Antrix Academy of Data Science

Helpline No. 8130-371-253
Email Id mail@antrixacademy.com
Website www.antrixacademy.com

Get Certified with Microsoft*
Exam 98-381 : Introduction to Programming Using Python

Microsoft Technology Associate

Course: Data Science - Master Program

Duration: 5 Months (Weekend)

Microsoft Technology Associate Certificate Voucher would to be given to every participant

Python	Statistical Fundamentals	Machine Learning	R	Hadoop & PySpark	MS SQL

Introduction to Python Programming

- Why do we need Python?
- Program structure in Python

Execution steps

- Interactive Shell
- Executable or script files.
- User Interface or IDE
- Introduction to Jupyter Editor

Data Types and Operations

- Numbers
- Strings
- List
- Tuple
- Dictionary
- Other Core Types

Statements and Syntax in Python

- Assignments, Expressions and prints
- If tests and Syntax Rules
- While and For Loops
- Iterations and Comprehensions

Functions in Python

- Function definition and call
- Function Scope
- Arguments
- Function Objects
- Anonymous Functions

File Operations

- Opening a file
- Using Files
- Other File tools

Data Analysis with pandas

- Using Series, DataFrame, Panels
- Data wrangling
- Sorting and filtering data
- Aggregate operations
- Analyzing time series
- Visualization with Pandas

Vectorizing Data in Numpy

- Creating Numpy arrays
- Common operations on matrices
- Using Analytics functions
- Views and broadcasting on Numpy arrays
- Optimizing performance by avoiding loops

Python: Data Manipulation – cleansing

- Cleansing Data with Python
- Data Manipulation steps(Sorting, filtering, duplicates, merging, appending, subsetting, derived variables, sampling, Data type conversions, renaming, formatting etc)
- Data manipulation tools(Operators, Functions, Packages, control structures, Loops, arrays etc)
- Python Built-in Functions (Text, numeric, date, utility functions)
- Python User Defined Functions
- Stripping out extraneous information
- Normalizing data
- Formatting data
- Important Python Packages for data manipulation (Pandas, Numpy etc)

Python: Accessing/Importing and Exporting Data

- Importing Data from various sources (Csv, txt, excel, access etc)
- Database Input (Connecting to database – MySQL, MS SQL, Oracle, Teradata)
- Viewing Data objects - subsetting, methods
- Exporting Data to various formats

Python: Data Analysis – Visualization

- Introduction exploratory data analysis
- Descriptive statistics, Frequency Tables and summarization
- Univariate Analysis (Distribution of data & Graphical Analysis)
- Bivariate Analysis(Cross Tabs, Distributions & Relationships, Graphical Analysis)
- Creating Graphs- Bar/pie/line chart/histogram/boxplot/scatter/density etc)
- Important Packages for Exploratory Analysis(NumPy Arrays, Matplotlib, Pandas and scipy.stats etc)

Machine Learning (Supervised Learning) - I

- Generalised Linear Models
  - Linear Regression
  - Ridge and Lasso Regression

    - Logistic Regression
- Classification
    - Random Forest
    - Decision Trees
    - Support Vector Machines
    - KNN
    - Naïve Bayes
    - Usage

Machine Learning (Unsupervised Learning) - II

- Clustering
    - K-Means
    - K Nearest Neighbours
    - Association Rule Learning
- Reinforcement Learning
    - Markov Decision
    - Monte Carlo Prediction

Introduction and Orientation

- Introduction to Data Science and R. Application and Uses case of R
- Introduction R/R-Studio - GUI
- Concept of Packages - Useful Packages (Base & other packages) in R

R Data Structure and its operation

- Variable & Value Labels – Date Values
- Data Types- Numeric, Integer, Factor, Boolean, Dates and Logical
- Vectors, Matrices, factors, Data frames, and Lists
- Importing Data from various sources
- Exporting Data to various formats)
- Viewing Data (Viewing partial data and full data)
- Missing Values
- Sequences of Numbers

Data Wrangling

- Data Manipulation steps- Sorting, Filtering, Duplicates, Merging, Appending, Sub-setting, Derived variables, Sampling, Data type conversions, renaming, formatting.
- Control Structures-if, If-else, Nested if-else
- Control Structures - Loops and advance loop functions
- R User Defined functions- Create your own functions
- R Operators
- Data Reshaping- Long to wide vice-versa
- Playing with Textual Data-Editing Textual data, regular expressions
- Data Aggregation and Summarization

Intro to Stats and Data Analysis

- Introduction exploratory data analysis (EDA)
- Descriptive statistics-Random sampling, Correlation, Central Limit Theorem, Variance Frequency Tables and summarization
- Univariate Analysis (Distribution of data & Graphical Analysis)
- Bivariate Analysis(Cross Tabs, Distributions & Relationships)

- Data Visualization
- Base Plotting System
- Exploratory data analysis using plots
- Univariate and Bi-variate plots
- Creating Graphs- Bar/pie/line chart/histogram/boxplot/scatter/density)

