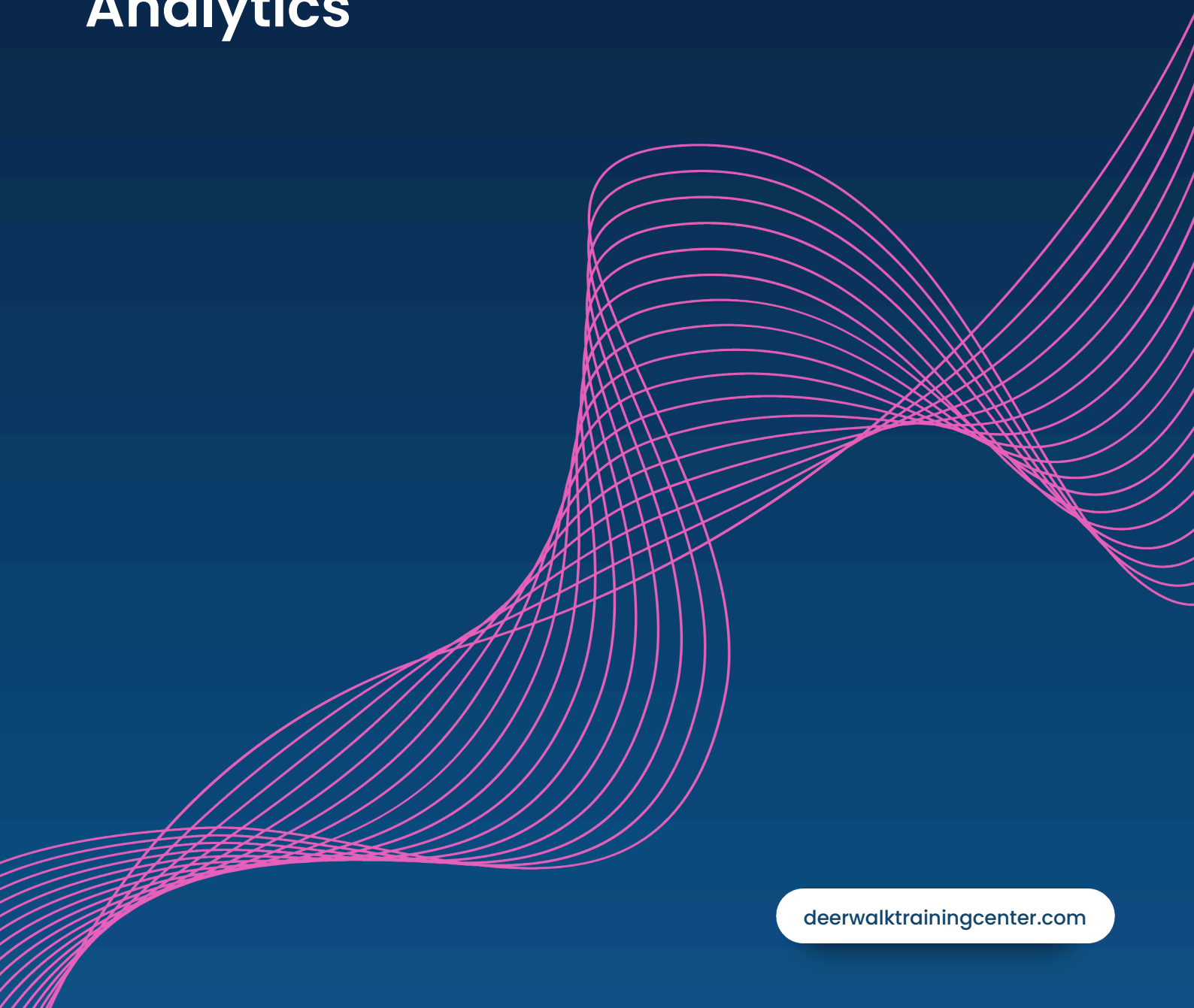
deerwalk
training center

DTA-OMT-105

Big Data Programming and Hadoop Analytics



deerwalktrainingcenter.com

Program Information

Nature of the Course
Lecture + Hands-on Labs

Total Hours per Day
64 hours

Course Duration
8 Weeks

Course Summary

This syllabus details the structure, content, and requirements for the Hybrid Big Data Programming and Hadoop Analytics course. The design emphasizes hands-on practice and real-world project experience within a blended learning environment.

Completion Criteria

After fulfilling all of the following criteria, the student will be deemed to have finished the module:

- Has attended 90% of all classes held.
- Has received an average of 80% on all assignments
- Has received an average of 60% in assessments.
- The tutor believes the student has grasped all of the concepts and is ready to go on to the second module.

Required Textbooks

Recommended resources include Hadoop & Spark official documentation, a GitHub repository with lab materials, cloud platform free tier access (for cloud modules), and recommended textbooks and research papers.

Prerequisites

- Basic understanding of the Linux command line.
- Fundamental programming knowledge (Python/Java preferred).
- Familiarity with SQL concepts.
- Basic understanding of data structures.

