ANALYTIXLABS

Introduction

A comprehensive, job-oriented training program crafted by experts

Introduction

Know your Trainer: Chandra Mouli Kotta Kota

Chief Data Scientist

Areas of Expertise:

- Big Data Analytics
- Business Analytics
- Machine Learning
- Marketing Analytics
- Risk Analytics
- Operation Analytics
- Digital Analytics
- Business Intelligence

Chandra Mouli Kotta Kota is a former Business Consultant/Data Scientist and has worked with prestigious companies like McKinsey, Citigroup(E-serve), Genpact in the past 10 years. He has worked for clients across the globe and is an expert in Business and Big Data Analytics.

Professional Experience

- Worked on Marketing Analytics(CSI, CLM & Pricing), Risk Analytics(Credit Risk), Operation Analytics and Digital Analytics with focus on Retail/E-Commerce, Banking, Insurance, Telecom and Media clients in Asia, Australia, Europe, and United States.
- Hands on experience in development of Marketing & CRM Models (Acquisition, LTV Models, Cross Sell & Upsell, Attrition and MROI Models), Pricing Models(price & promo) and Credit Risk Models (PD Models for Credit Cards, Consumer Loans and Insurance Portfolios)
- Hands on expertise in Big data and Multivariate analytical techniques including classical & machine learning algorithms including regression, instance based, regularization, Decision tree, Bayesian, clustering, Association rules, ANN, Deep learning and ensemble algorithms
- Have used different statistical flat forms like SAS/SAS EM, R, Python, SPSS, SPSS Modeler, Hadoop, Tableau, Salesforce, Sql, Excel, VBA .
- Trained and coached several client teams and various individuals on advanced analytics, Big data analytics tools and techniques as part of part of capability building programs

Academic Credentials

- Master of Science(Mathematics/Statistics): IIT-Madras, Chennai

Advance Big Data Science is a combination of two strong analytics projects – Data Science using Python AND Certified Big Data Expert

Crafted by team of experts and maintains a balance between theoretical concepts and practical applications

Advance Big Data Science is a comprehensive program with the following modules, weekly assignments and case studies

Timing: 6 hours per weekend live training (Saturday & Sunday 3 hours each) + Practice

Module 1	***Python & Machine Learning – 60 hours + Practice exercises*** Basic data handling, data manipulation, descriptive analytics and visualization, Advanced Analytics & Machine Learning
Module 2	***Hadoop – 45 hours + Practice Exercises*** Big Data Processing and analyzing using Hadoop & Hadoop Ecosystem
Module 3	***Spark – 15 hours + Practice exercises*** Working on Spark and connecting to Hive and Python/Scala for descriptive and predictive analytics using Machine Learning

Data Science using Python

Total Duration: 60 hours + Practice

Introduction to Data Science

- What is Data Science?
- Data Science Vs. Analytics vs. Data warehousing, OLAP, MIS Reporting
- Relevance in industry and need of the hour
- Type of problems and objectives in various industries
- How leading companies are harnessing the power of Data Science?
- Different phases of a typical Analytics/Data Science projects

Python: Introduction & Essentials

- Overview of Python- Starting Python
- Introduction to Python Editors & IDE's(Canopy, pycharm, Jupyter, Rodeo, Ipython etc…)
- Custom Environment Settings
- Concept of Packages/Libraries - Important packages(NumPy, SciPy, scikit-learn, Pandas, Matplotlib, etc)
- Installing & loading Packages & Name Spaces
- Data Types & Data objects/structures (Tuples, Lists, Dictionaries)
- List and Dictionary Comprehensions
- Variable & Value Labels – Date & Time Values
- Basic Operations - Mathematical - string - date
- Reading and writing data
- Simple plotting
- Control flow
- Debugging
- Code profiling

Python: Accessing/Importing and Exporting Data

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

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: 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)

Python: Basic statistics

- Basic Statistics - Measures of Central Tendencies and Variance
- Building blocks - Probability Distributions - Normal distribution - Central Limit Theorem
- Inferential Statistics -Sampling - Concept of Hypothesis Testing
- Statistical Methods - Z/t-tests (One sample, independent, paired), Anova, Correlation and Chi-square

Machine Learning -Predictive Modeling – Basics

- Introduction to Machine Learning & Predictive Modeling
- Types of Business problems - Mapping of Techniques
- Major Classes of Learning Algorithms -Supervised vs Unsupervised Learning,
- Different Phases of Predictive Modeling (Data Pre-processing, Sampling, Model Building, Validation)
- Overfitting (Bias-Variance Trade off) & Performance Metrics
- Types of validation(Bootstrapping, K-Fold validation etc)

Python: Machine Learning in Practice

- Linear Regression
- Logistic Regression
- Segmentation - Cluster Analysis (K-Means)
- Decision Trees (CHAID/CART/CD 5.0)
- Artificial Neural Networks(ANN)
- Support Vector Machines(SVM)
- Ensemble Learning (Random Forest, Bagging & boosting)
- Other Techniques (KNN, Naïve Bayes, LDA/QDA etc)
- Important Packages for Machine Learning (Sci Kit Learn, scipy.stats etc)

Advance Big Data Science using Python-Hadoop-Spark (2/3)

Total Duration: 105 hours + Practice

These are the standard topics and can be modified as per requirement

Introduction to Big Data

- 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

Hadoop(Big Data) Eco-System

- 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 Data 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 2.x Cluster Architecture
- A Typical enterprise cluster – Hadoop Cluster Modes

Hadoop-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 (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

