SISTEMA PARA GENERAR GRÁFICAS A PARTIR DE LOGS TCPDUMP USANDO HADOOP

Ángel Stalin Cruz Palaquibay
Pedro Alfredo Torres Arellano

Descripción general

- El Problema
- Motivación
- Objetivos
- Metodología del proyecto
- Resultados y Conclusiones
- Recomendaciones

EL PROBLEMA

El Problema

- Capacidad de procesamiento y análisis de logs tcpdump del orden de Gigabytes en programas tradicionales.

MOTIVACIÓN

Motivación

- La importancia de analizar logs tcpdump como fuente de conocimiento frente a ataques de red.

- Conocer y utilizar herramientas que implementen programación paralela para analizar grandes cantidades de datos.

OBJETIVOS

Objetivos

- Desarrollar una solución al problema utilizando herramientas que implemente programación paralela.
- Analizar el uso de los servicios de web de Amazon en la implementación de la solución.

METODOLOGÍA

Diseño de la Solución

Pre – Procesamiento Data set

Procesamiento masivo con MapReduce

Generación de Gráficas a partir resultados proceso anterior

Pre – Procesamiento Data set

Pre – Procesamiento Data set

- Formato para los archivos de texto generados

octubre - Bloc de notas

Archivo Edición Formato Ver Ayuda

```
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:10,src_ip=124.114.130.149,dst_ip=200.9.176.182,src_port=45023,dst_port=21,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:10,src_ip=200.9.176.182,dst_ip=124.114.130.149,src_port=21,dst_port=45023,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:11,src_ip=124.114.130.149,dst_ip=200.9.176.182,src_port=45023,dst_port=21,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:11,src_ip=200.9.176.182,dst_ip=124.114.130.149,src_port=21,dst_port=45023,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:11,src_ip=124.114.130.149,dst_ip=200.9.176.182,src_port=45023,dst_port=21,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:11,src_ip=200.9.176.182,dst_ip=124.114.130.149,src_port=21,dst_port=45023,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:12,src_ip=124.114.130.149,dst_ip=200.9.176.182,src_port=45023,dst_port=21,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:12,src_ip=200.9.176.182,dst_ip=124.114.130.149,src_port=21,dst_port=45023,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:12,src_ip=124.114.130.149,dst_ip=200.9.176.182,src_port=45023,dst_port=21,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:12,src_ip=200.9.176.182,dst_ip=124.114.130.149,src_port=21,dst_port=45023,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:13,src_ip=124.114.130.149,dst_ip=200.9.176.182,src_port=45023,dst_port=21,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:13,src_ip=200.9.176.182,dst_ip=124.114.130.149,src_port=21,dst_port=45023,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:13,src_ip=124.114.130.149,dst_ip=200.9.176.182,src_port=45023,dst_port=21,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:13,src_ip=200.9.176.182,dst_ip=124.114.130.149,src_port=21,dst_port=45023,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:13,src_ip=124.114.130.149,dst_ip=200.9.176.182,src_port=45023,dst_port=21,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:13,src_ip=200.9.176.182,dst_ip=124.114.130.149,src_port=21,dst_port=45023,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:14,src_ip=124.114.130.149,dst_ip=200.9.176.182,src_port=45023,dst_port=21,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:14,src_ip=200.9.176.182,dst_ip=124.114.130.149,src_port=21,dst_port=45023,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:14,src_ip=124.114.130.149,dst_ip=200.9.176.182,src_port=45023,dst_port=21,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:14,src_ip=200.9.176.182,dst_ip=124.114.130.149,src_port=21,dst_port=45023,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:15,src_ip=124.114.130.149,dst_ip=200.9.176.182,src_port=45023,dst_port=21,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:15,src_ip=200.9.176.182,dst_ip=124.114.130.149,src_port=21,dst_port=45023,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:15,src_ip=124.114.130.149,dst_ip=200.9.176.182,src_port=45023,dst_port=21,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:15,src_ip=200.9.176.182,dst_ip=124.114.130.149,src_port=21,dst_port=45023,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:16,src_ip=124.114.130.149,dst_ip=200.9.176.182,src_port=45023,dst_port=21,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:16,src_ip=200.9.176.182,dst_ip=124.114.130.149,src_port=21,dst_port=45023,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:16,src_ip=124.114.130.149,dst_ip=200.9.176.182,src_port=45023,dst_port=21,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:16,src_ip=200.9.176.182,dst_ip=124.114.130.149,src_port=21,dst_port=45023,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:16,src_ip=124.114.130.149,dst_ip=200.9.176.182,src_port=45023,dst_port=21,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:16,src_ip=200.9.176.182,dst_ip=124.114.130.149,src_port=21,dst_port=45023,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:17,src_ip=124.114.130.149,dst_ip=200.9.176.182,src_port=45023,dst_port=21,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:17,src_ip=200.9.176.182,dst_ip=124.114.130.149,src_port=21,dst_port=45023,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:1:17,src_ip=124.114.130.149,dst_ip=200.9.176.182,src_port=45023,dst_port=21,protocolo=TCP
Anio=2008,Mes=octubre,Dia=28,Hora=8:20:10,src_ip=200.9.176.182,dst_ip=200.9.176.129,src_port=null,dst_port=null,protocolo=ARP
Anio=2008,Mes=octubre,Dia=28,Hora=8:20:10,src_ip=200.9.176.129,dst_ip=200.9.176.182,src_port=null,dst_port=null,protocolo=ARP
Anio=2008,Mes=octubre,Dia=28,Hora=8:20:10,src_ip=200.9.176.182,dst_ip=200.9.176.5,src_port=32772,dst_port=53,protocolo=UDP
```

