Big Data Specialization

Course brief

How big is BIG? Become a big data expert through an intensive training program customised across various levels designed specifically for you. It will make participants solve real-time problems with huge datasets. Through this intensive program we aim to train the participants in a way that they are prepared to appear for International Certifications as mentioned below:

- Hortonworks Data Platform Certified Developer:Java (HDPCD:Java)
- Hortonworks Data Platform Certified Developer :Spark(HDPCD:Spark)
- Hortonworks Data Platform Certified Developer (HDPCD)
- CCA Spark from Cloudera

Course Structure:

There are 4 modules in BigData Specialization.

- DataScience with R
- Hadoop with MapReduce using Java
- Hadoop Ecosystem (Sqoop, Flume, Pig, Hive)
- Spark with Scala

Course Structure:

There are 4 modules in BigData Specialization.

- DataScience with R
- Hadoop with MapReduce using Java
- Hadoop Ecosystem (Sqoop, Flume, Pig, Hive)
- Spark with Scala

Syllabus

Subjects	Duration
Data Science with R	6 weeks (M-W-T)
Hadoop Map Reduce	6 weeks (M-W-T)
Hadoop Ecosystem (Sqoop, Flume, Pig, Hive)	6 weeks (M-W-T)
Spark with Scala	6 weeks (M-W-T)

Data Science with R

Curriculum

Introduction to Data Science

Introduction- Definition - DS in various fields - Examples - Impact of Data Science - Major Activities - Toolkit - Data Scientist - Compare with others - Data Science Team

Learning Outcomes:

- Understanding Data Science and related fields
- Be able to identify major activities of data science for the given problem
- Understanding role of Data Scientist and how it differs from a data engineer and a data analyst.
- Be able to choose deployment model for organization
- Understand how to create a data science team.

R Basics

Introduction to R : What is R - Data Science with other languages - Features of R - Environment - R at a glance. Basics of R(Series & Ctrl Statements): Assignment - Modes - Operators - special numbers - Logical values - Basic Functions - Generating data sets - Control Structures Vectors:Definition- Declaration - Generating - Indexing - Naming - Adding & Removing elements - Operations on Vectors - Recycling - Special Operators - Functions for vectors - Missing values - NULL values - Filtering & subsetting. Exercises.

Learning Outcomes:

- Understand the how R differs from other languages
- Be able to write R scripts for given problem
- Be able to generate series
- Be able to handle data in vectors and get required results from given data sets

Descriptive Statistics

Descriptive Statistics: Introduction - Descriptive Statistics - Central Tendency - Variability - Mean - Median - Range - Variance - Summary-Exercises Graphics : : Introduction - Types - Packages - Basic graph - Histograms - Stem Leaf Graph - Box Plots - Scatter Plots - Bar Plots.

Learning Outcomes:

- Understand the importance of statistics, types and statistics in real world
- Be able to find the central tendency, summary of given data sets
- Understand the importance of graphical output and various graphs
- Be able to plot various graphs for the given data set.
- Implement Descriptive Statistics in R.

Data Structures in R

Arrays: Creating Arrays - Dimensions & Naming - Indexing & Naming - Functions on Arrays Matrices : Creating Matrices - Adding rows/columns - Removing rows/columns - Reshaping - Operations - Special functions Listss: Creating - Naming - Accessing elements - Adding - Removing - Special Functions - Recursive Lists Data frames: Creating - Naming - Accessing - Adding - Removing - Special functions - Merging Exercises Functions: Creating - Functions on Function Object - Scope of Variables - Accessing Global Environment - Closures - Recursion - Creating New Binary Operator

Learning Outcomes:

- Understand various data structures in R
- Be able to choose suitable data structure for the given data set
- Be able to retrieve the required result from the given data set
- Be able to solve the problems by creating functions
- Be able to merge and split the data sets
- Be able to apply statistics on various data structures

