E-commerce Data Pipeline for Customer Behavior Analysis

Problem Statement: An e-commerce company wants to analyze customer behavior on their website to gain insights into customer preferences, purchasing patterns, and overall user experience. They need to build a data pipeline that collects, processes, and analyzes data from various sources to generate meaningful insights for decision-making. As a data engineer, your task is to design and implement the data pipeline to support this analysis.

Dataset:

The dataset for this project will consist of the following sources:

- Website Clickstream Data: Contains information about customer activities on the website, including page views, clicks, timestamps, and user IDs.
- Purchase Data: Contains details of customer purchases, such as product IDs, quantities, prices, and transaction IDs.
- Customer Data: Includes customer demographics, such as age, gender, location, and preferences.

Project Steps and Source Code:

Data Collection:

- Set up data collection tools (e.g., web trackers, log files) to capture website clickstream data.
- Integrate with the e-commerce platform's APIs to fetch real-time purchase data.
- Import customer data from a relational database or flat files.

Data Ingestion:

- Create an ingestion process to receive and store the raw data from different sources.
- Use tools like Apache Kafka or Apache Nifi for real-time data ingestion.
- Store the data in a data lake or distributed file system (e.g., HDFS).

Data Processing:

- Design ETL (Extract, Transform, Load) processes to cleanse and transform the raw data.
- Implement data quality checks and filtering to ensure data integrity.
- Perform data enrichment by joining different datasets based on common keys.
- Utilize Apache Spark for distributed data processing and transformation.
- Apply data modeling techniques (e.g., data normalization, denormalization) as per the analysis requirements.

Data Storage:

- Choose a suitable database or data warehouse (e.g., Apache Hive, Apache HBase) for storing processed data.
- Create optimized tables and partitions for efficient querying and analysis.
- Ensure data security and privacy measures are in place.

Data Analysis and Visualization:

- Use SQL queries or Spark SQL to extract relevant insights from the processed data.
- Perform customer segmentation based on demographics, purchase history, and website interactions.
- Generate reports, dashboards, and visualizations using tools like Apache Superset, Tableau, or Power BI.