Procesamiento masivo con MapReduce

- JNetStream es una librería de Java que nos permite capturar y enviar paquetes.
- JFreeChart es una librería de Java que facilita la generación de gráficos.
- Hadoop es una plataforma que nos permite desarrollar aplicaciones distribuidas la cual nos provee escalabilidad.

Procesamiento masivo con MapReduce

- S3 (Simple Storage Service).- Es el servicio que ofrece Amazon para almacenar y recuperar cualquier cantidad de datos en cualquier momento y desde cualquier lugar en la web.

- Con este servicio almacenamos datos de entrada, programa MapReduce en un archivo jar y los resultados obtenidos luego de procesarlos.

Procesamiento masivo con MapReduce

- EC2 (Elastic Compute Cloud).- Este servicio de Amazon provee los recursos computacionales necesarios para correr nuestra aplicación MapReduce.

- Es aquí donde se ejecutan los procesos de map y reduce en la data que es recuperada desde S3.

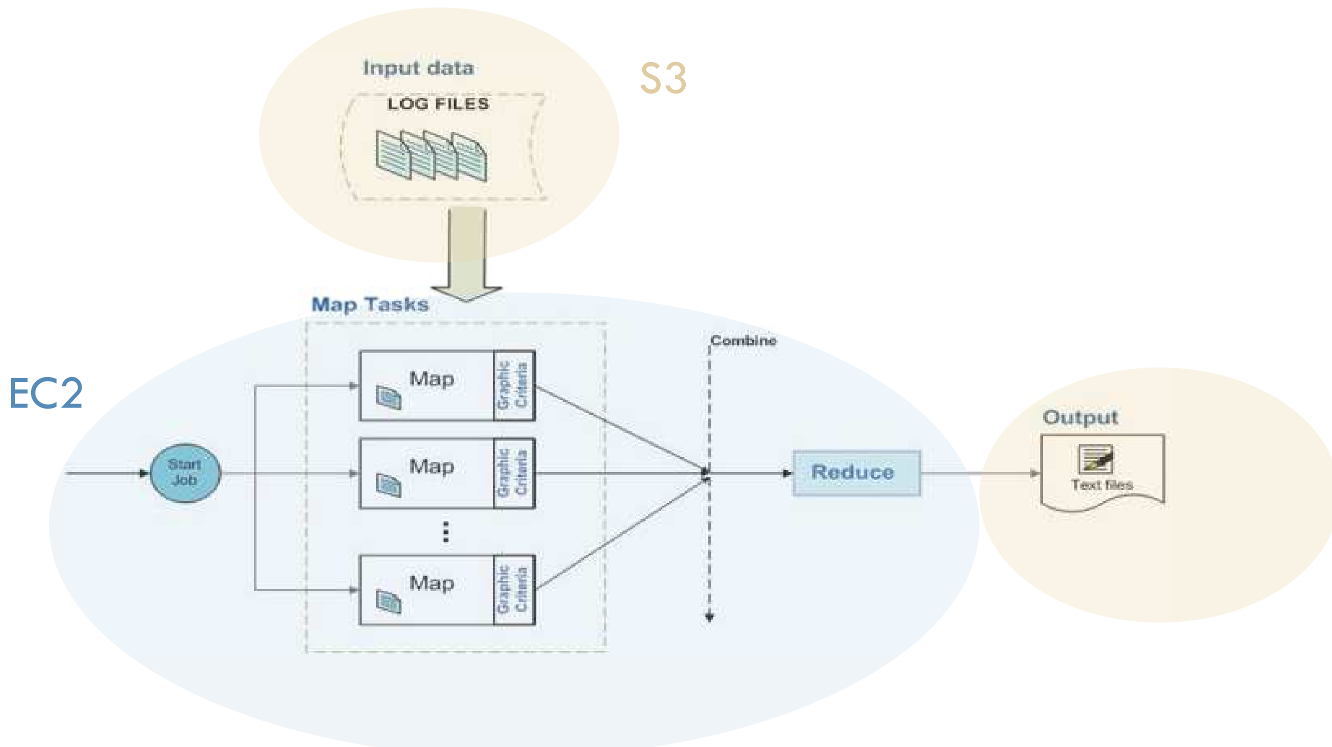
Procesamiento masivo con MapReduce

- Elastic MapReduce.- Es un servicio web que permite procesar grandes cantidades de datos, de una forma ordenada (pasos) a manera de algoritmo donde utiliza la infraestructura de Amazon EC2 y Amazon S3.

- Esto lo utilizamos para realizar los pasos de copiado del S3 a HDFS y visceversa y procesamiento MapReduce.

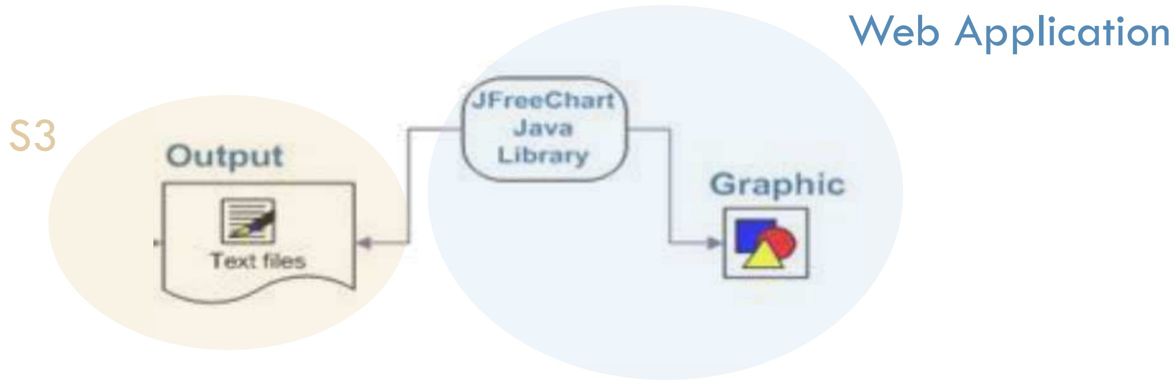
Procesamiento masivo con MapReduce

- Sucede en EC2 de Amazon.
- Datos son tomados del S3 de Amazon
- Resultados almacenados en el S3 de Amazon



