Flume и Morphlines трансформация потоков данных без строчки кода

Epam
Low Level Programming Department
Denis Pynkin

23 августа 2014 г.

1 Apache Flume

2 Cloudera Morphlines

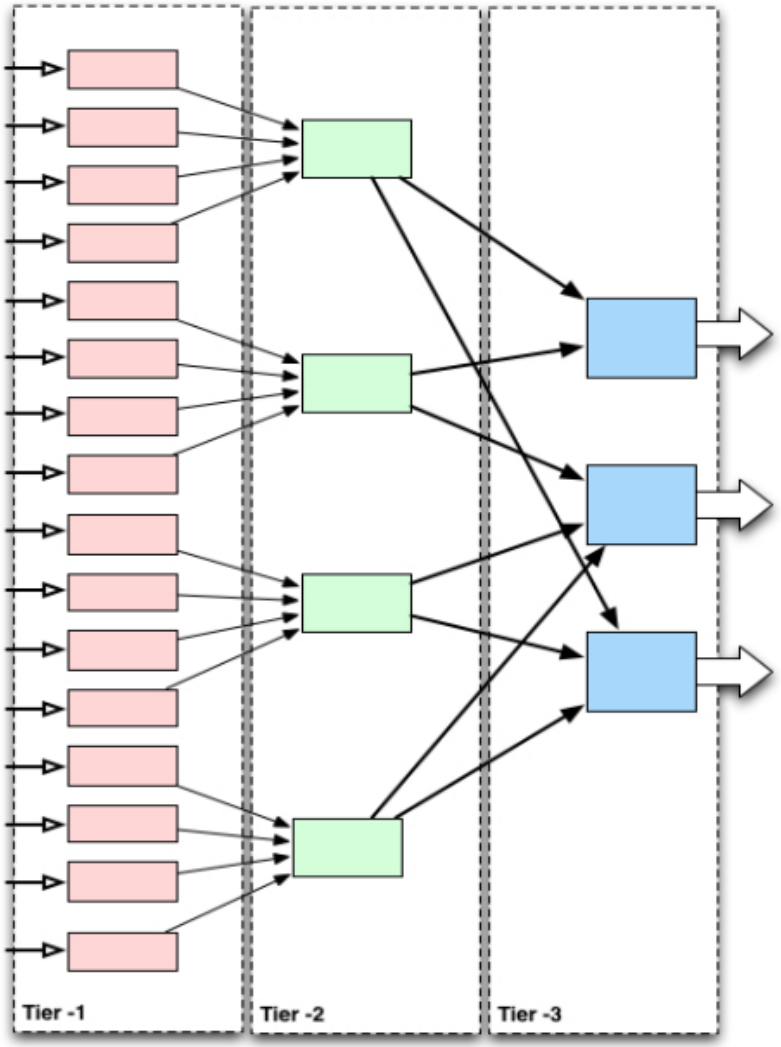
3 Синтетический пример использования

Apache Flume

Flume

Это распределенная система для эффективного сбора, агрегирования и перемещения больших потоков данных из множества разных источников в централизованное хранилище данных.

Apache Flume



Apache Flume

Достоинства Flume

- масштабируемость
- селективная и динамическая маршрутизация событий
- низкие задержки
- высокая пропускная способность
- декларативное описание

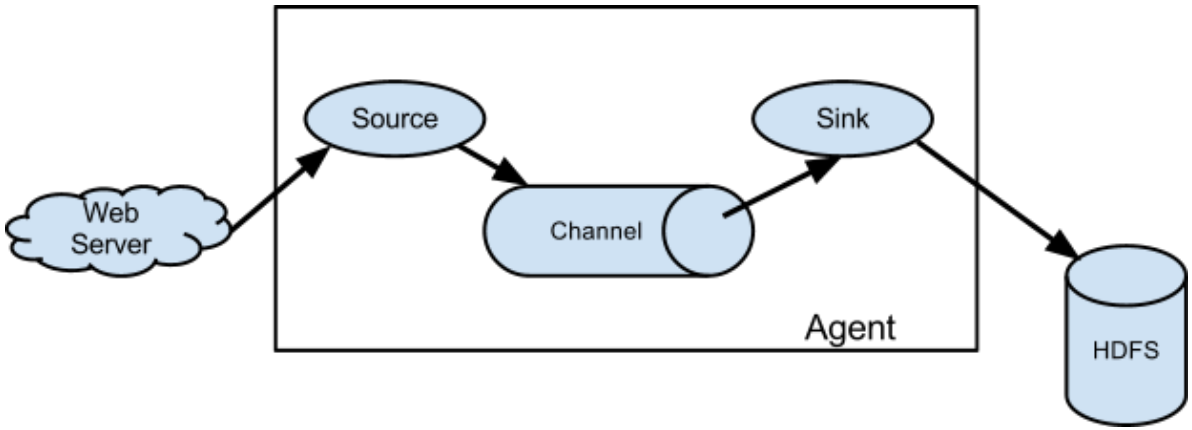
Конкуренты?

VS

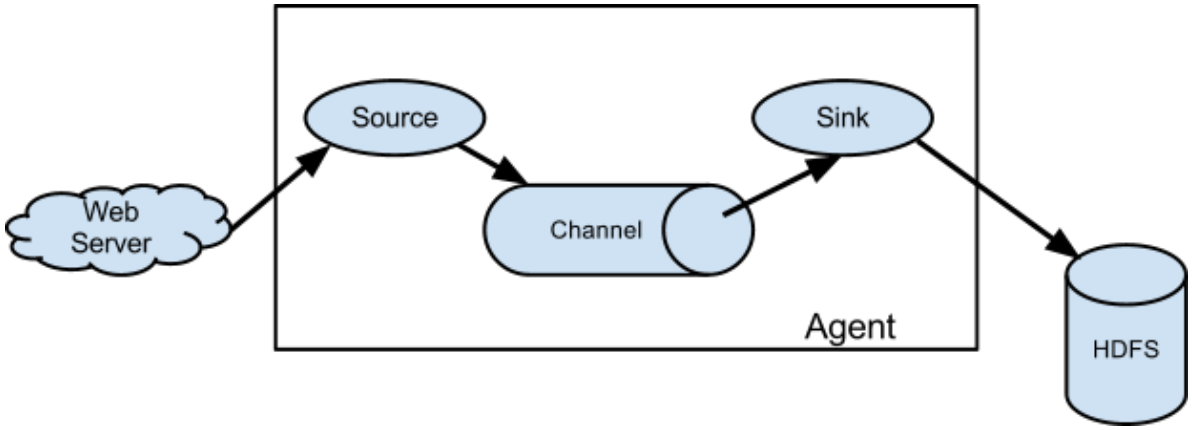
Модель данных

Flume Event

Определен, как элемент потока данных, состоящий из собственно данных и набора строковых атрибутов.

