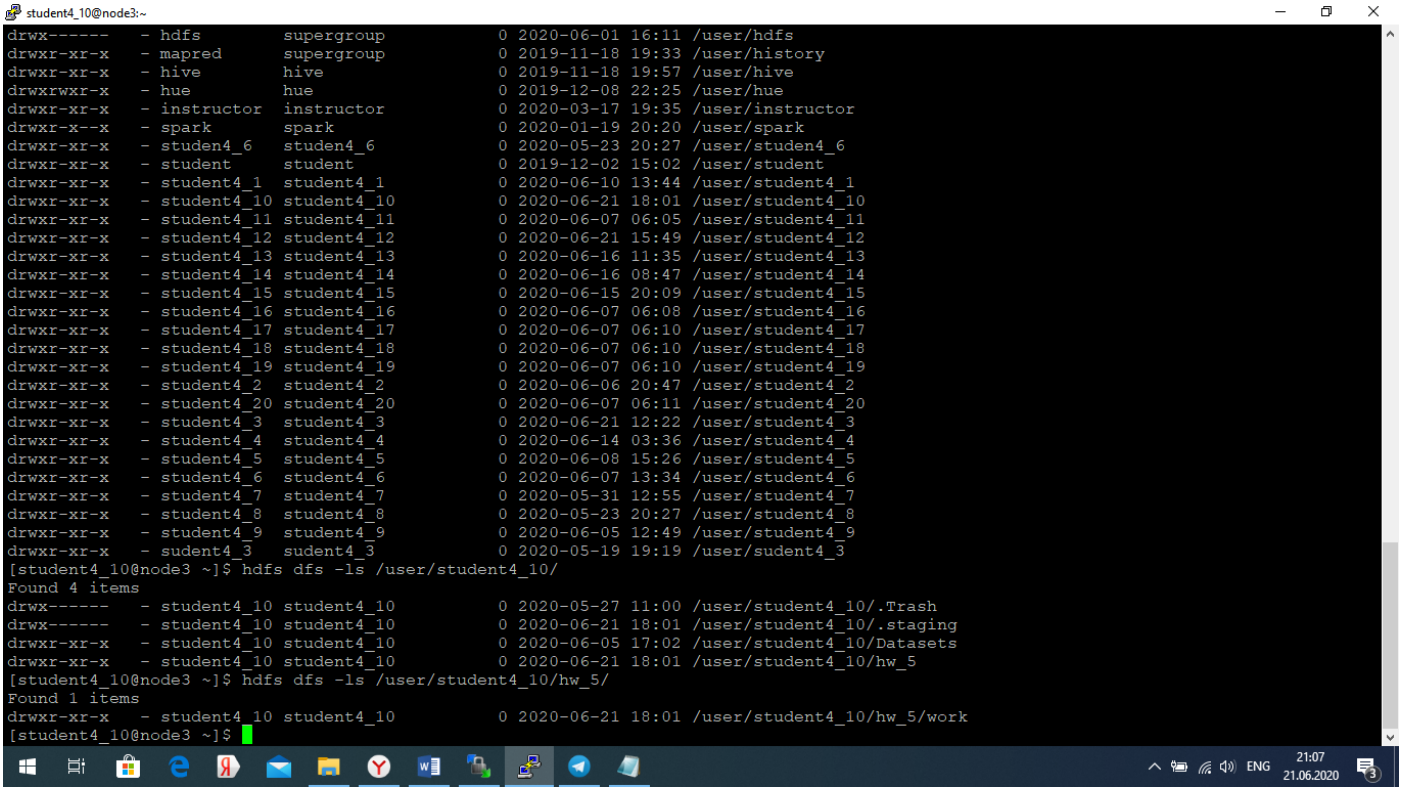


sqoop list-tables --connect jdbc:postgresql://node3.novalocal/pg_db --username exporter --password exporter_pass

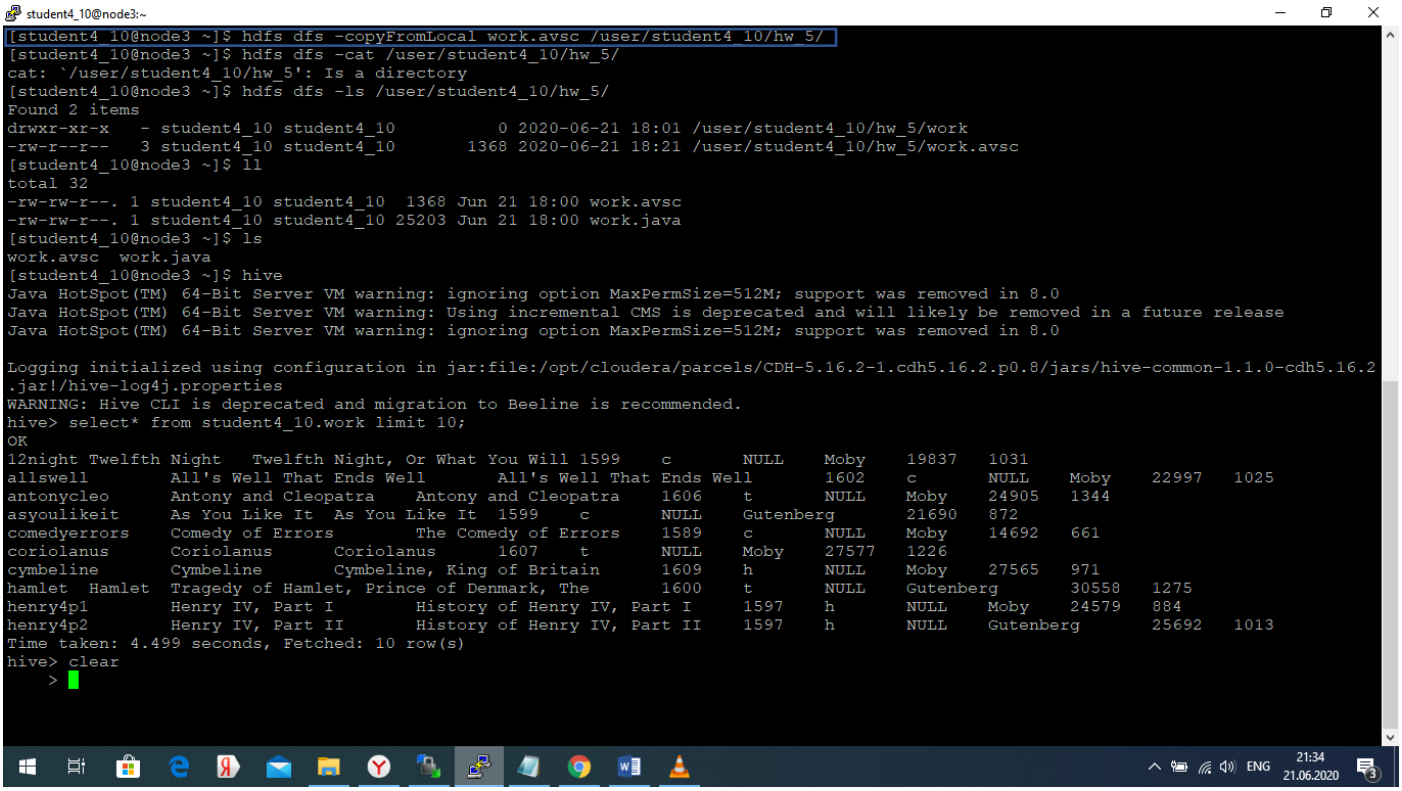
Скопируем Таблицу Work в локальную папку



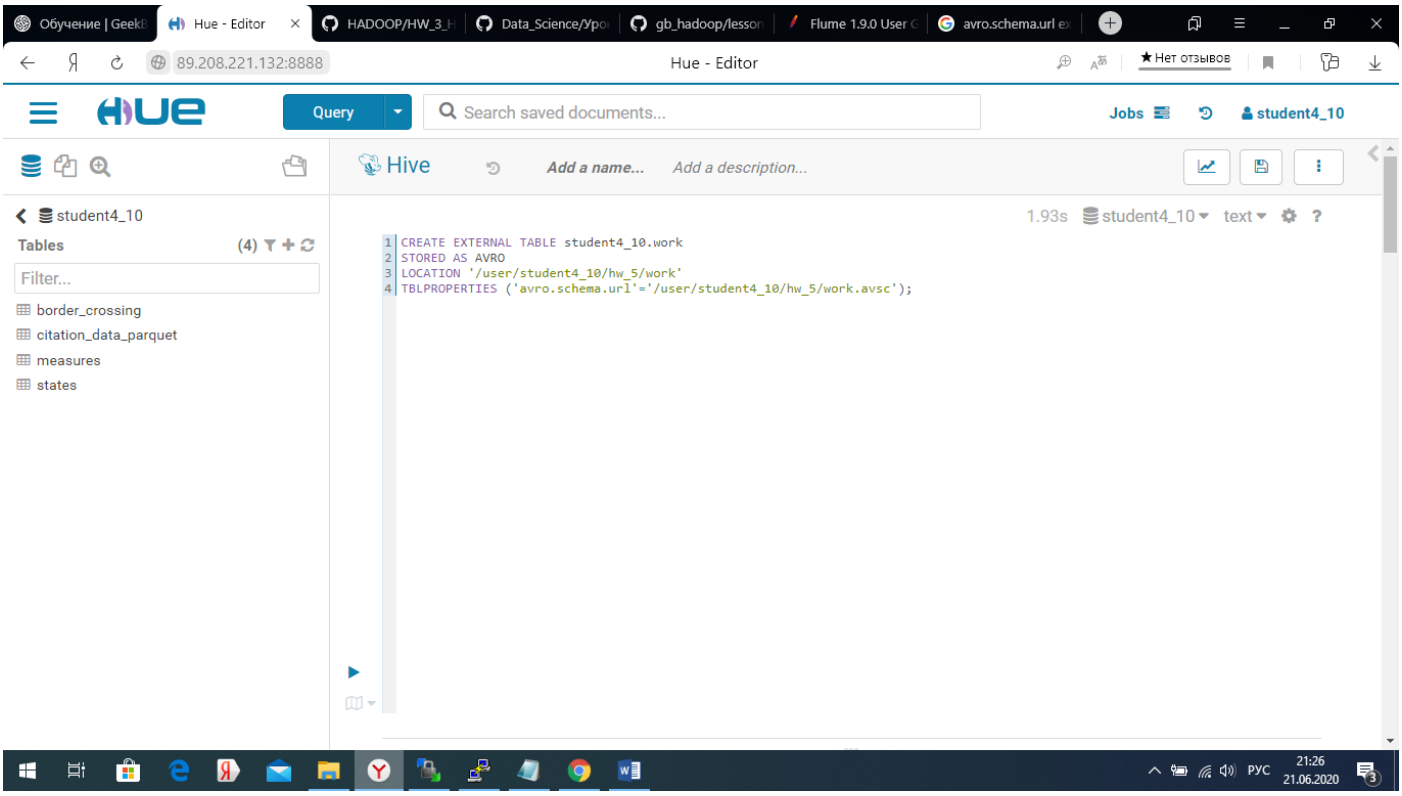
Проверим папку /user/student4_10

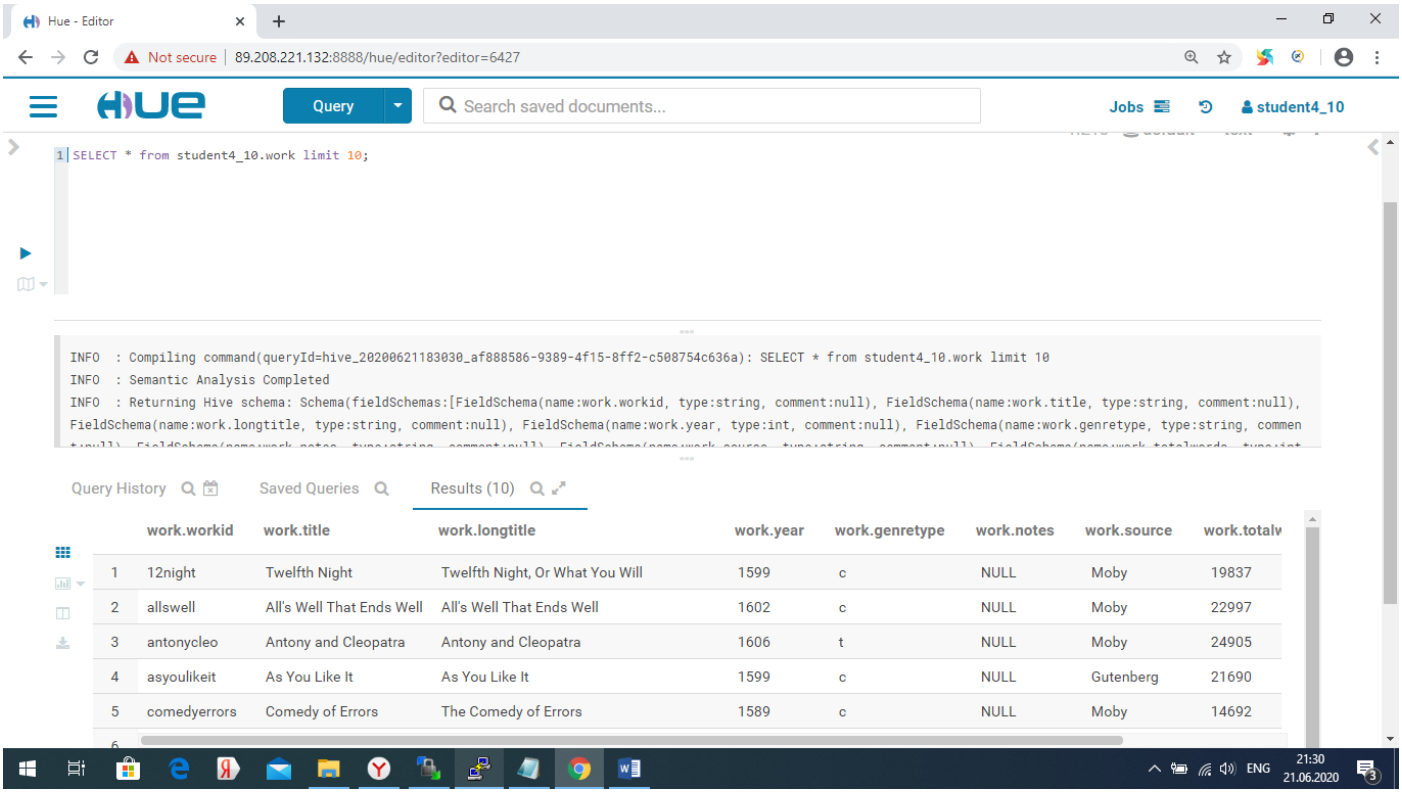


Скопируем схему структуры таблицы с локальной директории через команду COPYFROMLOCAL