Regression

Linear Regression: : Inferential Statistics - Types of Learning - Linear Regression- Simple Linear Regression - Coefficients - Confidence Interval - RSE - R2 - Implementation in R - lm - functions on lm - predict - Plotting - fitting regression line Exercises Multiple Linear Regression: Introduction- comparison with simple linear regression - Correlation Matrix - F Statistic - Response vs Predictors - Deciding important variable - Model fit - Predictions Generating a model - Interactive terms - Non Linear Transformations - Anova - lm with polynomial Exercises Classification & Logistic Regression : Classification - Examples - Logistic Regression Definition - Estimating coefficients - Predictions - Multiple Logistic Regression - More than 2 response classes - Implementation in R - glm - predict Exercises

Learning Outcomes:

- Understand Inferential Statistics, types and regression concepts
- Understand the population and sample for the given data set.
- Be able to understand how to fit a model for the given data set.
- Be able to find the relation between response and predictors.
- Be able to predict the values for given data set based on sample data set

Advanced Analysis

classification: : Linear Discriminant Analysis - Quadratic Discriminant Analysis - K-Nearest Neighbors- Exercises Support Vector Machines: Maximal margin classifier - Support Vector Classifier - Support vector machine - SVM with more than 2 classes - Exercises Neural Networks : Introduction - Nodes & Weights - Layered Architecture - Learning Rule - Implementation in R - Normalizing data - Creating training data sets - Fitting Neural Network - neuralnet - Plotting NN - Predictions - Denormalize - MSE - Exercises

Learning Outcomes:

- Understand the classification problems and multiple predictors
- Be able to solve the problem using LDA, KNN, QDA.
- Understand Support Vector Machines and its importance.
- Fit a Neural Network for the given data set.

Instructors

Mr. P.V.N.Balarama Murthy

Data Science with R

Mr. P.V.N.Balarama Murthy, is an M.Tech(CSE) having over 10 years of teaching and technical training experience. He is specialist in Data Science and Bigdata. He has experience in deploying hadoop clusters. As technical trainer, he has trained a number of people in C,C++, Java, Oracle, Hadoop (Administration, Development with MR, PIG, Hive, Flume, Sqoop) and Data Science with R. He has guided to his credit 15+ students to get Hortonworks certifications for Hadoop.

A dedicated, resourceful and result oriented instructor that he is, it is helping shape up careers of students.

Ms. Jyothi SanjeevaMani

Data Science with R

Ms. Jyothi SanjeevaMani has over 15 years of satisfying teaching and technical training experience. She is a Research Scholar of Big Data Analytics from a reputed university. As a technical trainer she trained many students in industry oriented subjects like C, C++, Java, MySQL, Oracle (SQL, PL/SQL), Python, Linux, Openstack, BigData - Hadoop(MapReduce, Pig, Hive, Sqoop, Flume), Data Science with both Python and R.

She is an Asst.Professor with the Department of IT at The Keshav Memorial Institute of Technology (KMIT).

She is a dedicated, resourceful and a result oriented instructor, who strives to help students change marginal grades into good grades.

Hadoop Map Reduce

Curriculum

Bigdata Concepts

Introduction Data, Storage, Bigdata, Distributed environment, Hadoop introduction History, Environment, Benefits - Subprojects HDFS, Map-Reduce, PIG, Hbase, Hive, Zoo-Keeper, SQOOP, Mahout, MongoDB, Hadoop DB

Learning Outcomes:

- Understand big data, challenges, distributed environment.
- Know hadoop and sub projects.

HDFS

Hadoop Architecture : Overall Architecture-NameNode - Datanode Fault Tolerance - Read&Write operations - Interfaces(Command line interface, JSP, API) - HDFS Shell - FS Shell Commands - Java API Programs

Learning Outcomes:

- Acquire knowledge of HDFS components , Namenode, Datanode.
- Acquire knowledge of storing and maintaining data in cluster, reading and writing data to/from cluster.
- Be able to maintain files in HDFS
- Be able to access data from HDFS through java program

Basic Map-Reduce

Map-Reduce Introduction - Map-Reduce Architecture - Yarn Architecture - Basic M-R Programs - Detailed description of M-R Methods and exercises -

Learning Outcomes:

- Understand Map-Reduce paradigm and Yarn Architecture.
- Analyze a given problem in map-reduce pattern.
- Be able to write Basic Map-Reduce Programs.

