Government of Pakistan

National Vocational and Technical Training Commission

Prime Minister Hunarmand Pakistan Program,
"Skills for All"

Course Contents/ Lesson Plan

Course Title: Big Data Analytics
Duration: 6 Months

Plot no. 38, Kirthar Road, H-9 Islamabad
051-9044250

Trainer Name	
Course Title	**Big Data Analytics**
Objective of Course	**Employable skills and hands on practice for Web Development, Graphic Designing and Mobile App Development** The main goal of this course is to help students learn, understand, and practice big data analytics and machine learning approaches, which include the study of modern computing big data technologies and scaling up machine learning techniques focusing on industry applications. Mainly the course objectives are: conceptualization and summarization of big data and machine learning, trivial data versus big data, big data computing technologies, machine learning techniques, and scaling up machine learning approaches.
Learning Outcome of the Course	The students learning outcomes are designed to specify what the students will be able to perform after completion of the course: • Ability to identify the characteristics of datasets and compare the trivial data and big data for various applications. • Ability to select and implement machine learning techniques and computing environment that are suitable for the applications under consideration. • Ability to solve problems associated with batch learning and online learning, and the big data characteristics such as high dimensionality, dynamically growing data and in particular scalability issues. • Ability to understand and apply scaling up machine learning techniques and associated computing techniques and technologies. • Ability to recognize and implement various ways of selecting suitable model parameters for different machine learning techniques. • Ability to integrate machine learning libraries and mathematical and statistical tools with modern technologies like Apache Spark.
Course Execution Plan	Total Duration of Course: **6 Months (26 Weeks)** Class Hours: **5 Hours per day** Theory: **20% Practical: 80%** Weekly Hours: **25 Hours Per week** Total Contact Hours: **650 Hours**

Companies Offering Jobs in the respective trade	1. Upwork 2. Freelancer 3. Fiverr 4. Government Institutes 5. Software Houses 6. Companies all over the world are offering its jobs as they want to know the trends of market
Job Opportunities	Upskilling in Big Data and Analytics field is a smart career decision. According to Allied Market Research, the globalmarket of only Hadoop/Spark will reach $84.6 Billion by 2021 and there is a shortage of 1.4-1.9 million Hadoop/Spark data analysts in the U.S. alone. Here is selection of specialist opportunities in your area: • Big Data Architect (Average Salary: 124000$ / Annum) • Big Data Engineer (Average Salary: 117000$ / Annum) • Big Data Developer (Average Salary: 88500$ / Annum)
No of Students	25
Learning Place	Classroom / Lab
Instructional Resources	**Development Platform:** • https://github.com/ , • https://spark.apache.org/, • https://www.edureka.co/apache-spark-scala-certification-training, • https://www.youtube.com/watch?v=iP1wOSsKjW8&list=PLS1QulWo1RIahlYDqHWZb81qsKgEvPiHn, • https://stackoverflow.com/ **Learning Material:** • https://spark.apache.org/docs/latest/api/python/index.htmlhttps://www.youtube.com/watch?v=9mELEARcxJo&list=PL9ooVrP1hQOGyFc60sExNX1qBWJyV5IMb • https://www.youtube.com/watch?v=Uct_EbThV1E&list=PLZ7s-Z1aAtmIbaEj_PtUqkqdmI1k7libK • https://www.edureka.co/apache-spark-scala-certification-training • https://www.youtube.com/watch?v=wjfeGxqAQOY&list=PLrjkTql3jnm-CLxHftqLgkrZbM8fUt0vn • https://www.youtube.com/watch?v=iP1wOSsKjW8&list=PLS1QulWo1RIahlYDqHWZb81qsKgEvPiHn

Scheduled Week	Module Title	Learning Units	Remarks
Week 1	➢ Introduction	• **Motivational Lecture** • **Course Introduction** • **Success stories** • **Job market** • **Course Applications** • **Institute/work ethics** • Discussion on Python and its market position. • Motivation regarding learning aspects of this course • Setting up environment for Python. • Installation of Anaconda • What is Big Data? • Characteristics of Big Data • The Impact of Big Data • Big Data - Beyond the Hype, Big Data Examples, Sources of Big Data • Big Data Adoption, The Big Data and Data Science • The Big Data Platform, Big Data and Data Science. Skills for Data Scientists	
Week 2	**Module -1** **Chapter 1.1-**	**Overview of DBMS** • Components of DBMS • Database Architecture • Types of Database Model • ER Model: Basic Concepts • ER Model: Creating ER Diagram • The Extended ER Model • Codd's 12 rule of RDBMS • Basic Concepts of RDBMS • Types of Database key • Introduction to Normalization **Basic SQL** • SQL Introduction • Create query • Alter query • Truncate, Drop and Rename query • All DML command • All TCL Command • All DCL Command • WHERE clause • SELECT query • LIKE clause	

