Introduction to Big Data

By University of California San Diego

About this Course

Interested in increasing your knowledge of the Big Data landscape? This course is for those new to data science and interested in understanding why Big Data Ecosystems has come to be. It is for those who want to become conversant with the terminology and the core concepts behind big data problems, applications, and systems. It is for those who want to start thinking about how Big Data might be useful in their business or career. It provides an introduction to one of the most common frameworks, Hadoop, that has made big data analysis easier and more accessible – increasing the potential for data to transform our world!

At the end of this course, you will be able to:

* Describe the Big Data landscape including examples of real world big data problems including the three key sources of Big Data: people, organizations, and sensors.

* Explain the V's of Big Data (volume, velocity, variety, veracity, valence, and value) and why each impacts data collection, monitoring, storage, analysis and reporting.

* Get value out of Big Data by using a 5-step process to structure your analysis.

* Identify what are and what are not big data problems and be able to recast big data problems as data science questions.

* Provide an explanation of the architectural components and programming models used for scalable big data analysis.

* Summarize the features and value of core Hadoop stack components including the YARN resource and job management system, the HDFS file system and the MapReduce programming model.

* Install and run a program using Hadoop!

This course is for those new to data science. No prior programming experience is needed, although the ability to install applications and utilize a virtual machine is recommended to complete the hands-on assignments.

Hardware Requirements:
2G Quad Core Processor (P7 or AMD V support recommended), 64 bit, (8) 8 GB RAM, (2) 20 GB disk free. How to find your hardware information: (Windows): Open System by clicking the Start button, right-clicking Computer, and then clicking Properties; (Mac): Open Overview by clicking on the Apple menu and clicking "About This Mac." Most computers with 8 GB RAM purchased in the last 3 years will meet the minimum requirements.You will need a high speed internet connection because you will be downloading files up to 4 Gb in size.

Software Requirements:
This course relies on several open source software tools, including Apache Hadoop. All required software can be downloaded and installed free of charge. Software requirements include: Windows 7+, Mac OS X 10.10+, Ubuntu 14.04+ or CentOS 6+ Virtualbox 5+.

Show less

Taught by:
Ilkay Altintas, Chief Data Science Officer
San Diego Supercomputer Center

Taught by:
Amarnath Gupta, Director, Advanced Query Processing Lab
San Diego Supercomputer Center (SDSC)

Basic Info	Course 1 of 6 in the Big Data Specialization
Commitment	3 weeks of study, 5-6 hours/week
Language	English, **Subtitles:** Arabic, French, Bengali, Ukrainian, Chinese (Simplified), Greek, Italian, Portuguese (Brazil), Vietnamese, Dutch, Korean, Onya, German, Pashto, Urdu, Russian, Thai, Indonesian, Swedish, Turkish, Azerbaijani, Spanish, Dari, Hindi, Japanese, Kazakh, Persian, Hungarian, Polish
How To Pass	Pass all graded assignments to complete the course.
User Ratings	★★★★½ Average User Rating 4.6

Syllabus

Module 1

Welcome

Welcome to the Big Data Specialization! We're excited for you to get to know us and we're looking forward to learning about you!

2 videos, 2 readings

1. **Video:** Welcome to the Big Data Specialization
2. **Reading:** By the end of this course you will be able to...
3. **Reading:** Optional: Watch this fun video about the San Diego Supercomputer Center!
4. **Video:** Tell us about yourself and learn about your classmates
5. **Discussion Prompt:** Let's Discuss: Why are you taking this class?

Show less

Module 2

Big Data: Why and Where

Data - it's been around (even digitally) for a while. What makes data "big" and where does this big data come from?

