Prácticas BigData

1. Lanzar un proceso en Streaming con Python

- En este caso vamos a crear un programa Python que recupere de forma aleatoria parte de las filas del fichero.
- Le llamamos “rand.py” por ejemplo

```
#!/usr/bin/env python
import sys, random
for line in sys.stdin:
        if (random.randint(1,100) <= int(sys.argv[1])):
                print line.strip()
```

- Lo ejecutamos con un reducer de 1

```
hadoop jar /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.9.0.jar -D mapred.job.reducers=1 -input /practicas/cite75_99.txt -output /resultado11 -mapper 'rand.py' -file rand.py

packageJobJar: [random.py, /tmp/hadoop-unjar2385850152370180720/] [] /tmp/streamjob2584922200311003370.jar tmpDir=null
18/01/07 15:26:33 INFO client.RMProxy: Connecting to ResourceManager at nodo1/192.168.56.101:8032
18/01/07 15:26:33 INFO client.RMProxy: Connecting to ResourceManager at nodo1/192.168.56.101:8032
18/01/07 15:26:35 INFO mapred.FileInputFormat: Total input files to process : 1
18/01/07 15:26:35 INFO net.NetworkTopology: Adding a new node: /default-rack/192.168.56.123:50010
18/01/07 15:26:35 INFO net.NetworkTopology: Adding a new node: /default-rack/192.168.56.103:50010
18/01/07 15:26:35 INFO net.NetworkTopology: Adding a new node: /default-rack/192.168.56.132:50010
18/01/07 15:26:35 INFO net.NetworkTopology: Adding a new node: /default-rack/192.168.56.125:50010
18/01/07 15:26:35 INFO mapreduce.JobSubmitter: number of splits:2
18/01/07 15:26:36 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
18/01/07 15:26:36 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1515325737805_0010
18/01/07 15:26:37 INFO impl.YarnClientImpl: Submitted application application_1515325737805_0010
18/01/07 15:26:37 INFO mapreduce.Job: The url to track the job: http://nodo1:8088/proxy/application_1515325737805_0010/
18/01/07 15:26:37 INFO mapreduce.Job: Running job: job_1515325737805_0010
```