		• ORDER BY clause • Group BY clause • Having clause • DISTINCT keyword • AND & OR operator • DIVISION operator **Advanced SQL** • SQL Constraints • SQL function • SQL Join • SQL Alias • SQL SET operation • SQL Sequences • SQL Views	
Week 3	**Chapter 1.2-**	• Types of IDE(s) and IDE that will be used in the duration of this course. e.g. Spyder, Jupyteretc • Hello World Program “Print Command” • Keyword Types • Expressions and Variables • Input Method • Conditions and Branching • Loops	
Week 4	**Chapter 2.1**	• String Operations • Lists and Tuples • Sets • Dictionaries • Reading and Writing files • Functions • Objects and Classes	
Week 5	**Chapter 2.2**	• Introduction with Numpy • Numpy one dimensional Array • Numpytwo-dimensional Array • Numpy Array Operations	
Week 6	**Chapter 3.1**	• Descriptive Statistics • Data Manipulation • Data Wrangling	
Week 7	**Chapter 3.2**	• Working with Pandas • Descriptive Statistics with Pandas • Group by with Python	

		• Data Manipulation with Pandas	
Week 8	**Chapter 4**	• Data Wrangling with Pandas • Discussion regarding exam	
Week 9	**Chapter 5.1**	• Introduction to Matplotlib • Basic Plotting with Matplotlib • Line Plots • Area Plots • Histograms	
Week 10	**Chapter 5.2**	• Bar Charts • Pie Charts • Box Plots • Scatter Plots • Word Cloud	
Week 11	**Chapter 6.1**	• What is Spark and what is its purpose? • Components of the Spark unified stack • Resilient Distributed Dataset (RDD) • Scala and Python overview	
Week 12	**Chapter 6.2**	• Understand how to create parallelized collections and external datasets • Work with Resilient Distributed Dataset (RDD) operations • Utilize shared variables and key-value pairs	
Week 13	**Chapter 6.3**	• Describe and run some Spark examples • Pass functions to Spark • Create and run a Spark standalone application	
Week 14	**Chapter 6.4**	• Understand and use the various Spark libraries	
Week 15	**Mid-Term Assignment**		
Week 16	**Chapter 7** Apache Shark Next-Generation Big Data Framework	• Apache Spark Next-Generation Big Data Framework • History of Spark • Why we should prefer spark? • Introduction to Apache Spark • Components of Spark • Application of In-memory Processing • Hadoop Ecosystem vs Spark • Advantages of Spark • Spark Architecture	

		• Spark Cluster in Real World • Demo: Running a Scala Programs in Spark Shell • Demo: Setting Up Execution Environment in IDE • Demo: Spark Web UI • Key Takeaways • Knowledge Check • Practice Project: Apache Spark Next-Generation Big Data Framework	
Week 17	**Chapter 8** Spark Core Processing RDD	• Introduction to Spark RDD • RDD in Spark • Creating Spark RDD • Pair RDD • RDD Operations • Demo: Spark Transformation Detailed Exploration Using Scala Examples • Demo: Spark Action Detailed Exploration Using Scala • Caching and Persistence • Storage Levels • Lineage and DAG • Need for DAG • Debugging in Spark • Partitioning in Spark • Scheduling in Spark • Shuffling in Spark • Sort Shuffle • Aggregating Data with Paired RDD • Demo: Spark Application with Data Written Back to HDFS and Spark UI • Demo: Changing Spark Application Parameters • Demo: Handling Different File Formats • Demo: Spark RDD with Real-world Application • Demo: Optimizing Spark Jobs • Key Takeaways • Knowledge Check • Practice Project: Spark Core Processing RDD	
Week 18	**Chapter 9** Spark SQL Processing DataFrames	• Spark SQL Processing DataFrames • Spark SQL Introduction • Spark SQL Architecture • Dataframes • Demo: Handling Various Data Formats	

		• Demo: Implement Various Dataframe Operations • Demo: UDF and UDAF • Interoperating With RDDs • Demo: Process Dataframe Using SQL Query • RDD vs Dataframe vs Dataset • Practice Project: Processing Dataframes • Key Takeaways • Knowledge Check • Practice Project: Spark SQL - Processing Dataframes	
Week 19	**Chapter 10.1** **Part 1** Spark Mlib Modelling BigData with Spark	• Spark Mlib Modeling Big Data With Spark • Role of Data Scientist and Data Analyst in Big Data • Analytics in Spark • Machine Learning • Supervised Learning • Demo: Classification of Linear SVM • Demo: Linear Regression With Real World Case Studies • Unsupervised Learning Demo: Unsupervised Clustering K-means	
Week 20	**Chapter 10.2** **Part 2** Spark Mlib Modelling BigData with Spark	• Reinforcement Learning • Semi-supervised Learning • Overview of Mlib • Mlib Pipelines • Key Takeaways • Knowledge Check • Practice Project: Spark Mlib - Modelling Big data With Spark	
Week 21	Employable Project/Assignment (6 weeks i.e. 21-26) in addition of regular classes. **OR** On job training (2 weeks)	• Guidelines to the Trainees for selection of students employable project like final year project (FYP) • Assign Independent project to each Trainee • A project based on trainee’s aptitude and acquired skills. • Designed by keeping in view the emerging trends in the local market as well as across the globe. • The project idea may be based on Entrepreneur. • Leading to the successful employment. • The duration of the project will be 6	

