



AFRICAN ADVANCED LEVEL TELECOMMUNICATIONS INSTITUTE (AFRALTI)

COURSE OUTLINE

Title:	Big Data Analytics (BDA)
Duration:	5 Days

OVERVIEW:

Big Data analytics allow organizations to make better business decisions by deriving value from vast amounts of untapped data. This course provides the knowledge and skills to build competitive strategies around data-driven insights. Further, you leverage an overall lifecycle approach using sophisticated data modeling within your Big Data analytics projects.

TARGET AUDIENCE:

This course is intended for technical and database professionals, managers, data analysts, data scientists, businesses analysts and database professionals. The course will be useful for those professional involved in forecasting and trends management.

PRE-REQUISITE:

Programming and a background in statistics is helpful, but not required.

LEARNING OUTCOMES

- Optimize business decisions and create competitive advantage with Big Data analytics
- Perform predictive analysis using the appropriate data mining approach
- Use Python to create code that reads data from sensors and stores it in a SQL database.
- Visualize, clean, manipulate and integrate data sets.
- Learn fundamental principles of Big Data platforms like Hadoop.
- Use storytelling to present insights gained from extracted data.
- Derive business benefit from unstructured data
- Apply the data analytics life cycle approach to Big Data

COURSE CONTENTS:

Big Data & Analytics

- Chapter 0 Course Introduction
- Chapter 1 Data and the Internet of Things
- Chapter 2 Fundamentals of Data Analysis
- Chapter 3 Data Analysis
- Chapter 4 Advanced Data Analytics and Machine Learning
- Chapter 5 Storytelling with Data
- Chapter 6 Architecture for Big Data and Data Engineering

Course Outline

Course Introduction

Big Data and Analytics
A Global Community
More Than Just Information
Course Overview

Chapter 1: Data and the Internet of Things

- 1.0 Introduction
 - 1.0.1 Welcome
 - 1.0.1.1 Chapter 1: Data and the Internet of Things
- 1.1 Value of Data
 - 1.1.1 The Data Aspect of a Connected World
 - 1.1.1.1 Unique Role of Data
 - 1.1.1.2 What is Data?
 - 1.1.2 Data is Growing Exponentially
 - 1.1.2.1 Estimating Exponential Growth
 - 1.1.2.2 Growth of Data
 - 1.1.2.3 Lab – Demonstrate the Growth of Data with VNI
 - 1.1.3 Data Growth Changes Our Lives
 - 1.1.3.1 Data Growth Impact
 - 1.1.3.2 Business Example: Kaggle
 - 1.1.3.3 Social Example: DrivenData
 - 1.1.3.4 Environmental Example: Planetary Skin Institute
- 1.2 Data and Big Data
 - 1.2.1 Where Does Big Data Come From
 - 1.2.1.1 Defining Big Data
 - 1.2.1.2 Big Data Characteristics
 - 1.2.1.3 Sources of Big Data
 - 1.2.1.4 Real-World Examples of Big Data Sources
 - 1.2.2 Open Data and Private Data
 - 1.2.2.1 What is Open Data?
 - 1.2.2.2 What is Private Data?
 - 1.2.2.3 Lab – Explore Sources of Open Data
 - 1.2.2.4 Lab – Where is My Data?
 - 1.2.3 Structured and Unstructured Data

- 1.2.3.1 Structured Data
- 1.2.3.2 Unstructured Data
- 1.2.3.3 Activity – Compare Structured and Unstructured Data
- 1.2.4 Data at Rest and Data in Motion
 - 1.2.4.1 Nature of Data
 - 1.2.4.2 Data at Rest and Data in Motion
 - 1.2.4.3 Activity – Identify Data at Rest and Data in Motion
 - 1.2.4.4 Activity – What is Big Data?
- 1.3 Managing Big Data
 - 1.3.1 Evolution to Big Data
 - 1.3.1.1 How Did We Get Here?
 - 1.3.1.2 Big Data Infrastructure
 - 1.3.2 Basic Data Management Technologies
 - 1.3.2.1 Flat File Databases
 - 1.3.2.2 Relational Databases
 - 1.3.2.3 Distributed Data and Processing
 - 1.3.2.4 What is SQLite
 - 1.3.2.5 SQLite Features
 - 1.3.2.6 Lab – Demonstrate Spreadsheet Limitations in Data Analysis
 - 1.3.2.7 About Programming
 - 1.3.2.8 Labs – Introduction to PL-App
 - 1.3.2.9 Lab – Take the Python Challenge

