Tecnologías de Big Data
¿Dónde aprender más?

Versión 1.0

PROFESOR/A
Luis Reina (luis_reina@es.ibm.com)

INDICE

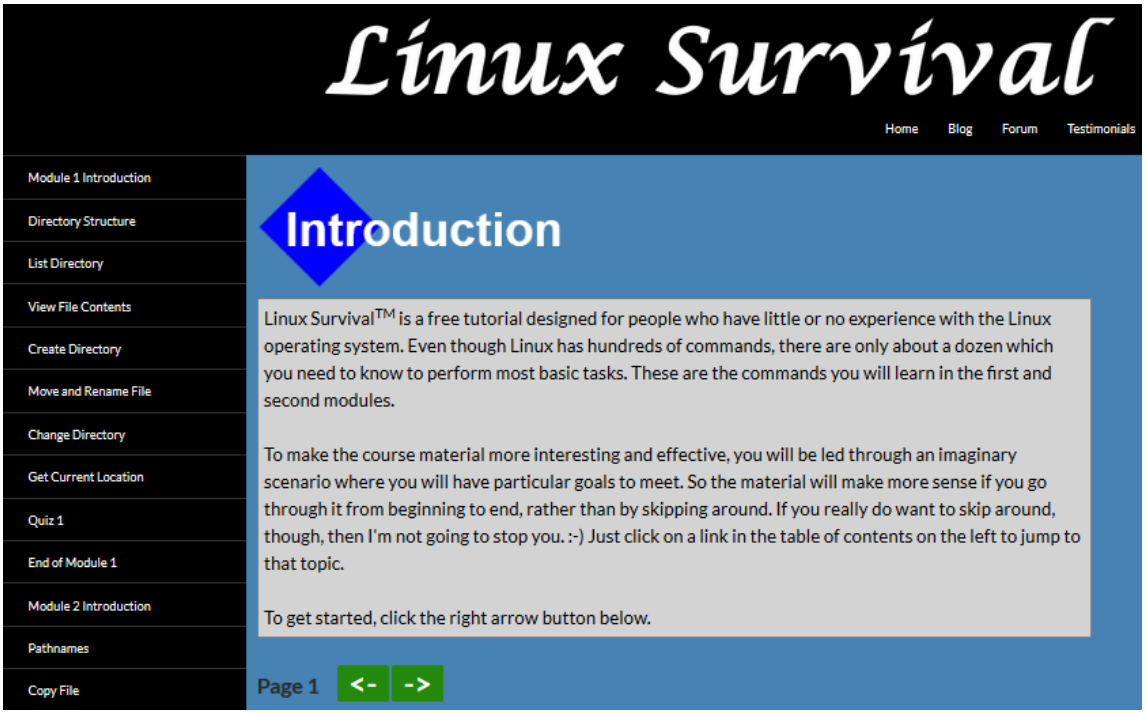
1. Prerrequisitos Importantes

1.1 Sistema Operativo Linux

Si hay que mencionar un sistema operativo cuando hablamos de Big Data la respuesta es **Linux**. Por lo tanto conocer, al menos a nivel de usuario, este sistema operativo es importante para entender bien y manejarse con las nuevas tecnologías de Big Data.

Realizad este mini curso:

http://linuxsurvival.com/linux-tutorial-introduction/



Id pulsando en las flechitas con el cuadrado **verde**.

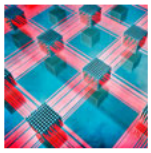
Podéis usar cualquiera de las máquinas virtuales que os he facilitado para practicar los comandos en un Linux "de verdad". Para abrir un terminal donde lancéis los comandos Linux pulsad el raton derecho en la consola y elegid "Open in Terminal".

1.2 Lenguaje SQL

El lenguaje por excelencia **de los datos** ha sido durante muchos años el SQL y seguirá siéndolo. Es un lenguaje mucho anterior al Big Data, pero muchas de las nuevas tecnologías lo incorporan (Big SQL, Impala, Spark SQL, Hive, etc.) y es muy importante conocerlo.

Este curso *online* es una buena introducción al SQL, os recomiendo que lo hagáis:

http://bigdatauniversity.com/courses/sql-course/



SQL Fundamentals

with **Yashavanthi Renukaswamy, Wagner Crivelini, Liviu Perniu, Priyanka Harish, Raul F. Chong**

Audience: Database Beginners

Time to complete: 5 hours

Available in: English

Enroll Now

About this SQL course

Most software written today relies on relational databases such as MySQL or DB2. Within this free SQL course, you'll learn the basics of the relational database model and the SQL language using DB2 Express-C, the free version of IBM DB2 database server. You will learn SQL and how to create, read, update and delete data from a database. This SQL tutorial is aimed at beginners, but it will give you enough information to get you started working with databases.

