

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 the Big Data Era 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 necessary to complete the hands-on assignments.

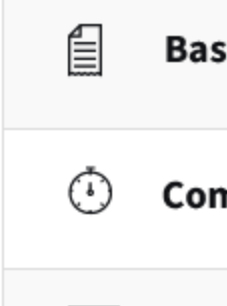
Hardware Requirements:

(A) Quad Core Processor (VT-x or AMD-V support recommended), 64-bit; (B) 8 GB RAM; (C) 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.

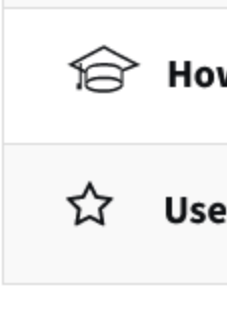
Software Requirements:

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, Oriya, 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	★★★★☆ 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


- Video:** [Welcome to the Big Data Specialization](#)
- Reading:** By the end of this course you will be able to...
- Reading:** Optional: Watch this fun video about the San Diego Supercomputer Center!
- Video:** Tell us about yourself and learn about your classmates
- 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

- Video:** [What launched the Big Data era?](#)
- Video:** Applications: What makes big data valuable
- Discussion Prompt:** Let's Discuss: What application area interests you?
- Video:** Example: Saving Lives with Big Data
- Video:** Example: Using Big Data to Help Patients
- Video:** A Sentiment Analysis Success Story: Meltwater helping Danone
- Reading:** Did you know?: 25 facts about big data
- Reading:** Slides: What Launched the Big Data Era?
- Reading:** Slides: Applications: What Makes Big Data Valuable?
- Reading:** Slides: Saving Lives With Big Data
- Reading:** Slides: Using Big Data to Help Patients
- Video:** Getting Started: Where Does Big Data Come From?
- Video:** Machine-Generated Data: It's Everywhere and There's a Lot!
- Video:** Machine-Generated Data: Advantages
- Video:** Big Data Generated By People: The Unstructured Challenge
- Video:** Big Data Generated By People: How is It Being Used?
- Video:** Organization-Generated Data: Structured but often siloed
- Video:** Organization-Generated Data: Benefits Come From Combining With Other Data Types
- Video:** The Key: Integrating Diverse Data
- Discussion Prompt:** Let's discuss: Who are you providing data to?
- Reading:** Extra Resources
- Reading:** Slides: Machine-Generated Data: It's Everywhere and There's a Lot!
- Reading:** Slides: Machine-Generated Data: Advantages
- Reading:** Slides: Big Data Generated By People: The Unstructured Challenge
- Reading:** Slides: Big Data Generated By People: How is it Being Used?
- Reading:** Slides: Organization-Generated Big Data: Structured But Often Siloed
- Reading:** Slides: Organization-Generated Big Data: Benefits
- 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 Vs". 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, 9 readings

- Video:** [Getting Started: Characteristics Of Big Data](#)
- Video:** Characteristics of Big Data - Volume
- Reading:** What does astronomical scale mean?
- Video:** Characteristics of Big Data - Variety
- Video:** Characteristics of Big Data - Velocity
- Video:** Characteristics of Big Data - Veracity
- Video:** Characteristics of Big Data - Valence
- Video:** The Sixth V: Value
- Reading:** A Small Definition of Big Data
- Discussion Prompt:** Practice: Writing Big Data questions
- Discussion Prompt:** Let's Discuss: Improving the Flamingo Game
- Reading:** Slides: Getting Started - Characteristics of Big Data
- Reading:** Slides: Characteristics of Big Data - Volume
- Reading:** Slides: Characteristics of Big Data - Variety
- Reading:** Slides: Characteristics of Big Data - Velocity
- Reading:** Slides: Characteristics of Big Data - Veracity
- Reading:** Slides: Characteristics of Big Data - Value
- Reading:** Slides: Characteristics of Big Data - Valence

[Show less](#)

 **Graded:** V for the Vs 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

- Video:** [Data Science: Getting Value out of Big Data](#)
- Video:** Building a Big Data Strategy
- Video:** How does big data science happen?: Five Components of Data Science
- Reading:** Five P's of Data Science
- Discussion Prompt:** Let's Discuss: Thinking more deeply about the Ps
- Video:** Asking the Right Questions
- Video:** Steps in the Data Science Process
- Video:** Step 1: Acquiring Data
- Video:** Step 2-A: Exploring Data
- Video:** Step 2-B: Pre-Processing Data
- Video:** Step 3: Analyzing Data
- Video:** Step 4: Communicating Results
- Video:** Step 5: Turning Insights into Action
- Discussion Prompt:** Let's Discuss: Building a Team
- Reading:** Slides: Getting Value Out of Big Data
- Reading:** Slides: Building a Big Data Strategy
- Reading:** Slides: The Five P's of Data Science
- Reading:** Slides: Asking the Right Questions
- Reading:** Slides: Steps in the Data Science Process
- Reading:** Slides: Step 1 - Acquiring Data
- Reading:** Slides: Step 2A-Exploring Data
- Reading:** Slides: Step 2B-Preprocessing Data
- Reading:** Slides: Step 3-Data Analysis
- Reading:** Slides: Step 4-Communicating Results
- 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.

 4 videos, 4 readings

- Video:** [Getting Started: Why worry about foundations?](#)
- Video:** What is a Distributed File System?
- Video:** Scalable Computing over the Internet
- Video:** Programming Models for Big Data
- Reading:** Slides: Getting Started-Why Worry About Foundations?
- Reading:** Slides: What is a Distributed File System?
- Reading:** Slides: Scalable Computing Over the Internet
- 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, 8 readings

- Video:** [Hadoop: Why, Where and Who?](#)
- Video:** The Hadoop Ecosystem: Welcome to the zoo!
- Video:** The Hadoop Distributed File System: A Storage System for Big Data
- Video:** YARN: A Resource Manager for Hadoop
- Video:** MapReduce: Simple Programming for Big Results
- Reading:** MapReduce in the Pasta Sauce Example
- Video:** When to Reconsider Hadoop?
- Video:** Cloud Computing: An Important Big Data Enabler
- Video:** Cloud Service Models: An Exploration of Choices
- Video:** Value From Hadoop and Pre-built Hadoop Images
- Reading:** Slides for Getting Started With Hadoop
- Reading:** Downloading and Installing Docker Desktop Instructions
- Reading:** Downloading Hands-On Materials
- Reading:** Basic terminal shell commands
- Reading:** Starting Hadoop
- Video:** Starting Hadoop
- Reading:** Run the WordCount program Instructions
- Video:** Run the WordCount program
- Discussion Prompt:** Let's Discuss: Map Reduce in your life
- 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,
[View More](#)

Peer-graded assignments


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

[View More](#)

Course 1 of 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