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





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

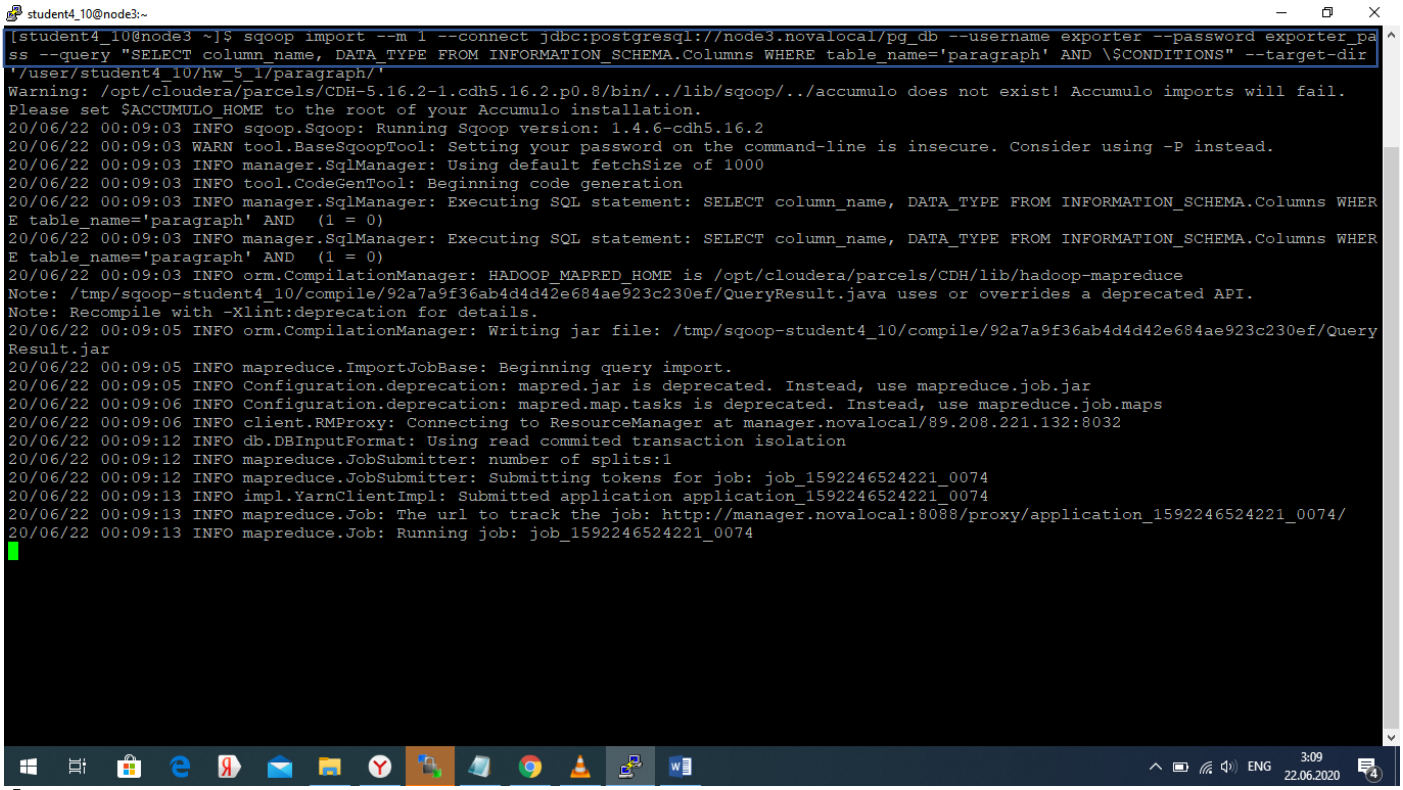
Поэтому через Sqoop проверим схему команды

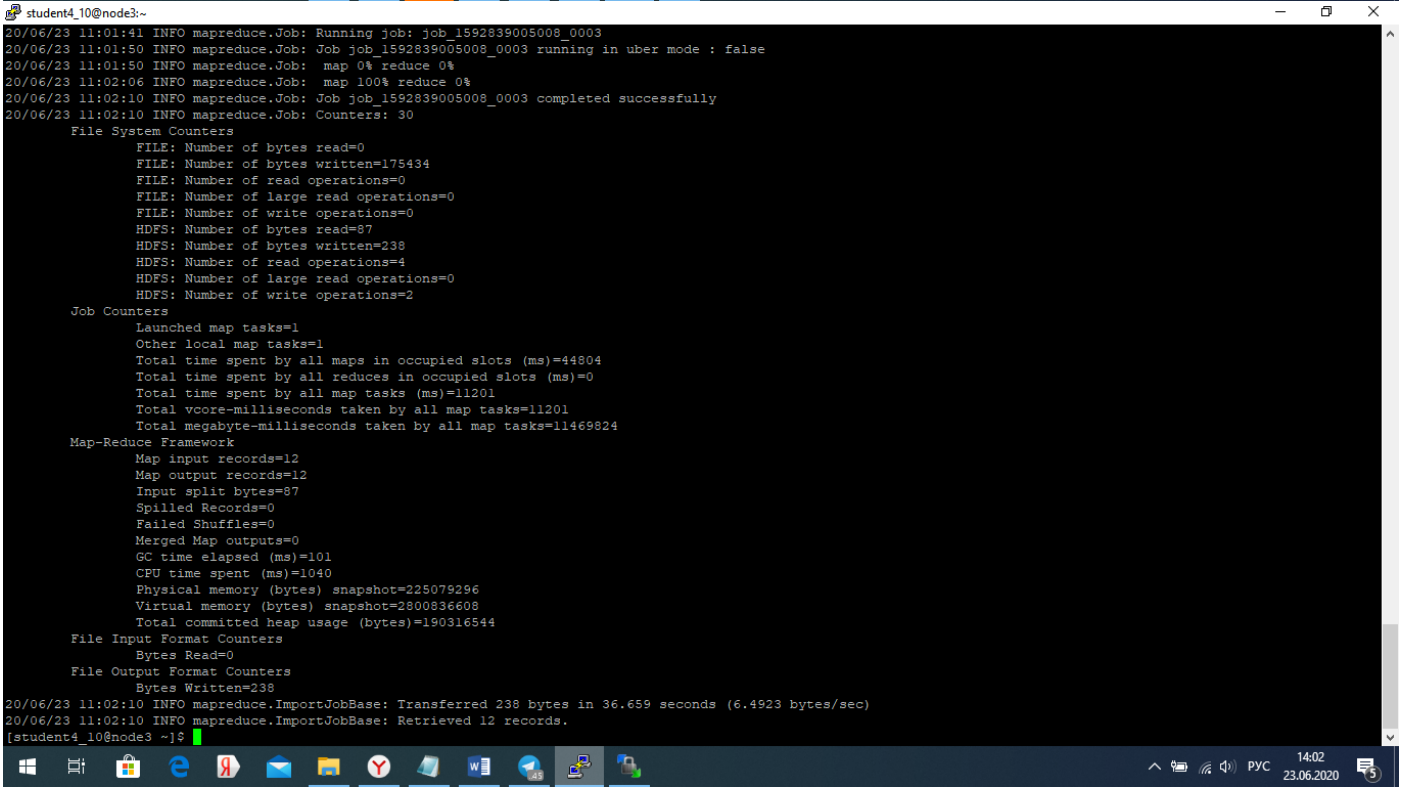
Схема для таблицы work

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

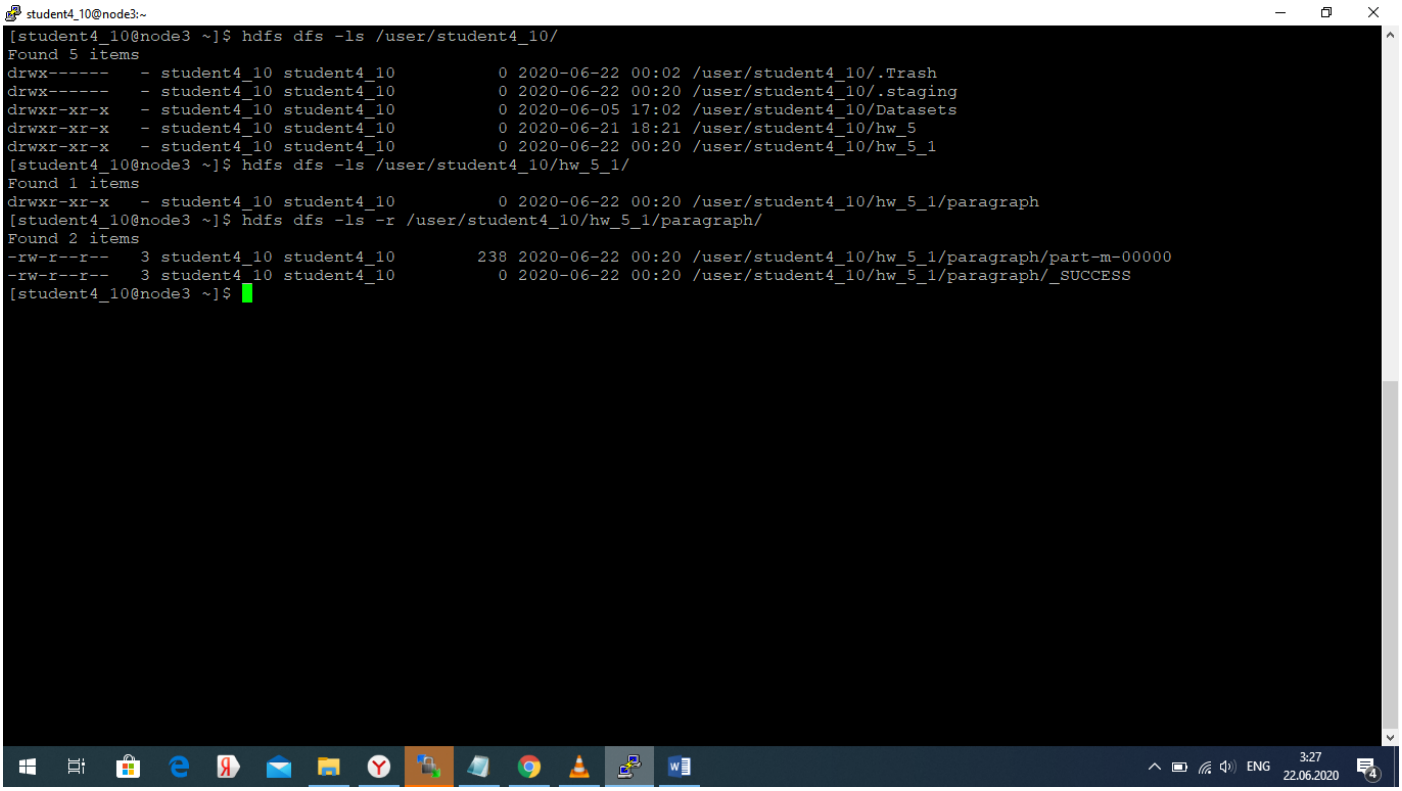
sqoop import --m 1 --connect jdbc:postgresql://node3.novalocal/pg_db --username exporter --password exporter_pass --query "SELECT column_name, DATA_TYPE FROM INFORMATION_SCHEMA.Columns WHERE table_name='paragraph' AND \$CONDITIONS" --target-dir '/user/student4_10/hw_5_1/**paragraph/'**

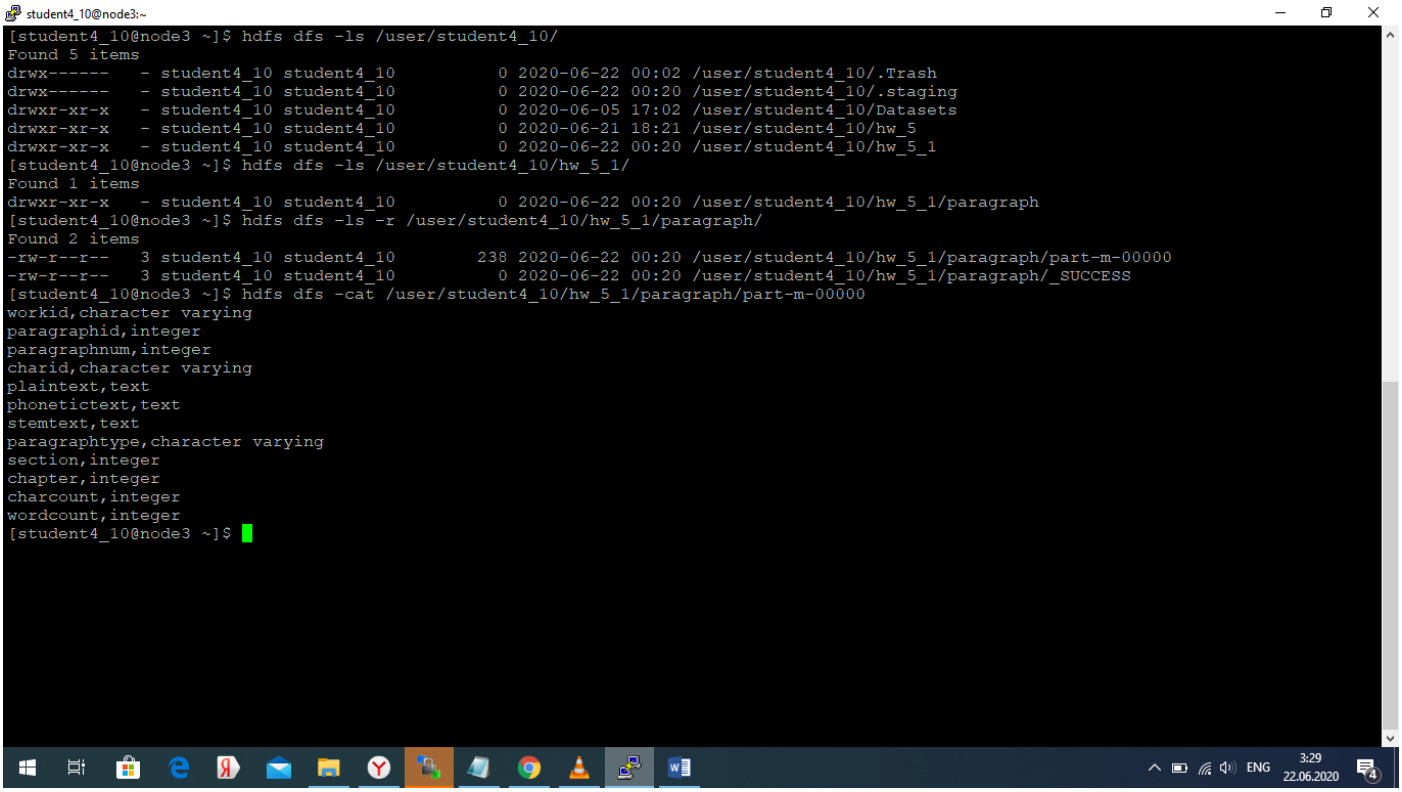




Добавил snapshot полного лога

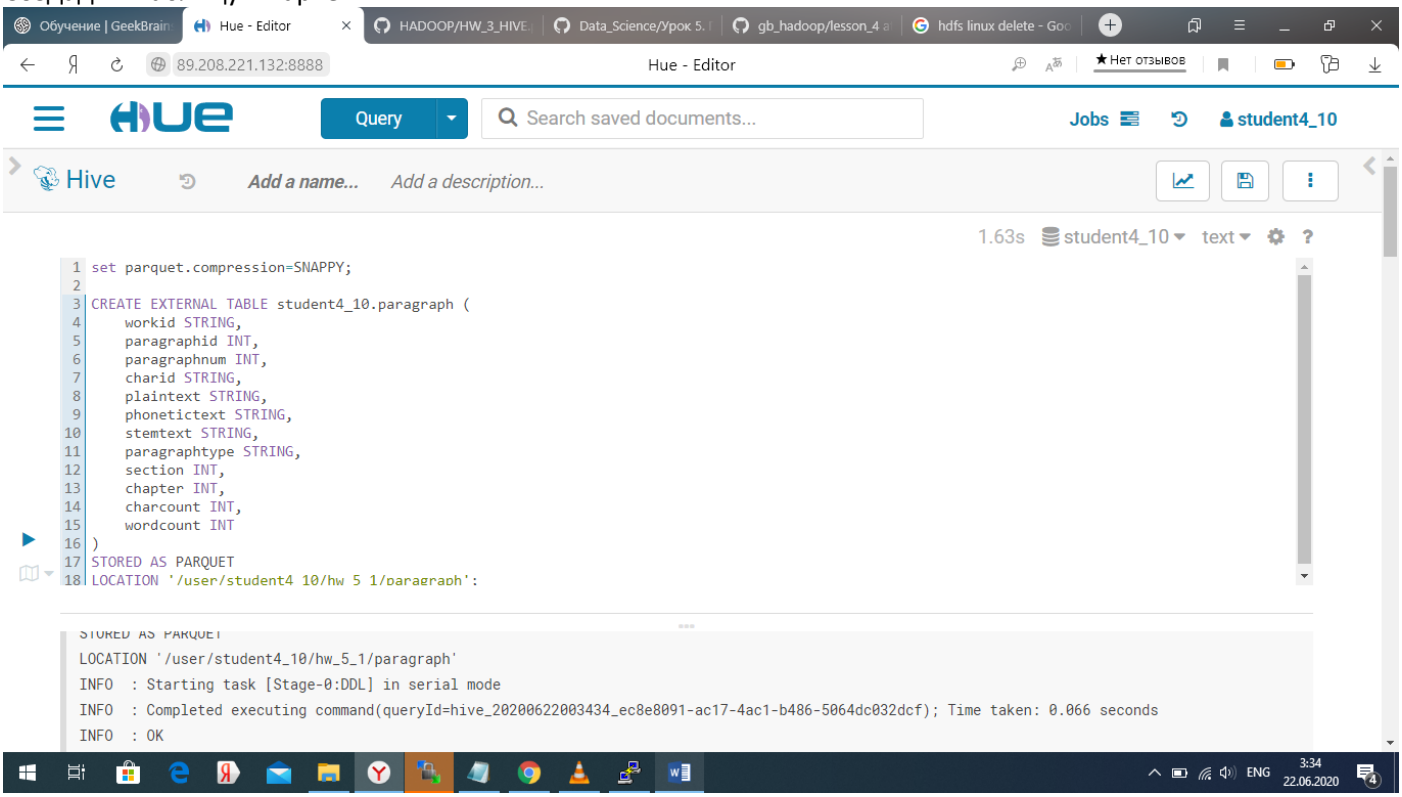
Проверим что импортировалось с базы PG_DATABASE (таблица параграф)





