

Big Data Integration and Processing

by University of California San Diego

About this Course

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

- *Retrieve data from example database and big data management systems
- *Describe the connections between data management operations and the big data processing patterns needed to utilize them in large-scale analytical applications
- *Identify when a big data problem needs data integration
- *Execute simple big data integration and processing on Hadoop and Spark platforms

This course is for those new to data science. Completion of Intro to Big Data is recommended. 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. Refer to the specialization technical requirements for complete hardware and software specifications.

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.

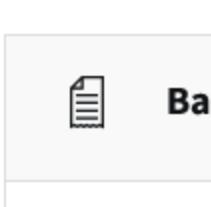
Software Requirements:

This course relies on several open-source software tools, including Apache Hadoop. All required software can be downloaded and installed free of charge (except for data charges from your internet provider). 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 3 of 6 in the Big Data Specialization
	Level	Beginner
	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, Hungarian, Polish
	How To Pass	Pass all graded assignments to complete the course.
	User Ratings	Average User Rating 4.4

Syllabus

Module 1

Welcome to Big Data Integration and Processing

Welcome to the third course in the Big Data Specialization. This week you will be introduced to basic concepts in big data integration and processing. You will be guided through installing Docker, downloading the data sets to be used for this course, and learning how to work with Jupyter notebooks.

3 videos, 5 readings

- Video:** [What is in this Course?](#)
- Video:** Summary of Big Data Modeling and Management
- Video:** Why is Big Data Processing Different?
- Discussion Prompt:** Getting to know you: Tell us about yourself and why you are taking this course.
- Reading:** Slides: Summary & Why Is Big Data Processing Different
- Reading:** Downloading and Installing Docker Desktop Instructions
- Reading:** Introduction to Jupyter Notebooks
- Reading:** Downloading Hands-On Materials
- Reading:** Basic terminal shell commands

[Show less](#)

Module 2

Retrieving Big Data (Part 1)

This module covers the various aspects of data retrieval and relational querying. You will also be introduced to the Postgres database.

5 videos, 2 readings

- Video:** [What is Data Retrieval? Part 1](#)
- Video:** What is Data Retrieval? Part 2
- Video:** Querying Two Relations
- Video:** Subqueries
- Reading:** Slides: What is Data Retrieval?
- Reading:** Querying Relational Data with Postgres
- Video:** Querying Relational Data with Postgres

[Show less](#)

Module 3

Retrieving Big Data (Part 2)

This module covers the various aspects of data retrieval for NoSQL data, as well as data aggregation and working with data frames. You will be introduced to MongoDB and Aerospike, and you will learn how to use Pandas to retrieve data from them.

5 videos, 3 readings

- Video:** [Querying JSON Data with MongoDB](#)
- Video:** Aggregation Functions
- Discussion Prompt:** Let's Discuss: MongoDB
- Video:** Querying Aerospike
- Reading:** Slides: Querying Data Part 2
- Reading:** Querying Documents in MongoDB
- Video:** Querying Documents in MongoDB
- Reading:** Exploring Pandas DataFrames
- Video:** Exploring Pandas DataFrames

[Show less](#)

Graded: Retrieving Big Data Quiz

Graded: Postgres, MongoDB, and Pandas

Module 4

Big Data Integration

In this module you will be introduced to data integration tools including Splunk and Datameer, and you will gain some practical insight into how information integration processes are carried out.

11 videos, 4 readings

- Video:** [Overview of Information Integration](#)
- Video:** A Data Integration Scenario
- Video:** Integration for Multichannel Customer Analytics
- Discussion Prompt:** Let's Discuss: Big Data Integration
- Reading:** Slides: Information Integration
- Video:** Big Data Management and Processing Using Splunk and Datameer
- Video:** Why Splunk?
- Video:** Connected Cars with Ford's OpenXC and Splunk
- Video:** Big Data Management and Processing using Datameer
- Reading:** Downloading Splunk Enterprise
- Video:** Installing Splunk Enterprise on Windows
- Video:** Installing Splunk Enterprise on Linux
- Reading:** Exploring Splunk Queries
- Video:** Exploring Splunk Queries
- Reading:** Optional: Instructions for Splunk Pivot Tutorial
- Video:** Optional: Creating Pivot Reports in Splunk

[Show less](#)

Graded: Information Integration - Quiz

Graded: Hands-On With Splunk

Module 5

Processing Big Data

This module introduces Learners to big data pipelines and workflows as well as processing and analysis of big data using Apache Spark.

