billups

Data Engineering Case Study

Welcome to the Data Engineering challenge!

Billups is the largest, independent, privately-held out-of-home (OOH) technology and managed services company in the U.S. Our mission is to transform the out-of-home experience through data science and technology — empowering brands to develop smarter and more accountable OOH campaigns.

The Billups Engineering team is a diverse team of data scientists, engineers and software developers working together to deliver value to our customers. The tasks in this case study have been designed to highlight a subset of the everyday task a data engineer would perform at Billups. We are not looking for a perfect solution, instead, we are more interested in your thought process and how you find and deliver solutions.

Submission deadline: 96 hours from receipt

Dataset

- Data Dictionary
- Historical_transactions.parquet (261 MB)
- Merchants.csv (39 MB)

Note: The merchants.csv and historical_transactions.parquet datasets are related by a common merchant_id column that exists in both files. These two files represent a subset of a commercial database and it is possible to have merchants without corresponding historical transactions. For historical transactions without a matching merchant, use the merchant_id as the name of the merchant.

Tasks

The questions are outlined below with the expected structure of the reports where applicable.

Question 1: Generate the top 5 merchants by purchase_amount for each month in the dataset, for each city in the dataset. For example:

Month	City	Merchant	Purchase Total	No of sales
Oct 2017	2	Merchant A	12,000,000	34000
Oct 2017	2	Merchant B	11,000,000	33000
Oct 2017	2	Merchant C	10,000,000	32000
Oct 2017	2	Merchant D	9,000,000	31000
Oct 2017	2	Merchant E	8,000,000	30000
Oct 2017	3	Merchant A	12,000,000	45000
Oct 2017	3	Merchant B	11,000,000	44000
Oct 2017	3	Merchant C	10,000,000	43000
Nov 2017	2	Merchant A	16,000,000	41000
Nov 2017	2	Merchant B	15,000,000	40000
Nov 2017	2	Merchant C	14,000,000	44000
Nov 2017	2	Merchant D	13,000,000	40000
Nov 2017	2	Merchant E	12,000,000	41000
Nov 2017	3	Merchant A	21,000,000	34000
Nov 2017	3	Merchant B	20,000,000	33000
Nov 2017	3	Merchant C	19,000,000	32000
Nov 2017	3	Merchant D	18,000,000	31000
Nov 2017	3	Merchant E	17,000,000	30000

Question 2: What is the average sale amount (purchase_amount) of each merchant in each state. Consider returning the merchants with the largest sales first:

Merchant	State ID	Average Amount
Merchant A	2	123,000,000
Merchant A	10	19,000,000
Merchant B	2	350,000
Merchant C	10	230,500

Question 3: Identify the top 3 hours where the largest amount of sales (purchase_amount) are recorded for each product category (category). I.e.

Product Category	Hour
A	1300
A	1400
A	1900
W	0800
W	1200
W	1900

Question 4: In which cities are the most popular merchants located. Is there a correlation between the location (city_id) and the categories (category) the merchant sells.

Note: Consider popularity in terms of the number of sales transactions of each merchant.

Question 5: A new merchant is coming in to do business and you have been assigned to give advice based strictly on the historical transactions. You are expected to provide a response to the following questions.

Note: Remember to state your assumptions if any.

a. Which cities would you advise them to focus on and why?
b. Which categories would you recommend they sell
c. Are there particular periods (months) that have interesting sales behaviors?
d. What hours would you recommend they open and close for the day?

e. Would you recommend accepting payments in installments? Assume a credit default rate of 22.9% per month.
   For this question, consider the “installments” header in the historical transactions and the impact it may have, if any, on merchant sales (merchant sales in terms of purchase_amounts). We are making a simplistic assumption that 25% of sales is gross profit to merchants, there are equal installments and everyone who defaulted did so after making half payment.

Cleaning:

1. Use the merchant_id as the merchant name where there is no corresponding merchant name for the merchant IDs in the historical table.
2. Don’t filter out records where the categories are null. You may replace null categories with the text “Unknown category” where applicable.

Submission format:

- You are required to use python as the programming language, using the pyspark interface for apache spark. It is highly recommended to use pyspark and pyspark functions and not spark sql when solving the problems

Submission method:

1. Your solution should contain both a report answering the questions/tasks as well as the code that produced the results
2. Upload your solution to a public github repository, send a link containing your solution to your recruiter

How it will be evaluated

Your submission will be evaluated on the following criteria in no particular order:

1. Correctness
2. Quality of the source code
3. Quality of presentations and results.

Submission deadline: 96 hours from receipt