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	★★★★½ 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

1. **Video:** What is in this Course?
2. **Video:** Summary of Big Data Modeling and Management
3. **Video:** Why is Big Data Processing Different?
4. **Discussion Prompt:** Getting to know you: Tell us about yourself and why you are taking this course.
5. **Reading:** Slides: Summary & Why Is Big Data Processing Different
6. **Reading:** Downloading and Installing Docker Desktop Instructions
7. **Reading:** Introduction to Jupyter Notebooks
8. **Reading:** Downloading Hands-On Materials
9. **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

1. **Video:** What is Data Retrieval? Part 1
2. **Video:** What is Data Retrieval? Part 2
3. **Video:** Querying Two Relations
4. **Video:** Subqueries
5. **Reading:** Slides: What is Data Retrieval?
6. **Reading:** Querying Relational Data with Postgres
7. **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 Aeropike, and you will learn how to use Pandas to retrieve data from them.

5 videos, 3 readings

1. **Video:** Querying JSON Data with MongoDB
2. **Video:** Aggregation Functions
3. **Discussion Prompt:** Let's Discuss: MongoDB
4. **Video:** Querying Aerospike
5. **Reading:** Slides: Querying Data Part 2
6. **Reading:** Querying Documents in MongoDB
7. **Video:** Querying Documents in MongoDB
8. **Reading:** Exploring Pandas DataFrames
9. **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

1. **Video:** Overview of Information Integration
2. **Video:** A Data Integration Scenario
3. **Video:** Integration for Multichannel Customer Analytics
4. **Discussion Prompt:** Let's Discuss: Big Data Integration
5. **Reading:** Slides: Information Integration
6. **Video:** Big Data Management and Processing Using Splunk and Datameer
7. **Video:** Why Splunk?
8. **Video:** Connected Cars with Ford's OpenXC and Splunk
9. **Video:** Big Data Management and Processing using Datameer
10. **Reading:** Downloading Splunk Enterprise
11. **Video:** Installing Splunk Enterprise on Windows
12. **Video:** Installing Splunk Enterprise on Linux
13. **Reading:** Exploring Splunk Queries
14. **Video:** Exploring Splunk Queries
15. **Reading:** Optional: Instructions for Splunk Pivot Tutorial
16. **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

1. **Video:** Big Data Processing Pipelines
2. **Video:** Some High-Level Processing Operations in Big Data Pipelines
3. **Video:** Aggregation Operations in Big Data Pipelines
4. **Video:** Typical Analytical Operations in Big Data Pipelines
5. **Discussion Prompt:** Let's Discuss: Big Data Pipelines in Your World
6. **Reading:** Big Data Processing Pipelines Slides
7. **Video:** Overview of Big Data Processing Systems
8. **Reading:** Big Data Workflow Management
9. **Video:** The Integration and Processing Layer
10. **Video:** Introduction to Apache Spark
11. **Video:** Getting Started with Spark
12. **Discussion Prompt:** Let's Discuss: Big Data Processing Systems
13. **Reading:** Slides for Big Data Processing Tools and Systems
14. **Reading:** WordCount in Spark
15. **Video:** WordCount in Spark
16. **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

1. **Video:** Spark Core: Programming in Spark using RDDs in Pipelines
2. **Video:** Spark Core: Transformations
3. **Video:** Spark Core: Actions
4. **Reading:** Slides for Module 5 Lesson 1
5. **Video:** Spark SQL
6. **Video:** Spark Streaming
7. **Video:** Spark MLLib
8. **Video:** Spark GraphX
9. **Discussion Prompt:** Let's Discuss: The Spark Ecosystem
10. **Reading:** Slides for Module 5 Lesson 2
11. **Reading:** Exploring SparkSQL and Spark DataFrames
12. **Video:** Exploring SparkSQL and Spark DataFrames
13. **Reading:** Analyzing Sensor Data with Spark Streaming
14. **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

1. **Reading:** Let's Analyze Soccer Tweets!
2. **Reading:** Expressing Analytical Questions as MongoDB Queries
3. **Reading:** Exporting Data from MongoDB to a CSV File
4. **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

Related Courses

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