MÁSTER

INFRAESTRUCTURAS COMPUTACIONALES PARA EL PROCESAMIENTO DE DATOS MASIVOS

PRÁCTICA DEL MÓDULO 2: Procesamiento paralelo basado en memoria con Apache Spark

2021-2022

Versión 1.0

Dr. Rafael Pastor Vargas
Dr. Agustín C. Caminero Herráez — Dr. Salvador Ros Muñoz

MÁSTER UNIVERSITARIO EN INGENIERÍA Y CIENCIA DE DATOS

Contenido

Introducción

En este documento se presenta el Trabajo Práctico (TP) del módulo 2 de la asignatura "INFRAESTRUCTURAS COMPUTACIONALES PARA EL PROCESAMIENTO DE DATOS MASIVOS", del "MÁSTER UNIVERSITARIO EN INGENIERÍA Y CIENCIA DE DATOS" de la UNED.

Este trabajo se realiza de forma individual. En las siguientes secciones se exponen los diferentes ejercicios que es necesario implementar para este trabajo.

Se proponen dos ejercicios diferentes. El primer ejercicio consiste en la implementación de un análisis exploratorio de datos mediante Spark (Dataframes y Spark SQL). El segundo ejercicio consiste en la implementación de un modelo predictivo empleando las librerías de ML de Spark.

ES NECESARIO REALIZAR EL EJERCICIO 1 DE FORMA OBLIGATORIA PARA APROBAR ESTE TRABAJO. EL SEGUNDO EJERCICIO PERMITE SUBIR LA NOTA HASTA CONSEGUIR LA MÁXIMA PUNTUACIÓN

El ejercicio 1 se valorará con un máximo de 8 puntos sobre 10, mientras que el ejercicio 2 valorará con un máximo de 2 puntos sobre 10.

La forma de evaluar el trabajo se hará en base a lo siguiente:

- Jupyter notebooks con el código realizado, listo para ser ejecutado en el entorno de desarrollo. Se debe incluir no solamente el código sino también las explicaciones necesarias, imágenes, … de forma que el notebook sea autocontenido. Al principio del notebook se debe indicar el nombre del/la estudiante.
- De forma optativa, se puede incluir una memoria explicativa. Esta memoria podrá incrementar la nota completa de cada ejercicio si aporta contenido significativo adicional (referencias, documentos adicionales, reflexiones, etc.)

Se valorará positivamente que el notebook o la memoria contenga un apartado final donde se explique la opinión del/la estudiante sobre este trabajo, puntos fuertes y/o débiles, recomendaciones para el futuro, así como una valoración general de este módulo.

Este material se deberá incluir en un fichero comprimido y enviado a través del curso virtual dentro de los plazos establecidos para su entrega. El nombre de dicho fichero comprimido deberá tener la estructura *TP2-ApellidosNombre.zip*, donde Apellidos y *Nombre* deben sustituirse por los valores correspondientes para el/la estudiante que realiza el envío (evitar el uso de acentos o símbolos).

A continuación, se detallan los ejercicios a completar.

Ejercicio 1: Análisis exploratorio con Spark.

Este ejercicio define uno de los escenarios habituales de uso de Spark en cuanto a realizar análisis exploratorio de los datos (Big Data) asociados a un dominio de aplicación. Se proporciona en el curso virtual una plantilla de ejecución de un notebook en la cual se debe rellenar las celdas correspondientes,

Para este caso concreto, se va a analizar un Dataset que está disponible en la web de repositorios de datos del IEEE: https://ieee-dataport.org/

En concreto, los datos están asociados a la identificación y caracterización de diferentes tipos de ciberataques que se denominan de denegación de servicio o DoS (Denial of Service). Tanto el Dataset, como la descripción de los campos del Dataset están disponibles en la siguiente dirección (aunque también se han dejado en el curso virtual):

https://ieee-dataport.org/documents/smart-defender-dataset

Si se desean más detalles sobre cómo se usa el Dataset desde el punto de vista de investigación, se puede acceder al siguiente enlace con la publicación asociada a dicho Dataset:

https://www.hindawi.com/journals/scn/2019/1574749/

Es un ejercicio guiado por lo que se debe usar el espacio de cada celda para dar la solución concreta a cada celda/apartado/cuestión. Sin embargo, se pueden introducir modificaciones si se estima necesario (añadir celdas adicionales, etc.). Hay una parte inicial de preparación y conexión del entorno y de los datos que es obligatorio ejecutar (concretamente la primera celda):