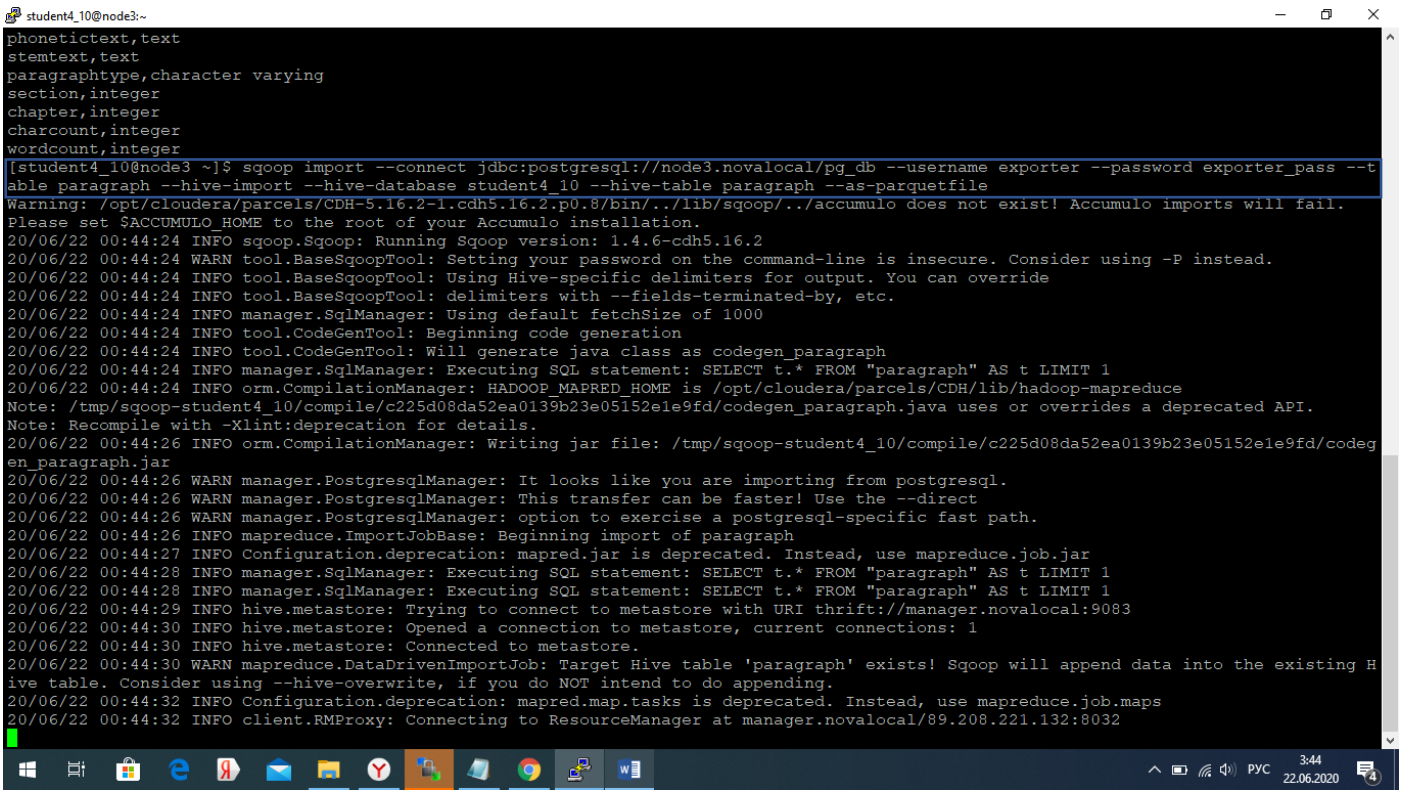
В схеме указаны колонки с типами данных

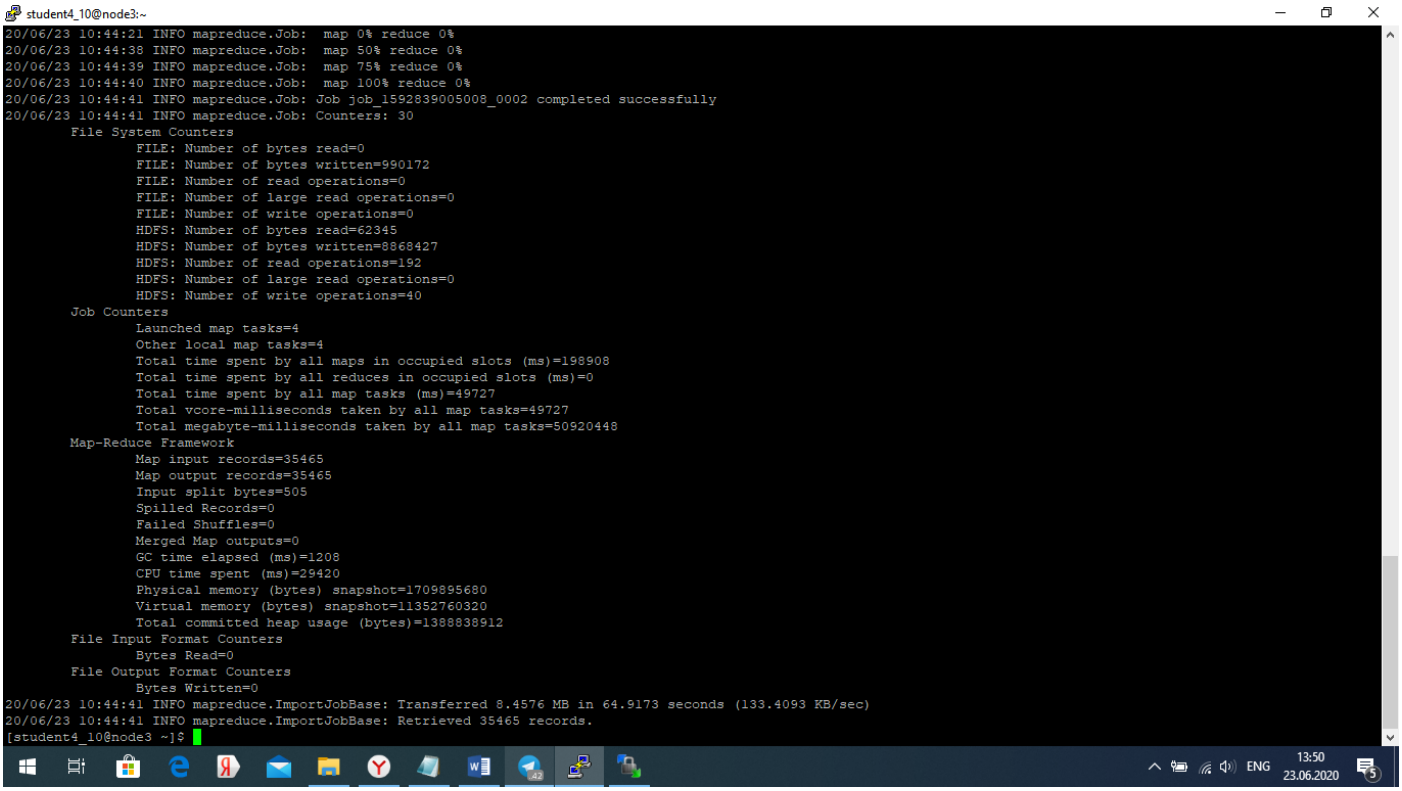
Создадим Таблицу в паркет



Импортируем данные в таблицу

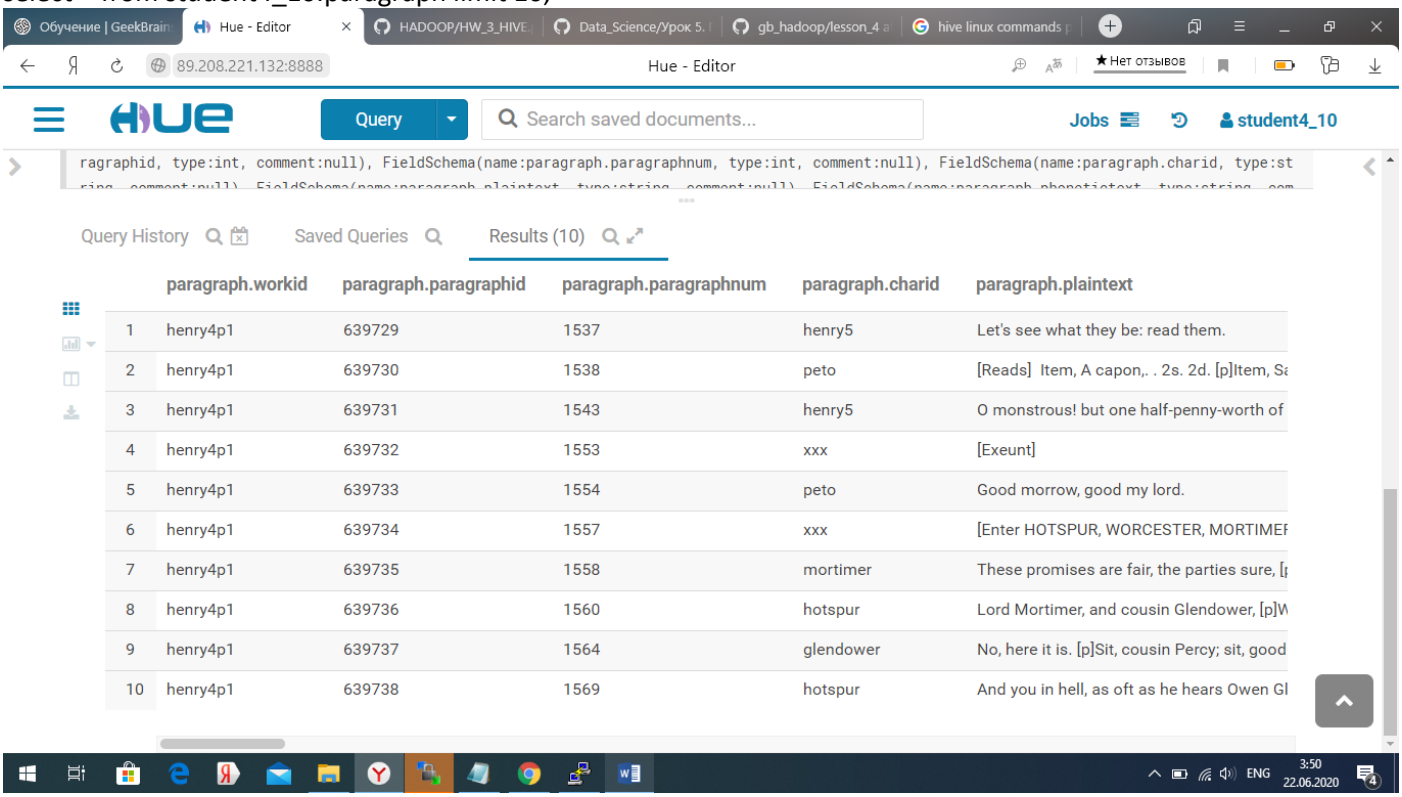
sqoop import --connect jdbc:postgresql://node3.novalocal/pg_db --username exporter --password exporter_pass --table paragraph --hive-import --hive-database student4_10 --hive-table paragraph_1 --as-parquetfile





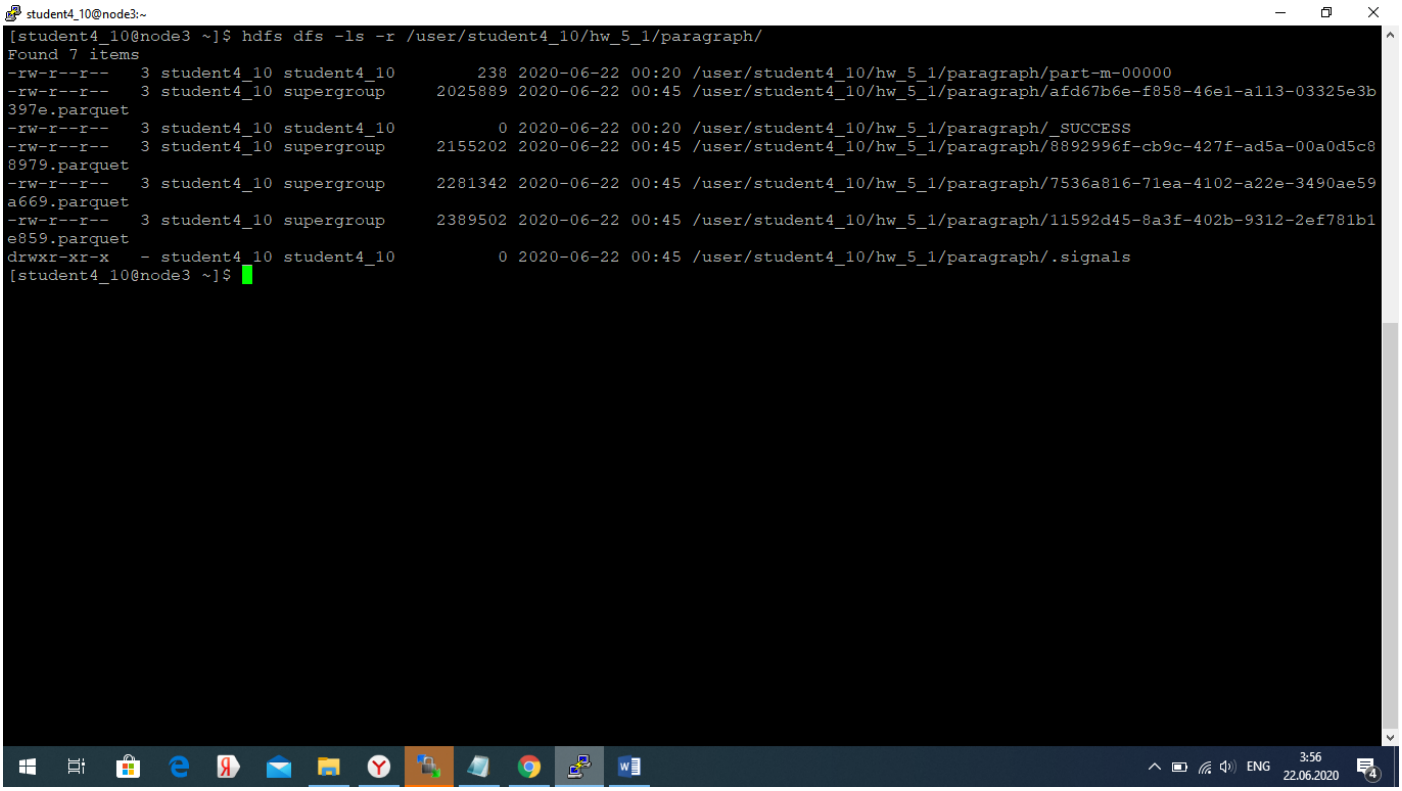
Скриншоты лога

Делаем проверку
select * from student4_10.paragraph limit 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, Sa
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, MORTIMEF
7	henry4p1	639735	1558	mortimer	These promises are fair, the parties sure, [p
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

