



## Welcome to Data Science and Big Data Analytics.

Copyright © 1996, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012 EMC Corporation. All Rights Reserved. EMC believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

THE INFORMATION IN THIS PUBLICATION IS PROVIDED "AS IS." EMC CORPORATION MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND WITH RESPECT TO THE INFORMATION IN THIS PUBLICATION, AND SPECIFICALLY DISCLAIMS IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Use, copying, and distribution of any EMC software described in this publication requires an applicable software license.

EMC2, EMC, Data Domain, RSA, EMC Centera, EMC ControlCenter, EMC LifeLine, EMC OnCourse, EMC Proven, EMC Snap, EMC SourceOne, EMC Storage Administrator, Acartus, Access Logix, AdvantEdge, AlphaStor, ApplicationXtender, ArchiveXtender, Atmos, Authentica, Authentic Problems, Automated Resource Manager, AutoStart, AutoSwap, AVALONidm, Avamar, Captiva, Catalog Solution, C-Clip, Celerra, Celerra Replicator, Centera, CenterStage, CentraStar, ClaimPack, ClaimsEditor, CLARiiON, ClientPak, Codebook Correlation Technology, Common Information Model, Configuration Intelligence, Configuresoft, Connectrix, CopyCross, CopyPoint, Dantz, DatabaseXtender, Direct Matrix Architecture, DiskXtender, DiskXtender 2000, Document Sciences, Documentum, elnput, E-Lab, EmailXaminer, EmailXtender, Enginuity, eRoom, Event Explorer, FarPoint, FirstPass, FLARE, FormWare, Geosynchrony, Global File Virtualization, Graphic Visualization, Greenplum, HighRoad, HomeBase, InfoMover, Infoscape, Infra, InputAccel, InputAccel Express, Invista, Ionix, ISIS, Max Retriever, MediaStor, MirrorView, Navisphere, NetWorker, nLayers, OnAlert, OpenScale, PixTools, Powerlink, PowerPath, PowerSnap, QuickScan, Rainfinity, RepliCare, RepliStor, ResourcePak, Retrospect, RSA, the RSA logo, SafeLine, SAN Advisor, SAN Copy, SAN Manager, Smarts, SnapImage, SnapSure, SnapView, SRDF, StorageScope, SupportMate, SymmAPI, SymmEnabler, Symmetrix, Symmetrix DMX, Symmetrix VMAX, TimeFinder, UltraFlex, UltraPoint, UltraScale, Unisphere, VMAX, Vblock, Viewlets, Virtual Matrix, Virtual Matrix Architecture, Virtual Provisioning, VisualSAN, VisualSRM, Voyence, VPLEX, VSAM-Assist, WebXtender, xPression, xPresso, YottaYotta, the EMC logo, and where information lives, are registered trademarks or trademarks of EMC Corporation in the United States and other countries.

All other trademarks used herein are the property of their respective owners.

© Copyright 2012 EMC Corporation. All rights reserved. Published in the USA.

Revision Date: February 17, 2012

Revision Number: MR-1CP-DSBDA.1.2.1



"DATA SCIENTISTS WILL BE THE ROCK STARS  
OF THE BIG DATA ERA."

Manage the  
Big Data  
Explosion



Are you ready for Big Data?

EMC<sup>2</sup> PROVEN PROFESSIONAL

Copyright © 2012 EMC Corporation. All Rights Reserved.

Introduction and Course Agenda 2

## Introduction and Course Agenda

The following topics are covered in this module:

- Overall course goal, objectives, and high-level flow
- Intended audience and expected background
- Classroom and lab environments

EMC<sup>2</sup> PROVEN PROFESSIONAL

Copyright © 2012 EMC Corporation. All Rights Reserved.

Introduction and Course Agenda 3

This module focuses on introducing the course, its goal and objectives, its structure, and the classroom and lab environments.

It also introduces students to other students and to the instructor(s).

## Overall Course Goal

- The goal of the *Data Science And Big Data Analytics Course* is for you to be able to ***immediately participate as a Data Science team member on big data and other analytics projects***

- ▶ *Data Scientist p-o-v*
- ▶ *Open*
- ▶ *Practical*



EMC<sup>2</sup> PROVEN PROFESSIONAL

Copyright © 2012 EMC Corporation. All Rights Reserved.