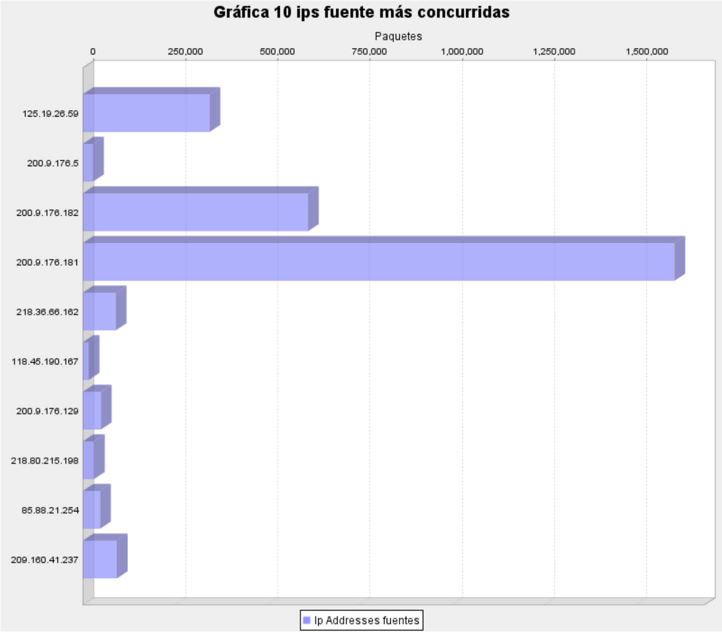
Generación de Gráficas a partir resultados proceso anterior

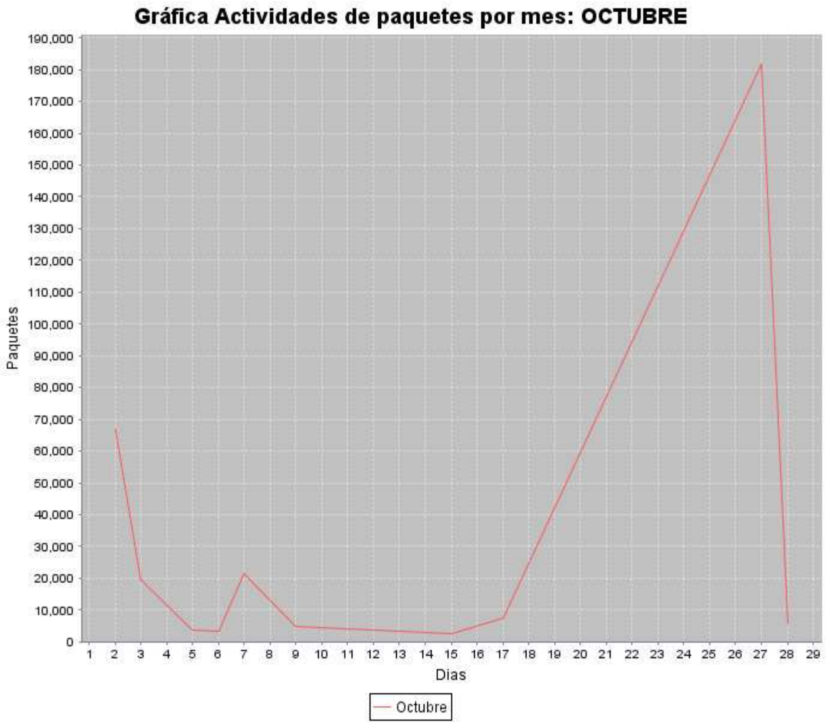
- Recuperación de resultados del S3 mediante aplicación web
- Uso de librería S3 para Java de Amazon
- Generar graficas con JFreeChart



