CodTech Internship Task - 1

Big Data Analysis using PySpark

Internship Report

Objective:

Perform analysis on a large dataset using PySpark to demonstrate scalability and derive insights.

Tools & Dataset Used

Tools Used:

- Apache Spark (PySpark)

- Google Colab / Jupyter Notebook

- NYC Yellow Taxi Trip Dataset (Large-scale public data)

- Python

Insights Generated

Insights Derived:

1. Average trip distance by payment type

2. Busiest pickup hours

3. Top 10 pickup locations

4. Passenger trends over time

5. Total revenue per day

Code Execution Summary

CodTech Internship Task - 1

Big Data Analysis using PySpark

PySpark Code Steps:

1. Import required libraries and start Spark session

2. Load and clean large CSV dataset

3. Perform analysis and create new columns (e.g. pickup hour)

4. Group and aggregate data to derive meaningful insights

5. Show final outputs and optionally save results

Final Deliverable

Deliverable:

- A complete PySpark script or notebook analyzing the dataset

- Includes data cleaning, aggregation, and insights

- Task completed successfully according to internship guidelines