Faculty of Computers and Artificial Intelligence
Cairo University

Third Year
Data Warehouse Assignment
2024

In this assignment you should solve the below **four** questions using **Microsoft SSIS**:

1. **[2 Marks]** Consume any **REST API** and load the response to the database. You don't have to load all the response fields, **3 or 4 is okay for me**.

   For example, this is a response from an API that is used to search for universities:


```
[
    {
        "state-province": null,
        "domains": [
            "mga.edu"
        ],
        "name": "Middle Georgia State College",
        "country": "United States",
        "web_pages": [
            "http://www.mga.edu/"
        ],
        "alpha_two_code": "US"
    },
    {
        "state-province": null,
        "domains": [
            "meu.edu.jo"
        ],
        "name": "Middle East University",
        "country": "Jordan",
        "web_pages": [
            "http://www.meu.edu.jo/"
        ],
        "alpha_two_code": "JO"
    },
    {…
    },
    {…
```


   You can create a database table named ***'University'*** with **three** columns: (name, country & alpha_two_code) and load those fields only.

2. **[2 Marks]** Implement **SCD type 4** for the below source table ***'Employee_Q2'***:

ID	Name	City	Email	Update_Date
1001	Ahmed	Cairo	ahmed@mail.com	20-04-2024
1002	Nehal	Giza	nehal@mail.com	20-04-2024
1003	Asem	Cairo	asem@mail.com	20-04-2024

Notes:

1. Create the necessary **two target tables** with the necessary **columns**, so that we have a target table that stores the latest version and a separate history table.
2. The SCD fields are **City and Email**.
3. Read the source data using **Incremental Load**.

3. **[3 Marks]** Load source data to a target table using **versioning** like below:

3.1 Source table *'Employee_Q3'*

ID	Name	City	Email	Schedule_Date
1001	Ahmed	Cairo	ahmed@mail.com	20-04-2024
1002	Nehal	Giza	nehal@mail.com	20-04-2024
1003	Asem	Cairo	asem@mail.com	20-04-2024

3.2 Target table after the first run on the same day

Emp_Key	ID	Name	City	Email	Insert_Date	Active_Flag	Version_No
1	1001	Ahmed	Cairo	ahmed@mail.com	20-04-2024	1	1
2	1002	Nehal	Giza	nehal@mail.com	20-04-2024	1	1
3	1003	Asem	Cairo	asem@mail.com	20-04-2024	1	1

3.3 Target table after the second run on the same day

Emp_Key	ID	Name	City	Email	Insert_Date	Active_Flag	Version_No
1	1001	Ahmed	Cairo	ahmed@mail.com	20-04-2024	0	1
2	1002	Nehal	Giza	nehal@mail.com	20-04-2024	0	1
3	1003	Asem	Cairo	asem@mail.com	20-04-2024	0	1
4	1001	Ahmed	Cairo	ahmed@mail.com	20-04-2024	1	2
5	1002	Nehal	Giza	nehal@mail.com	20-04-2024	1	2
6	1003	Asem	Cairo	asem@mail.com	20-04-2024	1	2

3.4 Target table after the first run on the next day (just change *Schedule_Date* in the source data to simulate the next day)

Emp_Key	ID	Name	City	Email	Insert_Date	Active_Flag	Version_No
1	1001	Ahmed	Cairo	ahmed@mail.com	20-04-2024	0	1
2	1002	Nehal	Giza	nehal@mail.com	20-04-2024	0	1
3	1003	Asem	Cairo	asem@mail.com	20-04-2024	0	1
4	1001	Ahmed	Cairo	ahmed@mail.com	20-04-2024	0	2
5	1002	Nehal	Giza	nehal@mail.com	20-04-2024	0	2
6	1003	Asem	Cairo	asem@mail.com	20-04-2024	0	2
7	1001	Ahmed	Cairo	ahmed@mail.com	21-04-2024	1	1
8	1002	Nehal	Giza	nehal@mail.com	21-04-2024	1	1
9	1003	Asem	Cairo	asem@mail.com	21-04-2024	1	1

Notes:

1. I have shown two runs only on the same day but during discussion, I can run **n** times and see the behavior of your solution.
2. Add **a new version** as long as you are running on the **same** day and close **all** old records.
3. Start from **version 1** again on the **next** day and so on.
4. **Don't** check for any change in the source data, load it as it is.

