Data Analysis on Shopping

Objective:

Online shopping website are getting more popular and had a huge increase in the past 10 years. To be able to easily save money and compare prices from website to website. Online resellers tent to sell at a lower price due to less overhead expenses. To shop while in the comfort of your own home, without having to step out of the door. In this project, we are working on E-commerce huge Datasets to extract some useful information so that the information can be used by a particular Organization for Maximum Business Profit, Customer Satisfaction, launching a new product.

Given Data:

Two Datasets are as below used for this project-

1. Customer Dataset is having 5 fields – Customer ID, Customer Name, LastName, Age, Profession.

2. Transaction Dataset is having 8 Fields – Transaction ID, Date, Customer ID, Amount, Hobby, Area of Interest, Locality, City, Payment Mode.

Project Description:

Use Case 1: Based upon the Amount:

* Shopper analysis all transaction of amount greater than 160 based upon that shopper give the hats off points to the customers.
* We are analyze the transaction range between 175 and 200 amount based upon that we are giving 20 percent discount of that product.

Uses Case 2: Based upon the Transactions

* Based upon the total count and total transaction we are identify the customer needs of the product .Based upon that we have to include the products in the shop.

We are decrease the count of the non-sell products and also increase the profit of the shop.

**Use Case 3: Based upon Maximum spend Transaction**

* **In this case we are finding the customer who has spent maximum amount in the shop.**

**We have to give the discount by 25 percent of the product and it helps to improve the reference of the customer as well.**

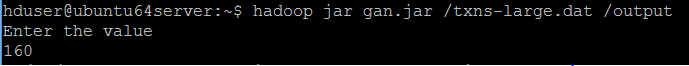
**Use Case 4: Based upon the Maximum amount transaction spend by the customer for the particular month**

* **We are give the coupons worth amount of rupees 2000 to the customers.**
* **We are announce that particular customer is the star of the month in our shop.**

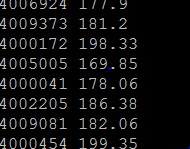
**Output:**

Task1: Find all the transaction where amt>160

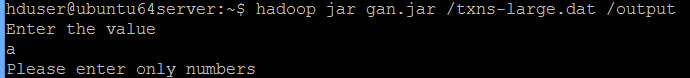
Input:



Output:

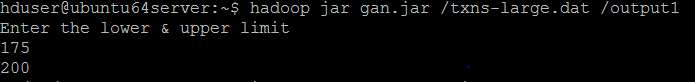


Data Validation:



Task2: Count all the transaction where amount is between 175 to 200.

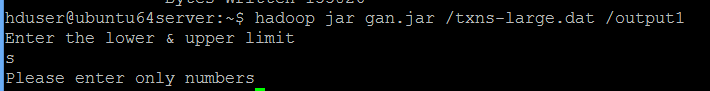
Input:



Output:



Data Validation:



Task3: Calculate the total sum and total count of all the transaction for each user id.

Calculate the average transaction value for each user id.

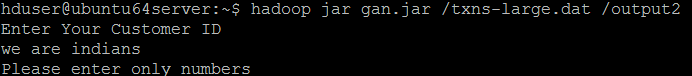
Input:



Output:



Data Validation:



Task4: Calculate total sales amt for each Month.

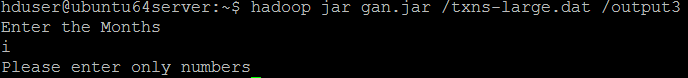
Input:



Output:



Data Validation:



Task5: Divide the file into 12 files, each file containing each month of data. For example. file 1 should contain data of january txn, file 2 should contain data of feb txn.

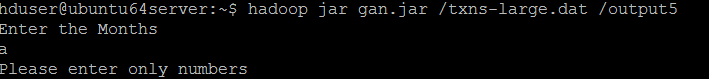
Input:



Output:

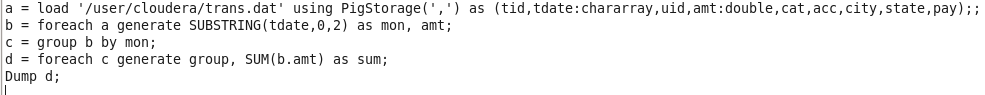


**Data validation:**

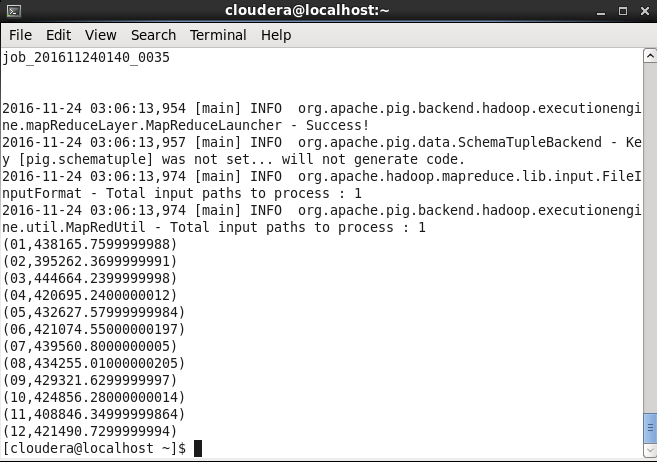


Task6: Sort the whole file on the basis of amt

Input:

