# CodTech Internship – Task 1: Big Data Analysis using PySpark

Intern Name: Shravya Bhandary

Tool Used: PySpark

Dataset Used: Online Retail Dataset (UCI Machine Learning Repository)

## Key Insights

* • The dataset contains approximately 50,000 transaction records.
* • United Kingdom had the highest number of transactions among all countries.
* • The top-selling product was 'WHITE HANGING HEART T-LIGHT HOLDER'.
* • Some transactions had negative quantities, which may indicate returns or cancellations.
* • PySpark handled the large dataset efficiently, showcasing its scalability and speed.

## Conclusion

This task demonstrated how PySpark can be used to analyze large-scale datasets. Using Spark DataFrames, we efficiently processed and summarized the retail data. This hands-on experience introduced scalable big data analysis techniques suitable for real-world applications.