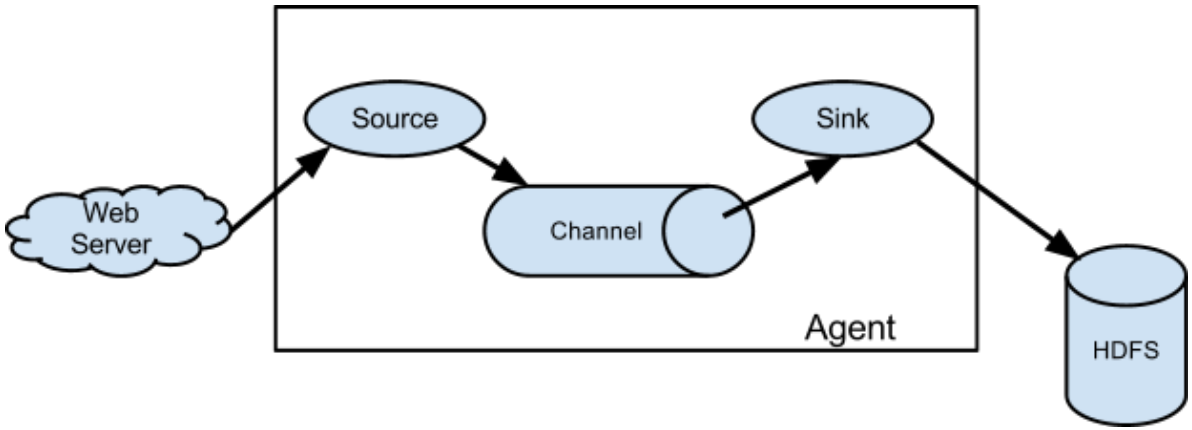


Flume: Source



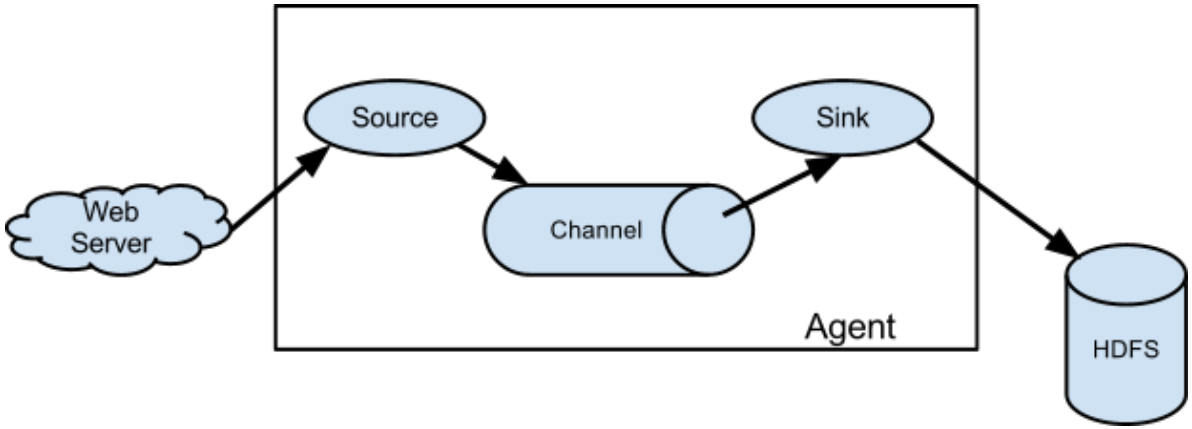
- AVRO
- Thrift
- JMS
- Exec
- NetCat
- Twitter
- Syslog
- HTTP
- **Custom**

