DATA SCIENCE & ANALYTICS

OVERALL COURSE OBJECTIVES:

You will be able to understand, analyze, and interpret Big Data landscapes by applying cutting-edge methods from machine learning and graph analytics. Utilizing prominent tools like Apache Spark, Jupyter Notebooks, and Python, this course will empower you to work with diverse and dynamic data sources efficiently while creating intuitive visualizations and models. The acquired skills will allow you to derive path-breaking insights from complex data and communicate them effectively through impactful presentations and reports.

LEARNING OUTCOMES: On successful completion of the course the students shall be able to:

1. Analyze and understand Big Data by applying techniques from machine learning and graph analytics.
2. Leverage major systems for data analysis such as Spark, Hortonworks, Cloudera, and MapR while developing proficiency in distributed file systems and map-reduce.
3. Apply advanced statistical measures and data visualization technologies to reveal patterns within large volumes of data.
4. Utilize different data management systems for handling dynamic, georeferenced, or large-scale data sources, with abilities to execute continuous integration or processing.
5. Conduct exploratory data analysis including cleaning, wrangling and visualizing from different data sources with programming languages such as python and R.
6. Create compelling reports and presentations that communicate findings and insights derived from complex data sets and machine learning models.

	Cloud Computing Applications, Part 2: Big Data and Applications in the Cloud
	Intro to Analytic Thinking, Data Science, and Data Mining
Data Science Fundamentals – I	Introduction to Data Science in Python
	What is Data Science?
Data Science Fundamentals – II	Tools for Data Science
	Data Analysis with Python
	Data Visualization with R
	Predictive Modeling, Model Fitting, and Regression Analysis
Data Science Applications	Cluster Analysis, Association Mining, and Model Evaluation
	Introduction to Big Data
	Big Data Modeling and Management Systems
Big Data Foundations	Big Data Integration and Processing
	Machine Learning With Big Data
	Graph Analytics for Big Data
Big Data Analytics	Big Data - Capstone Project

COURSE CONTENT:

Module 1: Cloud Computing Applications, Part 2: Big Data and Applications in the Cloud [20 hours]

In this course we continue Cloud Computing Applications by exploring how the Cloud opens up data analytics of huge volumes of data that are static or streamed at high velocity and represent an enormous variety of information. Cloud applications and data analytics represent a disruptive change in the ways that society is informed by, and uses information.

Sub-Topics
Course Orientation
Module 1: Spark, Hortonworks, HDFS, CAP
Module 2: Large Scale Data Storage
Module 3: Streaming Systems
Module 4: Graph Processing and Machine Learning

Formative Assessments:
4 Graded quizzes

Module 2 : Intro to Analytic Thinking, Data Science, and Data Mining [7 Hours]

Welcome to Introduction to Analytic Thinking, Data Science, and Data Mining. In this course, we will begin with an exploration of the field and profession of data science with a focus on the skills and ethical considerations required when working with data. We will review the types of business problems data science can solve and discuss the application of the CRISP-DM process to data mining efforts. A brief overview of Descriptive, Predictive, and Prescriptive Analytics will be provided, and we will conclude the course with an exploratory activity to learn more about the tools and resources you might find in a data science toolkit.

Sub-Topics
Data Mining and an Overview of Data Analytics
Data Science in Business
Data Science: The Field and Profession
Solving Problems with Data Science

Formative Assessments:
2 Graded Quizzes

Module 3: Introduction to Data Science in Python [35 Hours]

This course will introduce the learner to the basics of the Python programming environment, including fundamental Python programming techniques such as lambdas, reading and manipulating csv files, and the numpy library. The course will introduce data manipulation and cleaning techniques using the popular Python pandas data science library and introduce the abstraction of the Series and DataFrame as the central data structures for data analysis, along with tutorials on how to use functions such as group by, merge, and pivot tables effectively. By the end of this course, students will be able to take tabular data, clean it, manipulate it, and run basic inferential statistical analyses.

Sub-Topic
Fundamentals of Data Manipulation with Python
Data Processing with Pandas
Answering Questions with Messy Data

Formative Assessments:
4 quizzes and 9 coding/lab assignments.

Module 4 : What is Data Science? [19 Hours]

This course is for everyone and teaches concepts like how data scientists use machine learning and deep learning and how companies apply data science in business. You will meet several data scientists, who will share their insights and experiences in data science. By taking this introductory course, you will begin your journey into this thriving field.