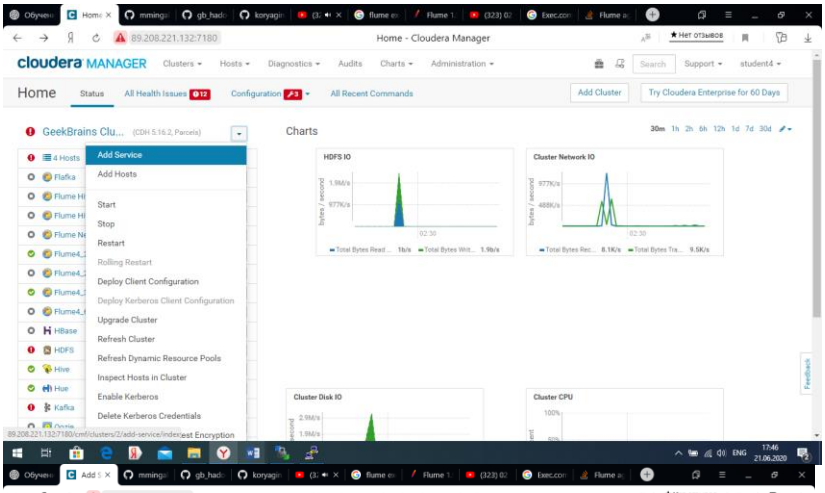


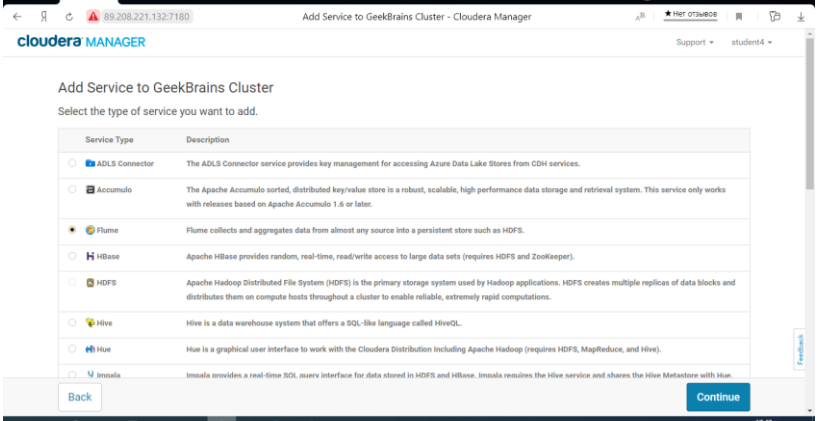
Точно таким способом можно выполнить работу и с Avro файлом

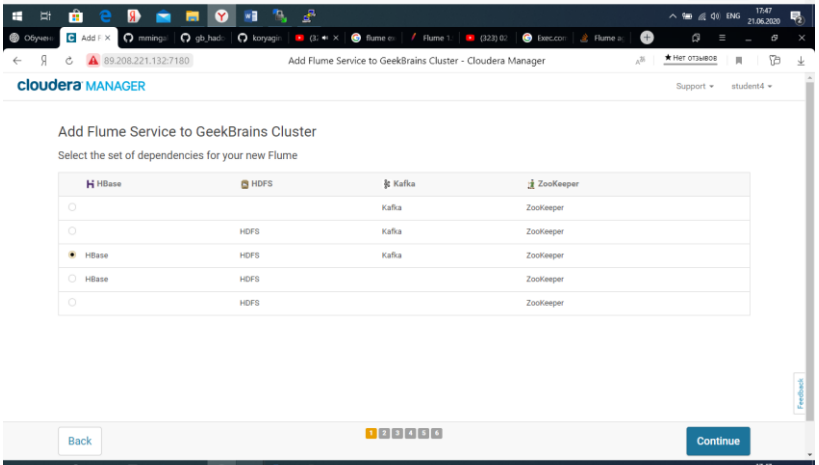
1.**Создать Flume-агент с именем, соответствующим имени своего пользователя (например Flume4_20)**
2. **Создать любой Flume поток используя Flume сервис соответствующего номера.**
- **Тип источника источник – exec**
- **Тип канала – memory**
- **Тип слива – hdfs**

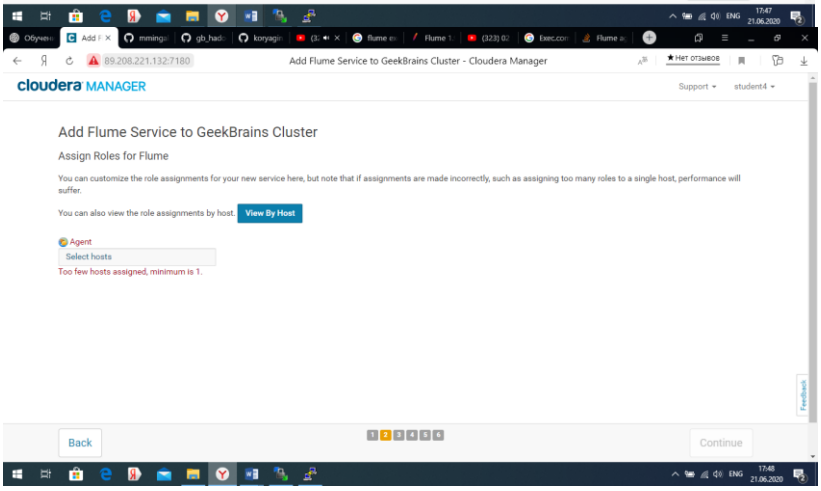
3. **Убедиться что данные поступают в слив.**
4. **Создать поверх данных в hdfs таблицу через которую можно просмотреть полученные данные.**
5. **[Продвинутый вариант] Сделать то-же самое используя несколько сливов в разные места, например в HDFS и в Hive одновременно**
6. **[Продвинутый** вариант] Повторить стандартный пример с выборкой сообщений из Twitter.

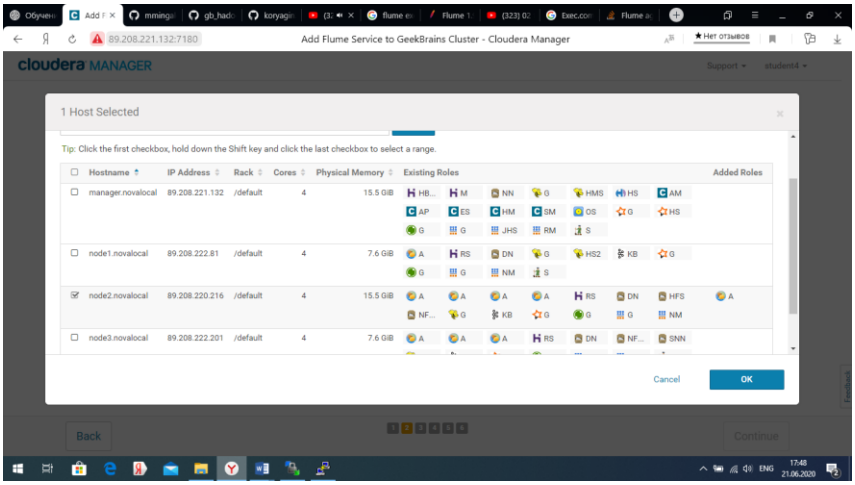
1.**Создать Flume-агент с именем, соответствующим имени своего пользователя (например Flume4_20)**

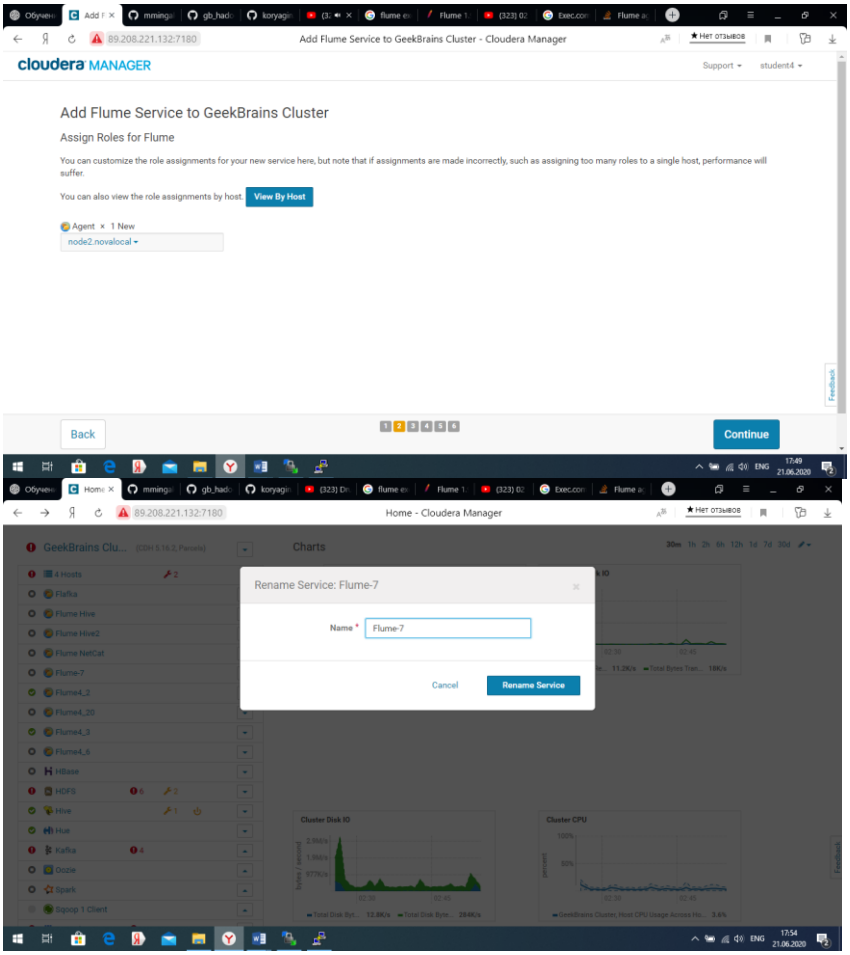

Home - Cloudera Manager
cloudera MANAGER
Home
Status
All Health Issues
Configuration
All Recent Commands
Add Cluster
Try Cloudera Enterprise for 60 Days
GeekBrains Clu...
Charts
Add Service
Add Hosts
Start
Stop
Restart
Rolling Restart
Deploy Client Configuration
Deploy Kerberos Client Configuration
Upgrade Cluster
Refresh Cluster
Refresh Dynamic Resource Pools
Inspect Hosts in Cluster
Enable Kerberos
Delete Kerberos Credentials
HDFS IO
Cluster Network IO
Cluster Disk IO
Cluster CPU

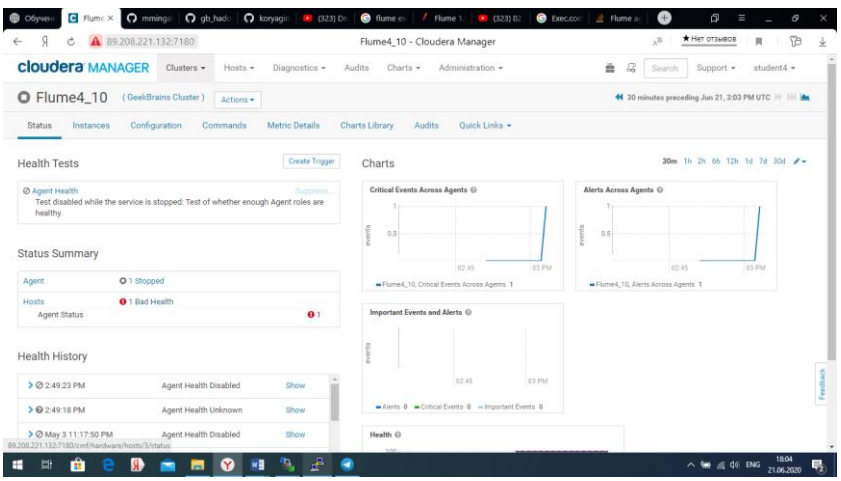

Add Service to GeekBrains Cluster - Cloudera Manager
cloudera MANAGER
Add Service to GeekBrains Cluster
Select the type of service you want to add.
Service Type
Description
ADLS Connector
The ADLS Connector service provides key management for accessing Azure Data Lake Stores from CDH services.
Accumulo
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.
Flume
Flume collects and aggregates data from almost any source into a persistent store such as HDFS.
HBase
Apache HBase provides random, real-time, read/write access to large data sets (requires HDFS and ZooKeeper).
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.
Hive
Hive is a data warehouse system that offers a SQL-like language called HiveQL.
Hue
Hue is a graphical user interface to work with the Cloudera Distribution Including Apache Hadoop (requires HDFS, MapReduce, and Hive).
Impala
Back
Continue


Add Flume Service to GeekBrains Cluster - Cloudera Manager
cloudera MANAGER
Add Flume Service to GeekBrains Cluster
Select the set of dependencies for your new Flume
HBase
HDFS
Kafka
ZooKeeper
Back
Continue

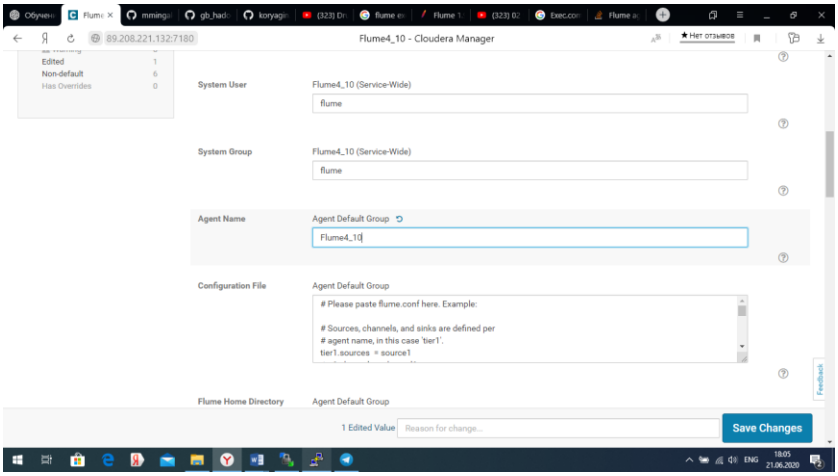

Add Flume Service to GeekBrains Cluster - Cloudera Manager
cloudera MANAGER
Add Flume Service to GeekBrains Cluster
Assign Roles for Flume
You can customize the role assignments for your new service here, but note that if assignments are made incorrectly, such as assigning too many roles to a single 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
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 GiB
node1.novalocal 89.208.222.81 /default 4 7.6 GiB
node2.novalocal 89.208.220.216 /default 4 15.5 GiB
node3.novalocal 89.208.222.201 /default 4 7.6 GiB
Cancel
OK
Back
Continue


Add Flume Service to GeekBrains Cluster - Cloudera Manager
cloudera MANAGER
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 × 1 New
node2.novalocal
Back
Continue
Home - Cloudera Manager
Charts
Rename Service: Flume-7
Name *
Flume-7
Cancel
Rename Service


Flume4_10 - Cloudera Manager
cloudera MANAGER
Flume4_10 (GeekBrains Cluster)
Actions
Status
Instances
Configuration
Commands
Metric Details
Charts Library
Audits
Quick Links
Health Tests
Create Trigger
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
Agent Status 1
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
Charts
Critical Events Across Agents
Alerts Across Agents
Important Events and Alerts
Health

BAD HEALTH!