13 videos, 13 readings

1. **Video:** What launched the Big Data era?
2. **Video:** Applications: What makes big data valuable
3. **Discussion Prompt:** Let's Discuss: What application area interests you?
4. **Video:** Example: Saving lives with Big Data
5. **Video:** Example: Using Big Data to Help Patients
6. **Video:** A Sentiment Analysis Success Story: Meltwater helping Danone
7. **Reading:** Did you know?: 25 facts about big data
8. **Reading:** Slides: What Launched the Big Data Era?
9. **Reading:** Slides Applications: What Makes Big Data Valuable?
10. **Reading:** Slides: Saving Lives With Big Data
11. **Reading:** Slides: Using Big Data to Help Patients
12. **Video:** Getting Started: Where Does Big Data Come From?
13. **Video:** Machine-Generated Data: It's Everywhere and There's a Lot!
14. **Video:** Machine-Generated Data: Advantages
15. **Video:** Big Data Generated By People: The Unstructured Challenge
16. **Video:** Big Data Generated By People: How Is It Being Used?
17. **Video:** Organization-Generated Data: Structured but often siloed
18. **Video:** Organization-Generated Data: Benefits Come From Combining With Other Data Types
19. **Video:** The Key: Integrating Diverse Data
20. **Discussion Prompt:** Let's discuss: Who are you providing data to?
21. **Reading:** Extra Resources
22. **Reading:** Slides: Machine-Generated Data: It's Everywhere and There's a Lot!
23. **Reading:** Slides: Machine-Generated Data: Advantages
24. **Reading:** Slides Big Data Generated By People: The Unstructured Challenge
25. **Reading:** Slides: Big Data Generated By People: How Is It Being Used?
26. **Reading:** Slides: Organization-Generated Big Data: Structured But Often Siloed
27. **Reading:** Slides: Organization-Generated Big Data: Benefits
28. **Reading:** Slides: The Key - Integrating Diverse Data

Show less

Graded: Why Big Data and Where Did it Come From?

Module 3

Characteristics of Big Data and Dimensions of Scalability

You may have heard of the "Big V's". We'll give examples and descriptions of the commonly discussed 5. But, we want to propose a 6th V and we'll ask you to practice writing big data questions targeting this V – value.

7 videos, 6 readings

1. **Video:** Getting Started: Characteristics of Big Data
2. **Video:** Characteristics of Big Data - Volume
3. **Reading:** What does astronomical scale mean?
4. **Video:** Characteristics of Big Data - Variety
5. **Video:** Characteristics of Big Data - Velocity
6. **Video:** Characteristics of Big Data - Veracity
7. **Video:** Characteristics of Big Data - Valence
8. **Video:** The Sixth V: Value
9. **Reading:** A Small Definition of Big Data
10. **Discussion Prompt:** Practice: Writing Big Data questions
11. **Discussion Prompt:** Let's Discuss: Improving the Flamingo Game
12. **Reading:** Slides: Getting Started - Characteristics of Big Data
13. **Reading:** Slides Characteristics of Big Data - Volume
14. **Reading:** Slides: Characteristics of Big Data - Variety
15. **Reading:** Slides: Characteristics of Big Data - Velocity
16. **Reading:** Slides: Characteristics of Big Data - Veracity
17. **Reading:** Slides: Characteristics of Big Data - Value
18. **Reading:** Slides: Characteristics of Big Data - Valence

Show less

Graded: V for the V's of Big Data

Module 4

Data Science: Getting Value out of Big Data

We love science and we love computing, don't get us wrong. But the reality is we care about Big Data because it can bring value to our companies, our lives, and the world. In this module we'll introduce a 5 step process for approaching data science problems.