Sub-Topics
Applications and Careers in Data Science
Data literacy for Data Science (Optional)
Data Science Topics
Defining Data Science and What Data Scientists Do

Formative Assessments:
8 Staff Graded Assessments

Module 5 : Tools for Data Science [18 Hours]

This course gives plenty of hands-on experience in order to develop skills for working with these Data Science Tools. With the tools hosted in the cloud on Skills Network Labs, you will be able to test each tool and follow instructions to run simple code in Python, R, or Scala. Towards the end the course, you will create a final project with a Jupyter Notebook. You will demonstrate your proficiency preparing a notebook, writing Markdown, and sharing your work with your peers.

Sub-Topics
IBM Watson Studio
Create and Share your Jupyter Notebook
Jupyter Notebooks and JupyterLab
Languages of Data Science
Overview of Data Science Tools
Packages, APIs, Datasets and Models
RStudio & GitHub

Formative Assessments:
1 Peer Review assignment, 5 Graded Quizzes & 1 Staff Graded Assessment

Module 6 : Data Analysis with Python [15 Hours]

In this course, You will learn how to import data from multiple sources, clean and wrangle data, perform exploratory data analysis (EDA), and create meaningful data visualizations. You will then predict future trends from data by developing linear, multiple, polynomial regression models & pipelines and learn how to evaluate them. In addition to video lectures you will learn and practice using hands-on labs and projects. You will work with several open source Python libraries, including Pandas and Numpy to load, manipulate, analyze, and visualize cool datasets. You will also work with scipy and scikit-learn, to build machine learning models and make predictions.

Sub-Topics
Data Wrangling
Exploratory Data Analysis
Final Assignment

Importing Data Sets
Model Development
Model Evaluation and Refinement

Formative Assessments:
1 Peer Review Assignment & 6 Staff Graded Assessments

Module 7 - Data Visualization with R [12 Hours]

In this course, you will learn the Grammar of Graphics, a system for describing and building graphs, and how the ggplot2 data visualization package for R applies this concept to basic bar charts, histograms, pie charts, scatter plots, line plots, and box plots. You will also learn how to further customize your charts and plots using themes and other techniques. You will then learn how to use another data visualization package for R called Leaflet to create map plots, a unique way to plot data based on geolocation data. Finally, you will be introduced to creating interactive dashboards using the R Shiny package. You will learn how to create and customize Shiny apps, alter the appearance of the apps by adding HTML and image components, and deploy your interactive data apps on the web.

Sub-Topics:

Module 1 - Introduction to Data Visualization
Module 2 - Basic Plots, Maps, and Customization
Module 3 - Dashboards
Module 4 - Final Assignment

Formative Assessments:
1 Peer Review Assignment & 7 Graded Quizzes

Module 8 - Predictive Modeling, Model Fitting, and Regression Analysis [4 Hours]

In this course, we will explore different approaches in predictive modeling, and discuss how a model can be either supervised or unsupervised. We will review how a model can be fitted, trained and scored to apply to both historical and future data in an effort to address business objectives. Finally, this course includes a hands-on activity to develop a linear regression model.

Sub-Topics

Data Dimensionality and Classification Analysis
Model Fitting
Predictive Modeling
Regression Analysis

Formative Assessments:
2 Graded Quizzes

Module 9 : Cluster Analysis, Association Mining, and Model Evaluation [4 Hours]

In this course we will begin with an exploration of cluster analysis and segmentation, and discuss how techniques such as collaborative filtering and association rules mining can be applied. We will also explain how a model can be evaluated for performance, and review the differences in analysis types and when to apply them.

Sub-Topics
Classification-Type Prediction Models
Cluster Analysis and Segmentation
Collaborative Filtering, Association Rules Mining (Market Basked Analysis)
Regression-Type Prediction Models

Formative Assessments:
2 Graded Quizzes

Module 10 : Introduction to Big Data [17 Hours]

This course is for those new to data science and interested in understanding why the Big Data Era has come to be. It is for those who want to become conversant with the terminology and the core concepts behind big data problems, applications, and systems. It is for those who want to start thinking about how Big Data might be useful in their business or career. It provides an introduction to one of the most common frameworks, Hadoop, that has made big data analysis easier and more accessible -- increasing the potential for data to transform our world!

