



ugr

Universidad
de Granada

Máster en Ciencia de Datos e Ingeniería de Computadores

Big Data II: Almacenamiento de datos masivos para procesamiento y análisis ETL

Parte I: Diseña un experimento ETL con Pig a tu medida

Autora

Lidia Sánchez Mérida

Contacto

lidiasm96@correo.ugr.es



Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

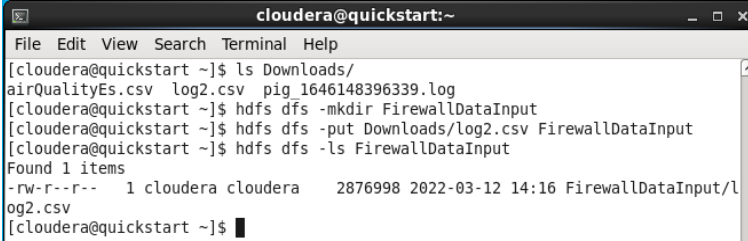
Granada, Marzo de 2022

Descripción de la base de datos

La base de datos seleccionada se denomina **Internet Firewall** y se encuentra disponible en el repositorio *UCI Machine Learning*, junto con un amplio conjunto de datasets que permiten realizar diversos experimentos para resolver problemas de regresión y clasificación. Este conjunto de datos procede de los registros generados por un *firewall* instalado en equipos informáticos de una universidad y data del año 2019. Dispone de un total de **65.532 registros y 12 columnas**, mayormente numéricas a excepción de la variable dependiente que se compone de cuatro clases: *allow*, *deny*, *drop*, *reset-both*. Una de las aplicaciones más interesantes consiste en **modelar el comportamiento del tráfico online** utilizando esta información para mejorar la toma de decisiones por parte de programas de seguridad, como los *firewalls*, a través de un clasificador que sea capaz de identificar qué acción es la más apropiada para cada situación. A partir de esta idea ha surgido un nuevo concepto conocido como NGFW (*Next-generation firewalls*) que hace referencia al nuevo estándar de seguridad que se lleva desarrollando e implementando en los últimos años. El principal objetivo consiste en construir **programas de seguridad proactivos** introduciendo técnicas de Aprendizaje Automático que les permitan aprender nuevas amenazas y comportamientos sospechosos para reducir el tiempo de respuesta en aplicar una solución automática y acorde a diversos ataques de ciberseguridad.

Experimentación con Pig

Al realizar esta práctica tras el ejercicio de Impala, ya disponemos del fichero **log2.csv** que contiene la base de datos explicada anteriormente dentro de la carpeta *Downloads*. A continuación construimos un directorio específico para trasladar este archivo al sistema de ficheros HDFS tal y como se muestra en la Figura 1.

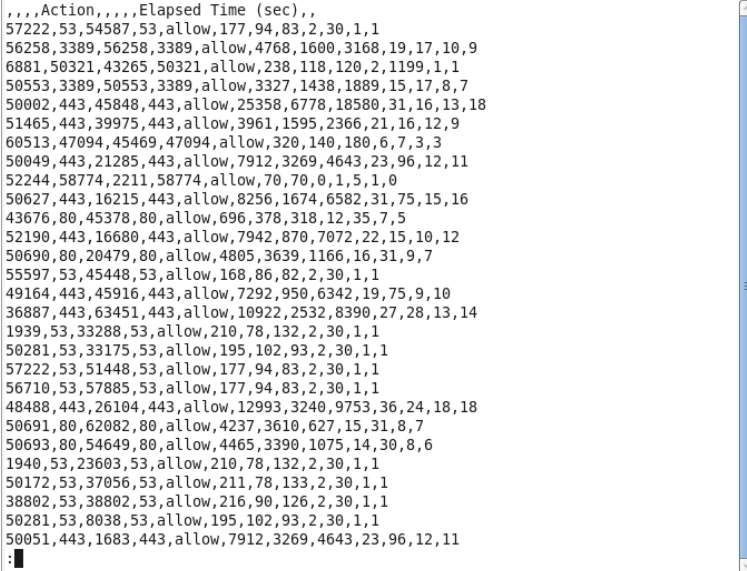


```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
[cloudera@quickstart ~]$ ls Downloads/  
airQualityEs.csv log2.csv pig 1646148396339.log  
[cloudera@quickstart ~]$ hdfs dfs -mkdir FirewallDataInput  
[cloudera@quickstart ~]$ hdfs dfs -put Downloads/log2.csv FirewallDataInput  
[cloudera@quickstart ~]$ hdfs dfs -ls FirewallDataInput  
Found 1 items  
-rw-r--r-- 1 cloudera cloudera 2876998 2022-03-12 14:16 FirewallDataInput/log2.csv  
[cloudera@quickstart ~]$
```

Figura 1. Creación de una carpeta en HDFS para trasladar el fichero de datos.