Flume: Channel



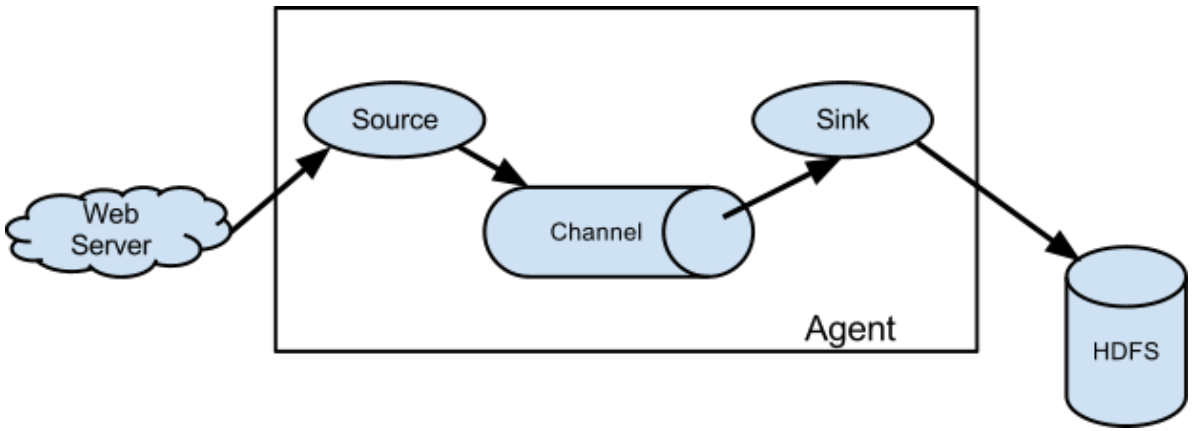
- Memory
- JDBS
- File
- **Custom**

Flume: Sink



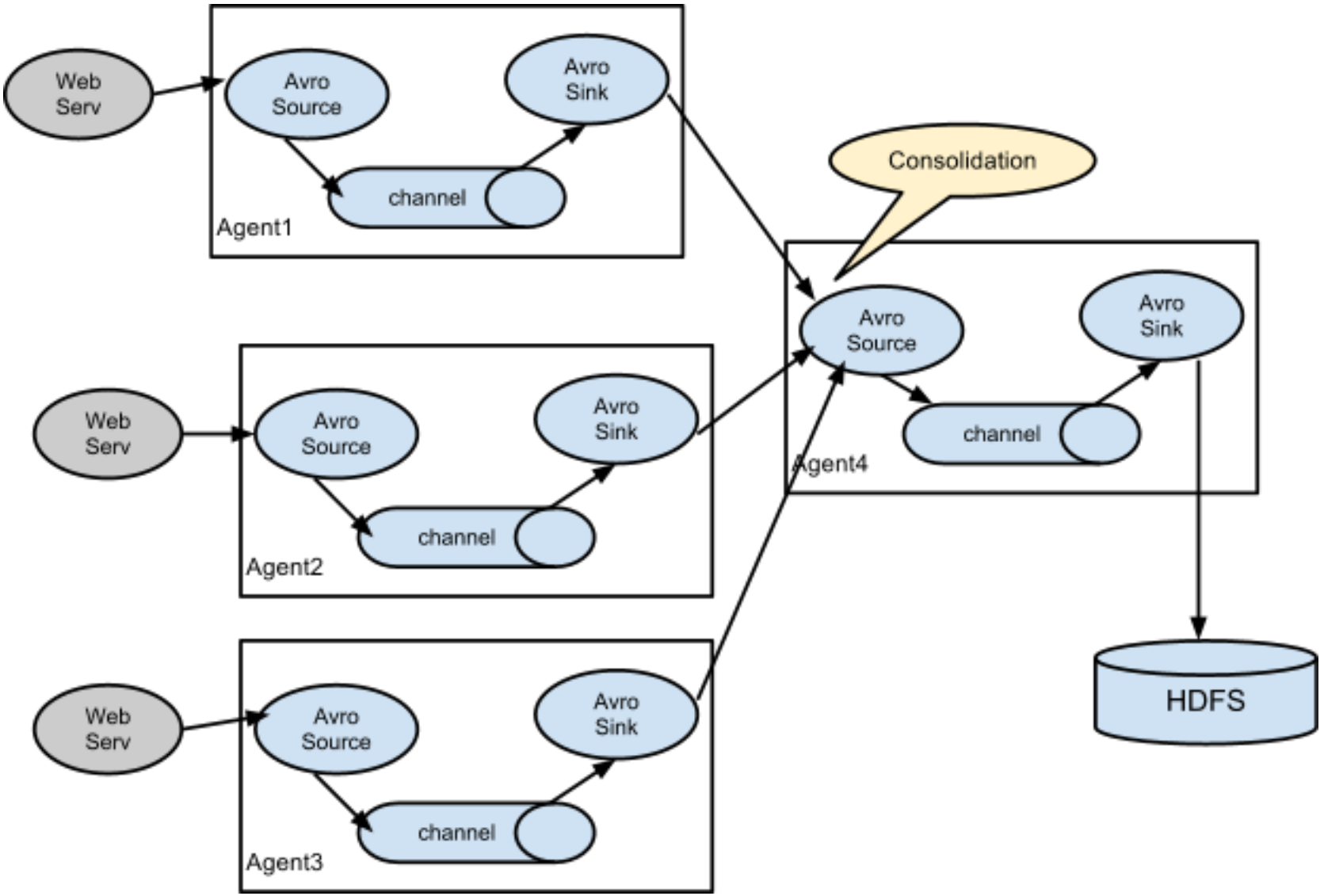
- AVRO
- Thrift
- Logger (stdout)
- File
- HDFS
- Null
- Solr
- Elastic Search
- **Custom**