Introduction to Big Data & Hadoop Ecosystem

- Introduction and relevance
- Uses of Big Data analytics in various industries like Telecom, E- commerce, Finance and Insurance etc.
- Problems with Traditional Large-Scale Systems
- Motivation for Hadoop
- Different types of projects by Apache
- Role of projects in the Hadoop Ecosystem
- Key technology foundations required for Big Data
- Limitations and Solutions of existing Data Analytics Architecture
- Comparison of traditional data management systems with Big Data management systems
- Evaluate key framework requirements for Big Data analytics
- Hadoop Ecosystem & Hadoop 2.x core components
- Explain the relevance of real-time data
- Explain how to use big and real-time data as a Business planning tool

Hadoop Cluster -Architecture - Configuration files

- Hadoop Master-Slave Architecture
- The Hadoop Distributed File System - Concept of data storage
- Explain different types of cluster setups(Fully distributed/Pseudo etc)
- Hadoop cluster set up - Installation
- Hadoop 2.x Cluster Architecture
- A Typical enterprise cluster – Hadoop Cluster Modes
- Understanding cluster management tools like Cloudera manager/Apache ambari

Hadoop Core Components - HDFS & MapReduce(YARN)

- HDFS Overview & Data storage in HDFS
- Get the data into Hadoop from local machine(Data Loading Techniques) - vice versa
- Map Reduce Overview (Traditional way Vs. MapReduce way)
- Concept of Mapper & Reducer
- Understanding MapReduce program Framework
- Develop MapReduce Program using Java (Basic)
- Develop MapReduce program with streaming API) (Basic)

Data Integration Using SQOOP & FLUME

- Integrating Hadoop into an Existing Enterprise
- Loading Data from an RDBMS into HDFS by Using Sqoop
- Managing Real-Time Data Using Flume
- Accessing HDFS from Legacy Systems

Data Analysis using PIG Data Analysis Using PIG

- Apache PIG - MapReduce Vs Pig, Pig Use Cases
- PIG's Data Model
- PIG Streaming

- Pig Latin Program & Execution
- Pig Latin : Relational Operators, File Loaders, Group Operator, COGROUP Operator, Joins and COGROUP, Union, Diagnostic Operators, Pig UDF
- Writing JAVA UDF's
- Embedded PIG in JAVA
- PIG Macros
- Parameter Substitution
- Use Pig to automate the design and implementation of MapReduce applications
- Use Pig to apply structure to unstructured Big Data

Data Analysis Using Hive

- Apache Hive - Hive Vs. PIG - Hive Use Cases
- Discuss the Hive data storage principle
- Explain the File formats and Records formats supported by the Hive environment
- Perform operations with data in Hive
- Hive QL: Joining Tables, Dynamic Partitioning, Custom Map/Reduce Scripts
- Hive Script, Hive UDF
- Hive Persistence formats
- Loading data in Hive - Methods
- Serialization & Deserialization
- Handling Text data using Hive
- Integrating external BI tools with Hadoop Hive

Processing Distributed Data with Apache Spark

- What is Spark
- Spark Ecosystem
- Spark Components
- What is Scala
- Why Scala
- SparkContext
- Spark RDD

SQL Overview

- Outlining SQL as the cornerstone of database activity
- Applying the ANSI/ISO standards
- Describing the fundamental building blocks: tables, columns, primary keys and foreign keys

Building the Database Schema

- Creating tables and columns
- Building tables with CREATE TABLE
- Modifying table structure with ALTER TABLE
- Adding columns to an existing table
- Removing tables with DROP TABLE

Protecting data integrity with constraints

- Guaranteeing uniqueness with primary key constraints
- Enforcing integrity with foreign key constraints
- Imposing business rules with check constraints
- Enabling and disabling constraints
- Removing constraints with ALTER TABLE

Improving performance with indexes

- Expediting data retrieval with indexes
- Recommending guidelines for index creation

Manipulating Data

- Modifying table contents
- Adding table rows with INSERT
- Changing row content with UPDATE
- Removing rows with DELETE

Applying transactions

- Atomic Consistent Isolated Durable (ACID) rules
- Controlling transactions with COMMIT and ROLLBACK

Writing Single Table Queries

- Retrieving data with SELECT
- Restricting rows with the WHERE filter
- Sorting the result with ORDER BY
- Handling NULL values in expressions
- Avoiding NULL value pitfalls in filter conditions

Querying Multiple Tables

- Applying the ANSI/ISO standard join syntax
- Matching related rows with INNER JOIN
- Including nonmatched rows with OUTER JOIN
- Creating a Cartesian product with CROSS JOIN

Combining results with set operators

- Stacking results with UNION
- Identifying matching rows with INTERSECT
- Utilizing EXCEPT to find nonmatching rows

Employing Functions in Data Retrieval

- Processing data with row functions
- Conditional formatting with the CASE expression
- Utilizing the CASE expression to simulate IF tests
- Dealing with NULL values

Performing analysis with aggregate functions

- Summarizing data using SUM, AVG and COUNT
- Finding the highest/lowest values with MAX and MIN
- Defining the summary level with GROUP BY
- Applying filter conditions with HAVING

Constructing Nested Queries

- Applying subqueries in filter conditions
- Correlated vs. noncorrelated subqueries
- Testing the existence of rows

Including subqueries in expressions

- Placing subqueries in the column list
- Creating complex expressions containing subqueries
- Handling subqueries that return no rows