Course Details

Week 1: — Foundations of Big Data and Hadoop

Topics:

- Introduction to Big Data & Analytics: Concepts & 5 Vs; Types of analytics (Descriptive, Predictive, Prescriptive)
- Hadoop Architecture & Ecosystem: HDFS, YARN, MapReduce overview; Hive, Pig, Spark introduction
- HDFS Basics & Operations: HDFS block storage, replication, commands; File permissions and quotas
- MapReduce Fundamentals: Workflow and lifecycle; Writing MapReduce programs
- Cluster Setup Basics: Multi-node cluster installation on Linux; Configuration files (core-site.xml, hdfs-site.xml, yarn-site.xml)

Hands-on Labs:

- Data visualization with Excel/Power BI
- Pseudo-distributed Hadoop setup
- HDFS file management
- WordCount with Python streaming
- Daemon setup and Web UI validation

Week 2: — Cluster Management & Data Processing

Topics:

- YARN Architecture & Scheduling: ResourceManager, NodeManager roles; Capacity vs. Fair Scheduler
- Hive for Data Warehousing: Hive architecture and metastore; HiveQL for querying
- Pig for Data Transformation: Pig Latin scripting basics; Data transformation workflows
- Cluster Monitoring Tools: Using Ambari/Cloudera Manager; Log analysis and health checks
- High Availability Setup: NameNode and YARN HA; Quorum Journal Manager

Hands-on Labs:

- Job submission and monitoring
- Table creation and aggregation
- Pig scripts for ETL tasks
- Simulated node failure and recovery
- Manual failover in 3-node cluster

Week 3: — Security, Ingestion & Spark Intro

Topics:

- Hadoop Security: Kerberos authentication; Apache Ranger for authorization
- Data Encryption: Data at rest vs. in transit encryption; HDFS encryption zones
- Data Ingestion with Sqoop: Import/export RDBMS data to/from HDFS; Sqoop connectors and optimization
- Data Ingestion with Flume: Log and streaming data ingestion; Flume agents and channels
- Introduction to Apache Spark: Spark vs. MapReduce; RDDs and transformations

Hands-on Labs:

- Simplified Kerberos + Ranger policy setup
- Configuring encryption
- MySQL to HDFS import
- Log streaming setup
- RDD operations in PySpark

Week 4: — 5 Spark Programming & SQL

Topics:

- Spark Core Deep Dive: RDD persistence and partitioning; Broadcast variables and accumulators
- Spark DataFrames & Datasets: Structured API introduction; DataFrame operations
- Spark SQL: Spark SQL architecture; Integrating Hive with Spark
- Spark MLlib Basics: Machine learning pipeline overview; Classification and regression examples
- Spark Streaming Intro: Micro-batch processing; DStreams and window operations

Hands-on Labs:

- Advanced RDD transformations
- DataFrame creation and querying
- SQL queries on Spark DataFrames
- Building a simple ML model
- Real-time word count example

Week 5: — Performance Tuning & Optimization

Topics:

- Hadoop Performance Tuning: Identifying bottlenecks (CPU, Memory, I/O, Network); HDFS and YARN parameter tuning
- Spark Performance Optimization: Partition tuning, serialization, memory management; Dynamic allocation and parallelism

- Troubleshooting Common Issues: Debugging failed jobs; Handling data skew and resource contention
- Backup & Recovery Strategies: Using distcp, snapshots, and fsck; Metadata recovery and rolling upgrades
- Rolling Restarts & Upgrades: Rolling vs. non-rolling upgrades; Daemon restart sequences

Hands-on Labs:

- Log analysis and parameter adjustment
- Tuning slow Spark jobs
- Fixing a “broken” cluster
- Backup between clusters
- Simulated rolling restart

Week 6: — Advanced Administration & Integration

Topics:

- Resource Isolation & Multi-tenancy: Linux cgroups and YARN node labels; Queue management and resource pools
- Advanced Hive & HBase Integration: Hive on Tez/Spark; HBase as a data source
- Oozie Workflow Scheduling: Defining workflows and coordinators; Scheduling recurring jobs

Hands-on Labs:

- Configuring multi-tenant queues
- Hive-HBase integration
- Creating an ETL workflow

Week 7: — Capstone Project Phase 1

Topics:

- Project Kickoff & Requirements: Defining the business problem and dataset; Architecture planning (HDFS, YARN, Hive, Spark)
- Data Pipeline Design: ETL workflow design; Tool selection (Sqoop, Flume, Spark)
- Data Ingestion & Cleaning: Ingesting structured and unstructured data; Data quality checks and cleaning
- Processing & Analytics: Implementing analytics logic in Spark/Hive; Building ML models if applicable
- Visualization & Dashboards: Connecting to BI tools (Tableau, Power BI); Creating dashboards and reports

Hands-on Labs:

- Architecture planning
- ETL workflow design
- Data cleaning and ingestion
- Implementing analytics logic
- Dashboard creation

Week : — Capstone Project Phase 1

Topics:

- Project Kickoff & Requirements: Defining the business problem and dataset; Architecture planning (HDFS, YARN, Hive, Spark)
- Data Pipeline Design: ETL workflow design; Tool selection (Sqoop, Flume, Spark)
- Data Ingestion & Cleaning: Ingesting structured and unstructured data; Data quality checks and cleaning
- Processing & Analytics: Implementing analytics logic in Spark/Hive; Building ML models if applicable
- Visualization & Dashboards: Connecting to BI tools (Tableau, Power BI); Creating dashboards and reports

Hands-on Labs:

- Architecture planning
- ETL workflow design
- Data cleaning and ingestion
- Implementing analytics logic
- Dashboard creation

Labs

Lab assignments will focus on the practice and mastery of contents covered in the lectures, and introduce critical and fundamental problem solving techniques to the students.

- Understand core Big Data concepts and the Hadoop ecosystem
- Set up and manage Hadoop clusters with HDFS and YARN
- Process and analyze large datasets using MapReduce, Hive, and Pig
- Build scalable data pipelines using Apache Spark (PySpark & Spark SQL)
- Ingest and manage data using Sqoop, Flume, and Kafka
- Design and implement an end-to-end Big Data project with analytics and visualization

deerwalk training center

Sifal, Kathmandu, Nepal
Phone: +977 - 01 - 5913021 | 4567153
Mobile: +977 - 9765355167 | 9860422021
Email: training@deerwalkcompware.com
Website: deerwalktrainingcenter.com