Flume: модификаторы

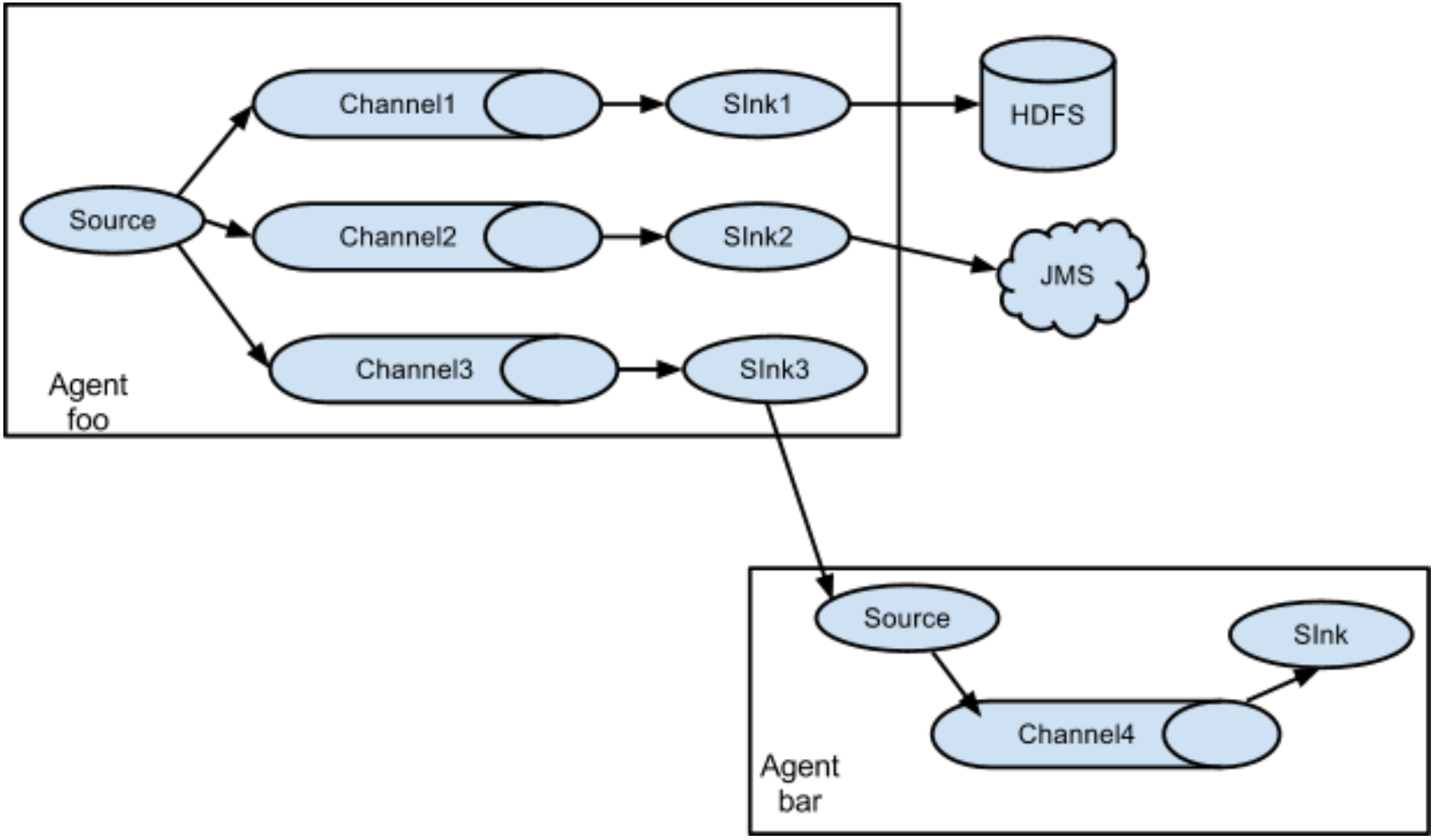


- Source Interceptor
- Channel Selector
- Sink Processor

Flume: топологии: fan-in



Flume: топологии: multiplexing/fan-out



1 Apache Flume

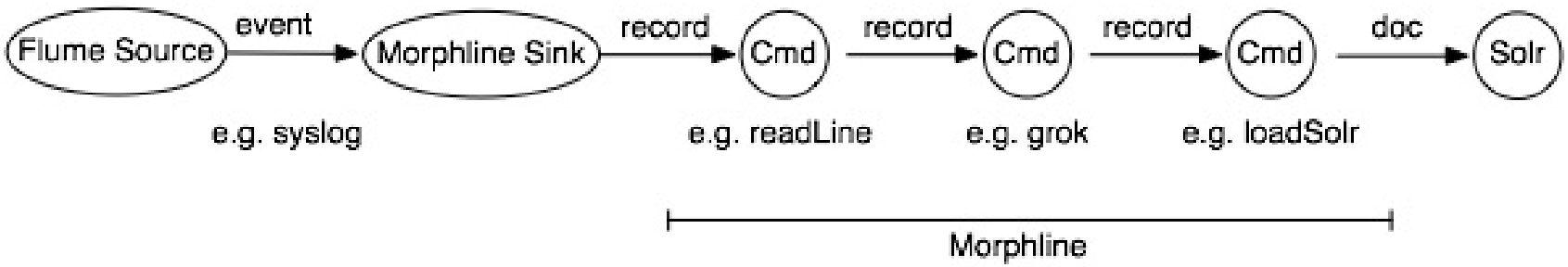
2 Cloudera Morphlines

3 Синтетический пример использования

Morphlines

Morphlines – это pipe

Pipe – это Morphlines



Morphlines

Конфигурационный файл, описывающий pipeline последовательных преобразований в парадигме ETL.

Модель данных

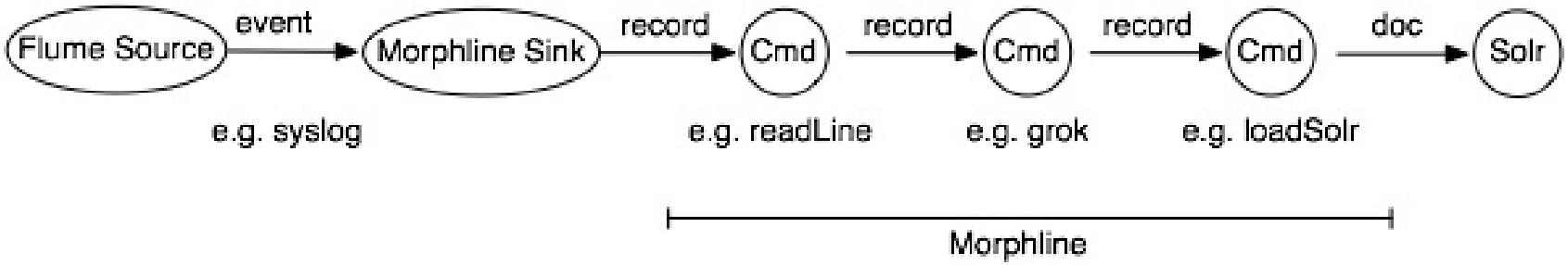
Концепция

Данные поступают в виде бесконечного (или хотя бы достаточно большого) потока записей

Record

Набор именованных полей, причем каждое поле может содержать от одного и более значений.
В качестве значения для поля можно использовать любой объект Java.