Chapter 2 Fundamentals of Data Analysis

- 2.0 Introduction
 - 2.0.1 Welcome
 - 2.0.1.1 Chapter 2: Fundamentals of Data Analysis
- 2.1 What is Data Analysis?
 - 2.1.1 Analytics Models
 - 2.1.1.1 Data is Everywhere
 - 2.1.1.2 CRISP-DM
 - 2.1.1.3 Data Analytics Tool Capabilities
 - 2.1.1.4 The Role of Python in Data Analysis
- 2.2 Using Big Data
 - 2.2.1 Why Analyze Big Data?
 - 2.2.1.1 Big Data and Decision Making
 - 2.2.1.2 Data, Information, Knowledge, and Wisdom
 - 2.2.1.3 Activity - Identify the Layers of the DIKW Pyramid
 - 2.2.2 Types of Data Analysis
 - 2.2.2.1 Descriptive Analytics
 - 2.2.2.2 Predictive Analytics
 - 2.2.2.3 Prescriptive Analytics
 - 2.2.2.4 Activity – Identify the Types of Data Analysis
 - 2.2.2.5 Lab – Basic Data Analysis
 - 2.2.3 Timely Analysis of Big Data
 - 2.2.3.1 The Role of Time in Data Analytics
 - 2.2.3.2 Traditional Analytics to Big Data Analytics
 - 2.2.3.3 Next Generation Analytics
 - 2.2.4 Data Analysis Lifecycle
 - 2.2.4.1 The Scientific Method

- 2.2.4.2 Business Value
- 2.2.4.3 Data Analysis Lifecycle Example
- 2.2.4.4 Activity – Identify the Data Analysis Lifecycle Elements
- 2.2.4.5 Lab – San Francisco Crime
- 2.3 Data Acquisition and Preparation
 - 2.3.1 Sources of Data
 - 2.3.1.1 Files
 - 2.3.1.2 Internet
 - 2.3.1.3 Sensors
 - 2.3.1.4 Databases
 - 2.3.2 Data Preparation
 - 2.3.2.1 Data Types and Formats
 - 2.3.2.2 Data Structures
 - 2.3.2.3 Extract, Transform, and Load Data
 - 2.3.2.4 Extracting Data
 - 2.3.2.5 Transforming Data
 - 2.3.2.6 Loading Data
 - 2.3.2.7 Lab – Preparing Data
- 2.4 Big Data Ethics
 - 2.4.1 What are the Ethical Concerns?
 - 2.4.1.1 Current and Future Regulations
 - 2.4.1.2 Big Data Ethics Scenarios
 - 2.4.1.3 Data Security
 - 2.4.1.4 Data Security in the Cloud
 - 2.4.1.5 Lab - Big Data Ethics
- 2.5 Preparation for Chapter 2 Internet Meter Labs
 - 2.5.1 Part 1
 - 2.5.1.1 Formatting Time and Date Data
 - 2.5.1.2 Reading and Writing Files
 - 2.5.1.3 Interacting with External Applications
 - 2.5.1.4 Lab - Internet Meter Data Analysis
 - 2.5.2 Part 2
 - 2.5.2.1 SQL
 - 2.5.2.2 Basic SQL Operations
 - 2.5.2.3 Working with Python and SQLite
 - 2.5.2.4 Lab – Working with Python and SQLite
 - 2.5.2.5 Lab – Internet Meter SQL
- 2.6 Summary

Chapter 3: Data Analysis

- 3.0 Introduction
 - 3.0.1 Welcome
 - 3.0.1.1 Chapter 3: Data Analysis
- 3.1 Analyzing Data
 - 3.1.1 Preliminaries
 - 3.1.1.1 Exploratory Data Analysis
 - 3.1.1.2 Analyzing IoT Data
 - 3.1.1.3 Observations, Variables, and Values
 - 3.1.1.4 Types of Variables
 - 3.1.1.5 Activity - Identify Data Analysis Terms

