Data Scientist – Technical Challenge

This is a technical challenge for Data Science team. You can develop with your favorite IDE or Notebook and send us the code files by mail or a repository link.

# Coding Challenge

1. Given an array of integers arr and an integer x, write a function that returns indices of the two numbers such that they add up to x. You can return the answer in any order.

To consider:

▪ Negative values possible

▪ Not previously sorted

▪ It fits in memory

1. Print all possible solutions if there are more than one.
2. Describe a solution for the case when the data does not fit in memory.

**Example 1:**

**Input:** arr = [2,7,11,15], x = 9

**Output:** [0,1]

**Explanation:** Because arr[0] + arr[1] == 9, we return [0, 1].

**Example 2:**

**Input:** arr = [3,2,4,3], x = 6

**Output:** [[1,2], [0,3]]

**Example 3:**

**Input:** arr = [3,3], x = 6

**Output:** [0,1]

**Constraints:**

* 2 <= arr.length <= 104
* -109 <= arr[i] <= 109
* -109 <= x <= 109

**Follow-up**: Can you come up with an algorithm that is less than O(n2) time complexity?

# Spark Challenge

**Exercise overview** Next exercise is about coding a simple ETL process using Spark. This exercise helps us to check your Spark level and 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+).

**Exercise goal** Attached 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 are 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 a unit test, how to test a full pipeline with an integration test, and how to release this job on production with data quality check?

# 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)

# Data Science Challenge

Attached you will find a document “holiday\_rental.csv” that includes information about holiday rentals including rental type, reviews, price, etc. We would like you to:

* Suggest the appropriate features and ML model to predict rental prices
* Using your preferred language train and test the model
* Identify whether the outliers can impact the model and in what way
* Re-train and test the model while addressing the outliers.