MÁSTER

INFRAESTRUCTURAS COMPUTACIONALES PARA EL PROCESAMIENTO DE DATOS MASIVOS

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

2025-2026

Versión 2.0

Dr. Rafael Pastor Vargas — Dr. Agustín C. Caminero Herráez

MÁSTER UNIVERSITARIO EN INGENIERÍA Y CIENCIA DE DATOS

Contenido

Introducción 2

Ejercicio 1: Track Big Data con PySpark en Datacamp.....3

Ejercicio 2: Spark y Machine Learning 5

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 en este trabajo.

Se proponen dos ejercicios diferentes. El primer ejercicio consiste en completar el track de Datacamp sobre Big Data con PySpark. El segundo ejercicio consiste en implementar un modelo predictivo utilizando las librerías de ML de Spark, siguiendo lo aprendido en el ejercicio 1.

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 5 puntos sobre 10, mientras que el ejercicio 2 se valorará con un máximo de 5 puntos sobre 10.

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

- Jupyter notebooks con el código realizado en el ejercicio 2, listo para ser ejecutado en el entorno de desarrollo. Se debe incluir no solo el código, sino también las explicaciones necesarias, imágenes, etc., de forma que el notebook sea autocontenido. Al principio del notebook se debe indicar el nombre del/la estudiante.
- Memoria explicativa del ejercicio 2. La composición de esta memoria se describe en el segundo apartado.

Se valorará positivamente que la memoria incluya un apartado final en el que se explique la opinión del/la estudiante sobre este trabajo (los dos ejericicios), los puntos fuertes y/o débiles, las recomendaciones para el futuro, así como una valoración general de este módulo.

Este material se deberá incluir en un fichero comprimido y enviar 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.

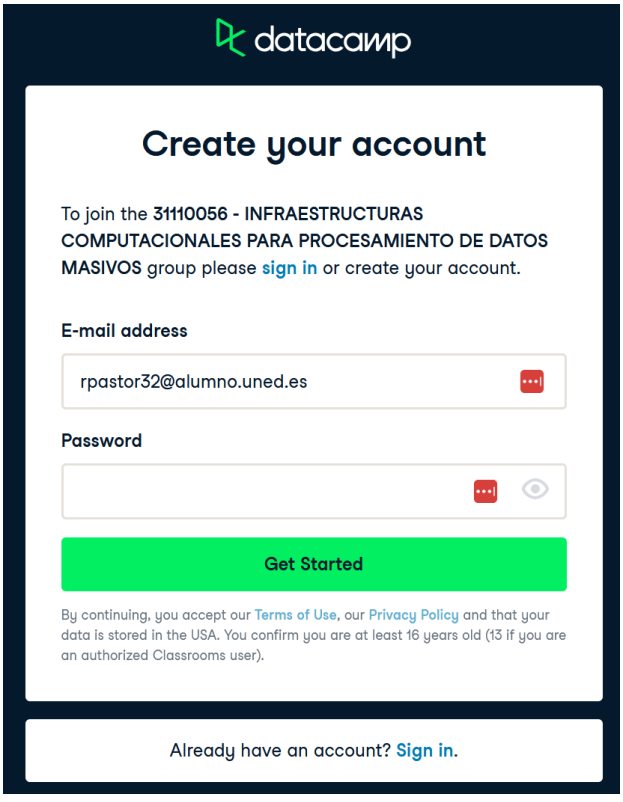
Ejercicio 1: Track Big Data con PySpark en Datacamp.

Este ejercicio se centra en el desarrollo de seis cursos específicos sobre el desarrollo de habilidades de Spark en la plataforma de aprendizaje DataCamp. El track incluye un proyecto adicional que aborda un ejemplo real. Este proyecto es optativo, pero es muy recomendable que se haga porque el segundo ejercicio se basa en un proyecto similar, pero definido por el estudiante en cualquiera de las posibilidades de algoritmos soportados por Spark.

