Digital Vidya

BIG DATA COURSE

Big Data Application Engineer/ Developer

Specialization in Apache Spark, Kafka, Airflow, HBase

In Exclusive Association with

A Govt. of India and Govt. of NCT Delhi Co. Initiative
Test Your Skills, Own Your Future

V skills CERTIFIED

21,347+ Participants | 10,000+ Brands | 1200+ Trainings | 45+ Countries
[Since 2009]

Digital Vidya is training partner for Google Partners Academy & Microsoft in India to train their partners

Big Data Application Engineer

With a high volume of data being produced daily, there is a huge demand for people with skills to manage, analyze and help organizations use this data effectively for data driven decision making. A growing field like this offers new exciting career opportunities for those who want in on the action.

This course is designed to address these opportunities for the role of Big Data Application Engineer. The primary responsibilities of this role include:

- Designing and maintaining fault-tolerance, highly distributed, and robust systems.
- Using the state of the art big data platforms to solve the business problems and help derive value from the data.
- Drive processing of hundreds of terabytes of unstructured and structured data.
- Enable machine learning/data mining systems to become more real-time and scale to large data ingestion systems.

Some of the core skills required for this role are:

1: Java/Scala/Python Coding Skills

2: Expertise in Big Data Platforms
Like Hadoop, Spark, Kafka etc.

3: Expertise in NoSQL Platform
Like HBase

4: Knowledge of Data Mining and Machine Learning Models.

Course Highlights

Who is this Course for

**Software Engineers, Developers, Testers, Quality Engineers,
Database Experts, Data Analysts, Java J2EE Developers, Python Developers**

Salient Features

- 3 Hrs/Week Live Instructor-Led Online Sessions
- 15 Days of Project Work
- Active Q/A Forum
- Class Labs/Home Assignment (10 hours/Week Learning Time)
- Govt. of India (Vskills Certified Course)
- Placement Support
- Personalised Training Program
- Lifetime Access To Updated Content and Videos
- Industry and Academia Faculty
- Top Big Data Tools Covered
- Specialize in Apache Spark, Kafka, Airflow, HBase
- Industry's Top Big Data Advisors

Course Advisors

Manas Garg heads the Analytics for Marketing at Paypal. He takes Data Driven Decisions for Marketing Success.

Manas Garg
Architect
PayPal

Shweta Gupta
Vice President, Tech.
Digital Vidya

Shweta Gupta has 19+ years of Technology Leadership experience. She holds a patent and number of publications in ACM, IEEE and IBM journals like Redbook and developerWorks.

Vishal is a Technology influencer and CEO of Right Relevance. (A platform used by millions for content & influencer discovery)

Vishal Mishra
CEO & Co-Founder
RIGHT RELEVANCE

Course Trainers

PRATEEK DUBEY

Prateek Dubey is a Bachelor of Computer Engineering Graduate from University of Pune. Currently, working with Royal Bank of Scotland (RBS) as a Senior Software Engineer - Data Engineer (Hadoop/ Spark/ AWS developer) to reshape the bank to achieve its vision of 2020. He's an expert in core Big Data Platforms - Cloudera distribution including Big Data implementation on Cloud - AWS. He is also a Freelance Hadoop Trainer and has trained several candidates in past on various Hadoop technologies like Hive/ Pig/ Impala/ Sqoop/ HBase/ Flume etc.

ROHIT KUMAR

Rohit Kumar is a Big Data researcher with publications in many prestigious International Conferences. He has 6 plus years experience in industry and expertise in various programming languages including Java, Scala, C++, Python, and Haskel. He works in variety of different database systems such as MySQL, Microsoft SQL, and Oracle Coherence and in many Big Data systems like Hadoop, Apache Spark, Apache Storm, Kafka, MongoDB.

Course Curriculum

Foundation Courseware

5 Weeks

Introduction to Big Data Storage

INTRODUCTION TO BIG DATA

What is Big Data (evolution)
Introduction to Big Data
Problems with Traditional Large-Scale Systems
Introduction to Distributed File Systems/ Computing
Big Data Solution Landscape
Industry Insight
Use Cases of Big Data Analytics
Big Data Technology Career Path
Cloudera Hadoop Docker Image Installation

INTRODUCTION TO HADOOP AND HDFS

Introduction to Hadoop
Limitations and Solutions of Traditional System
Motivation for Hadoop
History of Hadoop
Benefits of Hadoop
Hadoop Ecosystem

HADOOP ARCHITECTURE

Hadoop 1.x Core Components
Hadoop 2.x Core Components
Fundamentals of Hadoop
Hadoop Master-Slave Architecture
YARN for Resource Management
Different Types of Cluster Setups
Understanding Hadoop Configuration Files
Hadoop Security
HDFS Architecture

HADOOP FAULT TOLERANCE

Hands-On Exercise: HDFS Commands

Processing Framework

MAP REDUCE

Understanding Map Reduce
Map Reduce Overview
Data Flow of Map Reduce
YARN Map Reduce Detail Flow
Concept of Mapper & Reducer
Speculative Execution
Hadoop Fault Tolerance
Submission & Initialization of Map Reduce Job
Monitoring & Progress of Map Reduce Job

Data Storage

RDBMS, NOSQL DATABASE - HBASE

Introduction to NOSQL Databases
NOSQL v/s RDBMS
NOSQL Database Types
Introduction to HBase
HBase vs RDBMS
HBase Architecture
HBase Components

