**UNIT 1**

1. **Evolution of Big Data.**
2. **Introduction of Big Data.**
3. **Best Practices for Big Data Analytics.**
   1. **Best Practices.**
      1. Define Your goal.
      2. Choose the right tools
      3. Clean and preprocess your data.
      4. User machine learning Algorithm.
      5. Visualize yours results.
      6. Monitor and optimize.
      7. Ensure data security and privacy.
      8. Stay up-to-date with latest technologies.
4. **Big Data Characteristics.**
   1. **Characterizes:**
      1. Volume
      2. Velocity
      3. Varity
      4. Veracity
5. **Validating The Promotion of the Value of Big Data.**
6. **Big Data use cases.**
   1. **Uses cases:** 
      1. Retail and E-commerce
      2. Healthcare.
      3. Finance
      4. Manufacturing and Supply Chain.
7. **Perception and Quantification of Value.**
8. **Understanding Big Data Storage**
   1. **Challenges in Big Data Storage**
      1. Volume.
      2. Velocity.
      3. Variety.
      4. Veracity.
   2. **Technologies and Approaches for Big Data Storages.**
      1. Distributed File System.
      2. Cloud Storage.

***UNIT 2***

1. **General Overview of High-Performance Architecture**
2. **HDFS (Hadoop Distributed File System):**
   1. **Components:**
      1. Name Node.
      2. Data Node.
      3. Block
      4. Secondary Name Node.
      5. Client.
3. **MapReduce:**
   1. **MapReduce Frameworks (type) :**
      1. Vanilla MapReduce.
      2. Apache Sparks.
4. **YARN (Yet Another Resource Negotiator**
   1. **Types:** 
      1. Vanilla YARN.
      2. Kubernetes.
5. **Big Data Overview Analysis of data at Rest-Hadoop Analytics:**
6. **Limitation of existing distributed systems:**
   1. **Limitations:**
      1. Complexity.
      2. Latency.
      3. Programing Complexity.
      4. Limited Support for SQL.
      5. Storage Overheads.
      6. Community Support’s
7. **Hadoop Approach.**
8. **Hadoop Architecture.**
   1. **Architecture:**
      1. Hadoop Distributed File System (HDFS).
         1. NameNode
         2. DataNode.
      2. Resources Management.
         1. YARN (Yet Another Resources Negotiator).
      3. Data Storage.
         1. HDFS
      4. Data Processing.
         1. Map Reduce.
         2. Apache Spark.
         3. Apache Hive.
         4. Apache Pig.
         5. Apache Tez
         6. Data Visualization and Reporting.
         7. Security.
         8. Monitoring and Management.
         9. Metadata Management.
9. **Distributed File System: HDFS and GPFS.**
   1. **HDFS**
      1. **Type:**
         1. Vanilla HDFS.
         2. HDFS Federation**.**
   2. **GPFS**
      1. **Type:**
         1. Standard Spectrum Scale.
         2. Spectrum Scale for AI.
10. **Internal of Hadoop MR Engine:**
    1. **Type:**
       1. Apache Spark
       2. Apache Tez
11. **Hadoop Cluster Component.**
    1. **Component:**
       * 1. NameNode (Master Node).
         2. DataNode (Worker Node)
         3. Resource Manager (Master Node).
         4. NodeManager (Worker Node).
         5. JobTracker (Deprecated-Hadoop 1. ex).
         6. Gateway/Client.
         7. ZooKeeper.
12. **Hadoop Ecosystem**
    1. **Components**
       * 1. Hadoop Distributed File System.
         2. MapReduce.
         3. Apache Hive.
         4. Apache Pig.
         5. Apache Spark.
13. **Evaluation criteria for distributed Map Reduce runtimes**
    1. **Evaluation:**
       * 1. Fault Tolerance.
         2. Scalability.
         3. Performance.
         4. Support for Various Workloads.
         5. Security and Access Control**.**