Customize Key/Value from Map to Reduce

Rkey/value pairs - Different types of values from a mapper - GenericWritable - Custom values from mapper - Writable - Custom keys from Mapper - WritableComparable - Exercises

Learning Outcomes:

- Understand the key-value pairs from map to reduce
- Be able to design applications with custom value types
- Be able to design applications with custom key types
- Applications with Generic writable

Custom Input/Output files for Map-Reduce

Input format - FileInputFormat - Steps for Input - RecordReader - Custom FileInputFormat - Custom RecordReader - Exercise Output format - FileOutputFormat - RecordWriter - Custom FileOutputFormat - Custom RecordWriter

Learning Outcomes:

- Understand the input and output formats of map-reduce application.
- Be able to read different formats of files into map-reduce application.
- Be able to produce different formats of files from map-reduce application.

Process through Map to Reduce

Combiners - Partitioners - Secondary Sorting - Exercises

Learning Outcomes:

- Understand the process between map and reduce phases.
- Be able to optimize the performance of Map-Reduce application.
- Be able to classify the output of map-reduce application.
- Be able to use combiners in Map-Reduce application
- Be able to use Partitioners
- Be able to sort an additional key

Joins

Joins- various types - Reduce Side joins - Distributed Cache - Map-Side Join - Exercises

Learning Outcomes:

- Be able to take data from multiple data sets and join them.
- Be able to implement various joins in Map-Reduce.
- Be able to design applications with map-side joins.
- Be able to design application with reduce side join.

Instructors

Mr. P.V.N.Balarama Murthy

Hadoop Map Reduce

Mr. P.V.N.Balarama Murthy, is an M.Tech(CSE) having over 10 years of teaching and technical training experience. He is specialist in Data Science and Bigdata. He has experience in deploying hadoop clusters. As technical trainer, he has trained a number of people in C,C++, Java, Oracle, Hadoop (Administration, Development with MR, PIG, Hive, Flume, Sqoop) and Data Science with R. He has guided to his credit 15+ students to get Hortonworks certifications for Hadoop.

A dedicated, resourceful and result oriented instructor that he is, it is helping shape up careers of students.

Ms. Jyothi SanjeevaMani

Hadoop Map Reduce

Ms. Jyothi SanjeevaMani has over 15 years of satisfying teaching and technical training experience. She is a Research Scholar of Big Data Analytics from a reputed university. As a technical trainer she trained many students in industry oriented subjects like C, C++, Java, MySQL, Oracle (SQL, PL/SQL), Python, Linux, Openstack, BigData - Hadoop(MapReduce, Pig, Hive, Sqoop, Flume), Data Science with both Python and R.

She is an Asst.Professor with the Department of IT at The Keshav Memorial Institute of Technology (KMIT).

She is a dedicated, resourceful and a result oriented instructor, who strives to help students change marginal grades into good grades.

Hadoop Ecosystem (Sqoop, Flume, Pig, Hive)

Curriculum

Data Ingestion

Introduction - types of Data Ingestion - Ingesting Batch Data - Ingesting Streaming Data - Examples

Learning Outcomes:

Understanding Data Ingestion.

Sqoop

Introduction - Sqoop Architecture - Connect to MySQL database - Sqoop - Import - Export - Eval - Joins - exercises.

Learning Outcomes:

- Understand Sqoop architecture and uses
- Able to load real-time data from an RDBMS table/Query on to HDFS
- Able to write sqoop scripts for exporting data from HDFS onto RDMS tables

Flume

Introduction - Flume Architecture - Flume master - Flume Agents - Flume Collectors - creation of Flume configuration files - Examples - Exercises

Learning Outcomes:

- Understand Flume architecture and uses
- Able to create flume configuration files to stream and ingest data onto HDFS

Data transformation (PIG)

Introduction-Pig Data Flow Engine-Map Reduce Vs. Pig - Data Types-Basic Pig Programming-Modes of execution in PIG-Miscellaneous Commands - Group, Filter, Join, Order, Flatten, cogroup, Flatten, Illustrate, Explain - Parameter substitution-creating simple UDFs in Pig-Examples-Exercises.