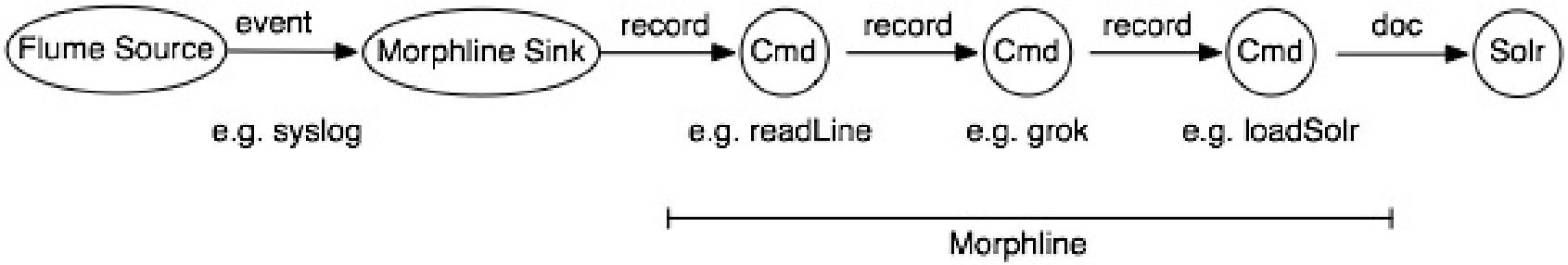
Morphlines



Morphlines – это фреймворк

Основная цель создания Morphlines – быстрая разработка приложений для обработки потоков данных в Hadoop с последующей записью результатов в Apache Solr, HBase, HDFS и прочие системы.

Morphlines



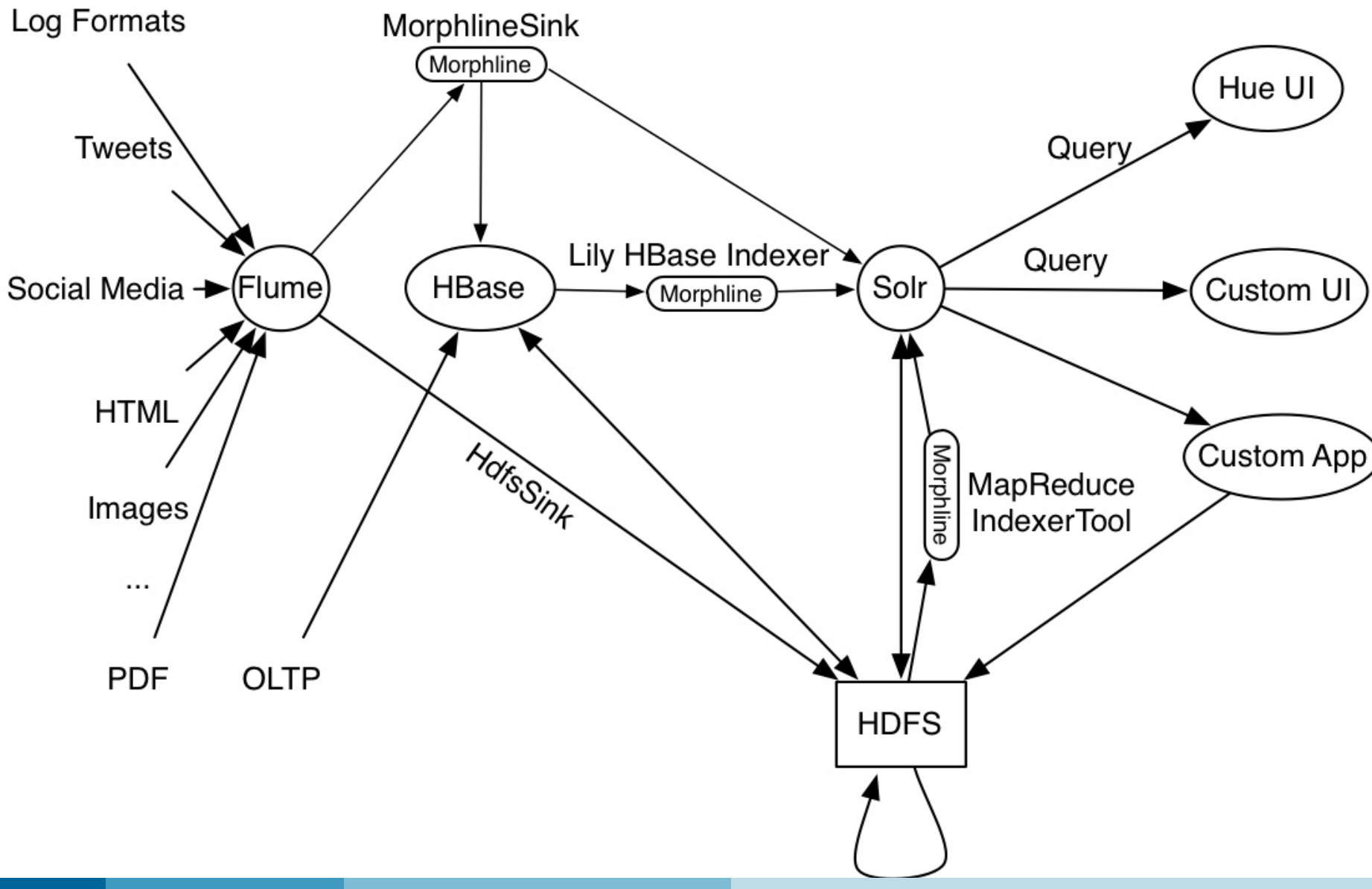
Интеграция с Flume

- Source interceptor
- Solr Sink

Интеграция

- любое Java приложение

Morphlines: пример интеграции



Описание преобразований

Конфигурационный файл

Последовательный список команд для преобразования данных, описанный в формате HOCON (Human Optimized Config Object Notation).
Возможность бранчевания с помощью команд pipe, if и tryRules.

