Task

You get two data samples which were generated from our website. However, in their current state, they are not suitable to be used right away by our Data Scientists and Data Analysts who want to focus on the model building without having to modify the data much further. Your task is now the following:

Overall Goal:

you should write a code which injects these data daily into a datalake in a proper structure.

This assignment sheet consists of three tasks. Please do your coding in any framework, engine or library that you think is sufficient. Then create a directoy in the root directory of your project and name it as DataLake. We will assume this as the data lake which you inject the data in it.

You don’t need to send us any data, code and documentation is enough.

There are two series of json files attached to the assignment with the following structures:

Visitors data:

visitor_id	visit_start	countPerday	country	first_hit_pagename	hits_avg	logged_in	region	registered	visits
1.7384862152964	2021-01-17 00:00:01	1	deu	Deutsch	3.7026402963000002	0	nw	false	1
4.9728034405323	2021-01-17 00:00:01	2	deu	Deutsch	8.6394940247	1	bw	false	1
3.74879066793689	2021-01-17 00:00:03	3	deu	null	28.3869089383	0	hh	false	1
2.963321959817	2021-01-17 00:00:13	4	deu	Deutsch	14.810561185200001	0	bw	false	1
3.696354652696	2021-01-17 00:00:15	5	nld	null	2.4684268642	0	li	false	1
8.610871244695	2021-01-17 00:00:16	6	deu	Select	16.0447746173	0	nw	false	1
2.635567038733	2021-01-17 00:00:18	7	nld	Nederlands	7.4052805926000005	0	zh	false	1
3.18657594723	2021-01-17 00:00:19	8	swe	null	13.5763477531	0	c	false	1
2.4392429434	2021-01-17 00:00:20	9	deu	Deutsch	7.4052805926000005	0	nw	false	1
2.7316363074.	2021-01-17 00:00:22	10	deu	Deutsch	7.4052805926000005	0	nw	false	1

Searches:

visitor_id	date_time	flight_date_outbound	origin_out	destination_out	flight_date_inbound	origin_ret	destination_ret	segments
1.07025337	2021-01-13T08:04	2021-02-05	DUS	LHR	null	null	null	2
2.15651273356	2021-03-14T00:57:	2021-10-21	CGN	LIS	2021-10-28	LIS	CGN	1
3.1407870872	2021-02-10T05:50:	2021-02-10	HAM	DUS	null	null	null	1
4.5427698171	2021-04-18T01:43:	2021-12-01	CGN	PMI	null	null	null	1
2.06885241525	2021-02-28T00:08:	2021-08-01	STR	CTA	2021-08-21	CTA	STR	3
1.8530073154	2021-01-13T00:50:	null	CGN	SKG	null	SKG	CGN	null
3.2890909778	2021-04-14T00:53:	2021-04-18	DUS	ARN	null	null	null	1
5.02704150	2021-01-31T09:00:	2021-09-19	HAJ	PMI	2021-09-24	PMI	HAJ	1
1.0263343	2021-02-04T08:33:	2021-05-06	HAM	MUC	2021-05-07	MUC	HAM	2
3.9804266	2021-03-04T07:33:	2021-03-05	CGN	PMI	null	null	null	2
1.233530801927213...	2021-03-11T06:13:	2021-07-23	HAM	VIE	2021-07-26	VIE	HAM	1

You should write a script which can perform the following tasks:

Task 1: Data Ingestion

Your code should read all the files that are given in this assingment and then based on the timestamp of each file's name, write them into the datalake with your desired format. Be aware that the pipeline could be accidentally triggered multiple times in a day. Please write a short description why you chose the exported format.

Task 2: Preprocessing

For this task, the data you now have in the datalake should be processed and cleansed.

We are interested to know more about which changes you will perform, why and which format and structure you have chosen for this task.

Remember, these are some sample files, but every hour, our systems generate big amounts of data and every optimization could help us to optimize and increase the performance. You can take into consideration some dimension tables(enums) for some columns as extra points.

Task3: Reports

For this task we would like to have a simple report which shows number of searches per region, country and date (not date time). For this report you need, first to join the datasets. To make the join faster and more accurate, we assume that the region of a visitor in visitors dataset will not change on daily basis. Then, you should first get the latest entry of each visitor per day and later perform the join with searches dataset. *report smaple:*

date	country	region	count
2021-03-05	deu	bw	725
2021-04-12	esp	co	4
2021-01-27	deu	hh	398
2021-05-02	gbr	lnd	11
2021-05-08	usa	or	10

Task4: Pipeline architecture

Assuming you get multiple large datasets every 10 minutes, how do you automatize this task? Which tools, which strategy would you use? Please give us only a simple architecture, including the infrastructure of your desired system.

Bonus Task

- An implementation of the described architecture from task 4 would be considered as a bonus task
- Dockerizing the project, which everything can be triggered inside the docker, would be considered as a bonus.