Методы анализа данных

Практическая работа №3

Работа с Apache Flume

# Получение данных с помощью Apache Flume

В этой лабораторной работе вы запустите агент Flume чтобы собрать данные из разных источников и сохранить их в HDFS или локальной файловой системе.

## Простой перенос данных

Этот агент позволяет пользователю генерировать события и последовательно логгировать их в консоль. В следующей конфигурации определяется один агент под названием agent1.

### Создаем файл конфигурации

mkdir flume

cd flume

vi transfer.conf

### 

### Файл конфигурации агента Agent1

В agent1 есть источник, который слушает порт 3333 для получения данных, канал, который буферизирует данные событий в памяти и слив, который логгирует данные событий в консоль.

﻿agent1.sources = netcatSrc

agent1.channels = memChannel

agent1.sinks = log

# Связываем источник с каналом

agent1.sources.netcatSrc.channels = memChannel

# Связываем приемник с каналом

agent1.sinks.log.channel = memChannel

# Устанавливаем тип источника данных

agent1.sources.netcatSrc.type = netcat

# Устанавливаем IP-адрес, на котором источник будет слушать

agent1.sources.netcatSrc.bind = 0.0.0.0

agent1.sources.netcatSrc.port = 3333

agent1.sinks.log.type = logger

agent1.channels.memChannel.type = memory

agent1.channels.memChannel.capacity = 100

### Запустим Flume agent1

flume-ng agent -name agent1 --conf-file transfer.conf

### 

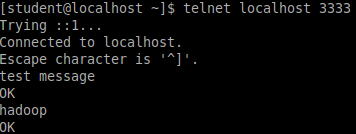
### Откроем ещё одно окно терминала и выполним команду telnet.

telnet localhost 3333

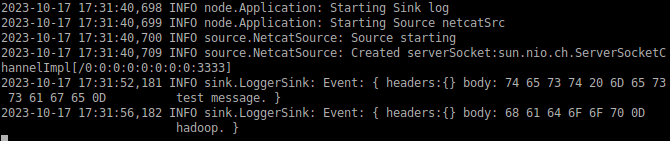
Typing whatever you want...

Hadoop

..



### Проверим, что сообщение, отправленное в telnet на шаге 4, вывелось в терминале, где был запущен агент flume на шаге 3.



Примечание: если telnet или порт закрыты неправильно, то, когда вы попытаетесь соединиться снова, может возникнуть ошибка о том, что порт уже открыт.

### telnet закрывается после нажатия клавиш ctrl + C или командой close после нажатия Ctrl+].

^] (ctrl+])

telnet> close

## Базовый перенос данных через буферную директорию

Agent2 будет использоваться для сохранения файлов, попадающих в буферную директорию, в локальную директорию.

### Создадим файл конфигурации

vi transfer\_spool.conf

### Файл конфигурации Agent2

﻿agent2.sources = dirSrc

agent2.channels = memChannel

agent2.sinks = fileSink

# Связываем источник с каналом

agent2.sources.dirSrc.channels = memChannel

# Связываем приемник с каналом

agent2.sinks.fileSink.channel = memChannel

# Устанавливаем тип источника данных как spoolDir

agent2.sources.dirSrc.type = spoolDir

# Устанавливаем директорию, в которой источник будет отслеживать файлы

agent2.sources.dirSrc.spoolDir = /home/student/flume/incoming

# Устанавливаем тип приемника данных

agent2.sinks.fileSink.type = file\_roll

agent2.sinks.fileSink.sink.directory = /home/student/flume/output

# Устанавливаем интервал для архивирования файлов

agent2.sinks.fileSink.sink.rollInterval = 0

agent2.channels.memChannel.type = memory

agent2.channels.memChannel.capacity = 100

### Создайте все необходимые папки (spool и output) самостоятельно.

### Запустим Agent2

flume-ng agent -name agent2 --conf-file transfer\_spool.conf

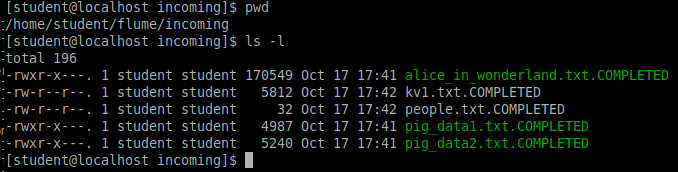
### Откроем еще одно окно терминала и скопируем несколько файлов в буферную директорию.