Конфигурационный файл

```
morphlines: [ {
  id: vsftpFileLog
  importCommands: [ "org.kitesdk.**" ]

  commands: [
    { readLine { charset : UTF-8 } }

    { addLocalHost {
      field : server
      useIP : false
    } }

    { generateUUID { field : id} }
  ] } ]
```

1 Apache Flume

2 Cloudera Morphlines

3 Синтетический пример использования

Т3

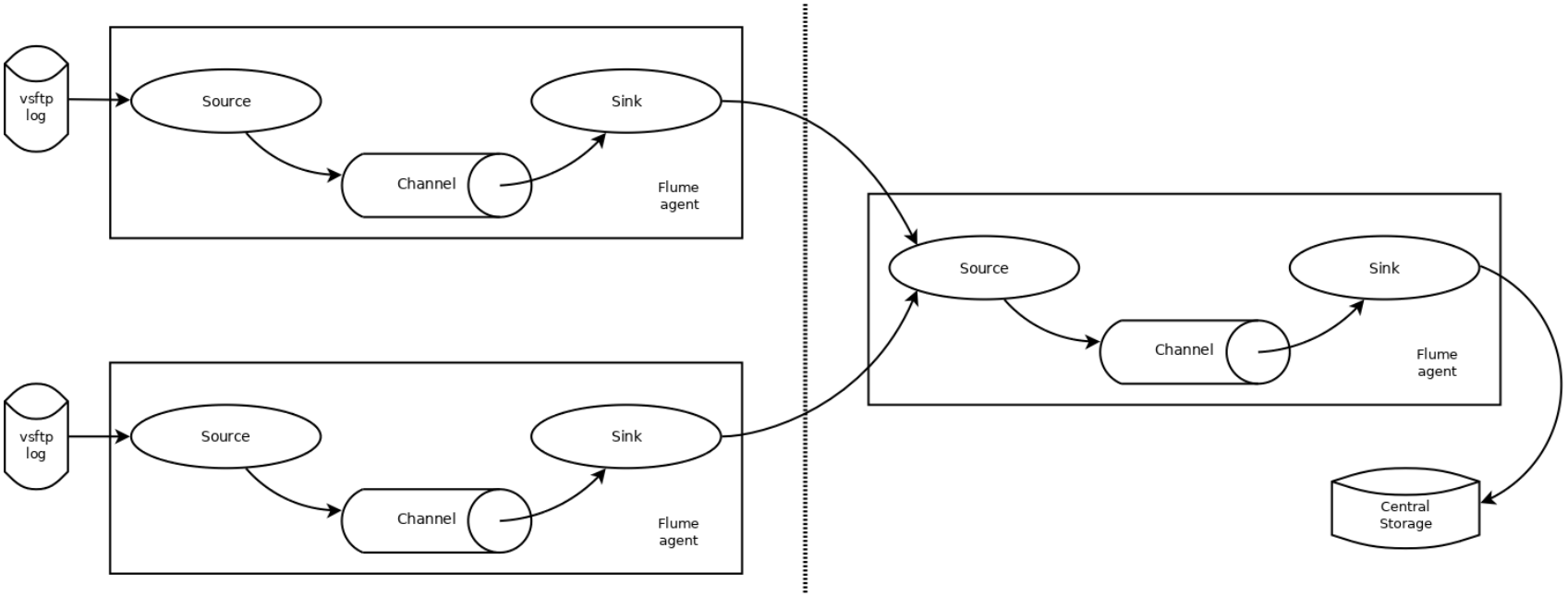
Разработать систему, собирающую статистику скачивания файлов со множества ftp-серверов и сохраняющую ее в файл в формате JSON, с указанием:

1. времени скачивания (в unix-time);
2. имени файла;
3. размера файла;
4. ip-адреса клиента;
5. имени сервера с которого скачали этот файл.

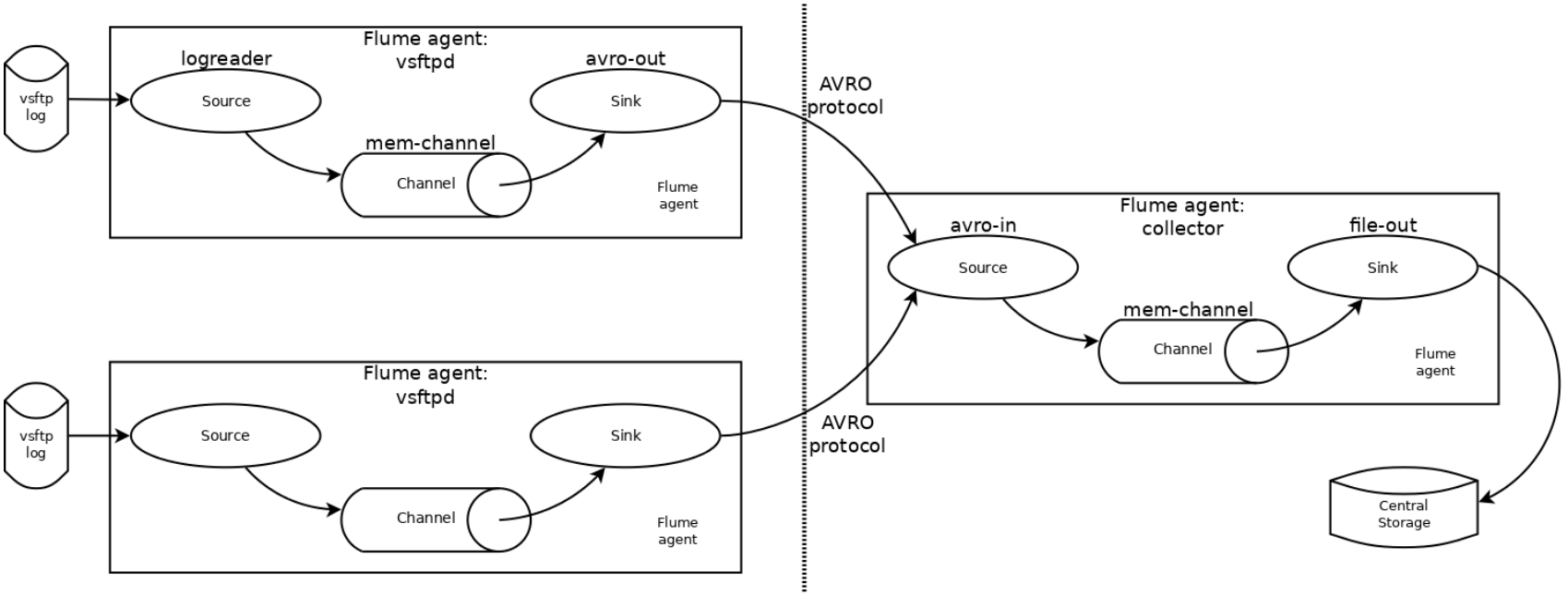
Формат строки в логе:

```
Tue Jul 22 19:15:23 2014 [pid 9388] [vsftpd] OK DOWNLOAD:
Client "10.6.136.54",
"/video/HighLoad master-class/out-1406031869.mkv",
189505527 bytes, 3909.30Kbyte/sec
```