- Introduction to Data Analysis Tools
- Apache PIG - MapReduce Vs Pig, Pig Use Cases
- Pig 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
- Data Analysis using PIG
- Writing UDF's
- Pig Macros
- Parametrization of Pig code
- Use Pig to automate the design and implementation of MapReduce applications
- Use pig for processing structure and unstructured data

Big Data Analytics using Hadoop (2/2)

Total Duration: 40 hours

These are the standard topics and can be modified as per requirement

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 Hive

Data Analysis using Impala

- Impala & Architecture
- How Impala executes Queries and its importance
- Hive vs. PIG vs. Impala
- Extending Impala with User Defined functions

Introduction to Other Ecosystem tools

- NoSQL Databases and Hbase
- Introduction to OOZIE

SPARK: Introduction

- Introduction to Apache Spark
- Streaming Data Vs. In Memory Data
- Map Reduce Vs. Spark
- Modes of Spark
- Spark Installation Demo
- Overview of Spark on a cluster
- Spark Standalone Cluster

Spark: Spark in practice

- Invoking Spark Shell
- Creating the Spark Context
- Loading a File in Shell
- Performing Some Basic Operations on Files in Spark Shell
- Building a Spark Project with sbt
- Running Spark Project with sbt
- Caching Overview
- Distributed Persistence
- Spark Streaming Overview(Example: Streaming Word Count)

Spark: Spark meets Hive

- Analyze Hive and Spark SQL Architecture
- Analyze Spark SQL
- Context in Spark SQL
- Implement a sample example for Spark SQL
- Integrating hive and Spark SQL
- Support for JSON and Parquet File Formats Implement Data Visualization in Spark
- Loading of Data
- Hive Queries through Spark
- Performance Tuning Tips in Spark
- Shared Variables: Broadcast Variables & Accumulators

Spark Streaming

- Extract and analyse data from twitter using spark streaming
- Comparison of Spark streaming with other streaming tools

Spark Graphx

- Overview of GraphX module in Spark
- Creating graphs with GraphX

Introduction to Machine Learning using Spark MLLib

- Understand Machine Learning framework
- Implement some of the ML Algorithms using spark MLLib

Big Data Project:

- Consolidate all the learnings
- Work on Big data project using Hadoop & Spark

Process: How it works?

1. **Class room/Live online session**
2. **The below information will be uploaded to learning Management System (LMS) after every class.**
   - ✓ Recording of the class
   - ✓ Software installation documents
   - ✓ Content Resources (Class presentations)
   - ✓ Reference books & links(if any)
   - ✓ Sample codes used in the class
   - ✓ Module wise Assignments & datasets
   - ✓ Industry- relevant project work
   - ✓ Class login-details will be updated on regular intervals
3. **Completion of module wise assignments & Projects**
4. **Evaluate Final project work along with assignments**
5. **Certification & placement assistance**

Note: Students can post your questions at forums in LMS and these can be answered by participants and instructor

Housekeeping rules

- ✓ LMS mails have been already shared with all for accessing class recordings and study materials
- ✓ All weekend recordings will be shared by Monday noon at most
- ✓ Periodic assignments and case studies will be shared and deadlines assigned
- ✓ Breaks can be taken only after completion of the Data Science using Python. Any break before that will be treated as a batch change
- ✓ Big Data foundation self-paced course (videos) will be added to LMS for who ever opted for Advanced Big Data Science or Certified Big Data using Hadoop

How to access LMS?

1. Click on LOGIN on our website https://www.analytixlabs.co.in/student/login
2. Type **YOUR email id** as the username
3. Password will be: **password you received from Analytixlabs** or **password which you registered on our website**
4. Click on "My Courses" > "Current Courses" from the options on the left hand side
5. You will now see the Certified Big Data Expert with option of

View classes - You can go to "Previous classes" where you can access all the recordings and Reference materials.

Please do not share your LMS details with anyone. For security reasons, your LMS access will be disabled incase you access this from multiple IPs.

If you face any challenge, please do let us know. We will be using this platform going forward for all the class recordings.

Course completion and career assistance

Course completion & Certification criteria

- You shall be awarded an AnalytixLabs certificate only post the submission and evaluation of mandatory course project work. These will be provided as a part of the training.
- There is no pass/fail for these assignments and projects . Our objective is to ensure that trainees get strong hands-on experience so that they are well-prepared for job interviews along with performance at their jobs.
- Incase the assignments and projects are not up-to-the-mark, trainees are welcome to take help and support for improvisation.
- While weekly schedule is shared with trainees for regular assignments, candidates get 3 months, post course completion, to submit their final assignment and projects.

What is included in career assistance?

- Post successful course completion, candidates can seek assistance from AnalytixLabs for profile building. A team of seasoned professionals will help you based on your overall education background and work experience. This will be followed by interview preparation along with mock interviews (if required)
- Job referrals are based on the requirements we get from various organizations, HR consultants and pool of AnalytixLabs' ex-students working in various companies.
- No one can truthfully provide job guarantee, particularly for good quality job profiles in Analytics. However, most of our students do get multiple interview calls and good career options based on the skills they learn during training. For this there will be continuous support from our side for as long as required.

Contact Us

Visit us on: http://www.analytixlabs.in/

For course registration, please visit: http://www.analytixlabs.co.in/course-registration/

For more information, please contact us: http://www.analytixlabs.co.in/contact-us/

Or email: info@analytixlabs.co.in

Call us we would love to speak with you: (+91) 7530818107

Join us on:

Twitter - http://twitter.com/#!/AnalytixLabs

Facebook - http://www.facebook.com/analytixlabs

LinkedIn - http://www.linkedin.com/in/analytixlabs

Blog - http://www.analytixlabs.co.in/category/blog/