16.0.2

Module 16 Roadmap

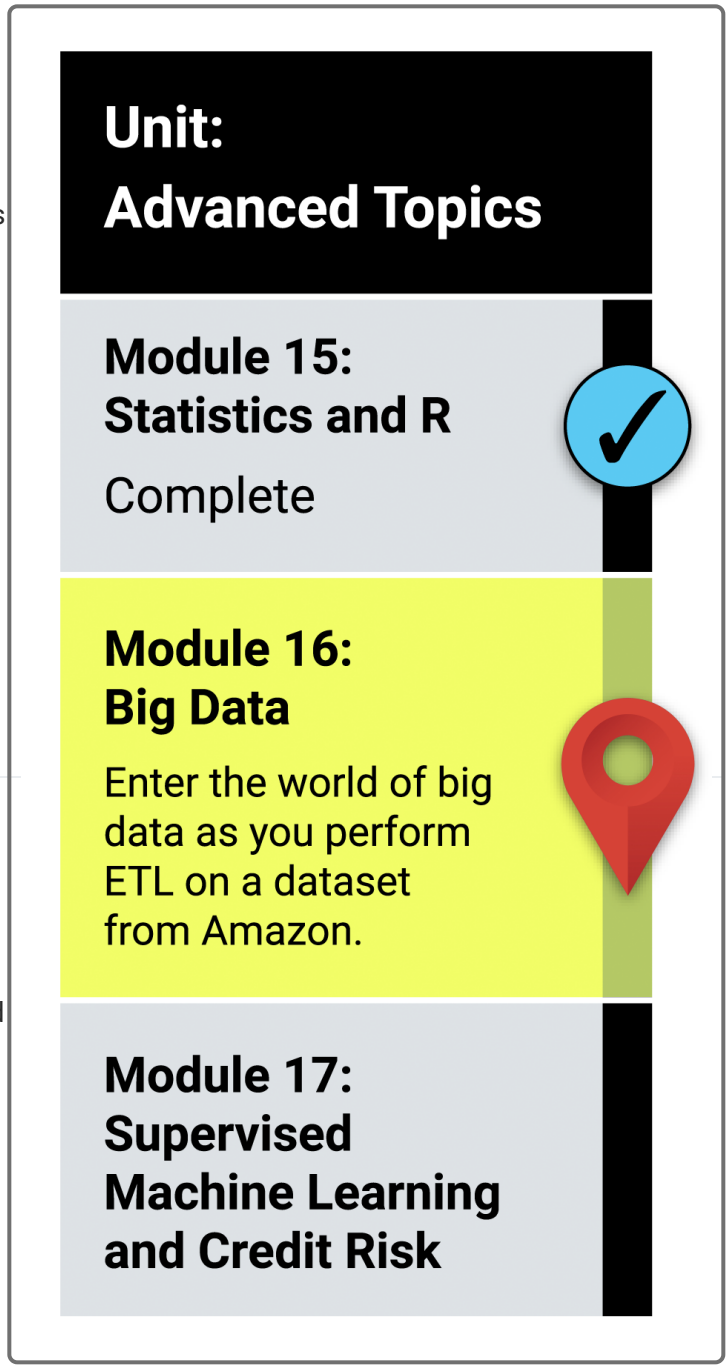
Looking Ahead

This week we'll cover what constitutes big data and how it's handled. We'll start by reviewing Hadoop and its ecosystem.

Within this big data and Hadoop context, we'll cover MapReduce and how it has improved the process for handling big data. We'll then move on to PySpark, which has become the leading technology for handling big data.

After diving into some of the technologies used with big data, we'll look at natural language processing (NLP) in relation to big data.

We'll close with an introduction to cloud services. Cloud services let us store large amounts of data at remote locations rather than locally, on top of many other services. This allows for more scalability and performance. We'll use the most popular cloud service available: Amazon Web Services (AWS).



What You Will Learn

By the end of this module, you will be able to:

- Define big data and describe the challenges associated with it.
- Define Hadoop and name the main elements of its ecosystem.
- Explain how MapReduce processes data.
- Define Spark and explain how it processes data.
- Describe how NLP collects and analyzes text data.

- Explain how to use AWS Simple Storage Service (S3) and relational databases for basic cloud storage.
- Complete an analysis of an Amazon customer review.

Planning Your Schedule

Here's a quick look at the lessons and assignments you'll cover in this module. You can use the time estimates to help pace your learning and plan your schedule:

- Introduction to Module 16 (15 minutes)
- Overview of Big Data (15 minutes)
- Using MapReduce to Process Data (30 minutes)
- Using Spark to Handle Large Datasets (30 minutes)
- Working with Spark DataFrames and Functions (2 hours)
- Natural Language Processing (1 hour)
- PySpark and Natural Language Processing (2 hours)
- Cloud Databases with Amazon Web Services (2 hours)
- Cloud Storage with S3 on AWS (2 hours)
- ETL in the Cloud (2 hours)
- Application (5 hours)