***Unit 3***

1. **Overview of Clustering K-means**
   1. **Steps:**
      * Assignment Sete.
      * Update Step.
      * Repeated.
   2. **Advantages.**
      * Simplicity.
      * Scalability.
      * Interpretability.
   3. **Limitations:**
      * Sensitivity to Initialization.
      * Determining the Number of Cluster (K).
      * Sensitivity to Outliers.
   4. **Use Cases**
      * Customer Segmentation
      * Image Segmentation.
      * Anomaly Detection.
      * Recommendation System.
2. **Determining the Number of Clusters**
   1. **Type:**
      * Elbow Method
      * Gap Statistics.
3. **Clustering**
   1. **Algorithm:**
      * K-mean Clustering.
      * Hierarchical Clustering.
      * DBSCAN (Density-Based Spatial Clustering of Application with Noise)
4. **Classification**
   1. **Key Components**
      * Training Data.
      * Features Extraction and Selection.
      * Model Training.
      * Model Evaluation.
      * Prediction or Inference.
5. **Segmentations**
   1. **Type:**
      * Demographic Segmentations.
      * Geographic Segmentations.
      * Behavioral Segmentation
6. **Linear Regression**
   1. **Key-Component.**
      * Dependent and Independent Variable.
      * Linear Relationship.
      * Model Evaluation.
   2. **Type:**
      * Simple Linear Regression.
      * Multiple Linear Regression.
7. **ML Search**
8. **Indexing.**
   1. **Type:**
      * Bree-Tree Indexing.
      * Hashing
      * Binary Search Indexing.
      * Bitmap Indexing.
9. **Created Inverted index using JAQL.**
   1. **Outlier/Component:**
      * Job Identification
      * Job Bundling Logic.
      * Resources Allocation and Scheduling.
      * Data Transfer Optimization.
      * Fault Tolerance and Error Handling.
      * Monitoring and Reporting.
   2. **Application**
      * Data Warehousing.
      * Recommendation Systems.
      * Fraud Detections
      * Sentiment Analysis.
10. **Classification.**
    1. **Components of Classification:**
       * Training Data
       * Feature Extraction and Selection
       * Model Training
       * Model Evaluation
       * Prediction or Inference
    2. **Common Classification Algorithms:**
       * Logistic Regression
       * Support Vector Machines (SVM)
       * Decision Trees and Random Forests
       * Naive Bayes Classifier
       * Neural Networks
11. **Classification: Decision Trees**
    1. **Decision Trees for Classification.**
       * Concept
       * Splitting Criteria
       * Tree Growing.
       * Pruning
       * Handling Categorical and Numerical Data.
       * Ensemble Methods.
    2. **Application**
       * Medical Diagnosis.
       * Email Spam Detection.
    3. **Components of a Decision Tree:**
       * Root Node
       * Internal Node
       * Leaf Node
       * Branches
    4. **Algorithm**
       * Selecting the Root Node.
       * Splitting Data
       * Recursive Splitting.
       * Stopping Criteria.
       * Handling Categorical and Numerical Data
       * Pruning
       * Assigning Class Labels
       * Prediction and Evaluation.
12. **Evaluating a Decision tree.**
    * + Confusion Matrix.
      + Accuracy.
      + Precision and Recall
      + Validation Set.

***UNIT 4***

1. **Introduction to stream computing**
   1. **Key element:**
      * Data Streams
      * Real-time Processing.
      * Scalability and Fault Tolerance.
      * Low Latency.
2. **Challenges/Limitations of Conventional Systems.**
   1. **Key Challenges**
      * Complex Data Analysis.
      * Data Security and Privacy.
      * Data Volume and Variety
      * High Maintenance Costs.
      * Processing speed.
      * Resources Utilization.
3. **Challenges to be solved - scalability, thread pooling etc:**
   1. **Scalability**
      * Load Balancing.
      * Vertical Scaling
      * Horizontal Scaling.
      * Distributed File Systems.
   2. **Thread Pooling.**
      * Thread Reuse
      * Thread Synchronization
      * Thread Pool Management.
      * Task Prioritization.
4. **Understand the challenges in handling streaming data from the real world and how to address those using stream computing:**
   * + Data Volume Management.
     + Data Variety Handling.
     + Data Velocity Processing.
     + Low Latency Processing.
5. **Benefit of stream computing in Big Data world**
   1. **Benefit.**
      * Real Time Analytics
      * Scalability and Flexibility.
      * Improved Customer Experience.
6. **Realtime Analytics Platform (RTAP)**
   1. **Key components:**
      * Data Ingestion and Integrations.
      * Event Detection and Alerting.
      * Data Visualization and Dashboards.
      * Machine Learning Integration
      * Data Security and Compliance.
7. **Real Time Sentiment Analysis**
   1. **Key Component:**
      * Data Collection
      * Natural Languages Processing
      * Sentiment Classification
      * Scalability and Performance
      * Contextual Understanding.
      * Visualization and Reporting
      * Application Integration.