Flume – Hands-On

Objetivo: O objetivo deste hands-on é utilizar o software Flume para ingestão de dados no HDFS.

Exercício

1. Faça o download de Apache Flume (https://flume.apache.org/download.html).

```
$ cd /vagrant
$ wget -c http://ftp.unicamp.br/pub/apache/flume/1.8.0/apache-flume-1.8.0-bin.tar.gz
```

2. Descompacte o arquivo

```
$ tar -zxvf apache-flume-1.8.0-bin.tar.gz
$ export FLUME_HOME=/vagrant/apache-flume-1.8.0-bin
```

3. Crie um arquivo de configuração para o Flume (flume-conf.properties) com o conteúdo abaixo.

```
agent1.sources = source1
agent1.sources.source1.type = exec
agent1.sources.source1.command = tail -F /tmp/events

agent1.sinks = sink1
agent1.sinks.sink1.type = hdfs
agent1.sinks.sink1.hdfs.path = /flume/events
agent1.sinks.sink1.hdfs.filePrefix = events-
agent1.sinks.sink1.hdfs.round = true
agent1.sinks.sink1.hdfs.roundValue = 10
agent1.sinks.sink1.hdfs.roundUnit = minute

agent1.channels = channel1
agent1.channels.channel1.type = memory
agent1.channels.channel1.capacity = 1000
agent1.channels.channel1.transactionCapacity = 100
agent1.channels.channel1.byteCapacityBufferPercentage = 20

agent1.sources.source1.channels = channel1
agent1.sinks.sink1.channel = channel1
```

4. Instale o pacote wamerican-insane

```
$ sudo apt install wamerican-insane
```

5. Instale o python2.7 na máquina local

```
$ sudo apt install python2.7 python-all
```

6. Crie um script Python (gen_events.py) com o conteúdo abaixo.

```
from random import randint
from datetime import datetime
word_file = "/usr/share/dict/american-english-insane"
WORDS = open(word_file).read().splitlines()
while True:
  currtime = datetime.now()
  words = ""
  for i in range(0,9):
    words = words + WORDS[randint(0,400000)] + " "
  print currtime.strftime('%Y/%m/%d %H:%M:%S') + "\t" + words
```

7. Crie um diretório no HDFS para fazer a ingestão de dados usando o Flume

```
$ $HADOOP_HOME/bin/hdfs dfs -mkdir -p /flume/events
```

8. Copie os arquivos jar do Hadoop necessários para funcionamento do Flume

```
$ sudo cp $HADOOP_HOME/share/hadoop/common/*.jar $FLUME_HOME/lib
$ sudo cp $HADOOP_HOME/share/hadoop/common/lib/*.jar $FLUME_HOME/lib
$ sudo cp $HADOOP_HOME/share/Hadoop/hdfs/*.jar $FLUME_HOME/lib
```

9. Abra 3 terminais

10. No primeiro terminal, execute o agente Flume

```
$ $FLUME_HOME/bin/flume-ng agent --conf $HADOOP_HOME/etc/hadoop --conf-file flume-conf.properties --name agent1
```

11. No segundo terminal, execute o script para geração dos eventos

```
$ python2.7 gen_events.py >> /tmp/events
```

12. No terceiro terminal, verifique os arquivos sendo criados no HDFS para capturar os eventos

```
$ $HADOOP_HOME/bin/hadoop fs -ls /flume/events
```

13. Pare o gerador de eventos usando Control+C

14. Verifique o conteúdo dos arquivos gerados no HDFS

15. Teste outras configurações de channel e sink (por exemplo, o sink Logger).