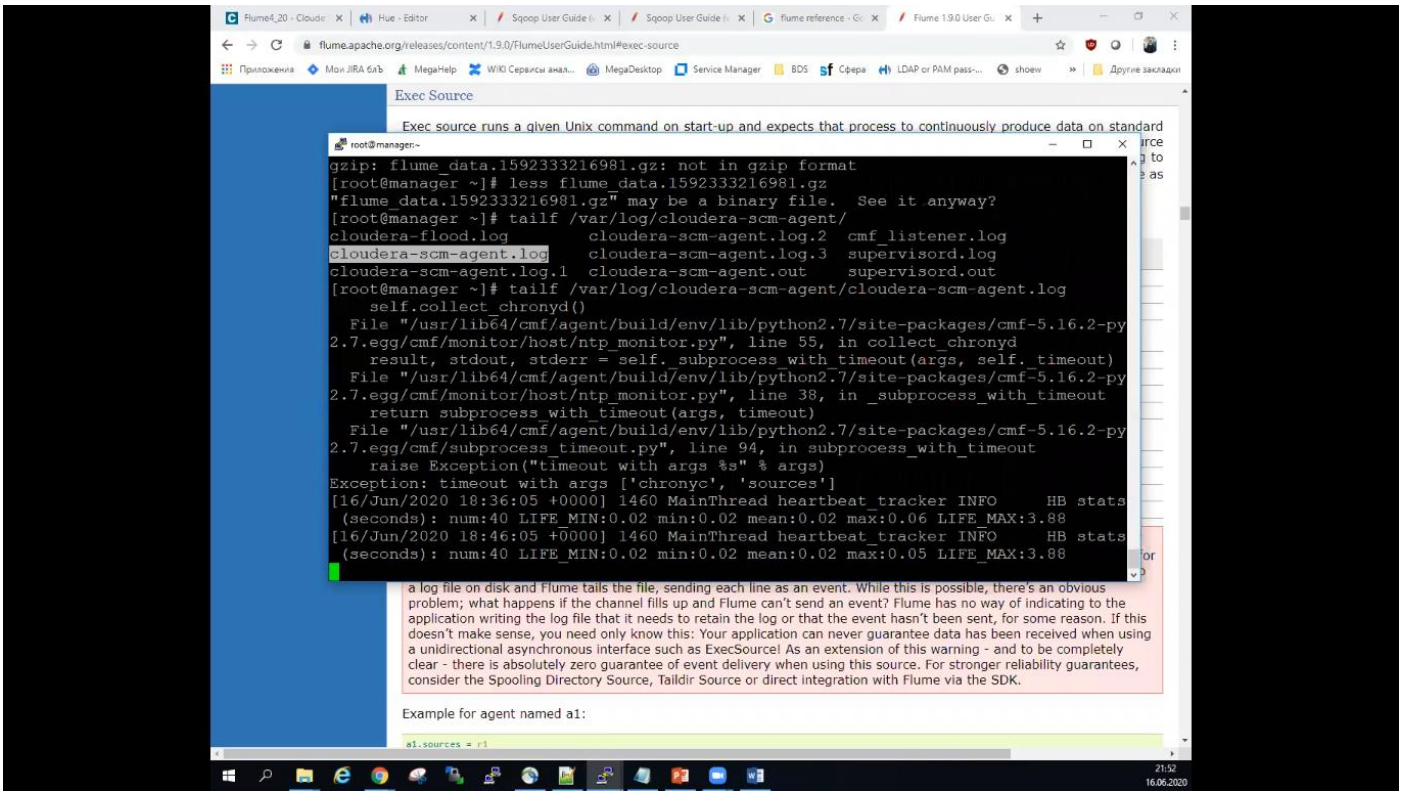


2. Создать любой Flume поток используя Flume сервис соответствующего номера.

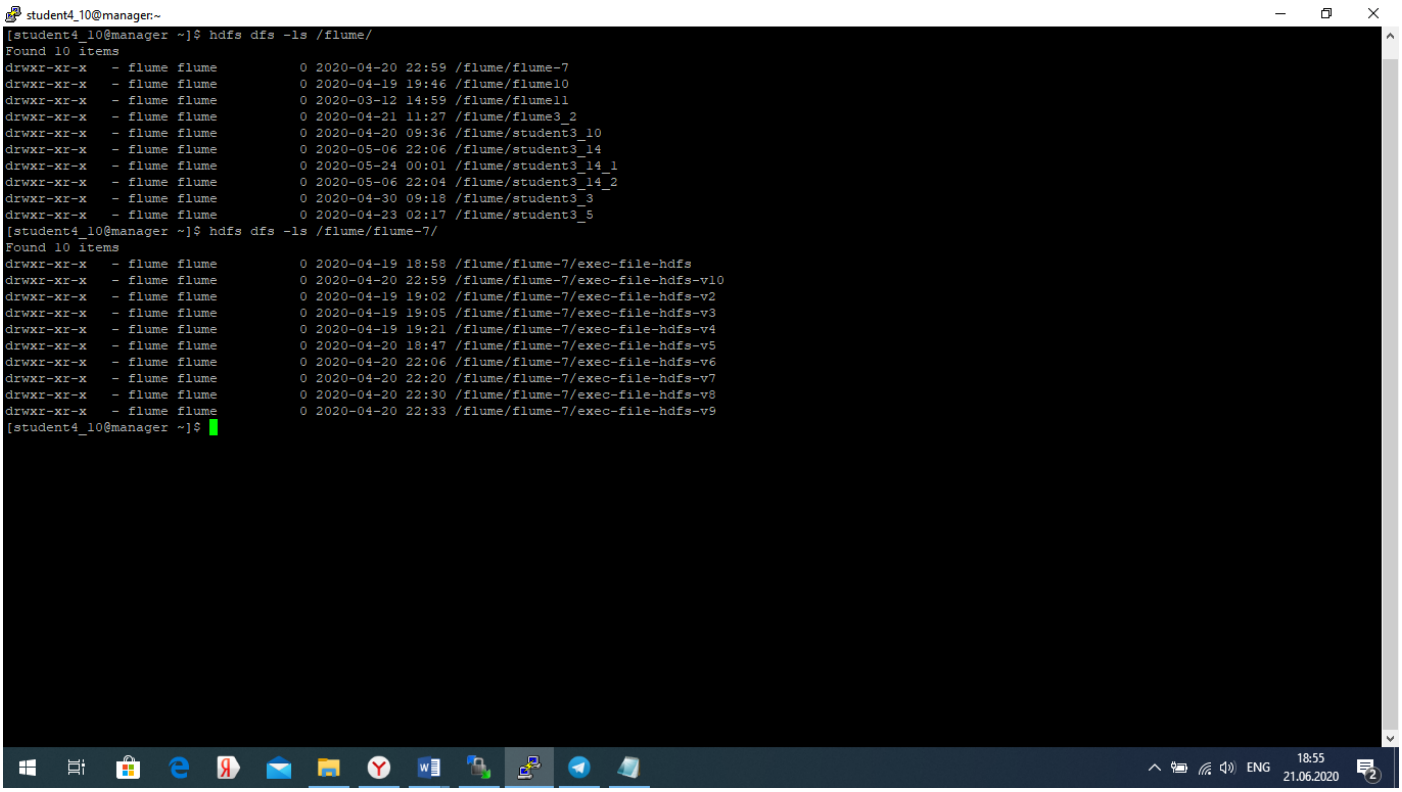
- **Тип источника источник – exec**
- **Тип канала – memory**
- **Тип слива – hdfs**

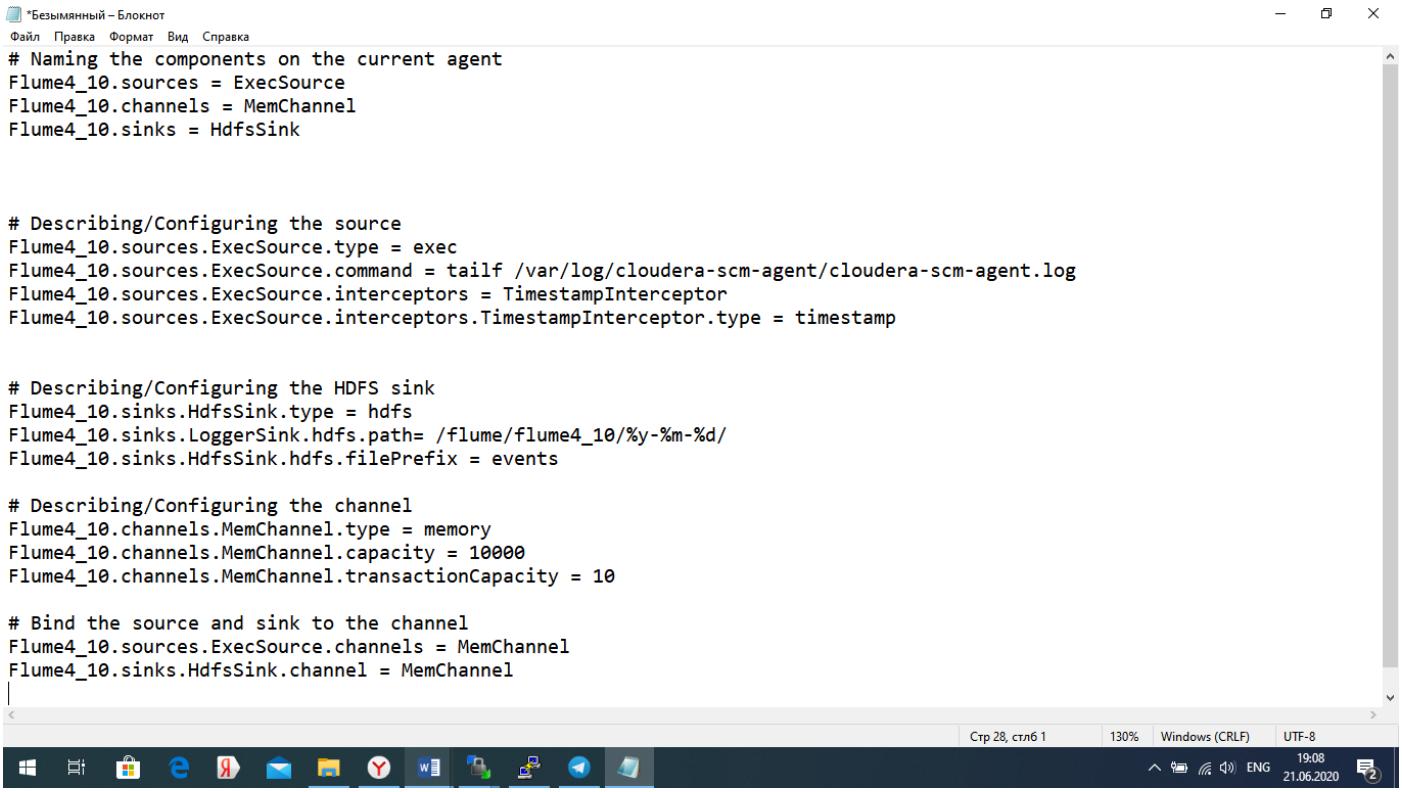
tailif /var/log/cloudera-scm-agent/cloudera-scm-agent.log

