Module 4 - RHadoop

RHadoop integrates R and Hadoop for large-scale data analysis. It includes packages like RHDFS (HDFS interaction), RHive (Hive queries in R), Rmr2 (MapReduce in R), and RHbase (HBase interface). The module explains data requirements design, data analytics lifecycle (problem identification, data preprocessing, performing analytics), and visualization using R packages like ggplot2 and rCharts.

Module 5 - Apache Spark

Apache Spark is a fast, in-memory data processing framework supporting batch, streaming, ML, and graph processing. Core components include Spark SQL, Spark Streaming, MLlib, and GraphX. The module covers Spark's architecture, benefits over MapReduce, and programming with PySpark and Scala shells.

Module 6 - Programming with RDD

RDDs (Resilient Distributed Datasets) are immutable collections enabling fault-tolerant parallel computation. Key features include lazy evaluation, transformations (map, filter), and actions (count, collect). It compares RDDs with DataFrames and Datasets, details narrow vs wide transformations, and includes examples like log mining and word count using PySpark.

Module 7 - Mining Data Streams

Data stream mining involves processing real-time, continuous data with memory and time constraints. Key topics include stream processing architecture, event vs processing time, concept drift, sliding windows, filtering, and statistical moment estimation. Applications include real-time monitoring, fraud detection, and predictive maintenance.

Module 8 - Case Studies

The case study demonstrates stock market prediction using MapReduce in R without RHadoop. It involves data acquisition from Yahoo Finance, preprocessing, and analytics using Hadoop Streaming. Visualization with ggplot2 in R helps identify stock movement patterns for informed

investment decisions.