Introduction and Course Agenda 4

Here is the primary goal of the course. To achieve it, the course content is focused on the point of view (p-o-v) of a Data Scientist, it teaches concepts and principles in an open, vendor-neutral manner so they can be applied in any technology environment, and it provides many hands-on labs for practical experience with coaching from the instructor(s).

## Intended Audience

- Individuals seeking to develop an understanding of Data Science from the perspective of a practicing Data Scientist:
  - ▶ **Managers of teams of business intelligence**, analytics, and big data professionals
  - ▶ Current **business and data analysts** looking to add big data analytics to their skills
  - ▶ Data and **database professionals** looking to exploit their analytic skills in a big data environment
  - ▶ **Recent college graduates** and graduate students looking to move into the world of data science and big data
  - ▶ Individuals seeking to take advantage of the EMC Proven™ Professional Data Scientist Associate (EMCDSA) certification

EMC<sup>2</sup> PROVEN PROFESSIONAL

Copyright © 2012 EMC Corporation. All Rights Reserved.

Introduction and Course Agenda 5

The audience for this course can come from a variety of backgrounds and be driven by different objectives.

## Expected Background

- Strong mathematical, quantitative capability
- Experience with statistical methods and basic proficiency with a statistical software package, such as R or RStudio, Minitab, Matlab, SAS, or SPSS
- Experience with the conditioning and management of business data including databases
- Basic programming skills, preferably including SQL



EMC<sup>2</sup> PROVEN PROFESSIONAL

Copyright © 2012 EMC Corporation. All Rights Reserved.

Introduction and Course Agenda 6

The lectures in this course assume students have a strong numerate background with some experience of statistical software packages and the conditioning of business data. The labs in this course assume some programming expertise (preferably with R or SQL).

## Course Objectives

Upon completion of this course, you should be able to:

- Immediately participate and contribute as a data science team member on big data and other analytics projects by:
  - ▶ Deploy a structured lifecycle approach to data science and big data analytics projects
  - ▶ Reframe a business challenge as an analytics challenge
  - ▶ Apply analytic techniques and tools to analyze big data, create statistical models, and identify insights that can lead to actionable results
  - ▶ Select optimal visualization techniques to clearly communicate analytic insights to business sponsors and others
  - ▶ Use tools such as R and RStudio, MapReduce/Hadoop, in-database analytics, and window and MADlib functions
- Explain how advanced analytics can be leveraged to create competitive advantage and how the data scientist role and skills differ from those of a traditional business intelligence analyst

EMC<sup>2</sup> PROVEN PROFESSIONAL

Copyright © 2012 EMC Corporation. All Rights Reserved.

Introduction and Course Agenda

7

Please review these objectives of the course.

## Please Briefly Introduce Yourself

- Name
- Company
- Work Location
- Role, and how analytics relates to it
- Analytics Expertise
- What you would like to achieve as a result of attending this course



EMC<sup>2</sup> PROVEN PROFESSIONAL

Copyright © 2012 EMC Corporation. All Rights Reserved.

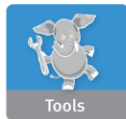
Introduction and Course Agenda 8

Please briefly introduce yourself to the other students and the instructor(s) so we can all better understand your current role, background, and motivation for attending this class.



## Course Modules and Navigation Icons

Data Science and Big Data Analytics	
1.	Introduction to Big Data Analytics
2.	Data Analytics Lifecycle + Lab
3.	Review of Basic Data Analytics Methods Using R + Labs
4.	Advanced Analytics - <i>Theory &amp; Methods</i> + Labs
5.	Advanced Analytics - <i>Technology &amp; Tools</i> + Labs
6.	The Endgame, or Putting it All Together + Final Lab



EMC<sup>2</sup> PROVEN PROFESSIONAL

Copyright © 2012 EMC Corporation. All Rights Reserved.

Introduction and Course Agenda 9

These is the progression through the training modules in this course. Most modules have one or more labs.

We will use the icons throughout the lectures and in the Student Resource Guide to assist in navigation.

