Introducción a Apache Spark

para empezar a programar el *big data*

Mario Macías
Mauro Gómez
Rubén Tous
Jordi Torres

EDITORIAL UOC

Director de la colección:

Diseño de la colección: Editorial UOC

Primera edición en lengua castellana: Noviembre 2015

Rambla del Poblenou 156, 08018 Barcelona
http://www.editorialuoc.com

Realización editorial: Oberta UOC Publishing, SL
Maquetación: Maria García
Impresión: Service Point FMI, S.A

ISBN: 978-84-9116-037-3
Depósito legal: B 25326-2015

Mario Macías
Doctor en Arquitectura de Computadores por la Universidad Politécnica de Cataluña, donde compagina su trabajo de profesor con el de investigador en el Barcelona Supercomputing Center-Centro Nacional de Supercomputación (BSC-CNS). Durante la última década ha trabajado en proyectos de investigación relacionados con *cloud computing* y eficiencia energética. Como actividad independiente, ha autopublicado dos libros de divulgación científica.

Mauro Gómez Parada
Graduado en Ingeniería Informática por la Universidad de Vigo. Actualmente está finalizando el máster de Ingeniería informática en la UPC (Universidad Politécnica de Barcelona) y trabaja como investigador sobre Spark con el grupo de Autonomic Systems del BSC-CNS. Como actividad independiente, ha colaborado en la creación de los Premios Galegos da Musica, nacidos en el año 2013 en Galicia.

Rubèn Tous
Doctor en Informática por la Universidad Pompeu Fabra. En la actualidad es profesor contratado doctor en el Departamento de Arquitectura de Computadores de la Universidad Politécnica de Cataluña e investigador colaborador en el Barcelona Supercomputing Center. Es experto en indexación, búsqueda y clasificación de información multimedia. Ha publicado más de cincuenta artículos de investigación en revistas y conferencias internacionales. Es coeditor de múltiples estándares de los grupos MPEG y JPEG de ISO, y ha sido codirector del Metadata Subgroup de JPEG.

Jordi Torres
Catedrático de la UPC y lidera un grupo de investigación en el BSC. Actualmente su investigación se centra en la convergencia de la computación de altas prestaciones con el *big data* y su aplicación a los retos que plantea la analítica del *big data* o la computación cognitiva. Dada su extensa carrera profesional en diferentes roles, también realiza actividades de consultoría y estrategia relacionadas con las tecnologías de próxima generación y su impacto, y actúa como experto para varias organizaciones y empresas o mentorizando a emprendedores. Una de sus pasiones es la divulgación científica, que lo ha llevado a escribir un par de libros, dar conferencias y colaborar con medios de comunicación como *La Vanguardia*. Mantiene un blog sobre tecnología en www.JordiTorres.eu.

El verdadero progreso es el que pone la tecnología al alcance de todos.

Henry Ford

Índice

Prólogo

Nos encontramos en un momento muy emocionante a la hora de trabajar en computación paralela y *big data*. El gran volumen de datos que hoy en día se genera en todos los campos de la industria y la ciencia está revolucionando la forma como interactuamos con las aplicaciones, creamos productos y estudiamos el mundo a nuestro alrededor. Al mismo tiempo, las herramientas necesarias para trabajar con estos datos se han vuelto más fáciles de usar que nunca, puesto que los desarrolladores las han hecho accesibles a más y más usuarios, requiriéndoles menos y menos esfuerzo para adoptarlas. Espero que Apache Spark termine siendo una de estas herramientas para ti, que te aporte un nuevo medio para trabajar con datos de manera fácil, potente, e incluso a veces divertida de usar.

Por ello estoy encantado de ver este primer libro sobre Spark escrito en lengua española –hasta ahora todos los libros eran en lengua inglesa–, escrito por un fantástico equipo de autores. Mario, Mauro, Rubén y Jordi son destacados miembros de la comunidad con gran experiencia en Spark y la computación paralela en general por sus investigaciones y desarrollos en Barcelona. Sin duda, han elaborado un libro completo y fácil de seguir, con muchos ejemplos, y no solo cubren los fundamentos de Spark, sino también las bibliotecas más utilizadas del ecosistema que conforma Apache Spark.

Espero que este libro sea solo una introducción a tu viaje al procesado paralelo de datos en el mundo *big data*. Las ideas aquí tratadas representan algunos de los mejores métodos ideados

para trabajar con datos hoy en día. El procesado avanzado de datos sigue siendo una de las áreas de investigación más activas dentro de las ciencias de la computación, y estoy seguro de que están por llegar muchas nuevas ideas de otros campos de la informática para abrirse paso dentro de este campo. Espero que Spark siga aportándote alguna de estas ideas y que este libro te permita empezar a aprender sobre esta nueva y emocionante área.

Matei Zaharia, CTO en Databricks y vicepresidente de Apache Spark

Prefacio

En el marco de los cambios tecnológicos que estamos viviendo en pleno siglo XXI, el crecimiento exponencial de la información disponible representará una transformación completa de la actual forma de vivir, trabajar y pensar. Y todo ello nos lleva a la siguiente gran tendencia que dominará la industria de las TIC en los próximos años, conocida como *big data*.

En este escenario, nos encontramos en un momento en el que Apache Spark aparece con un vigor sin precedentes en el universo de los programadores del *big data*, porque ya se vislumbra que va a ser una de las plataformas de gran impacto en los próximos años.

Este libro proporciona al lector una oportunidad para empezar a programar y manejar datos a través del ecosistema Apache Spark. Spark es actualmente uno de los paquetes de código abierto más importantes en el espacio del *big data* y por el que importantes empresas como IBM, SAP, Oracle o Amazon han apostado, siendo asimismo grandes contribuidoras.

Este libro de carácter introductorio, que puede utilizarse como texto de autoestudio o de soporte a cursos que requieran una introducción a Apache Spark, presenta una descripción del ecosistema de paquetes que conforman Apache Spark, así como sus características básicas, e incluye ejemplos de código en Python para que el lector pueda tener una comprensión de primera mano de algunas de las posibilidades de Apache Spark, probando estos ejemplos en su propio PC si así lo desea.

Pero no debemos olvidar que el proyecto Apache Spark está evolucionando muy rápidamente. Cuando empezamos a escribir

este libro, en junio de 2015, la versión 1.4.0 aún estaba en fase beta. En el momento de imprimirse, octubre de 2015, la versión 1.5 de Spark acaba de ver la luz.

En general, Apache Spark mantiene cuidadosamente la compatibilidad binaria y de fuente para interfaces de programación estables como las que se presentan en este libro introductorio. Por tanto, esperamos que los ejemplos básicos usados en el libro sigan funcionando en futuras versiones Apache Spark 1.x y si fuera necesario los iremos actualizando en la página web del libro.

Los autores esperamos que esta obra sirva para ayudar a que esta nueva tecnología, que sin duda va a marcar el futuro inmediato del *big data*, llegue a todos aquellos estudiantes (universitarios o de bachillerato), científicos y técnicos que quieran dar un paso más en la programación del *big data*. Y al igual que nos dice Matei en su prólogo, nosotros también deseamos que este libro sea solo el principio de un largo camino en este fascinante mundo que ofrece el análisis del *big data*.

Esperamos que el lector disfrute aprendiendo con este libro de la misma manera que nosotros, los autores, lo hemos hecho escribiéndolo.