- 3.1.2 Statistical Analysis
 - 3.1.2.1 What is Statistics?
 - 3.1.2.2 Populations and Samples
 - 3.1.2.3 Descriptive Statistics
 - 3.1.2.4 Inferential Statistics
 - 3.1.2.5 Statistics and Big Data
 - 3.1.2.6 Activity - Identify the Components of the IoT Analysis Pipeline
- 3.1.3 Characteristics of Samples
 - 3.1.3.1 Distributions
 - 3.1.3.2 Centrality
 - 3.1.3.3 Activity - Identify the Characteristics of Distributions
 - 3.1.3.4 Dispersion
- 3.1.4 Analysis Using Descriptive Statistics
 - 3.1.4.1 Using pandas
 - 3.1.4.2 Importing Data from Files
 - 3.1.4.3 Importing Data from the Web
 - 3.1.4.4 Descriptive Statistics in pandas
 - 3.1.4.5 Activity - Create Code for Each Function
 - 3.1.4.6 Lab - Descriptive Statistics in Python
- 3.1.5 Analysis Using Correlation
 - 3.1.5.1 Correlation vs. Causation
 - 3.1.5.2 Correlation Coefficient
 - 3.1.5.3 Correlation in pandas
 - 3.1.5.4 Appropriate Visualizations
 - 3.1.5.5 Lab - Correlation Analysis in Python
- 3.2 Preparation for Chapter 3 Internet Meter Lab
 - 3.2.1 Basic Analysis with pandas.
 - 3.2.1.1 Issues with Data Quality
 - 3.2.1.2 Dealing with Missing Data
 - 3.2.1.3 Converting Data Types
 - 3.2.1.4 Manipulating Dataframes
 - 3.2.1.5 Basic Data Statistics
 - 3.2.1.6 Lab - Internet Meter Visualization
- 3.3 Summary

Chapter 4: Advanced Data Analytics and Machine Learning

- 4.0 Introduction
 - 4.0.1 Welcome
 - 4.0.1.1 Chapter 4: Advanced Data Analytics and Machine Learning
- 4.1 Predictive Analytics
 - 4.1.1 Machine Learning
 - 4.1.1.1 Machine Learning: Looking Ahead
 - 4.1.1.2 What is Machine Learning?
 - 4.1.1.3 Types of Machine Learning Analysis
 - 4.1.1.4 A Machine Learning Process
 - 4.1.1.5 Common Applications of Machine Learning
 - 4.1.1.6 Interactive Activity - Identify Machine Learning Terms
 - 4.1.2 Regression
 - 4.1.2.1 Regression Analysis
 - 4.1.2.2 Linear Regression

- 4.1.2.3 Applications of Regression Analysis
- 4.1.2.4 Lab – Simple Linear Regression in Python
- 4.1.3 Classification
 - 4.1.3.1 Classification Problems
 - 4.1.3.2 Classification Algorithms
 - 4.1.3.3 Visualizing Classifications
 - 4.1.3.4 Applications of Classification
 - 4.1.3.5 Lab – Decision Tree Classification
- 4.2 Model Evaluation
 - 4.2.1 Validity and Reliability
 - 4.2.1.1 Issues in Using Analysis
 - 4.2.1.2 Validity
 - 4.2.1.3 Reliability
 - 4.2.2 Error in Analyses
 - 4.2.2.1 Error in Data Analytics
 - 4.2.2.2 Types and Sources of Measurement Error
 - 4.2.2.3 Errors in Predictive Analytics
 - 4.2.2.4 Interactive Activity - Identify the Types of Errors
 - 4.2.2.5 Lab – Evaluating Fit Errors in Linear Regression
 - 4.2.3 Evaluating Analytic Models
 - 4.2.3.1 Misleading Research
 - 4.2.3.2 Sensationalism in Research Findings
 - 4.2.3.3 Guidelines for Evaluating Results
- 4.3 Preparation for Chapter 4 Labs
 - 4.3.1 Regression and Prediction Lab
 - 4.3.1.1 Using scikit-learn for Regression Analysis
 - 4.3.1.2 Style Sheets for Plots
 - 4.3.1.3 Fitting the Data
 - 4.3.1.4 Lab - Internet Meter Linear Regression
 - 4.3.2 Lab - Internet Meter Anomaly Detection
 - 4.3.2.1 Plotting in 3D
 - 4.3.2.2 Interacting with a 3D Plot
 - 4.3.2.3 Detecting Anomalies
 - 4.3.2.4 Lab - Internet Meter Anomaly Detection
- 4.4 Summary

Chapter 5: Storytelling with Data

- 5.0 Introduction
 - 5.0.1 Welcome
 - 5.0.1.1 Chapter 5: Storytelling with Data
- 5.1 Building a Data Story
 - 5.1.1 Know Your Purpose
 - 5.1.1.1 Telling a Story
 - 5.1.1.2 Audience
 - 5.1.1.3 Business Value and Goal
 - 5.1.2 Proposition and Evidence
 - 5.1.2.1 Using Evidence
 - 5.1.2.2 Deductive Reasoning
 - 5.1.2.3 Inductive Reasoning

