

## Skills test for Junior Data Engineer job position

1.

The scenario:

Each day I get CSV files and I never know how clean or dirty the data really is, and if the column order is always the same. The only thing I know is that each CSV file has columns blogId, views, clicks (I don't know if this will be always the order of the columns).

The way I have approach this task is following: at the end of each day I put all the CSV files I got that day and I run the following Python Script. The script takes all CSV files, it puts them in right column order and adds the current date time and saves it in new CSV file. I also uploaded test CSV files so you can run the script in order to test it. Also I have written the comments if order to explain what each code line means.

```
import csv
import datetime
import os
from tabulate import tabulate

#define paths
csv_path = 'C:/Users/SOFIJA/PycharmProjects/SkillTest/jDataEngineer'
output_file = './out.csv'

#if you just want date and not time, put "%B %d, %Y"
current_date = datetime.datetime.now().strftime("%B %d %Y, %I:%M.%S %p")

#Load the output csv file or create an empty dataframe to store the new data
fields = ['blogId', 'views', 'clicks', 'current_date', 'csv']
old_csvs=set()
if os.path.isfile(output_file):
    out_csv = csv.DictReader(open(output_file))
    for row in out_csv:
        old_csvs.add(dict(row) ['csv'])
else:
    with open(output_file, 'a+', encoding='utf-8') as output:
        writer = csv.DictWriter(output, fieldnames=fields)
        writer.writeheader()

#iterate through csv files in the directory
for f in os.listdir(csv_path):
    #process file only if it's not already processed in a previous date and indexed
    in output_file
    if f.endswith('.csv') and f not in old_csvs:
        csv_file = os.path.join(csv_path,f)
        csv_data = csv.DictReader(open(csv_file))
        with open(output_file, 'a+', encoding='utf-8') as output:
            writer = csv.DictWriter(output, fieldnames=fields)
            for row in csv_data:
                row_clean = dict((k.lower().replace(' ', ''), v) for k,v in
dict(row).items())
                row_clean['current_date'] = current_date
                row_clean['csv'] = f
                #write row for output file
                writer.writerow(row_clean)

#print data
rows = []
```

```

out_csv = csv.DictReader(open(output_file))
for row in out_csv:
    rows.append(row)
print (tabulate(rows,headers=dict(zip(fields,fields)), tablefmt='orgtbl'))

```

The output will be following:

```

|  views | csv |  clicks |  blogid |  current_date |
|-----+-----+-----+-----+-----|
|    10 | 1.csv |      3 |      1 | December 01 2017, 10:57.00 AM |
|    14 | 2.csv |     15 |      1 | December 01 2017, 10:57.00 AM |
|     9 | 3.csv |      8 |      6 | December 01 2017, 10:57.00 AM |

```

The end result is a dataset that contains a date dimension that would allow presentation via a BI/Analytics tool to show blog traffic by day.

There are tools and software that does all work you need like:

1. We can use **Alteryx** - leader in the self-service data analytics movement with a platform that can prep, blend, and analyze all of your data, then deploy and share analytics at scale for deeper insights in hours. ([www.alteryx.com](http://www.alteryx.com))
2. We can use **Cloudera Impala** with **Apache Hadoop** ([www.cloudera.com](http://www.cloudera.com))

2.

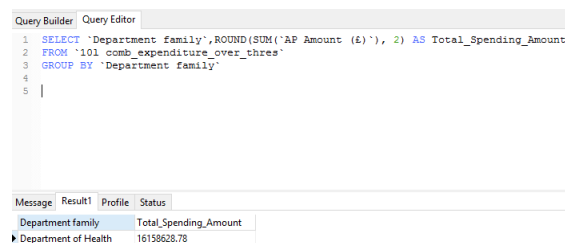
I have inserted the dataset is in CSV format into Navicat database and I did the queries using MySQL.

2a. Total spending (amount) by Department family.

```

SELECT `Department family`, ROUND(SUM(`AP Amount (£)`), 2) AS Total_Spending_Amount
FROM `10l comb_expenditure_over_thres`
GROUP BY `Department family`

```



The screenshot shows a query editor with the following SQL query:

```

1 SELECT `Department family`,ROUND(SUM(`AP Amount (£)`), 2) AS Total_Spending_Amount
2 FROM `10l comb_expenditure_over_thres`
3 GROUP BY `Department family`
4
5

```

Below the query editor, there is a table with the following data:

| Department family    | Total_Spending_Amount |
|----------------------|-----------------------|
| Department of Health | 16158628.78           |

2b. Total spending (amount) by Department family + Expense type.

```

SELECT `Department family`, `Expense Type`, ROUND(SUM(`AP Amount (£)`), 2) AS
Total_Spending_Amount
FROM `10l comb_expenditure_over_thres`
GROUP BY `Department family`, `Expense Type`

```

I used Tableau Public in order to get the table you have requested. Please go to the link [public.tableau.com/profile/sofija.milunov](https://public.tableau.com/profile/sofija.milunov)