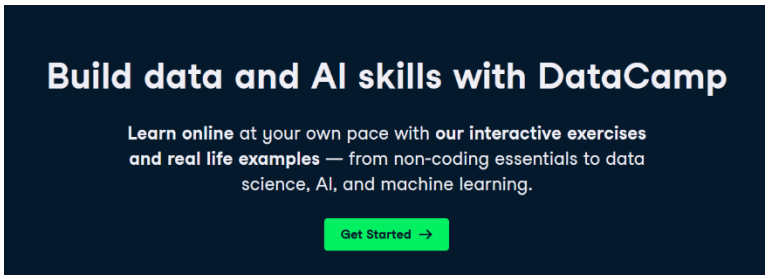
Se debe usar el usuario UNED (correo electrónico) para el alta en el grupo de trabajo de DataCamp. Para ello, se usará el siguiente enlace de invitación:

https://www.datacamp.com/groups/shared_links/bcfe4619f8a1391b83650905f476b624a3f3227772321c2bf0dd66729a1ea36d

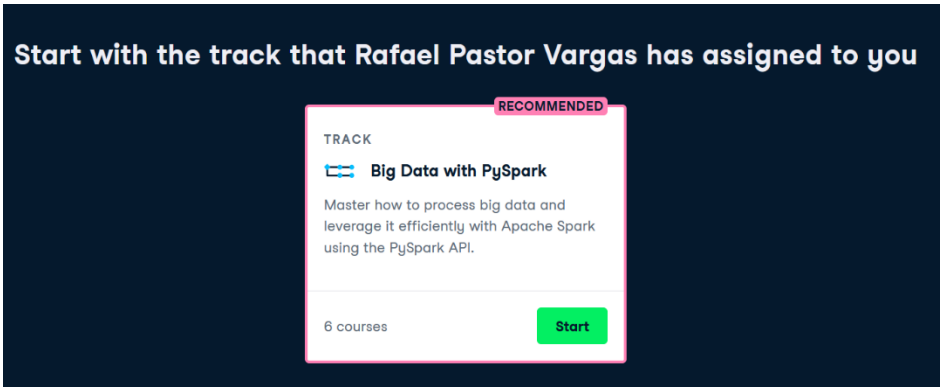
Una vez que se acceda a este enlace, se debe crear el usuario en el entorno con la cuenta UNED, asignar una contraseña al nuevo usuario y asignarlo al grupo de trabajo en DataCamp. Ese grupo de trabajo tiene una tarea (assignment) que contiene el trabajo a desarrollar. El seguimiento del desarrollo de los cursos es automático, por lo que el estudiante no debe capturar ninguna pantalla en ningún caso.



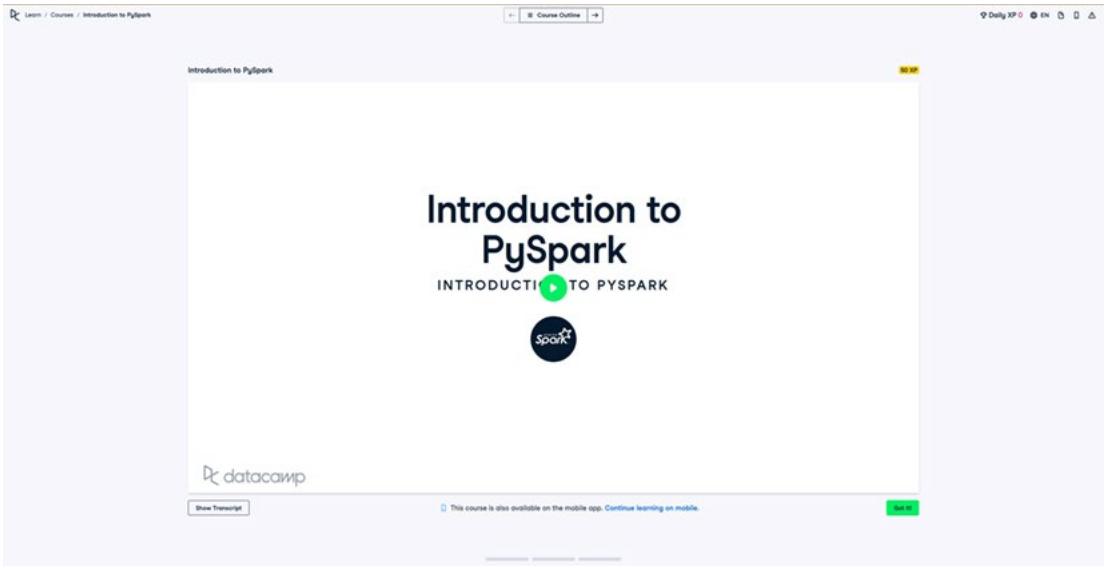
Una vez que se define la contraseña y se pulsa el botó, la siguiente pantalla muestra un aviso general sobre DataCamp. Se debe continuar pulsando el botón Get Started.



La siguiente pantalla muestra la asignación de la tarea realizada.

