

# Flume

Instalacja flume:

1. zaloguj się w cloudera managerze (admin / admin)
2. 'Flume-<twoj\_numer>', a następnie wejdź w zakładkę 'Configuration'
3. ustaw w polu 'HTTP Port' wartość 414<twoj\_numer>
4. Przekopiuj poniższą konfigurację do pola 'Configuration File'

```
tier1.sources = source1
tier1.channels = channel1
tier1.sinks = sink1

tier1.sources.source1.type = netcat
tier1.sources.source1.bind = 0.0.0.0
tier1.sources.source1.port = 99<twoj_numer>
tier1.sources.source1.channels = channel1
tier1.sources.source1.interceptors = i1 i2
tier1.sources.source1.interceptors.i1.type = org.apache.flume.interceptor.TimestampInterceptor$Builder
tier1.sources.source1.interceptors.i2.type = pl.isa.hadoop.LoggingInterceptor$Builder
tier1.sources.source1.interceptors.i2.logPrefix = LogPrefix
tier1.channels.channel1.type = memory

tier1.sinks.sink1.type = hdfs
tier1.sinks.sink1.channel = channel1
tier1.sinks.sink1.hdfs.path = /user/<login>/events/log_time=%y%m%d-%H%M
tier1.sinks.sink1.hdfs.filePrefix = events-
tier1.sinks.sink1.hdfs.fileType = DataStream
tier1.sinks.sink1.hdfs.writeFormat = Text

tier1.channels.channel1.capacity = 100
```

5. Dodaj do 'Plugin directories' wartość '/home/<login>/flume.plugins.d'
6. Kliknij 'save changes'
7. Utwórz na vm-cluster-node2 katalog '/home/<login>/flume.plugins.d/interceptors/lib' i wrzuć do niego zbudowanego jara z interceptorami.
8. Kliknij u góry na 'Actions'
9. Kliknij akcję 'Start'
10. uruchom generator transferów './generator.py adminsk01... 99<twoj\_numer>'
11. wylistuj katalog na hdfsie /user/<login>/events i zobacz czy pojawiają się logi transferów

Zadania:

1. Napisz interceptor, który odfiltruje przelewy na kwotę poniżej 200000, a pozostałe przepuści.
2. Zasymuluj opóźnienia w systemie dostarczania logów i zacznij generować przedwczorajsze logi poleceniem: `'./generator.py adminsk01... 9999 2'`. Napisz interceptor, który odkoduje datę z linii loga i ustawi odpowiedni timestamp w nagłówku. Do parsowania daty możesz skorzystać z klasy `SimpleDateFormat`. Sparsowana data ma metodę `getTime()`, która zwróci odpowiedni timestamp, który należy ustawić w nagłówku "timestamp".