LogoFamily

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Course Code/Course Title: DSO107 Introduction to Big Data** | | | | | |
| **Course Description:** The Introduction to Big Data course introduces students to Big Data on a conceptual level, and gives students exposure and practice with several skills and tools currently in use. These skills will be taught at a manageable level, and then scale up methods will be used to help students grasp the meaning and popularity of analyzing substantial amounts of data. Students will learn the foundational concepts of Big Data and will know how to move from Big Data basics to more business specific needs and requirements. | | | | | |
| **Quarter Credit Hours:**  3 | | **Course Length:**  40 hours | **Prerequisites:**  DSO102, DSO104, & DSO109 | | **Proficiency Exam**  Yes  No |
| **Theory Hours:** | 20 | **Course Start Date:**  **Course End Date:** | **Meeting Days/Times** | | |
| **Laboratory Hours:** | 20 |
| **Externship Hours:** | 0 |
| **Outside Hours:** | 10 |
| **Total Contact Hours:** | 40 |
| **Instructor Information:**  **Name:**  **Contact Number:**  **E-mail Address:**  **Office Hours:** | | | |  | |

**Required Resources:**

Ground based students are required to bring a laptop computer (either PC or MacBook) to class every day.

Online students are required to have a laptop or desktop computer with internet access.

Minimum: PC (Windows) or MacBook laptop. 4GB ram, 256GB HD, Core i5 or New M1/M1x/M2 Chipsets

Recommended: PC (Windows) or MacBook laptop. 8GB ram, 256GB SSD, Core i5 or New M1/M1x/M2 Chipsets

It is requirement that you are able to download programming resources to your laptop/desktop for this class.

LogoFamily

**Educational Objectives:**

## Upon successful completion of this course, students will be able to:

1. Explain the background and evolution of Big Data
2. Use the fundamentals of Hadoop
3. Hands on Hadoop experience
4. Learn the fundamentals of MapReduce
5. Hands on MapReduce experience
6. Learn how to scale up to large datasets using available tools
7. Learn how Hadoop and MapReduce utilize multiple computing clusters
8. Learn about other Hadoop technologies

**Course Outline**

**Course:**

**Week 1**

1. **Introduction:** Includes Introduction to Big Data, Python Review, Running Code in the Terminal for Windows Users, Running Code in the Terminal for Mac/Linux Users, Reading from Standard Input, Reading from Files (File IO)
2. **ETL & MapReduce:** Includes ETL & Map Reduce, Reading in Data, Reading in CSVs, Errors, MapReduce, Overall Goal, Create a Reduce File, Run the Map and Reduce Files Together, Counting the Types of Arrests, Key Terms
3. **Orchestration:** Includes Orchestrations, Streams, Crimes Data, Installing Packages for Windows, Installing Packages for Mac/Linux, Manager Set Up, Worker SetUp, Running the Files, Crime Analysis with MapReduce Using Orchestration, Monitor System Performance, Key Terms

**Week 2**

1. **Distribution:** Includes Partitioning Data, Multiple Workers, Distribution, Running the Files, Activity Monitoring, Process Scheduling, Load Balancing, Key Terms
2. **Amazon Web Service Set Up:** Includes Introduction to Spark, Introduction to Amazon Web Service, Accessing Your AWS Educate Starter Account, EC2 Setup, Connecting to EC2, Key Terms
3. **PySpark Set Up:** Includes PySpark Set Up, Installing Anaconda, Configuring Jupyter Notebook, Running Jupyter Notebook, Installing Additional Software, Installing Pip and Packages, Install Spark
4. **Using PySpark:** Includes Introduction, Windows: How to Reconnect to Your Instance, Mac/Linux: How to Reconnect to Your Instance, Using PySpark, Map(), Mapping a Dataset, ReduceByKey(), Filter(), SortBy(), Sample(), Distinct(), Union(), Key Terms

**Week 3**

1. **Hadoop:** Includes Introduction, What is Hadoop?, Key Terms
2. **Big Judgement:** Includes Failure Recovery, There’s Always An Exception
3. **Final Project**

**Workshops:**

|  |  |
| --- | --- |
| **Class: DSO107** | **Topic presented** |
| **Week 1 Workshop #1** | What is Big Data? (L1) |
| **Week 1 Workshop #2** | MapReduce (L2) |
| **Week 2 Workshop #1** | Orchestration and Distribution (L3, L4) |
| **Week 2 Workshop #2** | PySpark (L6, L7) |
| **Week 3 Workshop #1** | Hadoop (L8) |
| **Week 3 Workshop #2** | Practice Project (L10) |

**Points Distribution:**

* **L1 Hands On – 45 points:** Review Python fundamentals by searching for stop words and using the strip() function.
* **L2 Hands On – 45 points:** Utilize MapReduce functions in Python.
* **L3 Practice Hands On – 0 points:** Orchestrate the MapReduce function across multiple workers in Python.
* **L4 Practice Hands On – 0 points:** Distribute the MapReduce function across multiple workers in Python.
* **L7 Hands On – 45 points:** Utilize AWS to tap into PySpark and perform data transformations in Spark.
* **L8 Hands On – 45 points:** Assess the current big data job market.
* **L9 Practice Hands On – 0 points:** Comment the code and utilize try-except to improve functioning in Python.
* **L10 Final Project – 200 points:**  Orchestrate, distribute, and perform MapReduce on data in Python.

**Points Total:**

Professionalism, Attendance and Class Participation points 20 (5%)

Assignments/Hands-On/Homework: L1-9 Hands On total points180 (45%)

Projects/Competencies/Research: Final Project 200 (50%)

Total points: 400 (100%)

**Final Project:**

With the data given, create one manager and create two workers. Perform MapReduce to count accidents for each vehicle and what most common action prior to accident. Lastly, determine which action is most common.

**Course Evaluation Strategies (Methodologies):**

|  |  |
| --- | --- |
| Professionalism, Attendance and Class Participation\* | 5% |
| Assignments/Hands-On/Homework | 95% |
| Total | 100% |

## 

## 