Audience: Database Beginners**Time to complete**: 5 hours**Available in**: English

Course Syllabus

- Getting started
- Relational database concepts
- Working with database objects
- Reading data
- Inserting, updating, and deleting data
- Working with multiple tables

2. Hadoop

2.1 Curso Online: "Hadoop Fundamentals I"

http://bigdatauniversity.com/courses/hadoop-course/

Hadoop Fundamentals I

with **Warren Pettit, Asma Desai**

Audience: Hadoop Beginners

Time to complete: 4 Hours

Available in: English

Enroll Now

Hadoop Fundamentals I teaches you the basics of Apache Hadoop and the concept of Big Data. This Hadoop course is entirely free, and so are the materials and software provided. This is the third version of our most popular Hadoop course. Since Version 2 was published, several more detailed courses covering topics such as MapReduce, Hive, HBase, Pig, Oozie, and Zookeeper have been added. We recommend you start here and then dig deeper into the specific Hadoop technology you wish to learn more about.

Learn Hadoop

This Hadoop course is designed to give you a basic understanding of key Big Data technologies. In this Hadoop tutorial, we first begin with describing what Big Data is and the need for Hadoop to be able to process that data in a timely manner. This is followed by describing the Hadoop architecture and how to work with the Hadoop Distributed File System (HDFS) both from the command line and using the BigInsights Console that is supplied with InfoSphere BigInsights.

This Hadoop course was recently tested and updated for BigInsights Quick Start 4.1 (IBM's edition of Hadoop).

3. Spark

3.1 Curso Online: "Spark Fundamentals I"

http://bigdatauniversity.com/courses/spark-fundamentals/

Spark Fundamentals I

with **Henry Quach, Alan Barnes**

Audience: Data scientists, engineers, or anyone who is interested in learning about Spark.

Time to complete: 03:00

Available in: English

Go to class

Apache Spark is an open source processing engine built around speed, ease of use, and analytics. If you have large amounts of data that requires low latency processing that a typical Map Reduce program cannot provide, Spark is the alternative. Spark performs at speeds up to 100 times faster than Map Reduce for iterative algorithms or interactive data mining. Spark provides in-memory cluster computing for lightning fast speed and supports Java, Scala, and Python APIs for ease of development.

Spark combines SQL, streaming and complex analytics together seamlessly in the same application to handle a wide range of data processing scenarios. Spark runs on top of Hadoop, Mesos, standalone, or in the cloud. It can access diverse data sources such as HDFS, Cassandra, HBase, or S3.

Big Data University has been chosen by IBM as one of the issuers of badges as part of the IBM Open Badge program. Share your achievements through LinkedIn, Facebook, Twitter, and more!

Big Data University leverages the services of Pearson VUE Acclaim to assist in the administration of the IBM Open Badge program. **By enrolling into this course, you agree to Big Data University sharing your details with Pearson VUE Acclaim for the strict use of issuing your badge upon completion of the badge criteria.**

4. Ecosistema de Hadoop

4.1 Curso Oline: "Moving Data into Hadoop"

http://bigdatauniversity.com/courses/moving-data-into-hadoop-cognizant/

Enseña el uso de las herramientas Sqoop y Flume.

Moving Data into Hadoop (for Cognizant)

with **Glen Mules, Warren Pettit**

Audience: Hadoop Beginners

Time to complete: 4 hours

Available in: English

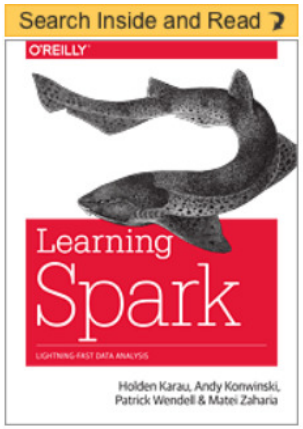
Go to class

This course describes techniques for moving data into Hadoop. There are a variety of ways to get data into Hadoop from simple Hadoop shell commands to more sophisticated processes. Several techniques are presented but two, Sqoop and Flume, are covered in greater detail.

5. Libros

5.1 Learning Spark

http://shop.oreilly.com/product/0636920028512.do



Larger Cover

Learning Spark

Lightning-Fast Big Data Analysis

By Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia

Publisher: O'Reilly Media
Final Release Date: January 2015
Pages: 276

4.0

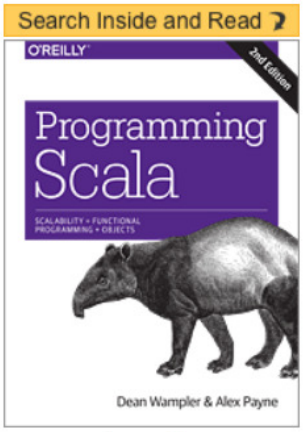
Read **26** Reviews | Write a Review

Data in all domains is getting bigger. How can you work with it efficiently? **Recently updated for Spark 1.3**, this book introduces Apache Spark, the open source cluster computing system that makes data analytics fast to write and fast to run. With Spark, you can tackle big datasets quickly through simple APIs in Python, Java,...

Full description

5.2 Programming Scala, 2nd Edition

http://shop.oreilly.com/product/0636920033073.d



Larger Cover

Programming Scala, 2nd Edition

Scalability = Functional Programming + Objects

By Dean Wampler, Alex Payne

Publisher: O'Reilly Media
Final Release Date: December 2014
Pages: 586

4.5

Read **2** Reviews | Write a Review

Get up to speed on Scala, the JVM language that offers all the benefits of a modern object model, functional programming, and an advanced type system. Packed with code examples, this comprehensive book shows you how to be productive with the language and ecosystem right away, and explains why Scala is ideal for today's highly...

Full description