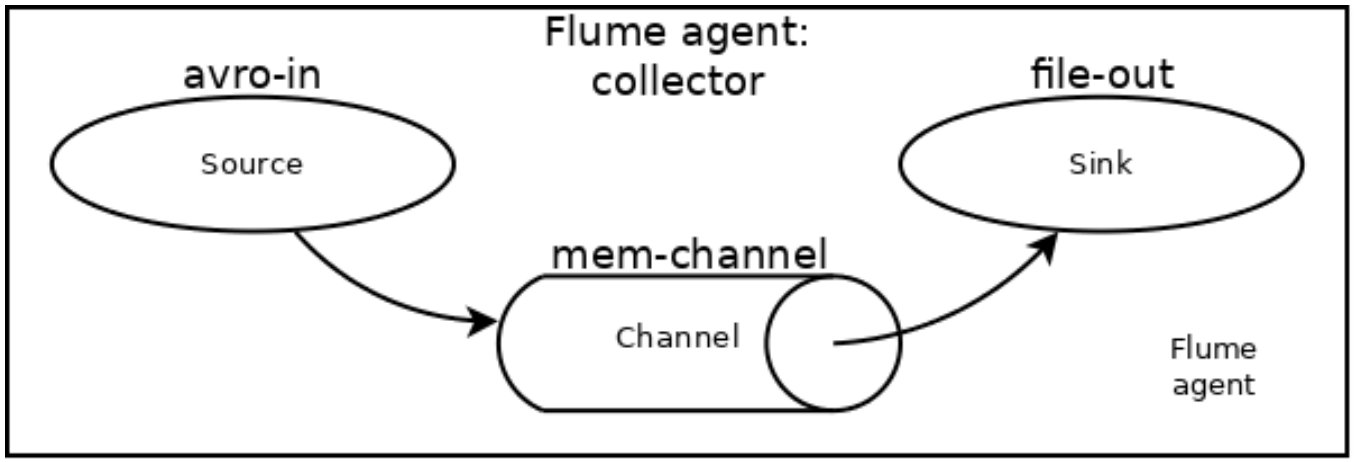
Архитектура



Архитектура: элементы



Описание агента 'collector'




```
collector.sources = avro-in
collector.channels = mem-channel
collector.sinks = file-out

collector.channels.mem-channel.type = memory
```

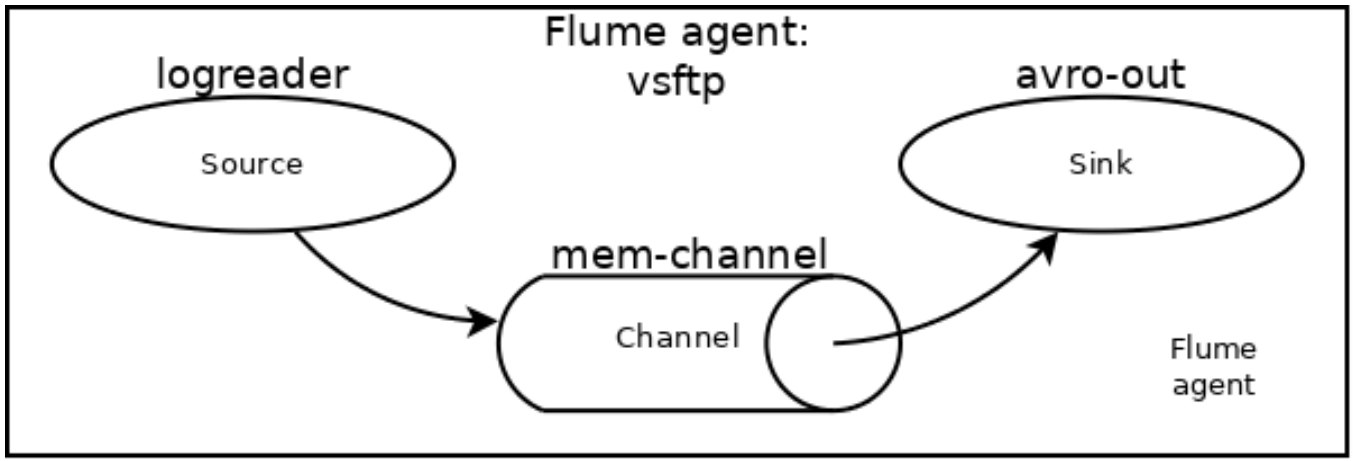
Описание агента 'collector': source

```
# get stream over avro channel
collector.sources.avro-in.type = avro
collector.sources.avro-in.channels = mem-channel
collector.sources.avro-in.bind = 0.0.0.0
collector.sources.avro-in.port = 5555
```

Описание агента 'collector': sink

```
# write stream to file every 5 minutes or 1000 events
collector.sinks.file-out.type = file_roll
collector.sinks.file-out.sink.rollInterval = 300
collector.sinks.file-out.batchSize = 1000
collector.sinks.file-out.sink.directory = /tmp/flume/log/
collector.sinks.file-out.channel = mem-channel
```

Описание агента 'vsftpd'




```
vsftpd.sources = logreader
vsftpd.channels = mem-channel
vsftpd.sinks = avro-out

vsftpd.channels.mem-channel.type = memory
```

