Analysis of Customer Purchase Behaviour Using Hadoop, Python, and SQL

Project Description: The objective of this project is to analyse customer purchase behaviour by processing and analysing a large dataset using Hadoop. The data will be stored in HDFS and processed using Python and SQL.

Project Steps:

1. Data Collection: Collect a large dataset of customer purchase transactions from a retail store or e-commerce website. The dataset should contain fields such as customer ID, purchase amount, purchase date, and product ID.
2. Data Preparation: Prepare the dataset by cleaning and transforming the data. This may include removing duplicate records, handling missing values, and converting data types.

# Python code to clean and transform the dataset

import pandas as pd

# Read the dataset

df = pd.read\_csv('customer\_transactions.csv')

# Remove duplicates

df.drop\_duplicates(inplace=True)

# Handle missing values

df.dropna(inplace=True)

# Convert data types

df['purchase\_date'] = pd.to\_datetime(df['purchase\_date'])

1. Data Loading: Load the cleaned dataset into HDFS.

# Hadoop command to load data into HDFS

hadoop fs -put customer\_transactions.csv /user/hadoop/customer\_transactions.csv

1. Data Processing: Use Hadoop MapReduce to process the dataset and generate insights on customer purchase behavior.

# Python code to define Map and Reduce functions

def mapper(record):

customer\_id, product\_id , purchase\_date, purchase\_amount = record.split(',')

yield customer\_id, float(purchase\_amount)

def reducer(customer\_id, purchase\_amounts):

total\_purchases = sum(purchase\_amounts)

yield customer\_id, total\_purchases

# Hadoop command to run MapReduce job

hadoop jar hadoop-streaming.jar \

-file mapper.py -mapper 'python mapper.py' \

-file reducer.py -reducer 'python reducer.py' \

-input /user/hadoop/customer\_transactions.csv \

-output /user/hadoop/customer\_purchases

1. Data Visualization: Use Python and SQL to visualize the insights generated by the data processing step.

# Python code to visualize customer purchase behavior

import matplotlib.pyplot as plt

import pyodbc

# Connect to SQL database

conn = pyodbc.connect('Driver={SQL Server};'

'Server=server\_name;'

'Database=db\_name;'

'Trusted\_Connection=yes;')

# Query data

cursor = conn.cursor()

cursor.execute('SELECT customer\_id, total\_purchases FROM customer\_purchases')

# Create a bar chart of total purchases by customer

data = cursor.fetchall()

df = pd.DataFrame(data, columns=['customer\_id', 'total\_purchases'])

plt.bar(df['customer\_id'], df['total\_purchases'])

plt.xlabel('Customer ID')

plt.ylabel('Total Purchases')

plt.show()

1. Conclusion: Draw conclusions from the analysis and make recommendations for the retail store or e-commerce website based on the insights gained.

Tools and Technologies Used:

* Hadoop: for storing and processing large amounts of data
* Python: for data processing and visualization
* SQL: for querying and analysing the data in Hadoop

Expected Outcome: At the end of this project, you should be able to analyse customer purchase behaviour and generate insights that can be used to improve the customer experience and increase revenue for the retail store or e-commerce website.