**Module 1: Introduction to Big Data**

* Definition and Characteristics of Big Data (Volume, Velocity, Variety, Veracity, Value)
* Importance of Big Data in Modern Businesses
* Overview of Big Data Ecosystem
* Traditional Data Processing vs. Big Data Analytics
* Use Cases of Big Data Analytics in Various Industries (Healthcare, Finance, Retail, etc.)

**Module 2: Big Data Infrastructure**

**2.1 Hadoop Ecosystem**

* Introduction to Hadoop Architecture
* Hadoop Distributed File System (HDFS)
  + Features and Benefits
  + File Structure and Block Concept
* MapReduce Programming Model
  + How MapReduce Works (Map, Shuffle, Reduce)
  + Writing and Executing MapReduce Jobs

**2.2 Apache Spark**

* Introduction to Spark and its Ecosystem
* Spark Core Concepts (RDDs, DataFrames, and DataSets)
* Spark SQL for Structured Data Processing
* Spark Streaming for Real-Time Data Processing
* Apache Spark MLlib for Machine Learning

**Module 3: Data Ingestion and ETL Processes**

* Data Sources and Data Ingestion Techniques
  + Structured Data: Relational Databases, CSV, JSON
  + Unstructured Data: Text, Images, Videos
* ETL Process Overview (Extract, Transform, Load)
* Tools for Data Ingestion
  + Apache Flume for Log Data
  + Apache Sqoop for RDBMS Integration
* Data Preprocessing and Cleaning Techniques

**Module 4: Data Storage and Management**

**4.1 Distributed Storage Systems**

* HDFS (Hadoop Distributed File System)
* Amazon S3 for Cloud Storage
* Google Cloud Storage

**4.2 NoSQL Databases**

* Introduction to NoSQL Concepts
* Key-Value Databases (Redis)
* Document Stores (MongoDB)
* Wide-Column Databases (Apache Cassandra)
* Graph Databases (Neo4j)

**Module 5: Big Data Analytics Frameworks**

* Batch Processing vs. Stream Processing
* Real-Time Analytics with Kafka and Spark Streaming
* Tools for Data Analysis
  + Apache Hive for SQL-Based Analysis
  + Pig for Procedural Scripting
* Data Workflow Management
  + Apache Airflow
  + Oozie

**Module 6: Machine Learning on Big Data**

* Introduction to Scalable Machine Learning
* Using MLlib for Big Data Machine Learning
* Implementing Supervised and Unsupervised Learning Algorithms
  + Regression, Classification, Clustering
* Hyperparameter Tuning and Model Evaluation in Spark
* Big Data Predictive Analytics Use Cases

**Module 7: Data Visualization and Reporting**

* Overview of Data Visualization Tools
  + Tableau, Power BI
  + Apache Superset
* Visualizing Big Data with Python (Matplotlib, Seaborn)
* Real-Time Dashboards
  + Integrating BI Tools with Big Data Frameworks
* Best Practices for Reporting and Storytelling with Data

**Module 8: Big Data Security and Governance**

* Data Privacy and Protection Mechanisms
* Data Encryption and Access Control in Big Data
* Compliance with GDPR, HIPAA, and Other Regulations
* Implementing Role-Based Access Control (RBAC) in Big Data Systems
* Auditing and Monitoring Big Data Environments

**Module 9: Hands-On Projects**

1. **Building a Data Lake**
   * Setting up a centralized data repository
   * Storing structured and unstructured data using Hadoop and S3
2. **Real-Time Analytics with Kafka and Spark Streaming**
   * Processing live data streams
   * Generating real-time insights (e.g., sentiment analysis)
3. **ETL Pipeline Implementation**
   * Ingesting data from multiple sources
   * Performing transformations and loading data into a data warehouse
4. **Developing a Big Data ML Model**
   * Predictive modeling on large-scale datasets using Spark MLlib
   * Evaluating and optimizing the model
5. **Visualizing Big Data Insights**
   * Creating dashboards using Tableau or Power BI
   * Presenting key findings from a dataset

**Module 10: Industry Applications and Case Studies**

* Retail Analytics: Customer Segmentation and Recommendation Systems
* Financial Analytics: Fraud Detection and Risk Management
* Healthcare Analytics: Predictive Models for Patient Care
* Social Media Analytics: Sentiment Analysis and Trend Prediction
* Manufacturing Analytics: Predictive Maintenance

**Module 11: Big Data Trends and Future Scope**

* Serverless Computing and Big Data
* Big Data in the Era of AI and IoT
* Role of Cloud Providers in Big Data Analytics (AWS, Azure, Google Cloud)
* Emerging Tools and Technologies
  + Delta Lake
  + Snowflake
  + Apache Iceberg