4. **[3 Marks]** We have a task to read data from an **attendance device** in a company and load this data to a target table in a better format with a **state** at the end of each record as follows:

State	Description
ebn el-shrka	Arrived on time (9 am) and worked more than 8 hours
mo7tram	Arrived on time and worked 8 hours
raye2	Arrived late but worked 8 hours
byst3bat	Arrived on time but worked less than 8 hours
msh mo7tram	Arrived late and worked less than 8 hours
no check out	No check-out record for the employee on that day

4.1 Sample from the source table *'Attendance_Device'*

ID	Employee_Id	Finger_Print_TS	In_Out
1	101	2024-03-12 09:00:00	in
2	101	2024-03-12 10:00:00	in
3	102	2024-03-12 09:00:00	in
4	103	2024-03-12 11:00:00	in
5	104	2024-03-12 09:15:00	in
6	105	2024-03-12 10:00:00	in
7	105	2024-03-12 11:00:00	in
8	105	2024-03-12 11:30:00	in
9	106	2024-03-12 09:10:00	In

10	107	2024-03-12 09:00:00	in
11	108	2024-03-12 09:00:00	in
12	101	2024-03-12 09:00:00	out
13	101	2024-03-12 17:00:00	out
14	101	2024-03-12 19:00:00	out
15	102	2024-03-12 17:00:00	out
16	103	2024-03-12 17:00:00	out
17	105	2024-03-12 10:00:00	out
18	105	2024-03-12 11:00:00	out
19	105	2024-03-12 18:00:00	out
20	106	2024-03-12 19:10:00	out
21	107	2024-03-12 14:00:00	out
22	108	2024-03-12 17:00:00	out
.	.	.	.
.	.	Different Day	.
.	.	.	.

4.2 Target table *'Employee_Attendance_Details'*

Att_Key	Emp_ID	Date	Time_In	Time_Out	Worked_Hours	State
1	101	2024-03-12	9:00	17:00	8	mo7tram
2	102	2024-03-12	9:00	17:00	8	mo7tram
3	103	2024-03-12	11:00	17:00	6	msh mo7tram
4	104	2024-03-12	9:15	null	null	no check out
5	105	2024-03-12	10:00	18:00	8	raye2
6	106	2024-03-12	9:10	19:10	10	ebn el-shrka
7	107	2024-03-12	9:00	14:00	5	byst3bat
8	108	2024-03-12	9:00	17:00	8	mo7tram
.	.	.	.	.	.	.
.	.	.	.	.	.	.

Assumptions:

1. The time of the **out** record from the device is always **>=** the time of the **in** record for the same employee and the same day.
2. No employee is allowed to check in **before 9 AM** (so don't worry about this case).
3. The device is not working properly and sometimes creates **in & out** records at the **same time** as in records **1 & 12**.
4. The employee may forget and check in or check out **multiple** times, in that case, load **min (check-in time)** in the ***Time_In*** column and **min(check-out time) that is >= max(check-in time)** in the ***Time_Out*** column.
5. You may find a **different** scenario other than the ones specified at the beginning of the question, in that case, set the state as ***undefined***.

General Notes:

1. No late submission will be accepted **for any reason**.
2. The team should consist of **TWO** students only.
3. **All** team members must attend the discussion. **ZERO** points will be given to the absentees.
4. During the discussion, I will test using **different** datasets, so be prepared for that.
5. If you have any questions, feel free to ask your TA during the **next lab** insha'Allah.
6. If you completed the entire assignment **without cheating**, wallahy be sure that you have **learned** a lot and you have done something **great**. Feel free to share your accomplishment on **LinkedIn** and I can assist you with that 😉

Where to submit?

1. Prepare a **zip** file that contains four folders, one for each problem, and name it as follows:
   DWH_Assignment_TA_Name_ID1_ID2.zip

   Ex: DWH_Assignment_Ahmed_Galal_20210011_20210022.zip

2. Upload the zip file to this Google Form: https://forms.gle/QRzaEFQeoLBWZYjs9

   The form will be closed on **Saturday, May 4th at 11:59 PM**.

Wishing you all the best ❤
Ahmed Galal