Big Data and Cloud Platforms

INTRODUCTION TO CLOUD PLATFORMS

Introduction to Cloud Computing
Cloud Computing Models
Understanding of Public, Private, Hybrid Cloud
Characteristics of Cloud Computing
Major Players in Market - AWS, Azure, Google Cloud
Overview of Amazon Web Services
Amazon Web Services Cloud Platform
Big Data on Cloud - Amazon EMR
Amazon Cloud Storage - S3
Adoption of AWS in Public and Private Sector

Big Data Application Engineer

15 Weeks

Prerequisite: Programming Language - Java, Python, RDBMS, SQL

Introduction and Spark Core

Introduction/Refresher to Big Data and Hadoop, Batch vs Real Time

DISTRIBUTED ARCHITECTURE BACKGROUND
What is Big Data Quick Intro
Basics of Distributed Architecture
Hadoop and MapReduce Intro (What was missing!!)
MapReduce Programing Exercise
Configuring Local Spark Setup

Steaming Platforms Introduction

STREAM PROCESSING BACKGROUND
Batch vs Stream
Basics of Stream Processing Architecture
Some famous Stream Processing Systems.
Spark Introduction
Spark Architecture
Deployment Architectures
Introduction to RDD

Introduction to Spark core

INVOKING SPARK SHELL
Creating the Spark Context
Loading a File in Shell
Performing Basic Operations on Files in Spark Shell
Local Mode
Spark Mode
Caching Overview
Distributed Persistence
Transformations in RDD
Actions in RDD
Loading Data in RDD
Saving Data Through RDD
Key-Value Pair RDD
Map Reduce and Pair RDD Operations

Spark Core Advanced

SPARK WEB UI.
Handling Other File Formats in Spark
File Formats
File Systems
Databases
Some Advanced Spark Programming
Broadcast Variables
Accumulators
Working on a Per-Partition Basis

Spark SQL

SPARK SESSION
Creating Data Frames
Data Frame Operations
Running SQL Queries Programmatically
Creating Datasets
Interoperating with RDDs
User-Defined Aggregate Functions
Data Sources
Generic Load/Save Functions
Manually Specifying Options
Run SQL on Files Directly
Saving to Persistent Tables
Bucketing, Sorting and Partitioning
JSON Datasets
Performance Tuning
Peak Behind the Hood: Catalyst Optimizer

Spark Advanced

Spark Application Development and Configurations Using Scala

OVERVIEW OF MAVEN
Building a Spark Project with Maven
Running Spark Project with Maven
Spark and Hadoop Integration-HDFS
Spark and Hadoop Integration-Yarn
Using Eclipse IDE for Spark Application Development
Dynamic Resource Allocation
Configuring Spark Properties

Spark Streaming

SPARK STREAMING ARCHITECTURE
Transformations in Spark Streaming
Fault tolerance in Spark Streaming
Checkpointing
Parallelism Level

Spark-Kafka use case

BASIC KAFKA
Kafka Consumer Integration
Kafka with Spark
Spark Twitter API Integration
Introduction to Airflow
Simple Use Case of Spark with Airflow

Spark with HBase and Hive

ADVANCED HBASE
Hbase Standalone
Spark Hbase
Hadoop Reading Data vs Hbase Reading Data Using Predicate Pushdown.
Hive Integration
How to Write your own API Integration with External Data Store

Spark Data Science Track

Introduction to Spark MLLib, Streaming and GraphX

QUICK RECAP OF SOME BASICS MATHS ON MATRIX MANUPULATIONS
Machine Learning with Spark
Data Types
Algorithms– Statistics
Classification and Regression
Clustering
Collaborative Filtering.
Spark GraphX
Spark Graph Frames Introduction

Capstone Project (3 Weeks)

With the advancement of social media, gathering and analyzing social media data for **Marketing and Trend** analysis is becoming quite popular. Among all the different social media platforms present currently, Twitter is one of the most popular platforms where people share their views actively. There are many companies which **mine and analyze Twitter data** for various purposes for example:

1. Voters sentiment analysis during elections by analyzing the polarity of tweets and the sentiments of people after a political event, many believe Twitter played a critical role in US 2016 Election.

2. Marketing of new products such as to understand when, where, and how consumers speak about purchasing your product or category, and track changes over time.

3. Evaluate campaign impact by assessing whether your latest creative campaign generate social buzz, and review which interest segments the campaign resonated with most.

Currently, approximately **500 million tweets** are generated daily!! That's more than 50 GB of data every day. Storing and analyzing it efficiently is one of the most challenging tasks in Twitter data analysis.

As part of the **Big Data Application Engineering Specialization**, we will look into following:

- **Solve interesting assignments** where we will analyze **Twitter** data
- **Explore 3 different ways** to analyze the **Twitter** data
  - Use Hadoop and Hive to store and analyze Twitter data using MapReduce Programming
  - Use Spark Core APIs to do a similar analysis and see the benefits of using Spark over MapReduce
  - Use the power of Spark-SQL APIs on doing twitter analysis using Data Frames
- **Create** a near real time analytical engine integrating **Twitter** Streaming API with **Apache Spark Streaming** and **Kafka** to do some interesting trend analysis

Tools Covered

Tools for Everyone

Tools for Big Data Application Engineer

Batch Options

Duration	20 Weeks
Batch Option	Weekend
Fee	INR 49,900 (+GST)

Certification

Certificate By

V skills CERTIFIED

Once you complete Certified Big Data Course, you are eligible to take the examination.

Exam Duration: 45 Minutes
Maximum Marks: 30
Passing Marks: 22
No Negative Marking

Number of Questions: 30
Exam Conducted Online
Attendance in Live Session: 75% (3 out of 4)
Mandatory Completion of all Assignments and Projects

Digital Vidya

Interested? Contact Us!

+91-84680-02880 info@digitalvidya.com

www.digitalvidya.com