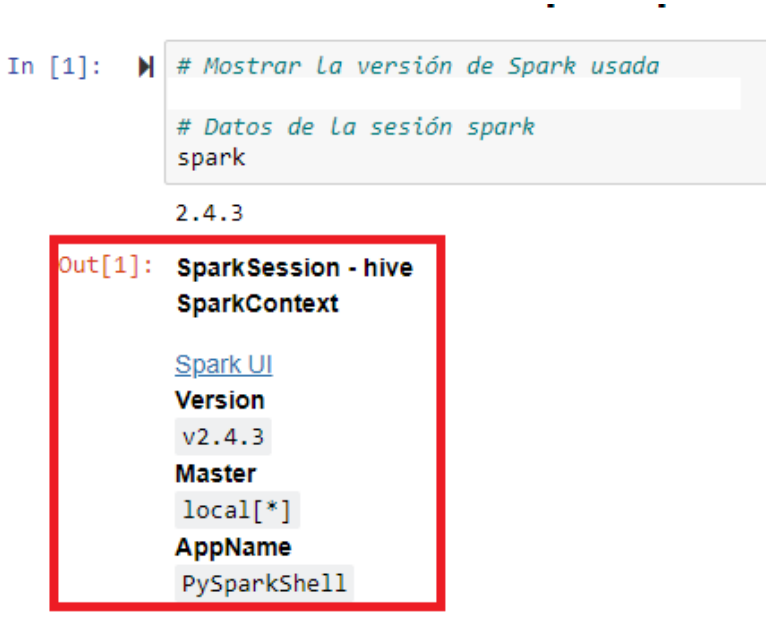


Figura 1. Entorno de ejecución.

El notebook consta de dos partes centradas en las operaciones con Dataframes y el uso de Spark SQL. En la primera parte se deben responder a una serie de cuestiones usando las operaciones/funciones/métodos del objeto Dataframe.

Figura 2. Parte 1 del ejercicio.

En la segunda parte se deben contestar a las mismas cuestiones, pero usando sentencias SQL para implementar la solución a la pregunta/cuestión.

La valoración de cada parte se hará teniendo en cuenta que la solución correcta de cada cuestión se puntuará con 0.35 puntos (son en total 16 cuestiones, por lo que es un total de 5,6 puntos) y que para el resto de puntuación (hasta 8 puntos) se valorará:

- Elegancia de la solución: uso del encadenamiento, simplicidad de las expresiones y verbosidad (traducción del término "verbose"), hasta 1.4 puntos.
- Agregación de elementos visuales que permitan una mejor interpretación/visualización de la exploración de datos, hasta 1 punto.

En la tabla 1 se muestran las soluciones que se deberían obtener al contestar a las cuestiones planteadas en el notebook (en ambas partes).

Q1

```
+-----------------+
|           Label1|
+-----------------+
|       http_flood|
|   http_slow_body|
|    tcp_fin_flood|
|  http_slow_range|
|           normal|
|        udp_flood|
|   http_slow_read|
|tcp_syn_ack_flood|
|    tcp_syn_flood|
|    tcp_ack_flood|
|http_slow_headers|
+-----------------+
```

Q2

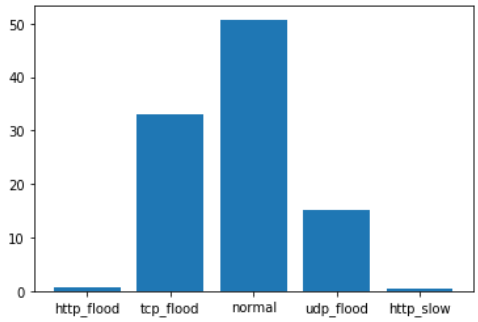
```
+----------+
|    Label3|
+----------+
|http_flood|
| tcp_flood|
|    normal|
| udp_flood|
| http_slow|
+----------+
```