- 5.1.2.4 Fallacies
- 5.1.2.5 Activity - Identify the Type of Reasoning
- 5.2 The Power of Visualization
 - 5.2.1 Pyplot
 - 5.2.1.1 Introduction to Pyplot
 - 5.2.1.2 Modifying the Default Style Inline
 - 5.2.1.3 Creating a Style Sheet
 - 5.2.1.4 Referencing a Style Sheet
 - 5.2.1.5 Matplotlib Style Sheets
 - 5.2.2 Plotly
 - 5.2.2.1 Introduction to Plotly
 - 5.2.2.2 Plotly Create
 - 5.2.2.3 Plotly Export
 - 5.2.2.4 Plotly in Offline Mode
 - 5.2.2.5 Explore the Power of Plotly
 - 5.2.3 Choosing the Right Visualization for the Job
 - 5.2.3.1 Common Types of Data Visualizations
 - 5.2.3.2 Line Charts
 - 5.2.3.3 Column Charts
 - 5.2.3.4 Bar Charts
 - 5.2.3.5 Pie Charts
 - 5.2.3.6 Scatter Plot
 - 5.2.3.7 Activity - Select the Best Visualization
 - 5.2.3.8 Lab - Visualizing Data in Excel
- 5.3 Preparation for Chapter 5 Labs
 - 5.3.1 Chapter 5 Labs
 - 5.3.1.1 Folium Library
 - 5.3.1.2 Folium Tilesets
 - 5.3.1.3 Modifying and Labeling a Map
 - 5.3.1.4 Lab - Advanced Data Visualization
 - 5.3.1.5 Lab - Internet Speed Compliance
- 5.4 Summary

Chapter 6: Architecture for Big Data and Data Engineering

- 6.0 Introduction
 - 6.0.1 Welcome
 - 6.0.1.1 Chapter 6: Architecture for Big Data and Data Engineering
- 6.1 Scaling Data Analytics
 - 6.1.1 The Cloud, the Fog and the Edge
 - 6.1.1.1 Edge Analytics and Cloud Analytics
 - 6.1.1.2 The Data Center
 - 6.1.1.3 Data Centers and Cloud Computing
 - 6.1.1.4 Data Center Structure
 - 6.1.1.5 Benefits of a Data Center
 - 6.1.1.6 Interactive Activity – Identify Data Center Benefits
 - 6.1.1.7 Issues in Data Center Security
 - 6.1.2 Virtualization
 - 6.1.2.1 What is Virtualization?
 - 6.1.2.2 Abstraction Layers
 - 6.1.2.3 Hypervisors

- 6.1.2.4 Containers
- 6.1.3 The Virtualized Data Center
 - 6.1.3.1 SaaS, PaaS, and IaaS
 - 6.1.3.2 Virtualized Data Storage
 - 6.1.3.3 Virtualized Network
 - 6.1.3.4 Lab – Install a Virtual Machine on a Personal Computer
- 6.2 Introduction to Data Engineering
 - 6.2.1 History of Data Engineering
 - 6.2.1.1 What is Data Engineering?
 - 6.2.1.2 The Role of the Data Engineer
 - 6.2.2 Big Data Systems
 - 6.2.2.1 Scalability with Big Data
 - 6.2.2.2 Availability with Big Data
 - 6.2.2.3 Fault Tolerance with Big Data
 - 6.2.3 What is Hadoop?
 - 6.2.3.1 How Hadoop Works: The HDFS
 - 6.2.3.2 How Hadoop Works: MapReduce
 - 6.2.3.3 The Evolution of Hadoop
 - 6.2.3.4 Video Tutorial - Big Data with Hadoop
- 6.3 The Big Data Pipeline
 - 6.3.1 Data Ingestion
 - 6.3.1.1 The Problem of Data Ingestion
 - 6.3.1.2 What Is Kafka?
 - 6.3.1.3 What Are the Advantage of Kafka vs Other Approaches?
 - 6.3.2 Data Storage
 - 6.3.2.1 The Problem of Data Storage
 - 6.3.2.2 What is Cassandra: NoSQL DB ?
 - 6.3.2.3 What Are the Advantages of Cassandra vs Hadoop for Storage?
 - 6.3.3 Compute
 - 6.3.3.1 The Problem of Computing Function
 - 6.3.3.2 What is Spark: Computing in Memory with RDDs
 - 6.3.3.3 What Are the Advantages of Spark vs MapReduce?
 - 6.3.4 The Lambda Architecture
 - 6.3.4.1 The Architecture: Batch Layer, Speed Layer, Serving Layer
 - 6.3.4.2 Lambda Architecture Example: Floating Bus Data
- 6.4 The Image Processing Labs
 - 6.4.1 Digital Images as Data
 - 6.4.1.1 Setting the Stage
 - 6.4.1.2 Lab - Install and Test the Raspberry Pi Camera
 - 6.4.1.3 Lab - Image Processing Change Detection
 - 6.4.1.4 Lab – Smile Detection
- 6.5 Summary

