DBDA.X400 - Introduction to Apache Spark with Scala - 3.0 units

Instructor: Hien Luu

Course Description

Apache Spark is a unified data analytics engine that can support batch, interactive, iterative, streaming, and graph processing use cases. The combination of elegant application programming interfaces (APIs) and a fast in-memory, general-purpose cluster computing system makes it an attractive option for companies to leverage for various data processing needs. Written in Scala, Apache Spark APIs are available in three programming languages: Scala, Java, and Python. This course, however, focuses on the API in Scala language, a functional  
programming language.

In this foundational course, we will explore Apache Spark, its architecture, and the execution model. We’ll start with a short introduction to Scala, its basic syntax, case class, and collection APIs. You’ll learn how to process data using DataFrame, Apache Spark’s structured data processing programming model that provides simple, powerful APIs. In addition to batch and iterative data processing, Apache Spark also supports stream processing, which enables companies to extract interesting and useful business insights in near real-time. The second half of the course covers stream processing capability and developing streaming applications with Apache Spark. We will briefly cover machine learning and how the Apache Spark MLlib component makes practical machine learning scalable and easy.

By the end of the course, you’ll have a good foundation in Scala language and a strong  
understanding of Apache Spark’s architecture, execution model, and programming model.  
You’ll be able to manipulate DataFrame through Apache Spark’s API and develop Apache Spark  
applications in Scala for batch, interactive, and stream processing applications. You will gain  
fundamental concepts in machine learning and be able to leverage MLlib library to build simple  
machine learning applications.

Prerequisite

|  |  |
| --- | --- |
| **Suggested:** | [**CMPR.X413Links to an external site.**](https://course.ucsc-extension.edu/modules/shop/index.html?action=section&OfferingID=5590711)**(Java Programming, Comprehensive)** |

**Skills Needed:** You should have prior object-oriented programming experience to learn Scala as this course offers only a short introduction to Scala.

Learning Outcomes

At the conclusion of the course, participants should be able to:

* Describe the Apache Spark’s architecture, execution model and programming model.
* Perform data processing by manipulating Apache Spark DataFrame APIs.
* Build batch and streaming data processing applications using Apache Spark
* Build small to medium Scala applications using Scala programming language.

Course Outline

|  |  |  |
| --- | --- | --- |
| **Week** | **Topics** | **Assignments** |
| 1 | * Introduction to Spark & execution model | * Assignment #1 |
| 2 | * Introduction to Scala Programming Language * A quick tutorial on using Databricks | * Assignment #2 |
| 3 | * Apache Spark Programming Model with DataFrame | * In-class exercises |
| 4 | * Apache Spark Programming Model with DataFrame | * In-class exercises |
| 5 | * Apache Spark SQL and Advanced Analytics Functions | * Assignment #3 |
| 6 | * Apache Spark Structured Streaming Processing #1 | * In-class exercises |
| 7 | * Apache Spark Structured Streaming Processing #2 | * Assignment #4 |
| 8 | * Introduction to Machine Learning | * In-class exercises |
| 9 | * Develop ML applications with Spark MLlib Library | * In-class exercises |
| 10 | * Option topic - Apache Kafka | * In-class exercises |

Required Tools and Materials

None

Recommended Tools and Materials

***Learning Spark: Lightning-Fast Big Data Analysis***, Holden Karau, et al., O'Reilly, 2015, ISBN-10: 1449358624, ISBN-13: 978-1449358624.

Performance Evaluation

|  |  |  |
| --- | --- | --- |
| **Criteria** | **Percentage** | **Description** |
| Assignment #1 | 20% | Read a paper about Apache Spark and answer questions. |
| Assignment #2 | 20% | Scala programming assignment |
| Assignment #3 | 30% | Perform data analysis using Spark DataFrame & SQL |
| Assignment #4 | 30% | Spark streaming application to perform twitter sentiment analysis |
| **Total:** | **100%** |  |

Grading

Letter grades (A through F) are the default options.  However, students have until the day before the course end date to change their grading preference to a Credit/No Credit Option.

Grading scale

|  |  |
| --- | --- |
| **Grade options** | **%** |
| **A** | >93 |
| **A-** | 90-92 |
| **B+** | 88-89 |
| **B** | 83-87 |
| **B-** | 80-82 |
| **C+** | 78-79 |
| **C** | 73-77 |
| **C-** | 70-72 |
| **D+** | 68-69 |
| **D** | 63-67 |
| **D-** | 60-62 |
| **F** | 59 and below |
| **Credit** | 60 and above |
| **No Credit** | 59 and below |

**\*For alternative grading options, students MUST** **contact**[**extensiongrades@ucsc.edu**](mailto:extensiongrades@ucsc.edu)**with the Alternative Grade Form.**

**Click Here to Review the**[**Grading and Credits WebsiteLinks to an external site.**](https://www.ucsc-extension.edu/info/policies/grading-and-credits-policy/)

UCSC Extension Policies:

Click here to view and print the [UCSC Extension Policies (PDF)Links to an external site.](https://file.ucsc-extension.edu/unexfiles/UNEX_Policies_Syllabus.pdf)