## Topics : Data Science and Big Data Analytics Course

Introduction to Big Data Analytics + Data Analytics Lifecycle	Review of Basic Data Analytic Methods Using R	Advanced Analytics – Theory and Methods	Advanced Analytics - Technology and Tools	The Endgame, or Putting it All Together + Final Lab on Big Data Analytics
Big Data Overview	Using R to Look at Data - Introduction to R	K-means Clustering	Analytics for Unstructured Data (MapReduce and Hadoop)	Operationalizing an Analytics Project
State of the Practice in Analytics	Analyzing and Exploring the Data	Association Rules	The Hadoop Ecosystem	Creating the Final Deliverables
The Data Scientist	Statistics for Model Building and Evaluation	Linear Regression	In-database Analytics – SQL Essentials	Data Visualization Techniques
Big Data Analytics in Industry Verticals		Logistic Regression	Advanced SQL and MADlib for In-database Analytics	+ Final Lab – Application of the Data Analytics Lifecycle to a Big Data Analytics Challenge
Data Analytics Lifecycle		Naive Bayesian Classifier		
		Decision Trees		
		Time Series Analysis		
		Text Analysis		

EMC<sup>2</sup> PROVEN PROFESSIONAL

Copyright © 2012 EMC Corporation. All Rights Reserved.

Introduction and Course Agenda 10

This slide outlines the modules of the course (at the top) and the individual lesson topics that comprise those modules (below).

The flow of the course is from left to right on the diagram.

Most modules have one or more labs associated with them for students to gain practical, hands-on experience, individually or in pairs, depending on the needs and constraints of each specific class.

## The Classroom Environment

- Locations
  - ▶ Restrooms
  - ▶ Cafeteria – coffee
  - ▶ Network / Phone Access
  - ▶ Smoking Area
  - ▶ First Aid
  - ▶ Water cooler - coffee
- Hours of Class
  - ▶ 8:30am – 5pm each days
  - ▶ Lunch typically 12:00
  - ▶ Approximately 60% lecture, 40% lab



EMC<sup>2</sup> PROVEN PROFESSIONAL

Copyright © 2012 EMC Corporation. All Rights Reserved.

Introduction and Course Agenda 11

The instructor will communicate the important facilities in the training center.

This course is intensive in lectures (50%) and labs (50%) and requires an 8:30 start each day.

## The Lab Environment

- Hardware:
  - ▶ VMWare Servers
  - ▶ Individual Virtual Machines
- Software – Open Source:
  - ▶ Data stored in Greenplum Community Edition Database (GPDB)
  - ▶ Access from desktop browsers
    - ▶ Microsoft & Apple Mac
  - ▶ Analytics via:
    - ▶ RStudio
    - ▶ PSQL interface for GPDB
    - ▶ Hadoop
    - ▶ MADlib



EMC<sup>2</sup> PROVEN PROFESSIONAL

Copyright © 2012 EMC Corporation. All Rights Reserved.

Introduction and Course Agenda 12

This slide describes the technical environment that supports execution of the labs in the course.

## Course Materials

- Student Reference Guide:
  - ▶ Lecture slides
  - ▶ Appendix:
    - ▶▶ References
    - ▶▶ Quick reference guides
      - LINUX
      - PSQL
      - R
- Student Lab Guide:
  - ▶ Lab instructions



EMC<sup>2</sup> PROVEN PROFESSIONAL

Copyright © 2012 EMC Corporation. All Rights Reserved.

Introduction and Course Agenda 13

Course materials include a hardcopy Student Reference Guide with References in an Appendix.

There is a separate Student Lab Guide with the detailed instructions for each lab in the course.

Softcopy of “codes” are available at the Students home directories (Locations detailed in the Lab guide) to enable copy/paste (rather than keying) of commands required for the labs.

## Classroom Etiquette

- **Limit usage of personal electronic devices**
  - ▶ Cell phones/PDAs (set to vibrate if possible)
  - ▶ Laptops (preferably closed during lecture)
  - ▶ If your phone rings, answer it as you step out of the classroom
- **Food and drink are allowed** in classroom
- **Inform the instructor** (and lab partner, if you have one) **of all absences** from classroom sessions
  - ▶ Excessive absences will be interpreted as non-attendance at the class
- Although we encourage collaboration during the class, ***please treat the data files, code and lab as intellectual property of EMC Education Services.***
  - ▶ Please do not redistribute without the consent of EMC Education Services



EMC<sup>2</sup> PROVEN PROFESSIONAL

Copyright © 2012 EMC Corporation. All Rights Reserved.

Introduction and Course Agenda 14

Please observe these simple rules to ensure you and your co-students gain the maximum learning from this course.