Generación de Gráficas a partir resultados proceso anterior

- Recuperación de resultados del S3 mediante aplicación web

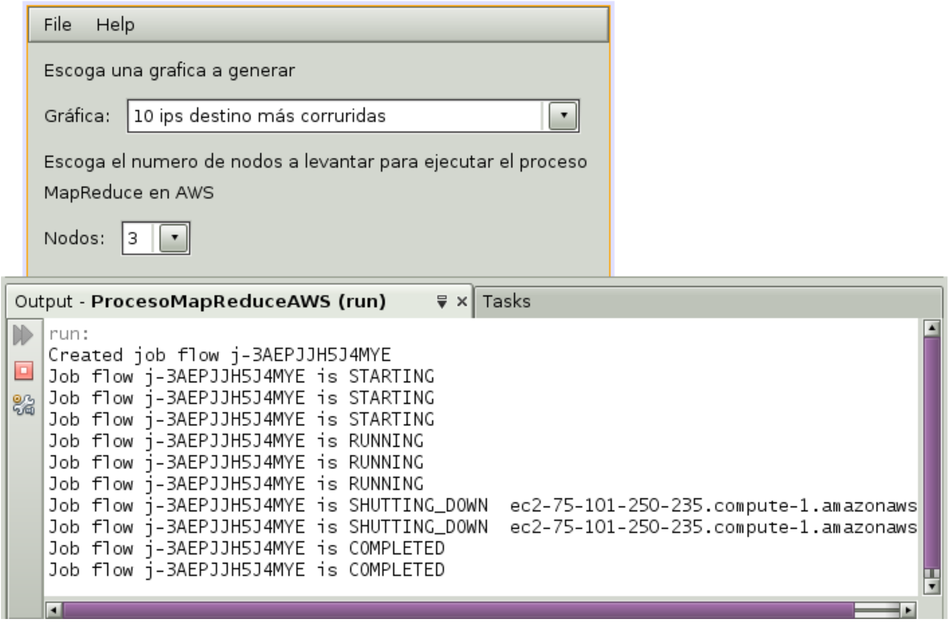




PRUEBAS

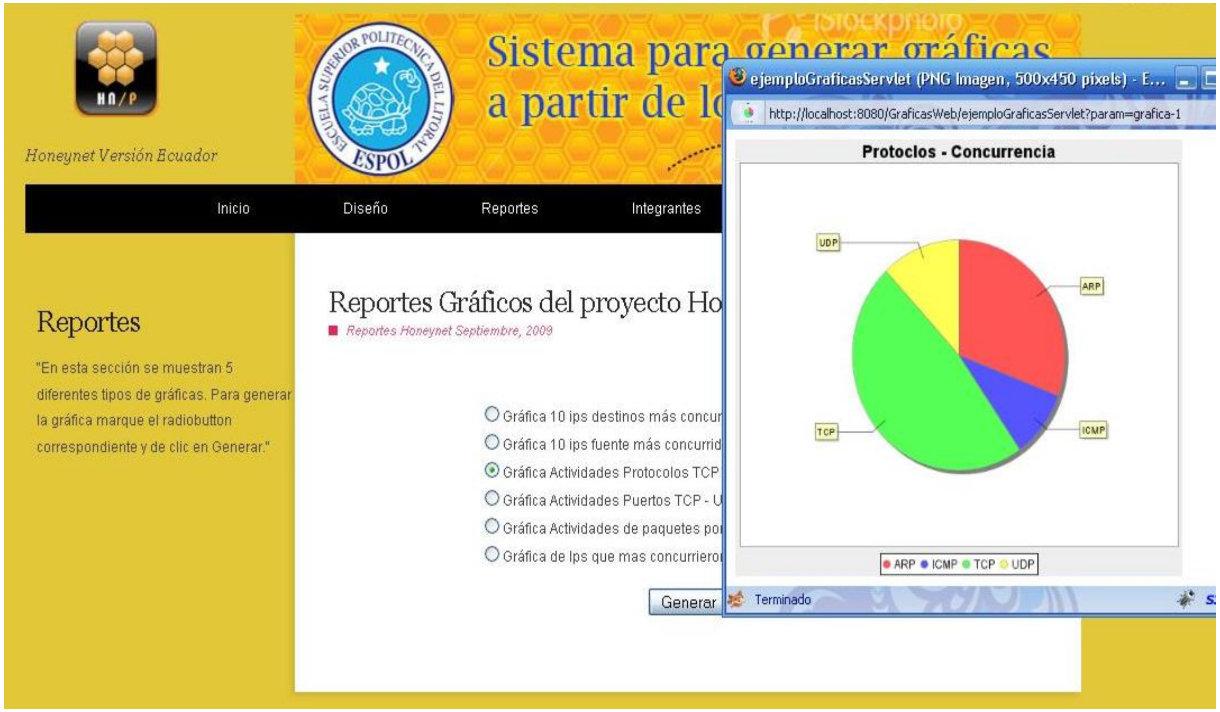
Pruebas

- Levantamiento de nodos en EC2
- Utilización de Elastic MapReduce



Pruebas

- Con la aplicación web recuperamos los resultados del S3 y mostramos la gráfica



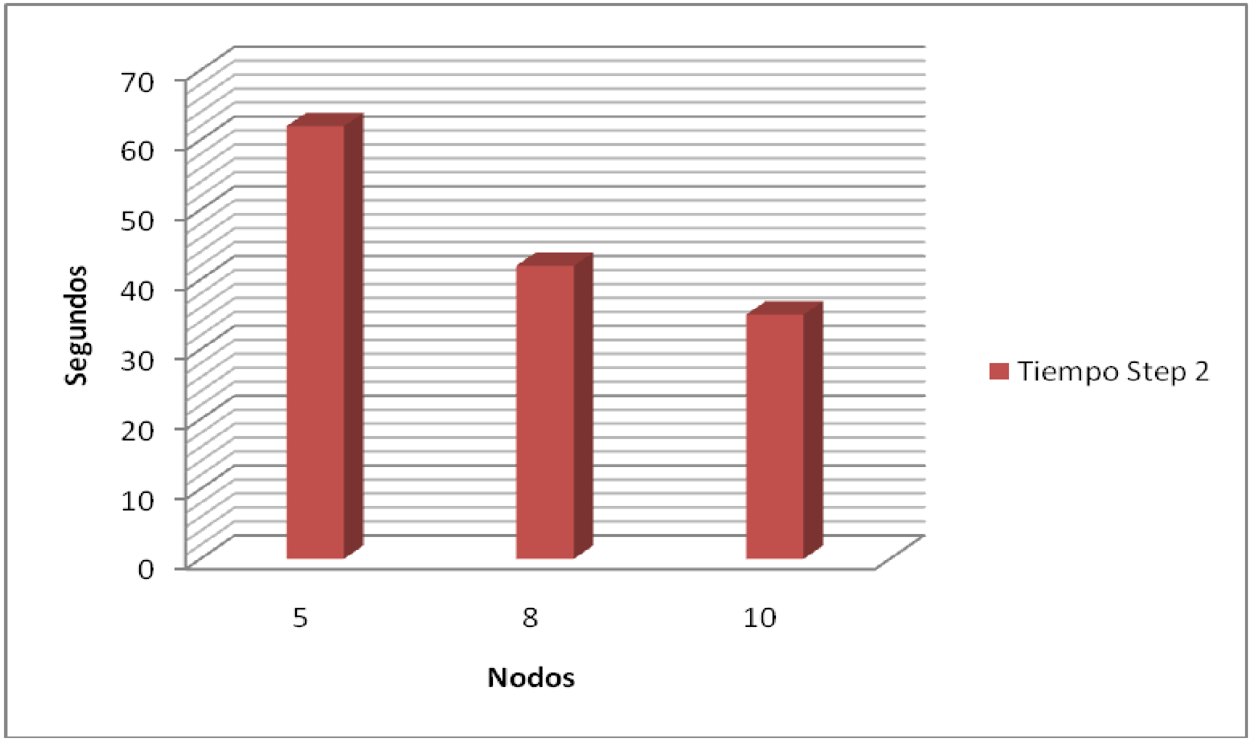
Tiempo de procesamiento de datos Wireshark vs MapReduce

	90MB	265MB	1GB
WIRESHARK	2 min	8min 18 seg	No fue posible
MAPREDUCE	47 seg	1 min 28 seg	6 min 3seg

Tiempos obtenidos sobre un data set de 1.4GB

	Tiempo			
No. Nodos	**Step 1 "S3 to HDFS"**	**Step 2 "MapReduce"**	**Step 3 "HDFS to S3"**	**Total**
5	66 minutos	62 segundos	21 segundos	67 min 23 seg
8	66 minutos	41 segundos	21 segundos	67 min 2 seg
10	65 minutos	35 segundos	23 segundos	65 min 58 seg

Tiempos obtenidos sobre un data set de 1.4GB



Conclusión

- Este sistema realiza un análisis en todo el data set, obteniendo información de mayor alcance que si lo hiciéramos con una herramienta común para este tipo de análisis.
- El uso de los servicios de Amazon como parte de la solución, fue de gran ayuda al proveernos los recursos computacionales para las pruebas del sistema.

PREGUNTAS