Invictus Engineers

DATA ENGINEERING COURSE CURRICULUM - 25 DAYS (5 WEEKS)

Invictus Engineers

3rd Floor, Surviba Towers, Plot No.247/3RT, Near Umesh Chandra Statue, Sanjeeva Reddy Nagar, Hyderabad, Telangana. 500038 **Email Us :** info@invictusengineers.com **Phone:** +91 99666 33097
www.invictusengineers.com

Bigdata & Hadoop:

1. Introduction

1.1 Big Data Introduction

- What is Big Data
- Bigdata Analytics
- Bigdata Challenges
- Technologies for Bigdata

1.2 Hadoop Introduction

- What is Hadoop?
- History of Hadoop
- Basic Concepts
- Future of Hadoop
- The Hadoop Distributed File System
- Anatomy of a Hadoop Cluster
- Breakthroughs of Hadoop
- Hadoop Distributions:
  - Apache Hadoop
  - Cloudera Hadoop (CDH)
  - Horton Networks Hadoop (HDP)
  - MapR Hadoop (mapR)
  - AWS - EMR

2. Hadoop Daemon Processes

- Hadoop Architecture
- High Availability
- **N**ame **N**ode

- **D**ata**N**ode
- **S**econdary **N**ame **N**ode
- **J**ob **T**racker/Resource Manager
- **T**ask **T**racker/Node Manager

3. HDFS (Hadoop Distributed File System)

- Blocks and Input Splits
- Data Replication
- Hadoop Rack Awareness
- Hadoop Cluster Architecture and Block Placement
- Accessing HDFS
  - CLI Approach
- HDFS basic file operations
- Basic Administration commands

4. Hadoop Installation Modes

- Local Mode
- Pseudo-distributed Mode
- Fully distributed mode

5. YARN

- What is YARN
- How YARN Works
  - Resource Manager
  - Node Manager
  - Application Master
  - Containers

- Advantages of YARN
- Fail Over Mechanizm

6. Writing a MapReduce Program

- Basic API Concepts
- The Driver Class
- The Mapper Class
- The Reducer Class
- The Combiner Class
- The Partitioner Class
- Examining a Sample MapReduce Program with several examples

7. Hive

- Hive concepts
- Hive architecture
- Managed tables and external tables
- Complex data types
- Partitioned tables
- Bucketed tables
- Joins in Hive
- Multiple ways of inserting data in Hive tables
- CTAS, views and alter tables
- Performance Tuning in Hive
- User defined functions in Hive
  - Hive UDF
  - Hive UDAF
  - Hive UDTF

8. Sqoop

- Sqoop concepts
- Sqoop architecture
- Connecting to RDBMS
- Internal mechanism of import/export
- Import data from Oracle/Mysql to Hive
- Export data to Oracle/Mysql
- Other Sqoop tools

8.1 Oozie

- Oozie concepts
- Oozie architecture
  - Workflow engine
  - Coordination Engine
- HPDL and XML for creating Workflows
- Nodes in Oozie
  - Action nodes
  - Control nodes
- Accessing Oozie jobs through CLI, and web console
- Creating workflows in Oozie for:
  - HDFS file operations scripts
  - MapReduce programs
  - Pig scripts
  - Hive scripts
  - Sqoop Imports/Exports

PySpark:

Introduction to Spark

- What is Spark?
- Spark Overview
- Setting up PySpark environment
- Using Spark Shell

Spark Basics

- Resilient Distributed Datasets (RDDs)
- Spark Context
- Spark Ecosystem
- In-Memory Computations in Spark

Working with RDDs

- Creating, Loading and Saving RDD
- Transformations on RDD
- Actions on RDD
- Key-Value Pair Transformation on RDDs
- RDD Partitioning
- RDD Persistence

Writing and Deploying on Cluster

- Spark Applications vs. Spark Shell
- Spark Runtime Architecture
  - Executors
  - Driver
  - Cluster Managers

- Creating Spark Context
- Building a Spark Application
- Deploying Spark Applications using Spark-Submit

Spark Job Execution

- RDD Lineage
- Jobs, Stages and Tasks
- Partition and Shuffles
- Data Locality
- Join with or without Partitioner, stages and tasks, etc
- Spark Web UI

Spark SQL

- Overview on Hive
- Spark SQL Architecture
- SparkSession in Spark SQL
- Working with DataFrames
- Integrating Spark SQL with Hive
- Integrating Spark SQL with JDBC Sources (MySQL)
- Handling CSV, JSON and Parquet File Formats
- Loading and Saving Data

Spark Streaming

Spark Streaming with DStreamAPI

- Spark Streaming Architecture
- Spark Streaming Transformations
  - Stateless
  - Stateful Transformations

- Rolling Window and Check pointing
- Integrating Spark with Kafka Streaming Data
- Structured Streaming
- Integrating Spark with Twitter Streaming Data
- Spark Streaming Performance Considerations

Structured Streaming

- Structured Streaming Overview
- Advantages of Structured Streaming
- Other Streaming system Vs. Structured Streaming
- Stateful operations
- Spark Structured Streaming
- Output Modes
- Spark Structured Streaming Example
- Window API
- Event Time
- Late Events
- Watermark

- Structured Streaming and Kafka integrations
- Integrating Spark with Twitter Streaming Data
- Spark Streaming Performance Considerations

Performance Tuning and Debugging

- Shared Variables: Broadcast Variables
- Shared Variables: Accumulators
- Common Performance Issues

- Performance Tuning Tips
- Spark WebUI
- Monitoring Driver and Executor Logs

AWS Data Engineering

1. **Big Data On Cloud**

- Introduction to Cloud - Models, Service Categories, AWS security, IAM
- AWS platform
- EC2
- S3
- Databases on AWS
- AWS EMR

Capstone Project:

Data Analytics in BFSI sector data - Spark and Hive Integration

2. Course Deliverables

- Workshop style coaching
- Interactive approach
- Course material
- Hands on practice exercises for each topic
- Quiz at the end of each major topic
- Tips and techniques on Certification Examinations
- Linux concepts and basic commands on demand