
Assignment: End-to-End Big Data Analytics Project Using Hadoop and Spark

Objective

To give participants hands-on exposure to Big Data technologies by building a small data processing and analytics pipeline. The assignment will help them:

- Understand Big Data concepts and Hadoop ecosystem basics.
- Perform operations on HDFS.
- Work with Spark Core (RDDs), Spark SQL (DataFrames), and MLlib for analytics and machine learning.

Problem Statement

"Retail Analytics Platform"

A retail company wants to analyze its historical sales data to gain insights into:

- Top-selling products and revenue generation.
- Customer purchasing patterns.
- Predict whether a customer is likely to make a repeat purchase.

You are required to build a **Big Data processing pipeline** to achieve this using **Hadoop and Spark**.

Assignment Tasks

Part 1 – Big Data and Hadoop Basics

1. **Theory Submission** (2–3 pages):
    - Explain **what Big Data is**, why it is important for enterprises, and list real-world use cases.

- Describe **Hadoop ecosystem components** – HDFS, YARN, MapReduce.
- Explain **HDFS architecture** and how data is stored/retrieved.

2. **HDFS Hands-On:**
   - Set up **HDFS** (local or pseudo-distributed mode).
   - Perform the following:
     - Create directories in HDFS.
     - Upload sample CSV files (`customers.csv`, `orders.csv`, `products.csv`) to HDFS.
     - List files, view file contents, and copy files from HDFS to local.
   - Submit **screenshots and command outputs**.

Part 2 – Spark Core (RDD Operations)

1. **Spark Setup:**
   - Install and configure **Apache Spark** in local mode.
2. **RDD Operations:**
   - Load sales data from HDFS.
   - Perform the following:
     - Transform raw data into **key-value pairs**.
     - Find **total sales per product** using `reduceByKey`.
     - Identify **top 5 customers by purchase value**.
     - Use **broadcast variables** for product reference data.
     - Use **accumulators** to count invalid records.
   - Submit **code and outputs**.

Part 3 – Spark SQL and DataFrames

1. **DataFrame Creation:**
    - Read CSV data (`customers`, `orders`, `products`) into Spark DataFrames.
2. **Spark SQL Queries:**
    - Register DataFrames as **temporary views**.
    - Write and execute Spark SQL queries:
        - List **customers with total spend > X**.
        - Get **monthly sales trends**.
        - Find **top-selling product category**.
3. **Multiple Data Formats:**
    - Save query results to **Parquet** and **JSON** formats in HDFS.

Part 4 – Spark MLlib

1. **Feature Engineering:**
    - Prepare a dataset with features like
    - customer age, frequency of purchases, average order value, etc.**MLlib Pipeline:**
    - Split dataset into train/test.
    - Build a **classification model** (e.g., Logistic Regression or Decision Tree) to predict if a customer will make a repeat purchase.
    - Evaluate model performance using **accuracy** or **F1-score**.
2. **Deliverables:**
    - Submit Jupyter Notebook or `.py` script with full pipeline.
    - Include brief documentation explaining each step.

Submission Deliverables

- A single compressed folder containing:
    1. **Theory write-up** (PDF/Word).
    2. **Screenshots of Hadoop HDFS operations.**
    3. **Spark RDD scripts and outputs.**
    4. **Spark SQL queries and results.**
    5. **MLlib notebook/script with explanation.**
- Naming convention: `BatchName_Assignment_RetailAnalytics.zip`