работает с этим лог файлом

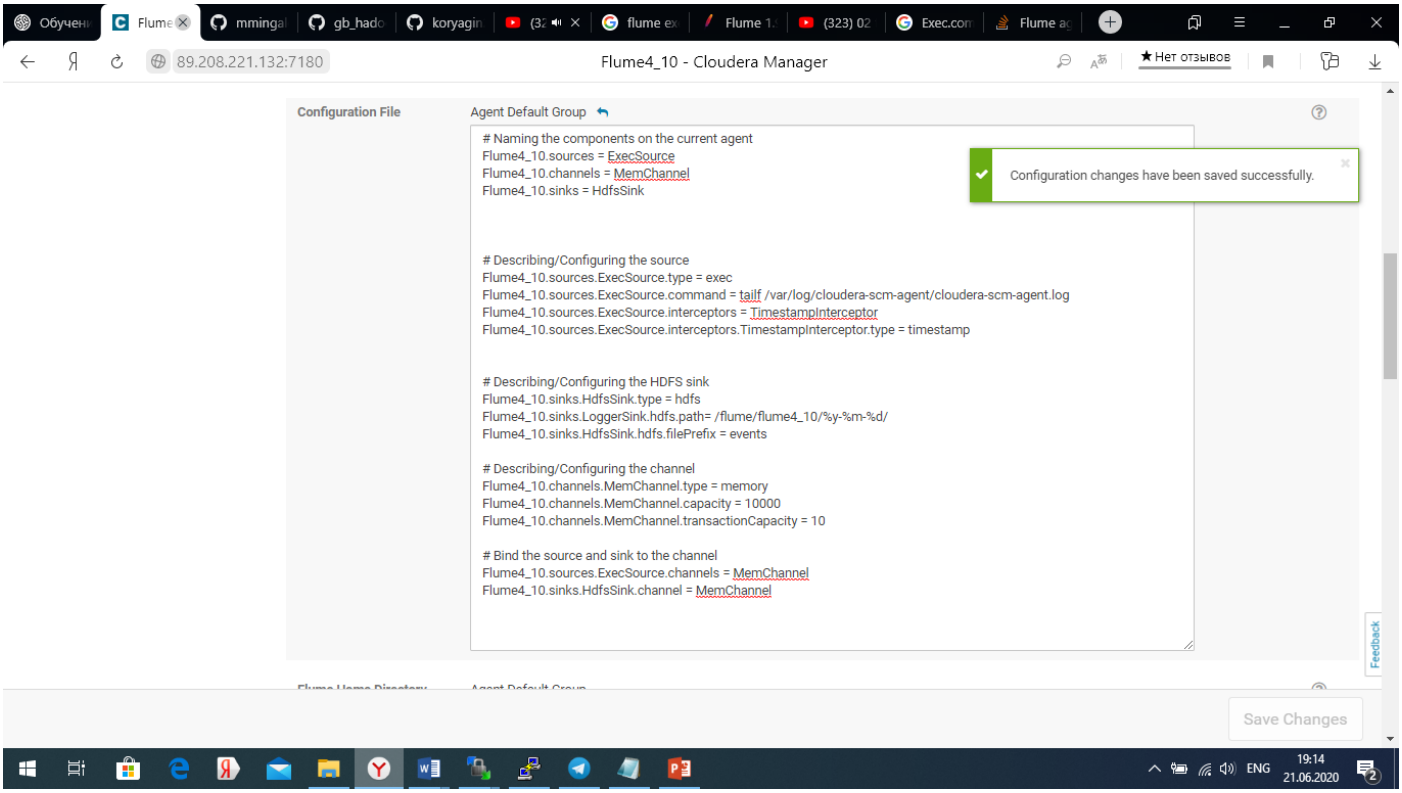


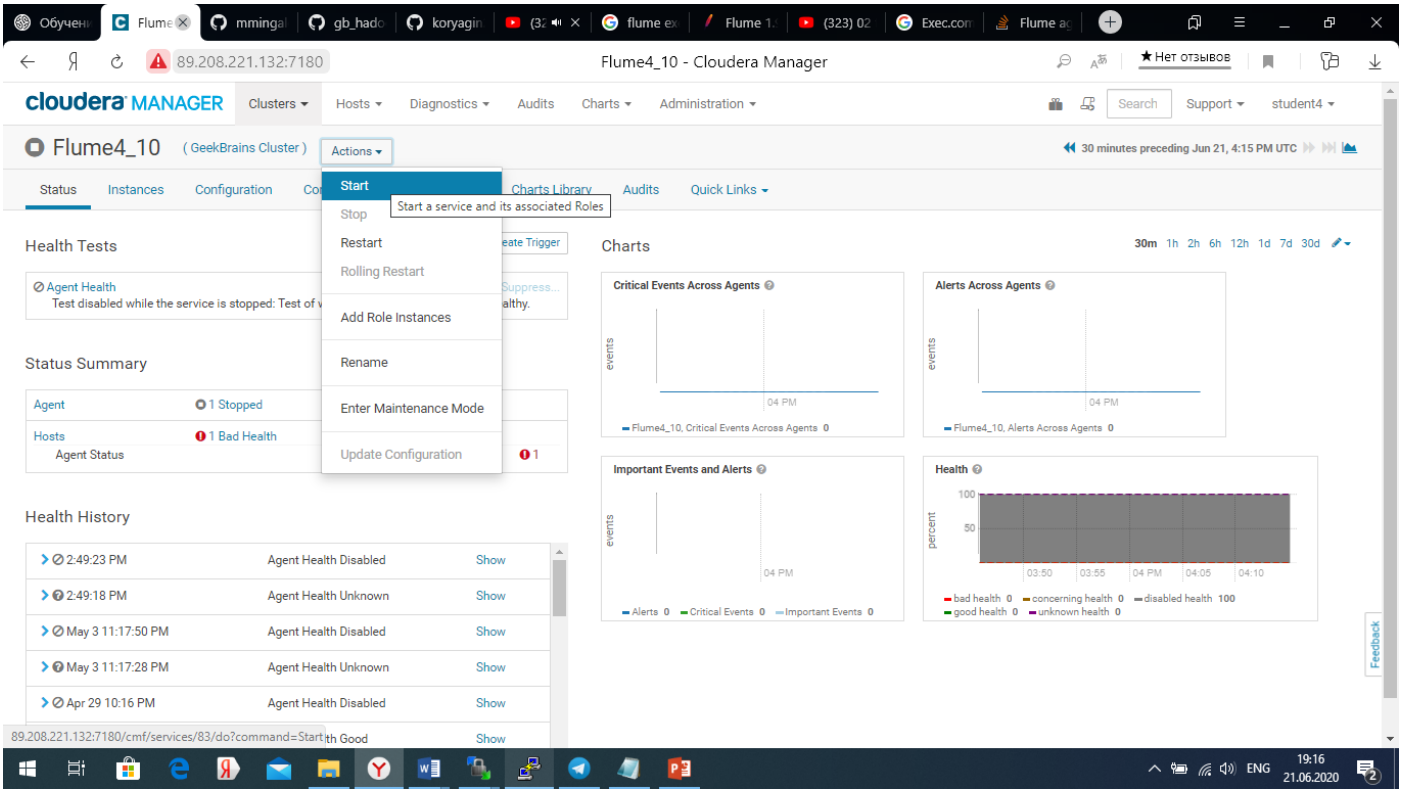
Писать буду в Flume папку !



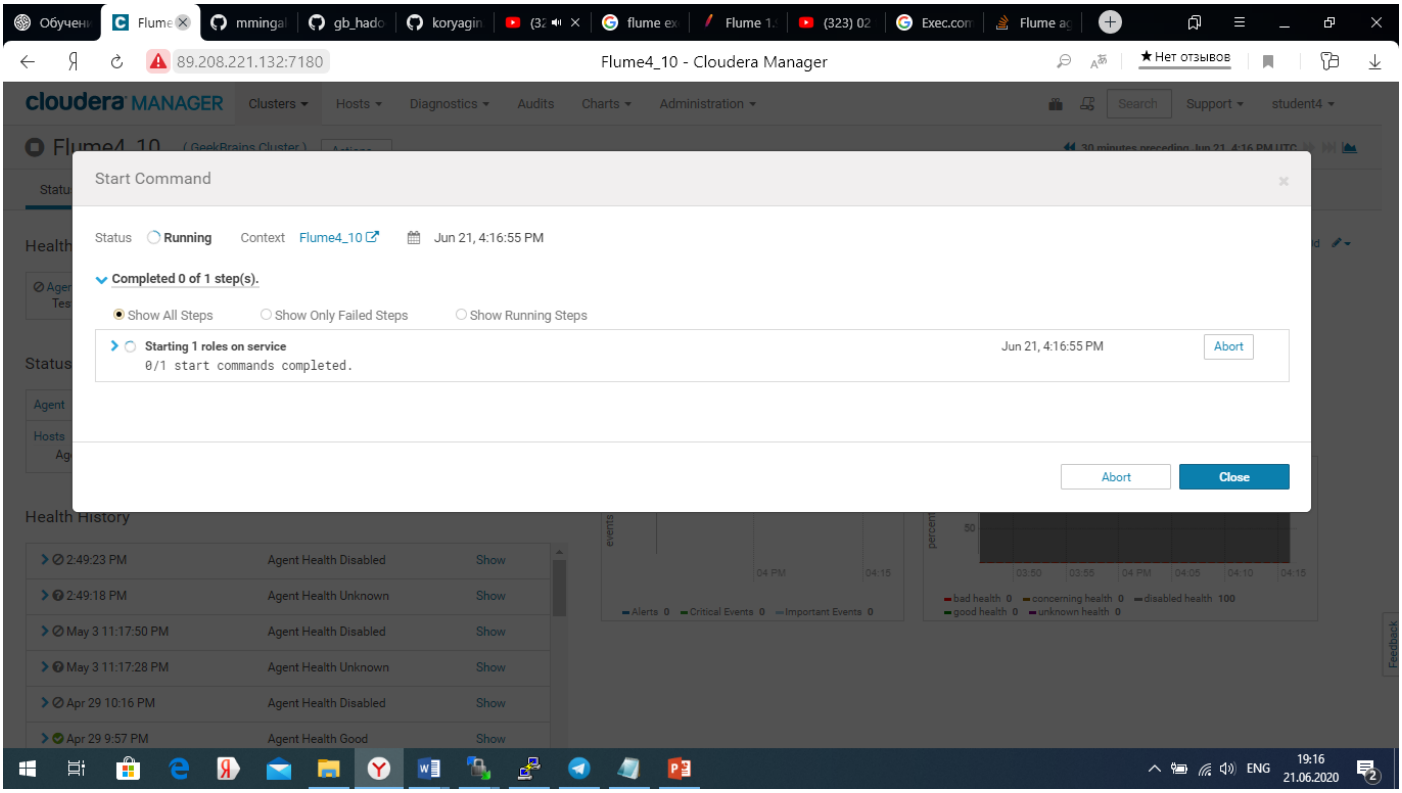


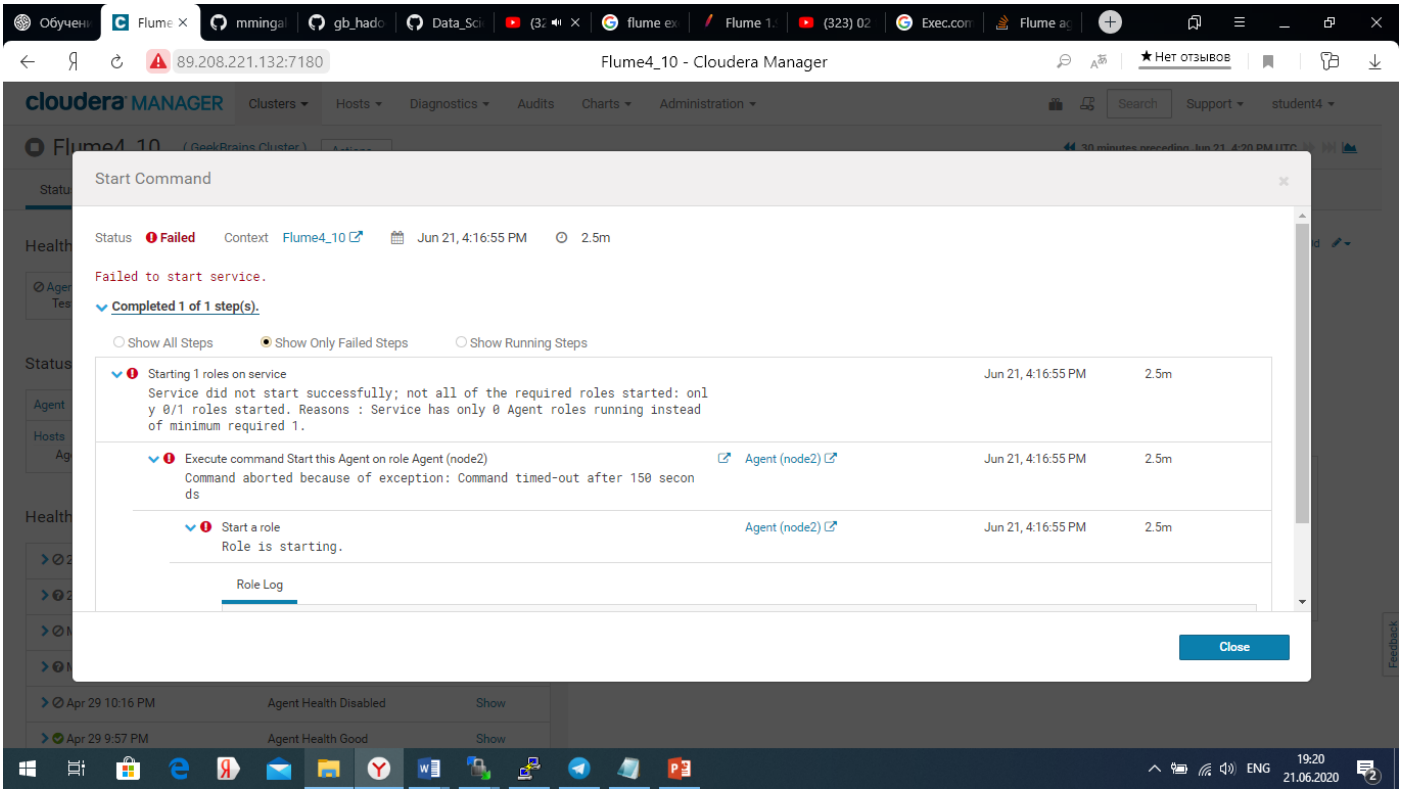
Код для configuration file


Flume4_10 - Cloudera Manager
89.208.221.132:7180
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
ENG
19:14
21.06.2020

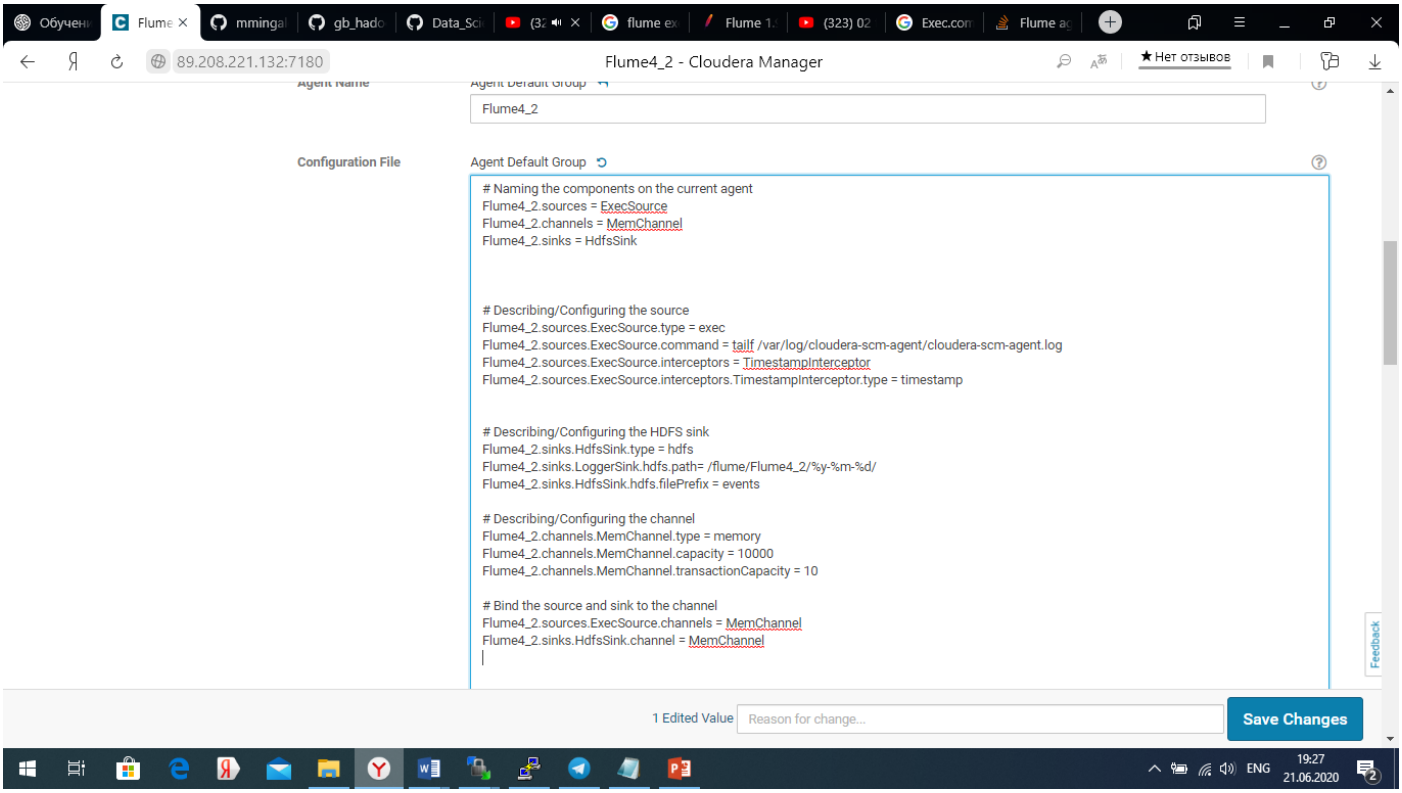

Flume4_10 - Cloudera Manager
89.208.221.132:7180
cloudera MANAGER
Clusters
Hosts
Diagnostics
Audits
Charts
Administration
Search
Support
student4
Flume4_10
(GeekBrains Cluster)
Actions
30 minutes preceding Jun 21, 4:15 PM UTC
Status
Instances
Configuration
Charts Library
Audits
Quick Links
Start
Stop
Restart
Rolling Restart
Add Role Instances
Rename
Enter Maintenance Mode
Update Configuration
Start a service and its associated Roles
Health Tests
Agent Health
Test disabled while the service is stopped: Test of
Status Summary
Agent
1 Stopped
Hosts
1 Bad Health
Agent Status
Health History
2:49:23 PM
Agent Health Disabled
Show
2:49:18 PM
Agent Health Unknown
May 3 11:17:50 PM
May 3 11:17:28 PM
Apr 29 10:16 PM
Charts
30m 1h 2h 6h 12h 1d 7d 30d
Critical Events Across Agents
Alerts Across Agents
Important Events and Alerts
Health
events
percent
04 PM
03:50
03:55
04:05
04:10
Flume4_10, Critical Events Across Agents 0
Flume4_10, Alerts Across Agents 0
Alerts 0
Critical Events 0
Important Events 0
bad health 0
concerning health 0
disabled health 100
good health 0
unknown health 0
Feedback
ENG
19:16
21.06.2020

BAD HEALTH !!!!