11 videos, 12 readings

1. **Video:** Data Science: Getting Value out of Big Data
2. **Video:** Building a Big Data Strategy
3. **Video:** How does big data science happen?: Five Components of Data Science
4. **Reading:** Five Ps of Data Science
5. **Discussion Prompt:** Let's Discuss: Thinking more deeply about the Ps
6. **Video:** Asking the Right Questions
7. **Video:** Steps in the Data Science Process
8. **Video:** Step 1: Acquiring Data
9. **Video:** Step 2-A: Exploring Data
10. **Video:** Step 2-B: Pre-Processing Data
11. **Video:** Step 3: Analyzing Data
12. **Video:** Step 4: Communicating Results
13. **Video:** Step 5: Turning Insights into Action
14. **Discussion Prompt:** Let's Discuss: Building a Team
15. **Reading:** Slides: Getting Value Out of Big Data
16. **Reading:** Slides: Building a Big Data Strategy
17. **Reading:** Slides: The Five Ps of Data Science
18. **Reading:** Slides: Asking the Right Questions
19. **Reading:** Slides: Steps in the Data Science Process
20. **Reading:** Slides: Step 1 - Acquiring Data
21. **Reading:** Slides: Step 2A: Exploring Data
22. **Reading:** Slides: Step 2B: Pre-Processing Data
23. **Reading:** Slides: Step 3 Data Analysis
24. **Reading:** Slides: Step 4 Communicating Results
25. **Reading:** Slides: Step 5 Turning Insights Into Action

Show less

Graded: Data Science 101

Module 5

Foundations for Big Data Systems and Programming

Big Data requires new programming frameworks and systems. For this course, we don't programming knowledge or experience - but we do want to give you a grounding in some of the key concepts.

6 videos, 4 readings

1. **Video:** Getting Started: Why worry about foundations?
2. **Video:** What is a Distributed File System?
3. **Video:** Scalable Computing over the Internet
4. **Video:** Programming Models for Big Data
5. **Reading:** Slides: Getting Started-Why Worry About Foundations?
6. **Reading:** Slides: What is a Distributed File System?
7. **Reading:** Slides: Scalable Computing Over the Internet
8. **Reading:** Slides: Programming Models for Big Data

Show less

Graded: Foundations for Big Data

Module 6

Systems: Getting Started with Hadoop

Let's look at some details of Hadoop and MapReduce. Then we'll go "hands on" and actually perform a simple MapReduce task using a Docker container. Pay attention - as we'll guide you in "learning by doing" in diagramming a MapReduce task as a Peer Review.

11 videos, 6 readings

1. **Video:** Hadoop: Why, Where and Who?
2. **Video:** The Hadoop Ecosystem: Welcome to the zoo!
3. **Video:** The Hadoop Distributed File System: A Storage System for Big Data
4. **Video:** YARN: A Resource Manager for Hadoop
5. **Video:** MapReduce: Simple Programming For Big Results
6. **Reading:** MapReduce in the Pasta Sauce Example
7. **Video:** When to Reconsider Hadoop?
8. **Video:** Cloud Computing: An Important Big Data Enabler
9. **Video:** Cloud Service Models: An Exploration of Choices
10. **Video:** Value From Hadoop and Pre-built Hadoop Images
11. **Reading:** Slides for Getting Started With Hadoop
12. **Reading:** Downloading and Installing Docker Desktop Instructions
13. **Reading:** Downloading Hands-On Materials
14. **Reading:** Basic terminal shell commands
15. **Reading:** Starting Hadoop
16. **Video:** Starting Hadoop
17. **Reading:** Run the WordCount program Instructions
18. **Video:** Run the WordCount program
19. **Discussion Prompt:** Let's Discuss: Map Reduce in your life
20. **Reading:** How do I figure out how to run Hadoop MapReduce programs?

Show less

Graded: Intro to Hadoop
Graded: Understand by Doing: MapReduce
Graded: Running Hadoop MapReduce Programs Quiz

View Less

How It Works

General

What do start dates and end dates mean?

Once you enroll,
More

Peer-graded assignments

Peer-graded assignments require you and your classmates to grade each other's work.
More

Course 1 of 1 Specialization

Unlock Value in Massive Datasets
Learn fundamental big data methods in six straightforward courses.

Big Data
University of California San Diego

Learn More

View the course in catalog

Related Courses

Big Data - Capstone Project
University of California San Diego

Graph Analytics for Big Data
University of California San Diego

Machine Learning With Big Data
University of California San Diego

Big Data Integration and Processing
University of California San Diego

Big Data Modeling and Management Systems
University of California San Diego