El siguiente paso consiste en abrir una sesión de Pig en un terminal para generar un nuevo flujo de datos a partir del archivo trasladado anteriormente especificando tanto los nombres de las columnas como sus tipos de datos. Posteriormente, mediante la

orden *store* almacenamos el flujo de datos recientemente creado en un directorio de HDFS para verificar que la ingesta de datos ha sido correcta, tal y como se muestra en la Figura 2.



```
,,,Action,,,,Elapsed Time (sec),,
57222,53,54587,53,allow,177,94,83,2,30,1,1
56258,3389,56258,3389,allow,4768,1600,3168,19,17,10,9
6881,50321,43265,50321,allow,238,118,120,2,1199,1,1
50553,3389,50553,3389,allow,3327,1438,1889,15,17,8,7
50002,443,45848,443,allow,25358,6778,18580,31,16,13,18
51465,443,39975,443,allow,3961,1595,2366,21,16,12,9
60513,47094,45469,47094,allow,320,140,180,6,7,3,3
50049,443,21285,443,allow,7912,3269,4643,23,96,12,11
52244,58774,2211,58774,allow,70,70,0,1,5,1,0
50627,443,16215,443,allow,8256,1674,6582,31,75,15,16
43676,80,45378,80,allow,696,378,318,12,35,7,5
52190,443,16680,443,allow,7942,870,7072,22,15,10,12
50690,80,20479,80,allow,4805,3639,1166,16,31,9,7
55597,53,45448,53,allow,168,86,82,2,30,1,1
49164,443,45916,443,allow,7292,950,6342,19,75,9,10
36887,443,63451,443,allow,10922,2532,8390,27,28,13,14
1939,53,33288,53,allow,210,78,132,2,30,1,1
50281,53,33175,53,allow,195,102,93,2,30,1,1
57222,53,51448,53,allow,177,94,83,2,30,1,1
56710,53,57885,53,allow,177,94,83,2,30,1,1
48488,443,26104,443,allow,12993,3240,9753,36,24,18,18
50691,80,62082,80,allow,4237,3610,627,15,31,8,7
50693,80,54649,80,allow,4465,3390,1075,14,30,8,6
1940,53,23603,53,allow,210,78,132,2,30,1,1
50172,53,37056,53,allow,211,78,133,2,30,1,1
38802,53,38802,53,allow,216,90,126,2,30,1,1
50281,53,8038,53,allow,195,102,93,2,30,1,1
50051,443,1683,443,allow,7912,3269,4643,23,96,12,11
```

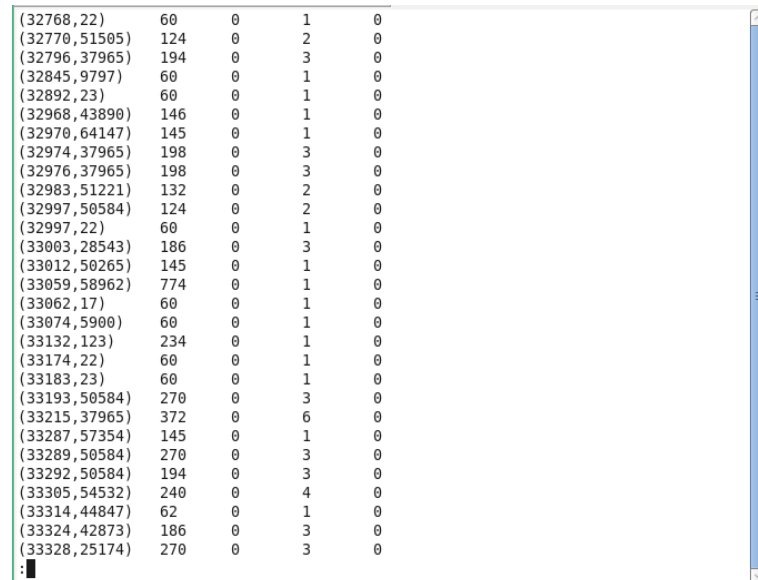
Figura 2. Información almacenada del flujo de datos a partir del fichero *log2.csv* en un directorio de HDFS.