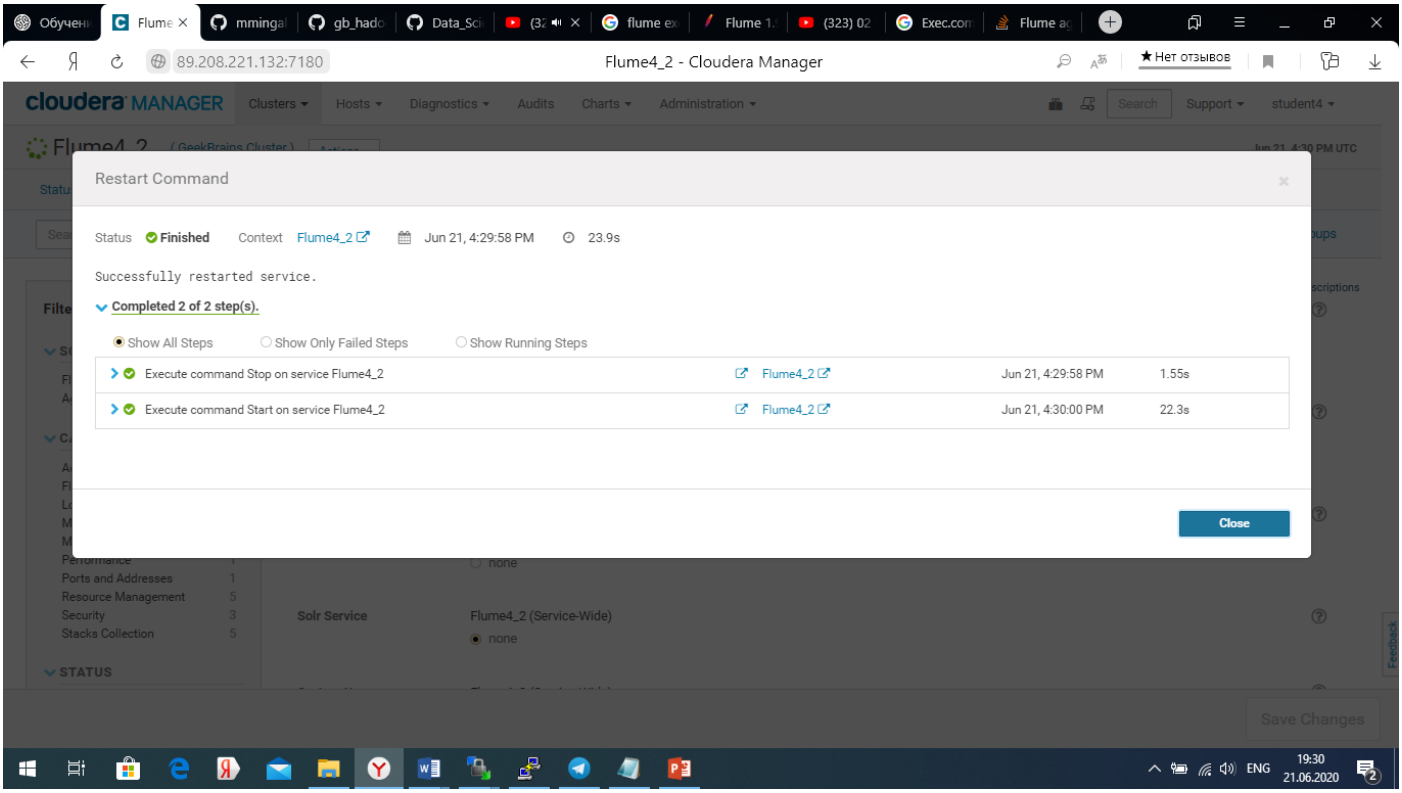


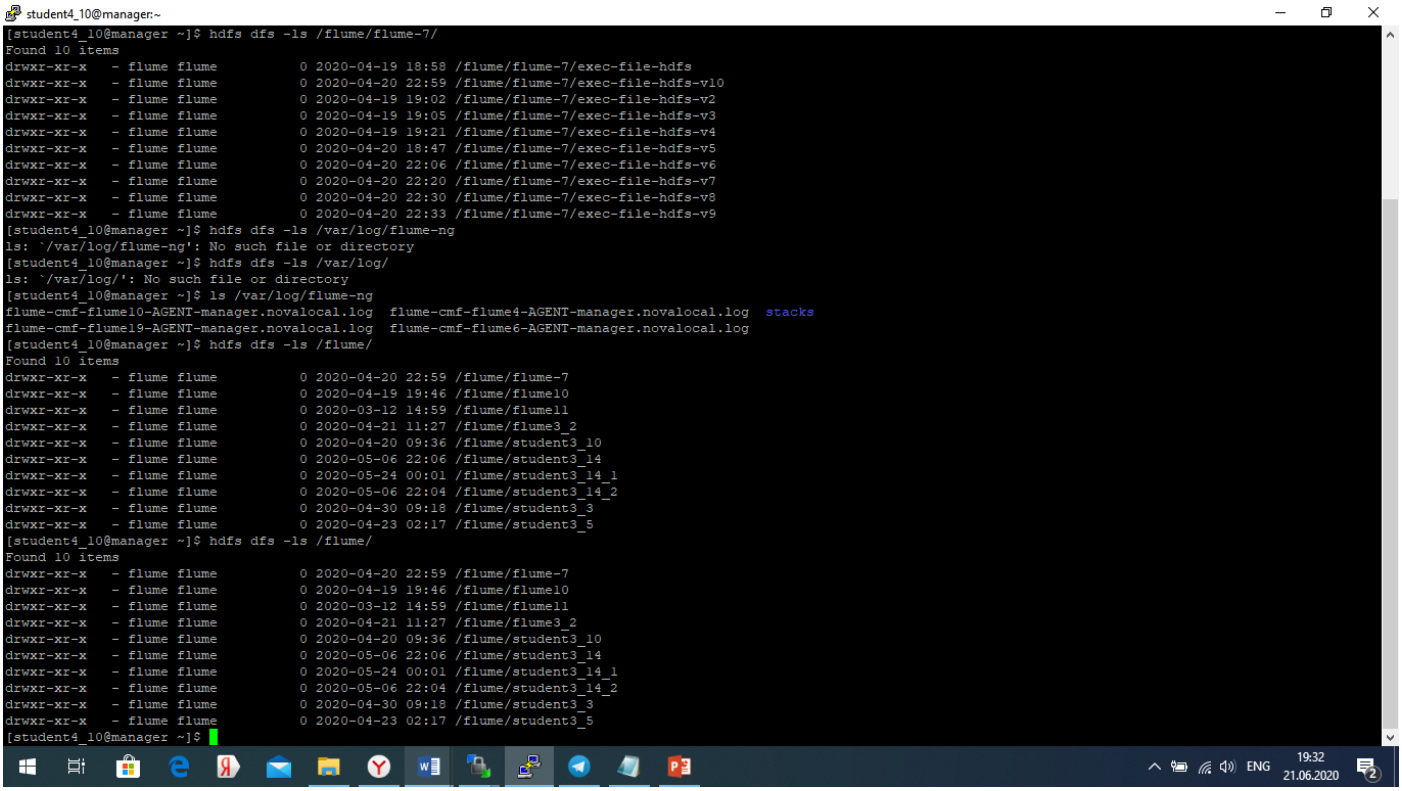


Я принял решение запустить с Flume 4_2 так как у него GOOD HEALTH



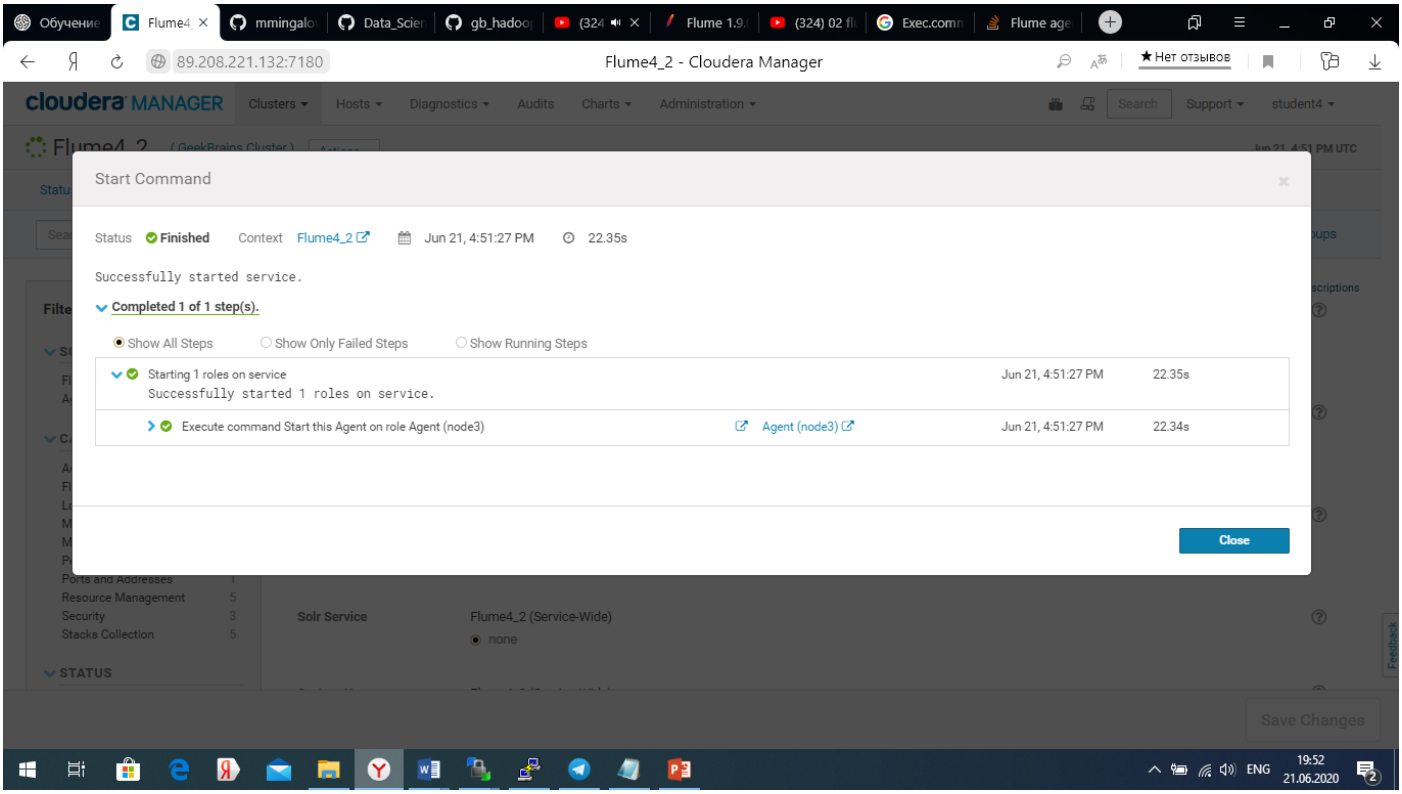
Save changes and restart

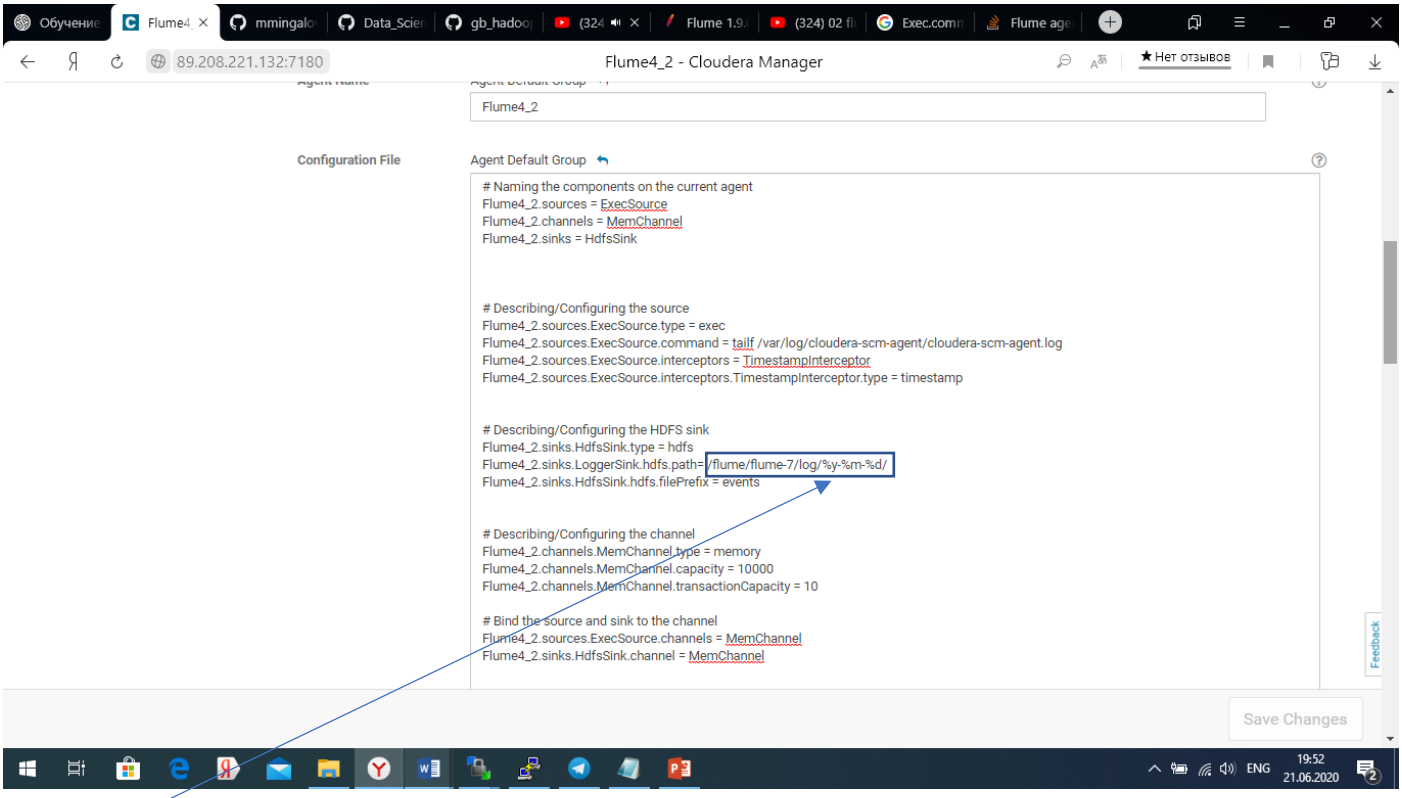




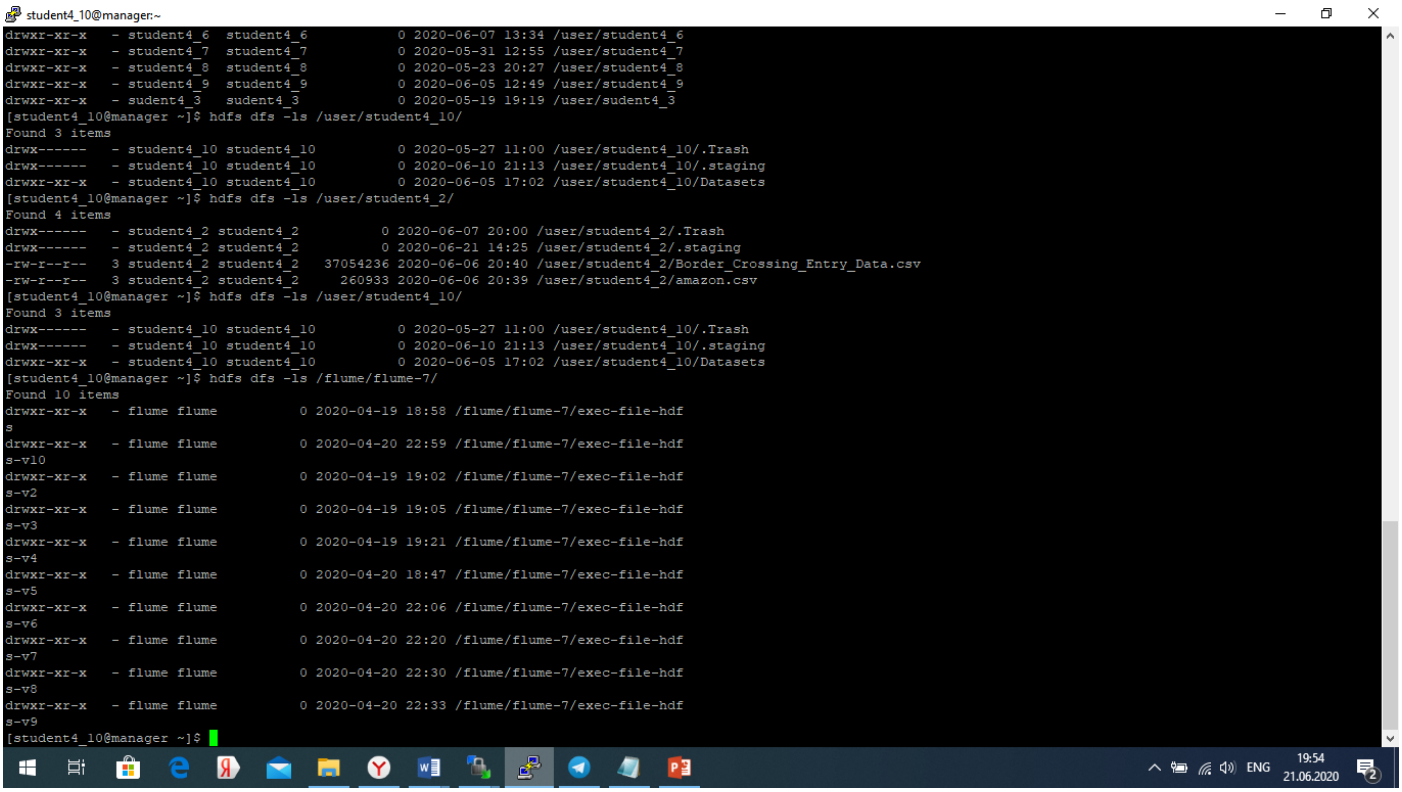
нету ничего

Попробую flume-7 в configuration file

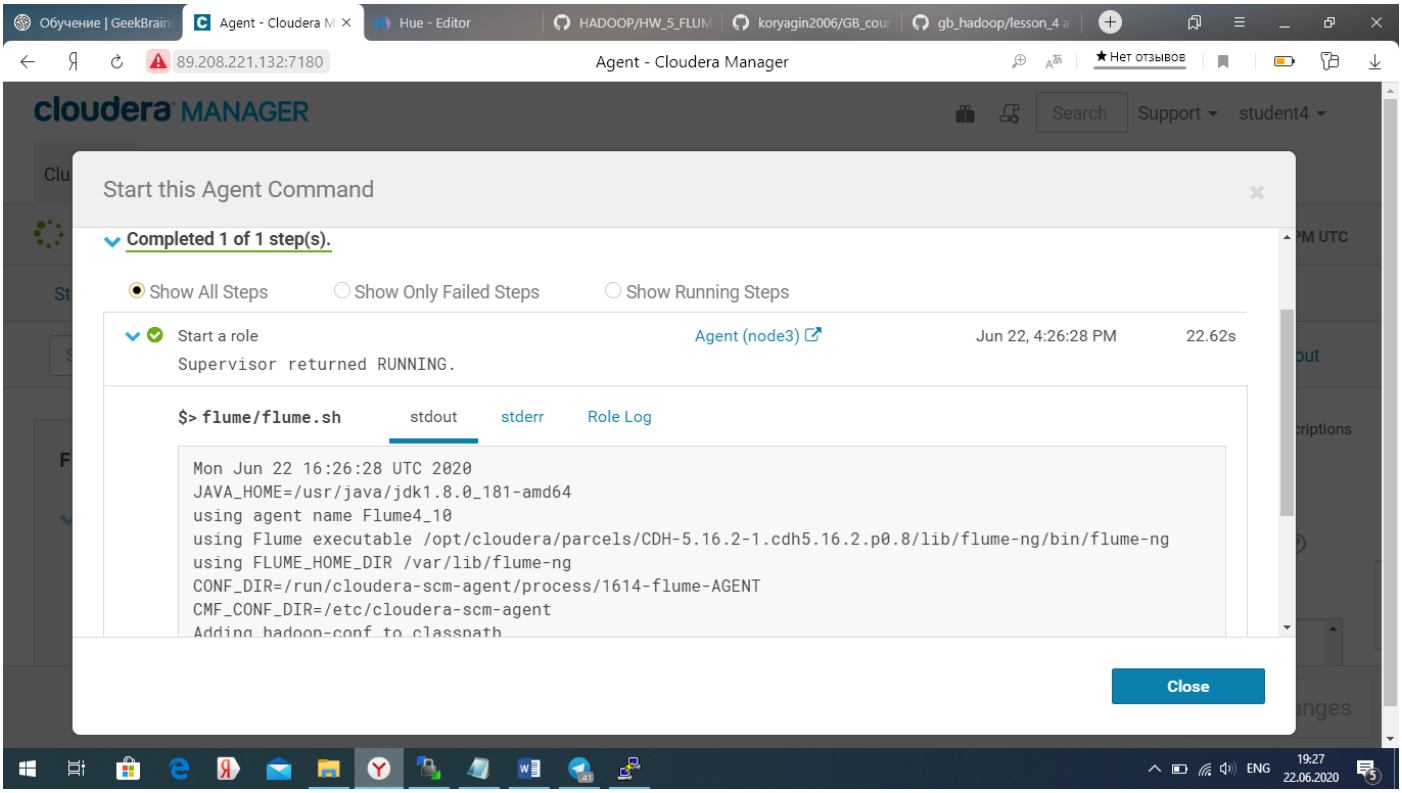


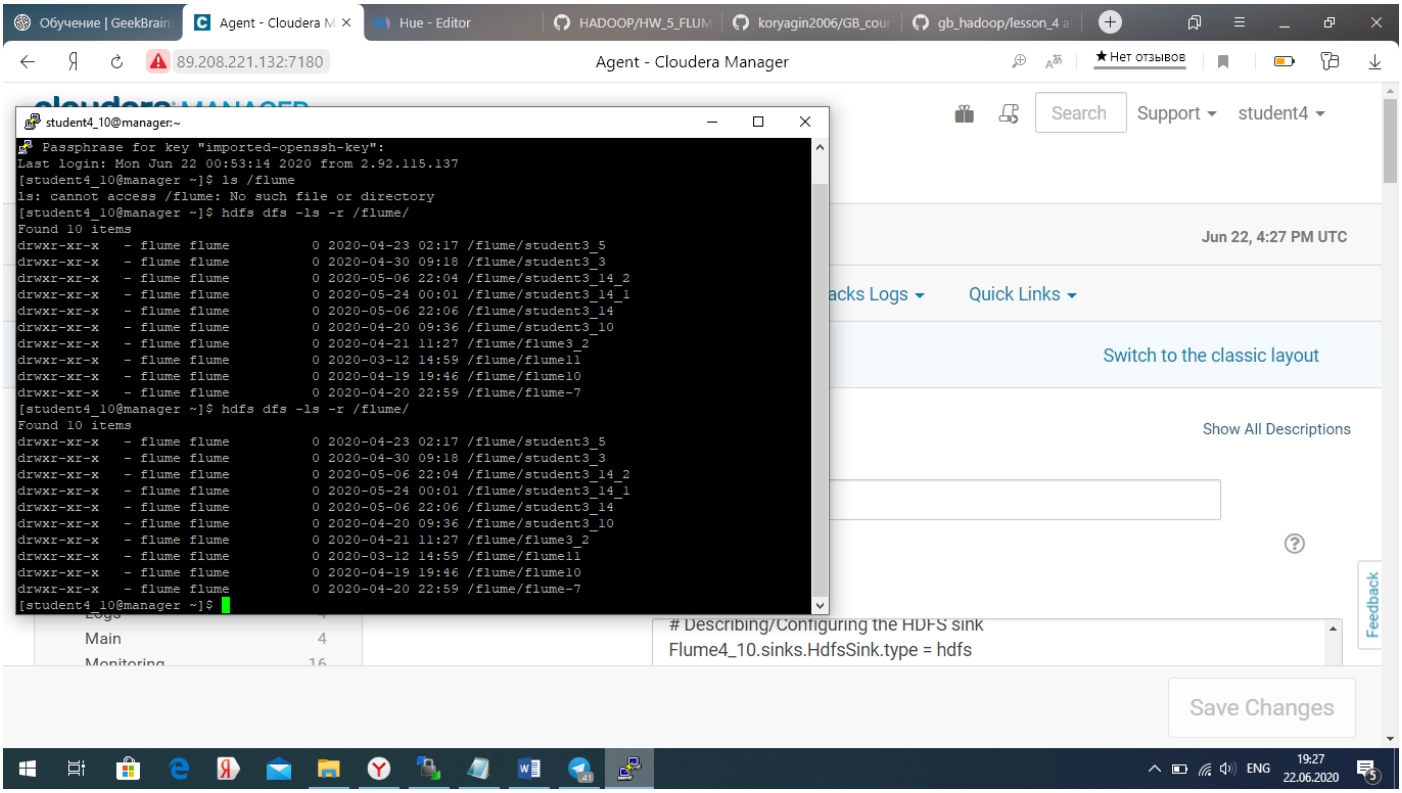


Flume-7



Не работает



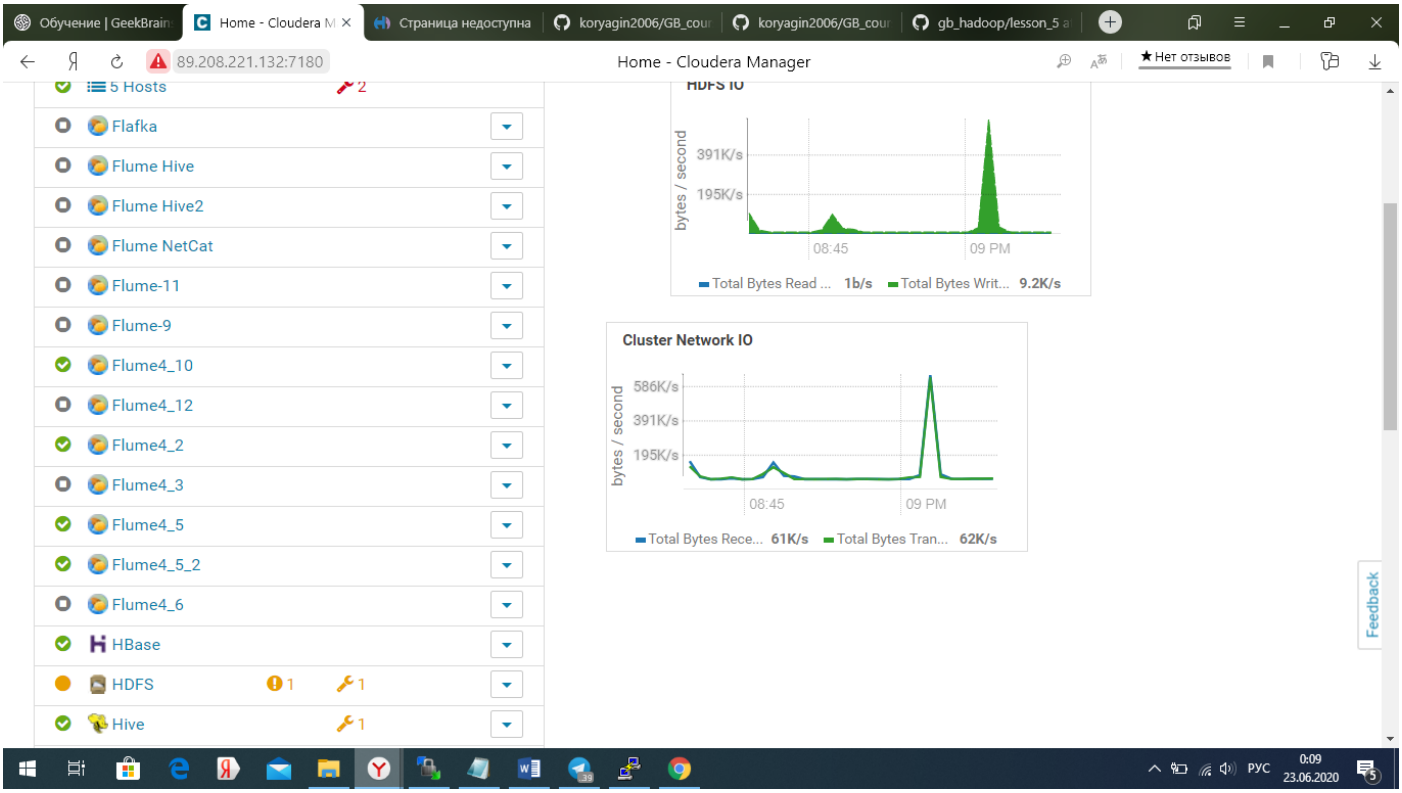


Запустил Flume 4_10 с node3.novalocal но в FLUME папке ничего не появилось

Вопрос

- Я написал правильный код

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


Обучение | GeekBrain
Home - Cloudera M
Страница недоступна
koryagin2006/GB_cour
koryagin2006/GB_cour
gb_hadoop/lesson_5
89.208.221.132:7180
Home - Cloudera Manager
Нет отзывов
5 Hosts
2
Flafka
Flume Hive
Flume Hive2
Flume NetCat
Flume-11
Flume-9
Flume4_10
Flume4_12
Flume4_2
Flume4_3
Flume4_5
Flume4_5_2
Flume4_6
HBase
HDFS
1
1
Hive
1
HDFS IO
bytes / second
391K/s
195K/s
08:45
09 PM
Total Bytes Read ... 1b/s
Total Bytes Writ... 9.2K/s
Cluster Network IO
bytes / second
586K/s
391K/s
195K/s
08:45
09 PM
Total Bytes Rece... 61K/s
Total Bytes Tran... 62K/s
Feedback
РУС
0:09
23.06.2020