Assignment 2 – Data Analysis with Pig & Hive

COS20028 Big Data Management and Analytics

Semester 2, 2021

Assignment Worth:	**15**% of total marks
Project:	Individual
Submission deadline	October 30th, 2021 (11.59PM)

CASE STUDY: Dual Core Inc.

Apache Pig makes it possible to write complex queries over big datasets using a language named Pig Latin. The queries are transparently compiled into MapReduce jobs and run with Hadoop. Like Pig, ***Apache Hive*** makes it possible to write complex queries over big datasets, but it has the advantage of using HiveQL, a language that is very easy to learn for people who are used to work with databases since it is very similar to SQL.

In this assignment, students are tasked to perform data analysis for Dualcore Inc. using Apache Pig & Hive with the raw data given.

Dataset & Data Model

Refer to **Data Model Reference** sheet for Dualcore Inc. in Canvas. Assignment 2. The dataset is available in the VM Image supplied in this unit. (Refer to week 09 lab content.)

Submission Requirements

Assignment Tasks

Question 1 – Data Exploration [1 mark]

Develop an ERD based on data model reference sheet given for Dualcore Inc. This is a process for data exploration and get familiarise with the dualcore dataset.

Question 2 – ETL with Pig [2 marks]

Note: Students are expected to complete Lab 9 before attempting following questions.

a) Load **customers** dataset from HDFS, extract their **first name** and **last name** then sort the result by first name in ascending order. Screenshot the result at the end of your code execution. [0.5 mark]

b) List down the name in the specified format (**fname** <space> **lname**) of customers who lives in "**Louisville**" city. Screenshot the result. Screenshot the result at the end of your code execution. [0.5 mark]

c) Group customers dataset based on their **state** and count number of customers in each state. Screenshot the result at the end of your code execution. [0.5 mark]

d) List down each **brand ONCE** in descending order from products dataset and store the result into HDFS directory named "/dualcore/products_brand". Include the output files in your submission. Screenshot the result at the end of your code execution. [0.5 mark]

Question 3 – Data Analysis with Pig [7 marks]

a) It seems like Dualcore Inc. advertising campaign was successful in generating new orders for the month May 2013. The sales manager is interested in knowing whether a customer, who buy tablet also buy other items. Prepare a Pig script to calculate the **max number of items contained in all orders** that contain the advertised tablet (product ID: 1274348) during the campaign period (1 May 2013 – 31 May 2013). Screenshot the result at the end of your code execution. [2 marks]

b) The customer service manager wishes to offer exclusive membership status for customers who make **more than 6 orders in year 2012**. Each of the eligible customer who made total purchase (in price) at least $5,000 but less than $7,000 will be categories as "Silver", $7,000 - $9,999.99 will be "Gold" while customer who made total purchase more than $10,000 will be categories as "Platinum", customers in the remaining range will be classified into "Unknow". Store the total number of customers in each category in an output directory named "dualcore/loyalty". [3 marks]

- You may use the script "loyalty_program.pig" in "/training_materials/analyst/exercises/disparate_datasets/bonus_02" folder as a guide to complete this task.
- Screenshot the total number of customers in each category.

c) Dualcore Inc. stored a total of 201,375 customer's records but not all customers had place orders before. The marketing manager wish to know which ten cities has the most customers who has placed order(s) before so he could plan where to organize roadshow for the next six months. Prepare a pig script ('top_ten_cities_customers.pig') to generate **top ten cities with most customers** and store the output in HDFS named 'dualcore/targeted_cities'. [2 marks]
Hint: there are various way to generate the same expected output.

- Screenshot the output with the top ten cities with the most customers category.

Question 4 – Data Analysis with Hive **[4 marks]**

Note: Students are expected to complete **Lab 10** before attempting following questions.

The Senior Strategic Director would like a quick review on the following queries. Using hive CLI / Hive Query Editor, execute an HiveQL to:

a) Identify how many brands have the USB-related product (Having USB as the term in the product name). What are those brands? Remember to remove duplicate brand. [1/2 marks]

b) Find the total number of orders 'Diana Lamb' has made before. Note that there are a few ways to get the same output. [1/2 marks]

c) Calculate total number of products for each brand in Dualcore. Sort the result by brand in descending order. [1/2 marks]

d) List down the details (brand, name and price) for the most low-cost product in each brand. [1/2 marks]

e) List down top five hot selling products and its brand in Dualcore Inc. [1/2 marks]

f) Calculate the **total revenue <u>in dollar</u>** for the month during the ad campaign period (1 May 2013 – 31 May 2013). [1/2 marks]

g) Breakdown the **total revenue <u>in dollar</u>** for the month during the ad campaign period (1 May 2013 – 31 May 2013) **by day**. [1 mark]

Question 5 – Data Management & Transformation with Hive **[1 mark]**

Note: Students are expected to complete **tasks above** before attempting following questions.

Copy over all "ElCheapo" products into a new table named 'elcheapo_products', view the table 'elcheapo_products' and then drop it. [1 mark]

Hints: *three queries expected.*