cd /home/student/flume/incoming

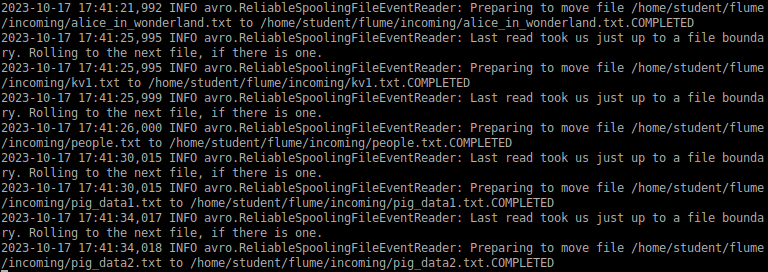
cp ~/Data/\*.txt .

vi hello.txt

This is test file for Flume.



### Вы должны увидеть сообщение о том, что pig\_data1.txt, pig\_data2.txt, alice\_in\_wonderland.txt и hello.txt скопированы в буферную директорию. Сообщение появится в терминале, где запущен Agent2.



Также вы можете проверить, что файл hello.txt был перенесен, открыв его утилитой vi. Перенесенный файл можно найти в директории output.

### Перенесенные файлы сохранены в виде файлов в директории output.

### 

## Использование Перехватчика (Interceptor)

Agent3 будет использоваться для вставки IP адреса хоста, где запущен агент, в заголовки события.

### Создаем файл конфигурации

vi interceptor.conf

### Конфигурационный файл Agent3

agent3.sources = netcatSrc

agent3.channels = memChannel

agent3.sinks = log

agent3.sources.netcatSrc.channels = memChannel

agent3.sinks.log.channel = memChannel

agent3.sources.netcatSrc.type = netcat

agent3.sources.netcatSrc.bind = 0.0.0.0

agent3.sources.netcatSrc.port = 3333

agent3.sinks.log.type = logger

agent3.channels.memChannel.type = memory

agent3.channels.memChannel.capacity = 100

#Устанавливаем перехватчик для источника

agent3.sources.netcatSrc.interceptors = i1

agent3.sources.netcatSrc.interceptors.i1.type = host

agent3.sources.netcatSrc.interceptors.i1.hostHeader = hostname

### 

### Запуск Agent3

flume-ng agent -name agent3 --conf-file interceptor.conf

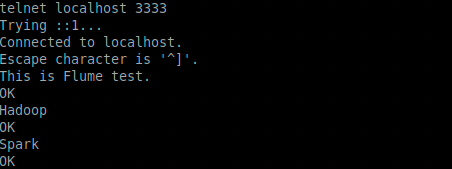
### Откроем ещё одно окно терминала и выполним команду telnet.

telnet localhost 3333

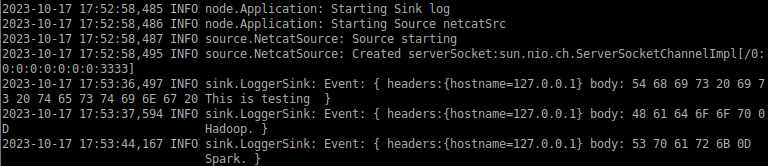
This is testing Flume with interceptor.

Hadoop

Spark



### Сообщение, отправленное в telnet на шаге 3.4, выводится с терминала, где агент был запущен на шаге 3.3, и подтверждается, что IP-адрес хоста, на котором в данный момент запущен агент, вставляется в заголовок события и передается дальше.



Замечание: IP-адрес - 127.0.0.1

### Удалим временную директорию, использовавшуюся для операций Flume.

$cd ~/flume

$rm -rf incoming output

### 

## Создайте новый поток данных «с нуля»

### Создайте файл конфигурации Flume со следующими настройками:

|  |  |
| --- | --- |
| Source | |
| Type | Netcat |
| Bind | localhost |
| Port | 11111 |
| Channel | |
| Type | Disk |
| Capacity | 1000 |
| transactionCapacity | 100 |
| Sink | |
| Type | logger |

### 

### Запустите агент

### В другом окне терминала запустите telnet и подключитесь к порту 11111. Напишите что-нибудь и убедитесь, что вы видите написанное в терминале, где запущен агент.

### 

### 

$telnet localhost 11111

...

Hello world! <ENTER>

OK