CODTECH Internship – Big Data Analysis Report

Student Name: Lavaraju ADARAPU

Institution: ESAIP Engineering School

Subject: Internship – Big Data (PySpark)

Submission Date: 26/07/2025

Project Title:

Taxi Trip Data Analysis using PySpark

Objective:

To analyze taxi trip data using PySpark by processing, cleaning, aggregating, and visualizing trends in average fare and trip volume over time. The goal is to demonstrate the use of big data tools on real-world-like data.

Dataset Used:

File Name: dummy_data.csv

Number of Rows: 1000

Fields Included:
- VendorID – Taxi vendor (1 or 2)
- tpep_pickup_datetime – Pickup timestamp
- tpep_dropoff_datetime – Drop-off timestamp
- passenger_count – Number of passengers
- trip_distance – Distance traveled (miles)
- fare_amount – Total fare charged

Tools & Technologies Used:

- Apache Spark (PySpark)
- Python
- Matplotlib
- Pandas (for plotting)
- Jupyter Notebook / VS Code
- Local File System (No Hadoop required)

Data Processing Steps:

1. Initialization:
   Spark session initialized in local mode using findspark.
2. Data Loading:
   CSV file read into a Spark DataFrame using .read.csv() with inferred schema.
3. Data Cleaning:
   Filtered out rows where:
   - trip_distance <= 0
   - fare_amount <= 0
   - Years not in the range 2009–2025
4. Transformation:
   Extracted year and month from tpep_pickup_datetime.
5. Aggregation:
   Grouped data by year and month and calculated average fare and trip count.

Analysis Output:

A monthly summary table was printed and saved as monthly_summary.txt.

Year	Month	Avg Fare	Trip Count
2019	01	12.30	85
2019	02	14.20	73
2020	06	13.80	95
2021	11	15.10	90
2023	12	16.45	91

Visualization:

Graph Created: monthly_summary.png
Line Plot: Average Fare per Month
Bar Plot: Number of Trips
X-Axis: Year-Month
Y-Axis: Fare ($) and Trips

Output Files:

dummy_data.csv – Input dummy dataset (1000 rows)
monthly_summary.txt – Monthly summary table (text format)
monthly_summary.png – Line+Bar graph (fare & trip count)

Conclusion:

This project demonstrated how PySpark can be used for end-to-end big data analysis — from reading large datasets, performing real-time transformations, to generating summaries and visualizations. The data pipeline was entirely local and did not require Hadoop or cloud resources.

Acknowledgements:

Special thanks to CODTECH and N. SANTHOSH KUMER

for guiding this project.