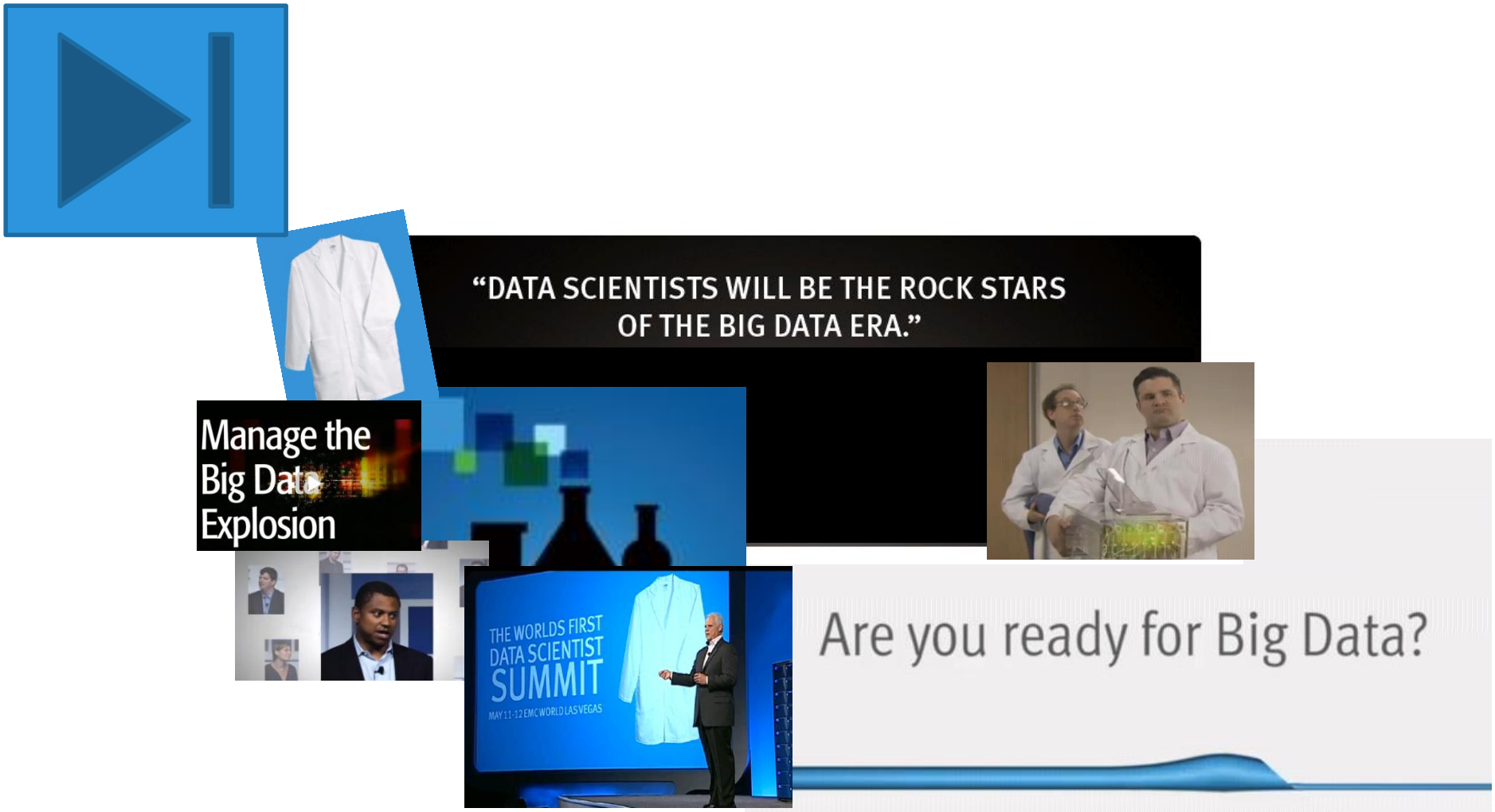
Data Science And Big Data Analytics Course

EMC^2 PROVEN PROFESSIONAL



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

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*

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

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

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:
  - Deploying a structured lifecycle approach to data science and big data analytics projects
  - Reframing a business challenge as an analytics challenge
  - Applying analytic techniques and tools to analyze big data, create statistical models, and identify insights that can lead to actionable results
  - Selecting optimal visualization techniques to clearly communicate analytic insights to business sponsors and others
  - Using 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

Please Briefly Introduce Yourself

- introduce yourself providing some background on your knowledge and experience in the following areas:

- statistics - how comfortable do you feel with probability and statistics, specifically Bayes Rule; probability distributions; hypothesis tests and linear regression? If not comfortable, how would you bring yourself up to speed in those areas so you can hit the ground running in this program? What resources would be helpful for you to get there?

- data management - how comfortable do you feel with databases and writing simple SQL queries? If not comfortable, how would you bring yourself up to speed in those areas so you can hit the ground running in this program? What resources would be helpful for you to get there?

- programming - how comfortable are you with writing code or scripts (using tools such as Python, Java, Perl, R, or even VBA within an excel spreadsheet?) If not comfortable, how would you bring yourself up to speed in those areas so you can hit the ground running in this program? What resources would be helpful for you to get there?

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 - *Theory & Methods* + Labs
5.	Advanced Analytics - *Technology & Tools* + Labs
6.	The Endgame, or Putting it All Together + Final Lab

Topics : Data Science and Big Data Analytics

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

The Classroom Environment

- Navigation
- Due dates
- How to contact your instructor

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

Course Materials

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

Classroom Etiquette

- Although we encourage collaboration during the class, ***please treat the data files, code and lab as intellectual property of EMC Education Services and SNHU.***
  - Please do not redistribute without the consent of EMC Education Services OR SNHU

Lab Exercise 1: Introduction to Data Environment

This first lab introduces the Analytics Lab Environment you will be working on throughout the course.

After completing the tasks in this lab you should be able to:

- Authenticate and access the Virtual Machine (VM) assigned to you for all of your lab exercises
- Locate data sets you will be working with for the course's labs
- Use meta commands and PSQL to navigate through the data sets
- Create sub-sets of the big data, using table joins and filters to analyze subsequent lab exercises