9 videos, 4 readings

- Video:** [Big Data Processing Pipelines](#)
- Video:** Some High-Level Processing Operations in Big Data Pipelines
- Video:** Aggregation Operations in Big Data Pipelines
- Video:** Typical Analytical Operations in Big Data Pipelines
- Discussion Prompt:** Let's Discuss: Big Data Pipelines in Your World
- Reading:** Big Data Processing Pipelines Slides
- Video:** Overview of Big Data Processing Systems
- Reading:** Big Data Workflow Management
- Video:** The Integration and Processing Layer
- Video:** Introduction to Apache Spark
- Video:** Getting Started with Spark
- Discussion Prompt:** Let's Discuss: Big Data Processing Systems
- Reading:** Slides for Big Data Processing Tools and Systems
- Video:** WordCount in Spark
- Video:** WordCount in Spark
- Discussion Prompt:** Let's Discuss: Word Count

[Show less](#)

Graded: Pipeline and Tools

Graded: WordCount in Spark

Module 6

Big Data Analytics using Spark

In this module, you will go deeper into big data processing by learning the inner workings of the Spark Core. You will be introduced to two key tools in the Spark toolkit: Spark MLlib and GraphX.

9 videos, 4 readings

- Video:** [Spark Core: Programming In Spark using RDDs in Pipelines](#)
- Video:** Spark Core: Transformations
- Video:** Spark Core: Actions
- Reading:** Slides for Module 5 Lesson 1
- Video:** Spark SQL
- Video:** Spark Streaming
- Video:** Spark MLlib
- Video:** Spark GraphX
- Discussion Prompt:** Let's Discuss: The Spark Ecosystem
- Reading:** Slides for Module 5 Lesson 2
- Reading:** Exploring SparkSQL and Spark DataFrames
- Video:** Exploring SparkSQL and Spark DataFrames
- Reading:** Analyzing Sensor Data with Spark Streaming
- Video:** Analyzing Sensor Data with Spark Streaming

[Show less](#)

Graded: More on Spark

Graded: SparkSQL and Spark Streaming

Module 7

Learn By Doing: Putting MongoDB and Spark to Work

In this module you will get some practical hands-on experience applying what you learned about Spark and MongoDB to analyze Twitter data.

4 readings

- Reading:** Let's Analyze Soccer Tweets!
- Reading:** Expressing Analytical Questions as MongoDB Queries
- Reading:** Exporting Data from MongoDB to a CSV File
- Reading:** Analyzing Tweets About Countries

[Show less](#)

Graded: Check Your Query Results

Graded: Check Your Analysis Results

[View Less](#)

How It Works

General

What do start dates and end dates mean?

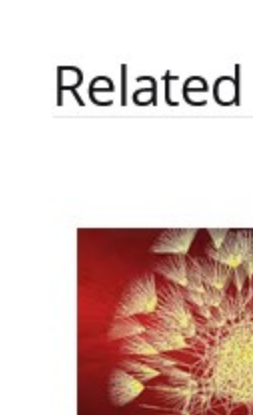
Once you enroll,

[More](#)

Course 3 of Specialization

Unlock Value in Massive Datasets

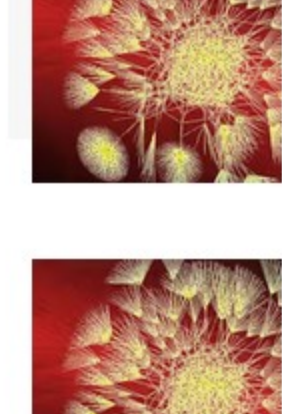
Learn fundamental big data methods in six straightforward courses.



Big Data
University of California San Diego

[Learn More](#)

[View the course in catalog](#)



Big Data - Capstone Project
University of California San Diego



Introduction to Big Data
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 Modeling and Management Systems
University of California San Diego