		weeks • Ideas may be generated via different sites such as: https://1000projects.org/ https://nevonprojects.com/ https://www.freestudentprojects.com/ https://technofizi.net/best-computer-science-and-engineering-cse-project-topics-ideas-for-students/ • Final viva/assessment will be conducted on project assignments. • At the end of session the project will be presented in skills competition • The skill competition will be conducted on zonal, regional and National level. • The project will be presented in front of Industrialists for commercialization • The best business idea will be placed in NAVTTC business incubation center for commercialization. --- **OR** **On job training for 2 weeks:** • Aims to provide 2 weeks industrial training to the Trainees as part of overall training program • Ideal for the manufacturing trades • As an alternate to the projects that involve expensive equipment • Focuses on increasing Trainee’s motivation, productivity, efficiency and quick learning approach.	
Week 22	**Chapter 11.1** **Part 1** Stream Processing Frameworks and Spark Streaming	• Streaming Overview • Real-time Processing of Big Data • Data Processing Architectures • Demo: Real-time Data Processing • Spark Streaming • Demo: Writing Spark Streaming Application • Introduction to DStreams • Transformations on DStreams • Design Patterns for Using Foreachrdd • State Operations • Windowing Operations • Join Operations Stream-dataset Join • Demo: Windowing of Real-time Data	

		Processing ● Streaming Sources Demo: Processing Twitter Streaming Data ● Structured Spark Streaming ● Use Case Banking Transactions ● Structured Streaming Architecture Model and Its Components ● Output Sinks	
Week 23	**Chapter 11.2** **Part 2** Stream Processing Frameworks and Spark Streaming	● Structured Streaming APIs ● Constructing Columns in Structured Streaming ● Windowed Operations on Event-time ● Use Cases ● Demo: Streaming Pipeline ● Practice Project: Spark Streaming ● Key Takeaways ● Knowledge Check ● Practice Project: Stream Processing Frameworks and Spark Streaming	
Week 24	**Chapter 12.1** **Part 1** Spark GraphX	• Spark GraphX • Introduction to Graph • GraphX in Spark • GraphX Operators • Join Operators • GraphX Parallel System • Algorithms in Spark	
Week 25	**Chapter 12.2** **Part 2** Spark GraphX	● Pregel API ● Use Case of GraphX ● Demo: GraphX Vertex Predicate ● Demo: Page Rank Algorithm ● Key Takeaways ● Knowledge Check ● Practice Project: Spark GraphX Project Assistance ● **Final Project Assessment**	
Week 26	Entrepreneurship and Final Assessment in project	• Job Market Searching • Self-employment • Freelancing sites • Introduction • Fundamentals of Business Development • Entrepreneurship • Startup Funding • Business Incubation and Acceleration • Business Value Statement • Business Model Canvas • Sales and Marketing Strategies • How to Reach Customers and Engage CxOs	

		• Stakeholders Power Grid • RACI Model, SWOT Analysis, PEST Analysis • SMART Objectives • OKRs • Cost Management (OPEX, CAPEX, ROCE etc.) • Final Assessment	

List of Machinery / Equipment

Sr. No	Name of item as per curriculum	Quantity physically available at the training location
1	Computers Minimum Corei5 • LCD Display 17” with built in speakers	25
2	DSL Internet Connection (Minimum 1 MB)	Available on every PC
3	**Accessories/Devices** • Connectors • Multimedia • Printer (NW printer) • Audio/visual aid • White Board • Pin Board • Flip Chart Board • Hard copy of Training Material • Mobile Phones	25 each
4	**Wires, data cables, power plugs, power supply**	For every PC
5	**UPS**	Available
6	**Generator / Solar Backup**	Available
7	**Air Conditioner (2 Tons)**	Available

1. **Software List**

Sr. No	Software Name
1.	MS Office 2016 (Installed on each PC)
2.	Operating System (Windows, Linux or other Operating Systems)
3.	Programming Languages including NetBeans, Android studio (Licensed
4.	Web Servers including IIS, Apache (Licensed software installed on each PC)
5.	Databases including MySQL, ERWIN (Licensed software installed on each PC)
6.	FTP Client including FileZilla, File Manager (Licensed software installed on each PC)
7.	Web hosting manager/control panel
8.	Web browser including Internet Explorer, Google Chrome, Mozilla Firefox, Netscape, Opera (installed on each PC)
9.	Firewall (each PC)
10.	Security scanning tools including Antivirus (each PC) Networking
11.	Required Software's: • Anaconda Jupyter • MySQL Database • MS Office • MS Visio • MySQL

2. **Minimum Qualification of Teachers / Instructor**

The qualification of teachers / instructor of this course should be minimum **of bachelors in Computer science with minimum 3 years of development experience** in relevant trade.

- Bachelors of Computers Science / Networks (Hons)

3. **Supportive Notes**

Teaching Learning Material

Books Name	**Author**
Python Crash Course	Eric Matthes

Big Data Analysis with Python	Ankit Shukla,Ivan Marin and Sarang VK
Big Data Course (Edureka Online Course)	