Una vez disponemos de un flujo inicial con la información disponible, a continuación comenzamos a plantear las diferentes partes de la consulta que deseamos realizar. Su objetivo consiste en conocer las **cantidades de paquetes y bytes enviados y recibidos a través de los puertos disponibles durante aquellas transacciones que se han clasificado como *deny***, es decir, acciones categorizadas como sospechosas y que por ende un programa de seguridad no debería de permitir su continuación. De este modo se pretende analizar e identificar la existencia de posibles patrones que nos permitan conocer las características de las operaciones maliciosas para su detección temprana y actuación inmediata. Para conseguir dicho resultado se han aplicado los siguientes operadores en el mismo orden en el que se presentan:

1. Filtramos los registros mediante el campo *Action* para obtener únicamente aquellos pertenecientes a la clase *deny*.
2. A continuación agrupamos las filas resultantes por los campos *SourcePort* y *DestinationPort* para obtener las parejas de puertos de entrada y salida de cada una de las transacciones clasificadas como denegadas.
3. Posteriormente realizamos una proyección sobre los datos disponibles para mostrar únicamente las parejas de puertos resultantes del paso anterior, además

de la suma del número de bytes y paquetes tanto enviados como recibidos en sendos casos.

4. Finalmente almacenamos el flujo de datos resultante en el sistema de ficheros HDFS con el objetivo de visualizar los resultados tal y como se muestra en la Figura 3.



(32768,22)	60	0	1	0
(32770,51505)	124	0	2	0
(32796,37965)	194	0	3	0
(32845,9797)	60	0	1	0
(32892,23)	60	0	1	0
(32968,43890)	146	0	1	0
(32970,64147)	145	0	1	0
(32974,37965)	198	0	3	0
(32976,37965)	198	0	3	0
(32983,51221)	132	0	2	0
(32997,50584)	124	0	2	0
(32997,22)	60	0	1	0
(33003,28543)	186	0	3	0
(33012,50265)	145	0	1	0
(33059,58962)	774	0	1	0
(33062,17)	60	0	1	0
(33074,5900)	60	0	1	0
(33132,123)	234	0	1	0
(33174,22)	60	0	1	0
(33183,23)	60	0	1	0
(33193,50584)	270	0	3	0
(33215,37965)	372	0	6	0
(33287,57354)	145	0	1	0
(33289,50584)	270	0	3	0
(33292,50584)	194	0	3	0
(33305,54532)	240	0	4	0
(33314,44847)	62	0	1	0
(33324,42873)	186	0	3	0
(33328,25174)	270	0	3	0

Figura 3. Subconjunto de datos del flujo resultante de la consulta en Fig.

Observando el flujo de datos obtenido tras ejecutar cada uno de los pasos de la consulta anterior, existen algunos aspectos a destacar.

- **Tanto el número de bytes como de paquetes recibidos es cero** en la gran mayoría de las operaciones clasificadas como no permitidas. Este hecho nos indica que en la mayoría de transacciones no se reciben datos desde una fuente externa, sino que es el propio cliente el que envía información. Una posible teoría es que los ordenadores de la universidad sean utilizados mayormente para realizar consultas en Internet, o bien que la propia red interna disponga de una serie de restricciones que limiten la bajada de información.
- El número de bytes enviados comprende un rango de valores más amplio a diferencia del **número de paquetes enviados que se encuentra entre 1 y 3**, a excepción de algunos casos. Controlar el tráfico entrante es una de las principales características de los programas de seguridad. Por ende, si el objetivo es mantener un ataque de ciberseguridad el mayor tiempo posible, una de las principales precauciones que pueden tomar los ciberdelincuentes es mantener el envío de un número bajo de paquetes para no causar ninguna sospecha.

- Si bien la mayoría de puertos origen se encuentran por encima del 1024, es decir, no se tratan de puertos protegidos, existen **numerosas operaciones no permitidas en los puertos 80, 443, 993 y 1024**, que se encuentran orientados a establecer conexiones con Internet y servicios cifrados de correos electrónicos. En ellos se pueden observar un **número más elevado de paquetes enviados** que en los restantes puertos.