Lab Activities

Chapter 1 Data and the Internet of Things

- 1.1 Value of Data
 - 1.1.2 Data is Growing Exponentially
 - 1.1.2.3 Lab – Demonstrate the Growth of Data with VNI
- 1.2 Data and Big Data
 - 1.2.2 Open Data and Private Data
 - 1.2.2.3 Lab – Explore Sources of Open Data
 - 1.2.2.4 Lab – Where is My Data?
- 1.3 Managing Big Data
 - 1.3.2 Basic Data Management Technologies
 - 1.3.2.6 Lab – Demonstrate Spreadsheet Limitations in Data Analysis
 - 1.3.2.8 Labs – Introduction to PL-App
 - 1.3.2.9 Lab – Take the Python Challenge

Chapter 2 Fundamentals of Data Analysis

- 2.2 Using Big Data
 - 2.2.2 Types of Data Analysis
 - 2.2.2.5 Lab – Basic Data Analysis
 - 2.2.4 Data Analysis Lifecycle
 - 2.2.4.5 Lab – San Francisco Crime
- 2.3 Data Acquisition and Preparation
 - 2.3.2 Data Preparation
 - 2.3.2.7 Lab – Preparing Data
- 2.4 Big Data Ethics
 - 2.4.1 What are the Ethical Concerns?
 - 2.4.1.5 Lab - Big Data Ethics
- 2.5 Preparation for Chapter 2 Internet Meter Labs
 - 2.5.1 Part 1
 - 2.5.1.4 Lab - Internet Meter Data Analysis
 - 2.5.2 Part 2
 - 2.5.2.4 Lab – Working with Python and SQLite
 - 2.5.2.5 Lab – Internet Meter SQL

Chapter 3 Data Analysis

- 3.1 Analyzing Data
 - 3.1.4 Analysis Using Descriptive Statistics
 - 3.1.4.6 Lab - Descriptive Statistics in Python
 - 3.1.5 Analysis Using Correlation
 - 3.1.5.5 Lab – Correlation Analysis in Python
- 3.2 Preparation for Chapter 3 Internet Meter Lab
 - 3.2.1 Basic Analysis with pandas.
 - 3.2.1.6 Lab – Internet Meter Visualization

Chapter 4 Advanced Data Analytics and Machine Learning

- 4.1 Predictive Analytics
 - 4.1.2 Regression

- 4.1.2.4 Lab – Simple Linear Regression in Python
- 4.1.3 Classification
 - 4.1.3.5 Lab – Decision Tree Classification
- 4.2 Model Evaluation
 - 4.2.2 Error in Analyses
 - 4.2.2.5 Lab – Evaluating Fit Errors in Linear Regression
- 4.3 Preparation for Chapter 4 Labs
 - 4.3.1 Regression and Prediction Lab
 - 4.3.1.4 Lab - Internet Meter Linear Regression
 - 4.3.2 Lab - Internet Meter Anomaly Detection
 - 4.3.2.4 Lab - Internet Meter Anomaly Detection

Chapter 5 Storytelling with Data

- 5.2 The Power of Visualization
 - 5.2.3 Choosing the Right Visualization for the Job
 - 5.2.3.8 Lab - Visualizing Data in Excel
- 5.3 Preparation for Chapter 5 Labs
 - 5.3.1 Chapter 5 Labs
 - 5.3.1.4 Lab - Advanced Data Visualization
 - 5.3.1.5 Lab - Internet Speed Compliance

Chapter 6 Architecture for Big Data and Data Engineering

- 6.1 Scaling Data Analytics
 - 6.1.3 The Virtualized Data Center
 - 6.1.3.4 Lab – Install a Virtual Machine on a Personal Computer
- 6.4 The Image Processing Labs
 - 6.4.1 Digital Images as Data
 - 6.4.1.2 Lab - Install and Test the Raspberry Pi Camera
 - 6.4.1.3 Lab - Image Processing Change Detection
 - 6.4.1.4 Lab – Smile Detection

For more information, please contact us on

Tel: +254 710 207 061, + 254 733 444 421

Email: training@afraiti.org