Project Summary

Member1 Name: Emaan Ayesha Bin Te Anwer (DS-W-018/2025)

Member2 Name: Saad Ahmed Khan (DS-W-017/2025)

Amazon Reviews Big Data Pipeline

The Amazon Reviews Big Data Pipeline is a comprehensive data engineering and analytics initiative that illustrates how artificial intelligence and big data frameworks can be integrated to derive meaningful insights from massive volumes of e-commerce information. The central goal of this project is to process, analyze, and interpret customer reviews from Amazon to evaluate consumer sentiment and support data driven business strategies.

The pipeline starts with the ingestion of a large dataset containing Amazon product reviews. Using Apache Spark (PySpark), the data undergoes cleaning and transformation to convert it into a well-structured format. This process involves renaming nested attributes, eliminating missing or irrelevant entries, and selecting essential columns such as product name, brand, category, review text, and rating. After preprocessing, the refined data is stored in Parquet format to ensure efficiency, scalability, and faster access during later stages of analysis.

In the subsequent stage, the preprocessed data is subjected to sentiment analysis powered by artificial intelligence. The implementation utilizes the Hugging Face Transformers library with a DistilBERT pretrained model to classify reviews as Positive or Negative based on their textual content. This step highlights the seamless incorporation of Natural Language Processing (NLP) models within a real-world big data pipeline.

The analyzed results are then structured and stored for visualization in tools such as Power BI or Tableau, providing an interactive view of customer behavior, product performance, and sentiment trends. These insights empower organizations to improve decision-making processes and enhance customer satisfaction.

This project showcases the entire data lifecycle from data acquisition to intelligent analysis demonstrating the effectiveness of combining AI techniques with big data systems. Its modular design ensures adaptability, allowing similar pipelines to be applied to other domains such as market research, brand reputation monitoring, and consumer behavior analysis.

Overall, the Amazon Reviews Big Data Pipeline exemplifies practical skills in data orchestration, ETL (Extract, Transform, Load) operations, machine learning integration, and analytical visualization.