Q3

49.26 %

Q4

	traffic	percentage
0	http_flood	0.762637
1	tcp_flood	32.9407
2	normal	50.7429
3	udp_flood	15.1516
4	http_slow	0.402198



Q5

```
+------------+
|ip_proto_int|
+------------+
|        17.0|
|         6.0|
+------------+
```

Q6

```
+------------+-----------------+
|ip_proto_int| sum(ip_len_mean)|
+------------+-----------------+
|        17.0|7064.346040678179|
|         6.0| 26867.1914740567|
+------------+-----------------+
```

Q7

```
+---------+------------------+
|   Label1|  sum(ip_len_mean)|
+---------+------------------+
|udp_flood|            6894.0|
|   normal|170.34604067817844|
+---------+------------------+
```

Q8	+----------------+------------------+ | Label1| sum(ip_len_mean)| +----------------+------------------+ | normal|11615.053998509025| | tcp_syn_flood| 5713.195544936026| |tcp_syn_ack_flood| 5273.157952268173| | tcp_fin_flood| 2000.0| | tcp_ack_flood| 2000.0| | http_flood| 90.88054677859463| | http_slow_range| 44.20993791090954| |http_slow_headers|44.157370741482964| | http_slow_read|43.283366733466934| | http_slow_body|43.252756179024715| +----------------+------------------+

Tabla 1. Soluciones al ejercicio de la primera parte.

Ejercicio 2: Spark y Machine Learning

En este ejercicio no se va a proporcionar ningún notebook guiado, sino que se propone al estudiante la realización de un ejercicio libre en el que se realice el entrenamiento de alguno de los modelos de Machine Learning a los que se da soporte en Spark:

https://spark.apache.org/docs/latest/ml-guide.html

En concreto se recomienda usar alguno de los que están en la categoría de Regresión y Clasificación:

https://spark.apache.org/docs/latest/ml-classification-regression.html

o en la categoría de Clustering:

https://spark.apache.org/docs/latest/ml-clustering.html

Se pueden usar ejemplos ya desarrollados, pero siempre se deberán referenciar en la entrega del notebook correspondiente. En caso de no hacerlo y comprobarse que existe ese ejemplo funcional en Internet, automáticamente se suspende esta parte y también la práctica completa. Si se usan ejemplos ya desarrollados se deben introducir mejoras en el entrenamiento del modelo, bien sea en las métricas de validación o en el propio modelo (por ejemplo, usar Gaussian Mixture Model en vez de Kmeans y comparar las dos aproximaciones, si es posible)

A modo de ejemplo, se pueden usar para la parte de Clustering estas referencias:

https://medium.com/rahasak/k-means-clustering-with-apache-spark-cab44aef0a16

https://rsandstroem.github.io/sparkkmeans.html

https://github.com/xsankar/global-bd-conf (esta referencia es del autor del libro "Fast data processing with Spark 2" y vienen ejemplos de varios algoritmos)

Para la valoración de esta parte, se usarán los siguientes criterios

- Notebook funcional, es decir, ejecutado con el entrenamiento del modelo y las predicciones correspondientes para el caso de los datos de test (hasta 0,5 puntos).
- Originalidad de la solución, en cuanto a algoritmo empleado y dominio de aplicación: ciberseguridad, retail, ehealth, finances, etc. (hasta 0,5 puntos).
- Elegancia de la solución: uso del encadenamiento, simplicidad de las expresiones y verbosidad (hasta 0,5 puntos).
- Extensión y agregación de elementos visuales que permitan una mejor interpretación/visualización de la exploración de datos (hasta 0,25 puntos). Aquí, en el caso de usar un ejemplo ya desarrollado, se valorará específicamente las extensiones para mejorar la comprensión del modelo desarrollado.
- Aplicabilidad real en el dominio de actuación de los datos, evitando pruebas de concepto con aplicabilidad no directa o evidente (hasta 0,25 puntos).

Como sugerencia, se recomienda buscar un ejemplo del ámbito de actuación del entorno profesional del estudiante. Esto facilitará el desarrollo y comprensión del modelo predictivo.