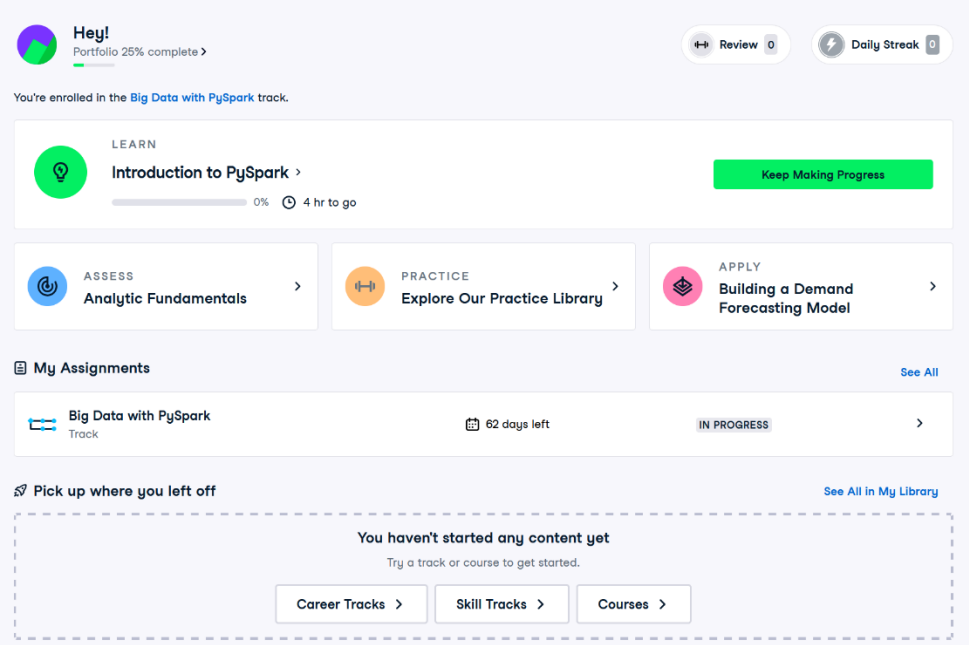


Una vez que se pulse el botón, se mostrará automáticamente el primer curso del track que integra la tarea a desarrollar. Los cursos se gestionan mediante una interfaz simple y la tarea consiste en completarlos todos.

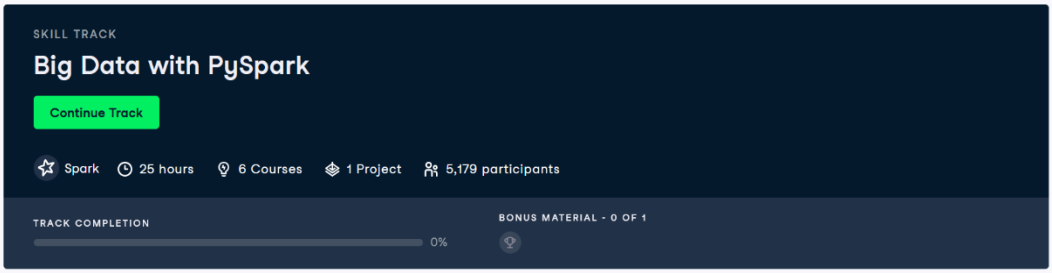


Se recomienda usar la opción Learn, situada en la parte superior izquierda de la pantalla, antes de comenzar el curso. Esta opción lleva al dashboard de aprendizaje de DataCamp y permite visualizar tanto el aprendizaje autónomo (os

permite realizar otros cursos/proyectos) así como las tareas asociadas al estudiante para esta asignatura.



Para empezar o volver a retomar la tarea, solo se debe pulsar en mis asignaciones (My Assignments), el nombre del track y se accederá al último punto del curso que se haya desarrollado.



En este ejercicio no hay que entregar nada más que incorporar en la memoria del segundo ejercicio comentarios e impresiones sobre el uso de la plataforma DataCamp.

Ejercicio 2: Spark y Machine Learning

En este ejercicio se propone al estudiante la realización de un ejercicio libre (proyecto) 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 por su nivel de dificultad básico/medio:

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

En todo caso, se valorará el nivel de complejidad del proyecto en la nota de este ejercicio, especialmente si se desarrollan proyectos complejos.

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 un ejemplo funcional en Internet, se suspende automáticamente 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 1,5 puntos).
- Originalidad de la solución, en cuanto a algoritmo empleado y dominio de aplicación: ciberseguridad, retail, ehealth, finances, etc. (hasta 2 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 0,5 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 y complejidad, evitando pruebas de concepto con aplicabilidad no directa o evidente (hasta 1,0 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.