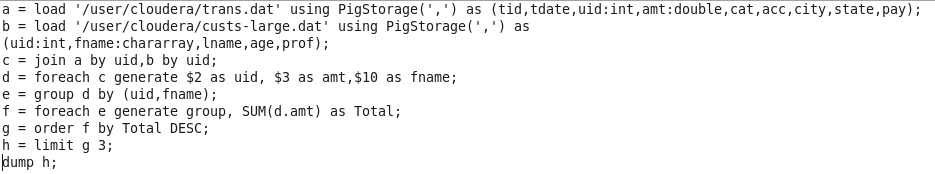


Output:

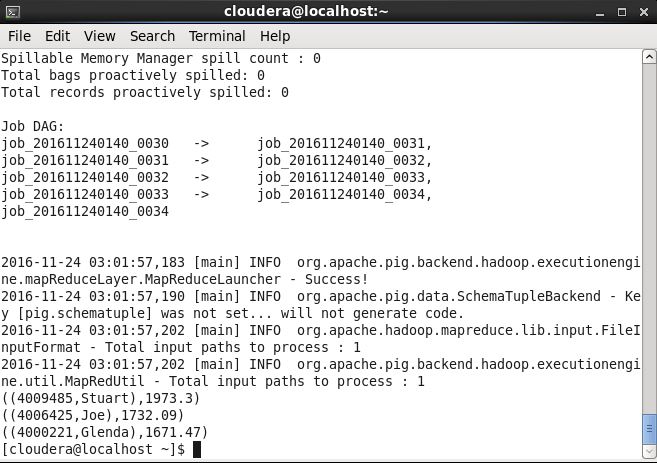


Task7: Find the name of top 3 spenders

Input:

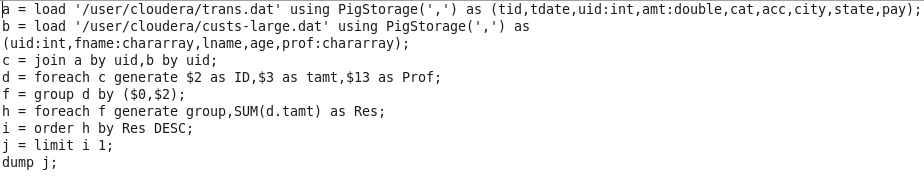


Output:

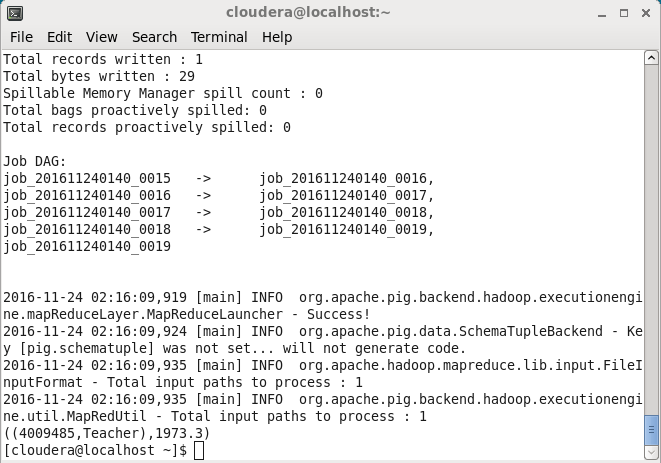


Task8: Find the name of user who has spend the maximum amount.

Input:



Output:



Technology used:

Apache Hadoop: Apache Hadoop an open-source software framework used for distributed storage and processing of very large data sets. It consists of computer clusters built from commodity hardware. Java

MapReduce Program: Hadoop MapReduce is a software framework for easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner.

Apache Hive: Apache Hive is data warehouse infrastructure built on top of Apache Hadoop for providing data summarization, ad-hoc query, and analysis of large datasets. It provides a mechanism to project structure onto the data in Hadoop and to query that data using a SQL-like language called HiveQL (HQL).

Apache Pig: Apache Pig is a high-level platform for creating programs that run on Apache Hadoop. The language for this platform is called Pig Latin. Pig can execute its Hadoop jobs in MapReduce.

SOFTWARE Used:

1) Virtual Box

2) Eclipse

3) Ubuntu Terminal (for MapReduce)

4) Cloudera OS (for HIVE) SYSTEM

REQUIREMENT:

* Minimum 8GB RAM.
* Next Generation Processor Chips like Intel I7.
* Windows 64 Bit System

Conclusion:

Clickstream data and monitoring online behavior can help optimize e-commerce sites. Without the assistance of big data, the sheer volume of clickstream data would be difficult to analyze. And retailers can incorporate other metrics such as social media shares, purchase history, and more to improve performance for e-commerce websites.

The conclusion for this project is to increase profit as well as customer satisfication of the organization.