Описание агента 'vsftpd': sink

```
# sink stream over avro channel
vsftpd.sinks.avro-out.type = avro
vsftpd.sinks.avro-out.channel = mem-channel
vsftpd.sinks.avro-out.hostname = 127.0.0.1
vsftpd.sinks.avro-out.port = 5555
```

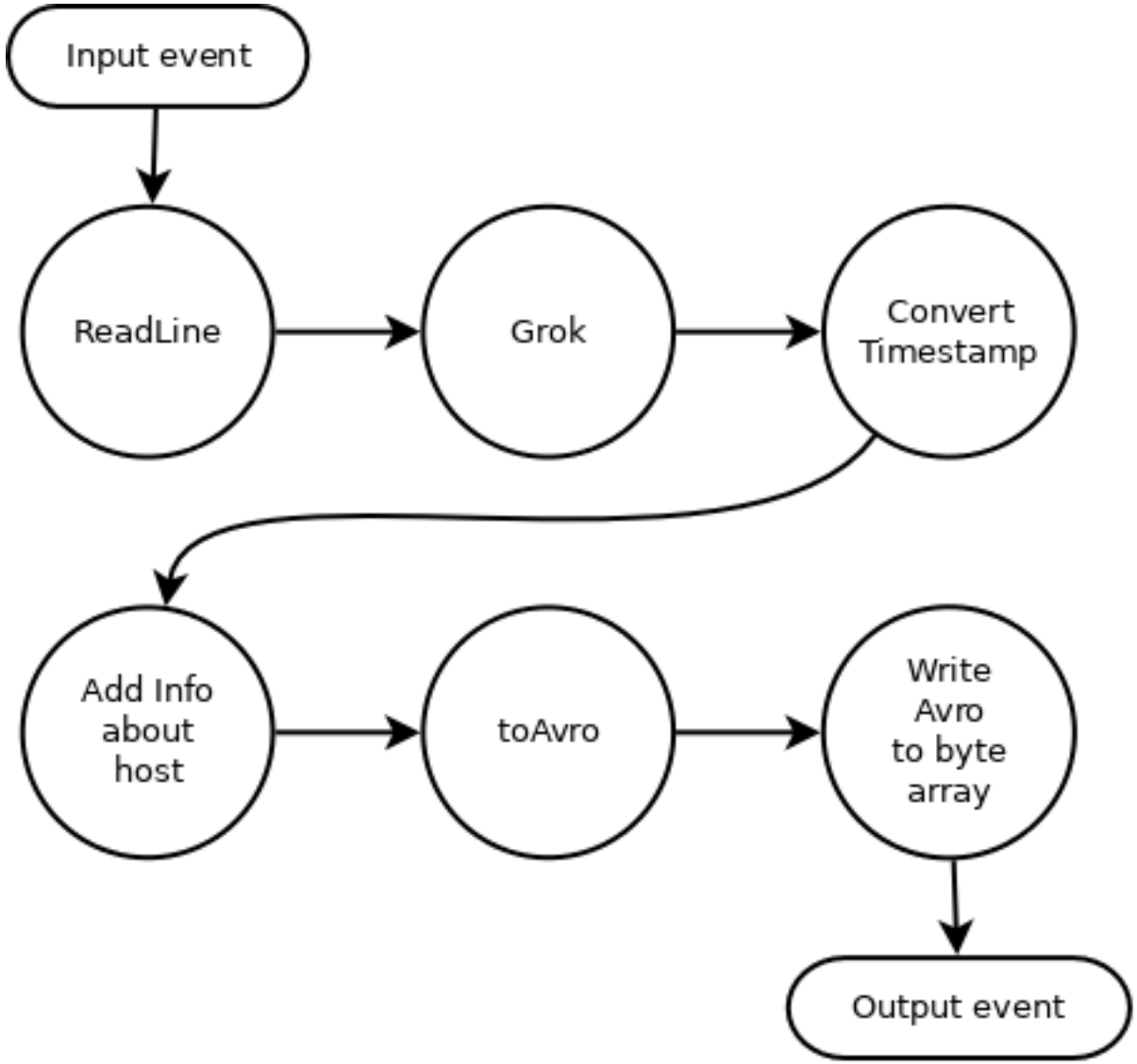
Описание агента 'vsftpd': source

```
# Read from log file
vsftpd.sources.logreader.type = exec
vsftpd.sources.logreader.command = tail -F /var/log/vsftpd.l
vsftpd.sources.logreader.channels = mem-channel
# expecting veeeeery loooong string
vsftpd.sources.logreader.deserializer = LINE
vsftpd.sources.logreader.deserializer.maxLineLength = 32768
```

Описание агента 'vsftpd': source: morphlines

```
# set interceptor for converting string to AVRO
vsftpd.sources.logreader.interceptors = morphline
# morphline interceptor config
vsftpd. . . .interceptors.morphline.type =
  org.apache.flume.sink.solr.morphline.MorphlineInterceptor$
vsftpd. . . .interceptors.morphline.morphlineFile =
  /tmp/flume/conf/morphlines.conf
vsftpd. . . .interceptors.morphline.morphlineId =
  vsftpFileLog
```

Morphline pipe



Morphline: разборка строк

```
{ grok {
  dictionaryString : """
TS %{DAY} %{MONTH} %{MONTHDAY} %{TIME} %{YEAR}
PATH (?>/(?>[\w\s_%!$@:.,-]+|\\.)*)+
  """
  expressions : {
    message : """
      %{TS:timestamp}.*Client "%{IP:ip}",
      "%{PATH:file}", %{INT:size}.*
"""
}}}
```

Morphline: преобразуем в AVRO

```
# Convert Morphline Record to AVRO-event according schema
{ toAvro {
 schemaString : """
 { "type": "record",
   "name": "ftpfile",
   "fields": [
     { "name": "timestamp","type": "long","default": -1 },
     { "name": "server","type": "string","default": "" },
     { "name": "ip", "type": "string", "default": "" },
     { "name": "file", "type": "string", "default": "" },
     { "name": "size", "type": "long",   "default": -1 }
   ]}
   """
} }
```

That's All Folks!

Вопросы ?

denis_pynkin@epam.com