Learning Outcomes:

- Understand Apache PIG , PIG Data Flow Engine
- Understand data types, data model, and modes of execution.
- Able to store the data from a Pig relation on to HDFS.
- Able to load data into Pig Relation with or without schema.
- Able to split, join, filter, and transform the data using pig operators
- Able to write pig scripts and work with UDFs.

Data Analysis (HIVE)

Introduction - Hive Architecture - Hive Vs. RDBMS- HiveQL and Shell - Data Types and Schemas - Hive Commands - Hive Tables: Managed Tables, External Tables, Partitions, bucketing - Joins - views - SortBy - distribute by - HCatalog-Using HCatStorer and HCatLoader- Examples - Exercises

Learning Outcomes:

- Understand the importance of Hive, Hive Architecture
- Able to create Managed, External, Partitioned and Bucketed Tables
- Able to Query the data, perform joins between tables
- Understand storage formats of Hive
- Understand Vectorization in Hive

Instructors

Mr. P.V.N.Balarama Murthy

Hadoop Ecosystem

Mr. P.V.N.Balarama Murthy, is an M.Tech(CSE) having over 10 years of teaching and technical training experience. He is specialist in Data Science and Bigdata. He has experience in deploying hadoop clusters. As technical trainer, he has trained a number of people in C,C++, Java, Oracle, Hadoop (Administration, Development with MR, PIG, Hive, Flume, Sqoop) and Data Science with R. He has guided to his credit 15+ students to get Hortonworks certifications for Hadoop.

A dedicated, resourceful and result oriented instructor that he is, it is helping shape up careers of students.

Ms. Jyothi SanjeevaMani

Hadoop Ecosystem

Ms. Jyothi SanjeevaMani has over 15 years of satisfying teaching and technical training experience. She is a Research Scholar of Big Data Analytics from a reputed university. As a technical trainer she trained many students in industry oriented subjects like C, C++, Java, MySQL, Oracle (SQL, PL/SQL), Python, Linux, Openstack, BigData - Hadoop(MapReduce, Pig, Hive, Sqoop, Flume), Data Science with both Python and R.

She is an Asst.Professor with the Department of IT at The Keshav Memorial Institute of Technology (KMIT).

She is a dedicated, resourceful and a result oriented instructor, who strives to help students change marginal grades into good grades.

Spark with Scala

Curriculum

Introduction to Scala

Introduction- Data types - variables - Control Structures-strings-classes-methods-objects

Advanced concepts

Traits, mixins, packages, lists, sets, maps, tuples

Introuction to Spark

RDDs

Spark SQL

Instructors

Mr. P.V.N.Balarama Murthy

Spark with Scala

Mr. P.V.N.Balarama Murthy, is an M.Tech(CSE) having over 10 years of teaching and technical training experience. He is specialist in Data Science and Bigdata. He has experience in deploying hadoop clusters. As technical trainer, he has trained a number of people in C,C++, Java, Oracle, Hadoop (Administration, Development with MR, PIG, Hive, Flume, Sqoop) and Data Science with R. He has guided to his credit 15+ students to get Hortonworks certifications for Hadoop.

A dedicated, resourceful and result oriented instructor that he is, it is helping shape up careers of students.

Ms. Jyothi SanjeevaMani

Spark with Scala

Ms. Jyothi SanjeevaMani has over 15 years of satisfying teaching and technical training experience. She is a Research Scholar of Big Data Analytics from a reputed university. As a technical trainer she trained many students in industry oriented subjects like C, C++, Java, MySQL, Oracle (SQL, PL/SQL), Python, Linux, Openstack, BigData - Hadoop(MapReduce, Pig, Hive, Sqoop, Flume), Data Science with both Python and R.

She is an Asst.Professor with the Department of IT at The Keshav Memorial Institute of Technology (KMIT).

She is a dedicated, resourceful and a result oriented instructor, who strives to help students change marginal grades into good grades.