Sub-Topics:

Big Data: Why and Where
Characteristics of Big Data and Dimensions of Scalability
Data Science: Getting Value out of Big Data
Foundations for Big Data Systems and Programming
Systems: Getting Started with Hadoop

Formative Assessments:
1 Peer Review Assignment & 6 Graded Quizzes

Module 11 : Big Data Modeling and Management Systems [13 Hours]

In this course, you will experience various data genres and management tools appropriate for each. You will be able to describe the reasons behind the evolving plethora of new big data platforms from the perspective of big data management systems and analytical tools. Through guided hands-on tutorials, you will become familiar with techniques using real-time and semi-structured data examples. Systems and tools discussed include: AsterixDB, HP Vertica, Impala, Neo4j, Redis, SparkSQL. This course provides techniques to extract value from existing untapped data sources and

discovering new data sources.

Sub-Topics
Big Data Management: The "M" in DBMS
Big Data Modeling
Big Data Modeling (Part 2)
Designing a Big Data Management System for an Online Game
Introduction to Big Data Modeling and Management
Working With Data Models

Formative Assessments:
1 Peer Review Assignment & 4 Graded Quizzes

Module 12 : Big Data Integration and Processing [18 Hours]

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.

Sub-Topics
Big Data Analytics using Spark
Big Data Integration
Learn By Doing: Putting MongoDB and Spark to Work
Processing Big Data
Retrieving Big Data (Part 1)
Retrieving Big Data (Part 2)
Welcome to Big Data Integration and Processing

Formative Assessments:
10 Graded Quizzes

Module 13 : Machine Learning With Big Data [22 Hours]

This course provides an overview of machine learning techniques to explore, analyze, and leverage data. You will be introduced to tools and algorithms you can use to create machine learning models that learn from data, and to scale those models up to big data problems.

Sub-Topics
Classification
Data Exploration
Data Preparation
Evaluation of Machine Learning Models
Introduction to Machine Learning with Big Data

Regression, Cluster Analysis, and Association Analysis

Formative Assessments:
11 Graded Quizzes

Module 14 : Graph Analytics for Big Data [13 Hours]

This course gives you a broad overview of the field of graph analytics so you can learn new ways to model, store, retrieve and analyze graph-structured data. After completing this course, you will be able to model a problem into a graph database and perform analytical tasks over the graph in a scalable manner. Better yet, you will be able to apply these techniques to understand the significance of your data sets for your own projects.

Sub-Topics
Computing Platforms for Graph Analytics
Graph Analytics
Graph Analytics Techniques
Introduction to Graphs
Welcome to Graph Analytics

Formative Assessments:
1 Peer Review Assignment & 6 Graded Quizzes

Module 15 : Big Data - Capstone Project [21 Hours]

In this culminating project, you will build a big data ecosystem using tools and methods form the earlier courses in this specialization. You will analyze a data set simulating big data generated from a large number of users who are playing our imaginary game "Catch the Pink Flamingo". During the five week Capstone Project, you will walk through the typical big data science steps for acquiring, exploring, preparing, analyzing, and reporting. In the first two weeks, we will introduce you to the data set and guide you through some exploratory analysis using tools such as Splunk and Open Office. Then we will move into more challenging big data problems requiring the more advanced tools you have learned including KNIME, Spark's MLLib and Gephi. Finally, during the fifth and final week, we will show you how to bring it all together to create engaging and compelling reports and slide presentations. As a result of our collaboration with Splunk, a software company focus on analyzing machine-generated big data, learners with the top projects will be eligible to present to Splunk and meet Splunk recruiters and engineering leadership.

Sub-Topics
Acquiring, Exploring, and Preparing the Data
Clustering with Spark
Data Classification with KNIME
Final Submission
Graph Analytics of Simulated Chat Data With Neo4j

Reporting and Presenting Your Work
Simulating Big Data for an Online Game

Formative Assessments:
5 Peer Review Assignments & 1 Graded Quiz

ASSESSMENT:

For summative assessments, Coursera will provide question banks for which exams can be conducted on the Coursera platform or the faculty will create their own assessments.

Note: If a Course or Specialization becomes unavailable prior to the end of the Term, Coursera may replace such Course or Specialization with a reasonable alternative Course or Specialization.