Data Engineer – Technical Challenge

This is a technical challenge for Data Engineering team. **You have 3 days to answer and send back your answers.** You can write your answers on paper and send us back photos or type them directly here. You can also develop with your favorite IDE or Notebook and send us the code files by mail or a repository link.

1. Coding Challenge

You can write code in your preferred language (Python, Scala, Java or R) Write a function reverse that:

1. Given an array of integers, find the two positions/indices that sum a specific value a. To consider:
   - There is always a solution. There is only one
   - Negative values possible
   - Not previously sorted
   - It fits in memory
2. Print all possible solutions if there more than one.
3. Describe a solution for the previous case when the data does not fit in memory.

2. Spark Challenge

Exercise overview Next exercise is about coding a simple ETL process using Spark. This exercise help us to check out your Spark level at the same time we analyze your coding style. Feel free to use any tool for develop (Notebook, IDE, paper…). You can use the Spark SDK of your choice (preferably Spark 2+) or any other distributed framework like Hadoop/MapReduce, Hive, Pig…

Exercise goal Attach to this document you'll find a “events.csv” file containing users’ actions. Each action has a timestamp and a possible value, either "open" or "close". We would like you to reduce data temporal granularity to 10 minutes, so that there is only one single row for each 10 minutes. Over this temporal aggregation count how many actions of each type there is per minute. After previous calculation, please compute the average number of actions each 10 minutes. Finally, we would like you to compute the top 10 minutes with a bigger amount of "open" action.

Can you do a proposal about how to test this job with unit test, how to test a full pipeline with a integration test and how to release this job on production with data quality check?

3. SQL Challenge

You can write the SQL query or the code necessary to produce the required results.

IMPRESSIONS

Product_id	click	date
1002313003	true	2018-07-10
1002313002	false	2018-07-10
…	….	…

PRODUCTS

Product_id	category_id	price
1002313003	1	10
1002313002	2	15
…	….	…

PURCHASES

Product_id	user_id	date
1002313003	1003431	2018-07-10
1002313002	1003432	2018-07-11
…	….	

1. Given an IMPRESSIONS table with product_id, click (an indicator that the product was clicked), and date, write a query that will tell you the click-through-rate of each product by month
2. Given the above tables write a query that depict the top 3 performing categories in terms of click through rate.
3. Click-through-rate by price tier (0-5, 5-10, 10-15, >15)

4. Data Architecture Challenge

We are managing parking lots that a client can check with a mobile app. An app can tell the drive if a parking is full or not. On entering/leaving the parking a client can scan QR/NFC code on entrance machines and the cost must be automatically charged when leaving parking. We are interested in monitoring when the parking is full or empty to modify prices accordingly. We also would like to create a predictive model that learns when a client is going to the parking to send him a push message informing how many places are left or if the parking is full.

1. What tracking events would you propose? What data model for event analysis? What technologies?
2. How would you design the Backend system? What data model for the Operational system? What technologies?
3. Explain how to combine the operational architecture with the analytical one?
4. Could you propose a process to manage the development lifecycle? And the test and deployment automation?