**Group members:**

**(e.g: Username, StudentID, Name, Parts responsible)**

# **Task 1**

Student ID used for publication type selection: davby024

Major type: Article (last digit is even)

Minor type: Book (4%6)+1 = 5

Code used to retrieve your data should be in extract.xql.

# **Task 2**

Start schema definitions (fact table must be the first one).

Fact\_Publication(publication\_id, details\_id, author\_id, publisher\_id, date\_id, links\_id, publication\_type , [publication\_count])

publication\_id (PK)

details\_id (SK)

author\_id (SK)

publisher\_id (SK)

time\_id (SK)

links\_id (SK)

Dim\_Details(details\_id, title, vol, nmbr, pgs)

details\_id (SK)

Dim\_Author(author\_id, author\_name, editor\_name)

author\_id (SK)

Dim\_Publisher(publisher\_id, publisher\_name, journal\_name)

publisher\_id (SK)

Dim\_Time(time\_id, year, month)

time\_id (SK)

Dim\_Links(links\_id, url\_str, ee, cdrom, cite)

links\_id (PK)

**Descriptions of usefulness dimension tables.**

Dim\_details: This table holds the key information relating to the content of the publication, and can be helpful for analysing how size, number of pages or volume number changes the way publications perform.

Dim\_Author: This table holds information of people who have played a part in contributing to the publication. It could be used in analysis to see how different editors, authors or combinations of the two perform differently in articles they write.

Dim\_Publisher: This table has the key details of the publisher and they journals they publish. It is helpful for analysing the way publishers and their journals affect the way publications perform. I.e. Some publishers, or specific journals they publish might consistently perform better.

Dim\_Time: This table is helpful for allowing users to analyse the dates and times when publications are published and how this changes their overall performance

Dim\_links: This table contains the key links to the publication, and it’s cites. This dimension serves to help analyse how the publication table performs in terms of its cites and how different urls could affect the way they perform.

Table creation statements (SQL)

CREATE TABLE Dim\_Details (

details\_id INT GENERATED AS IDENTITY PRIMARY KEY,

title VARCHAR(255),

vol VARCHAR(50),

nmbr VARCHAR(50),

pgs VARCHAR(50)

);

CREATE TABLE Dim\_Author (

author\_id INT GENERATED AS IDENTITY PRIMARY KEY,

author\_name VARCHAR(255),

editor\_name VARCHAR(255)

);

CREATE TABLE Dim\_Publisher (

publisher\_id INT GENERATED AS IDENTITY PRIMARY KEY,

publisher\_name VARCHAR(255),

journal\_name VARCHAR(255)

);

CREATE TABLE Dim\_Time (

time\_id INT GENERATED AS IDENTITY PRIMARY KEY,

yr INT,

mnth INT

);

CREATE TABLE Dim\_Links (

links\_id INT GENERATED AS IDENTITY PRIMARY KEY,

url\_str VARCHAR(255),

ee VARCHAR(255),

cdrom VARCHAR(255),

cite VARCHAR(255)

);

CREATE TABLE Fact\_Publication (

publication\_id INT GENERATED AS IDENTITY PRIMARY KEY,

details\_id INT,

author\_id INT,

publisher\_id INT,

time\_id INT,

links\_id INT,

publication\_type VARCHAR(50),

publication\_count INT,

FOREIGN KEY (details\_id) REFERENCES Dim\_Details(details\_id),

FOREIGN KEY (author\_id) REFERENCES Dim\_Author(author\_id),

FOREIGN KEY (publisher\_id) REFERENCES Dim\_Publisher(publisher\_id),

FOREIGN KEY (time\_id) REFERENCES Dim\_Time(time\_id),

FOREIGN KEY (links\_id) REFERENCES Dim\_Links(links\_id)

);

Table schemas (screenshots)

Fact\_Publication:



Dim\_Details:



Dim\_Author:



Dim\_Publisher:



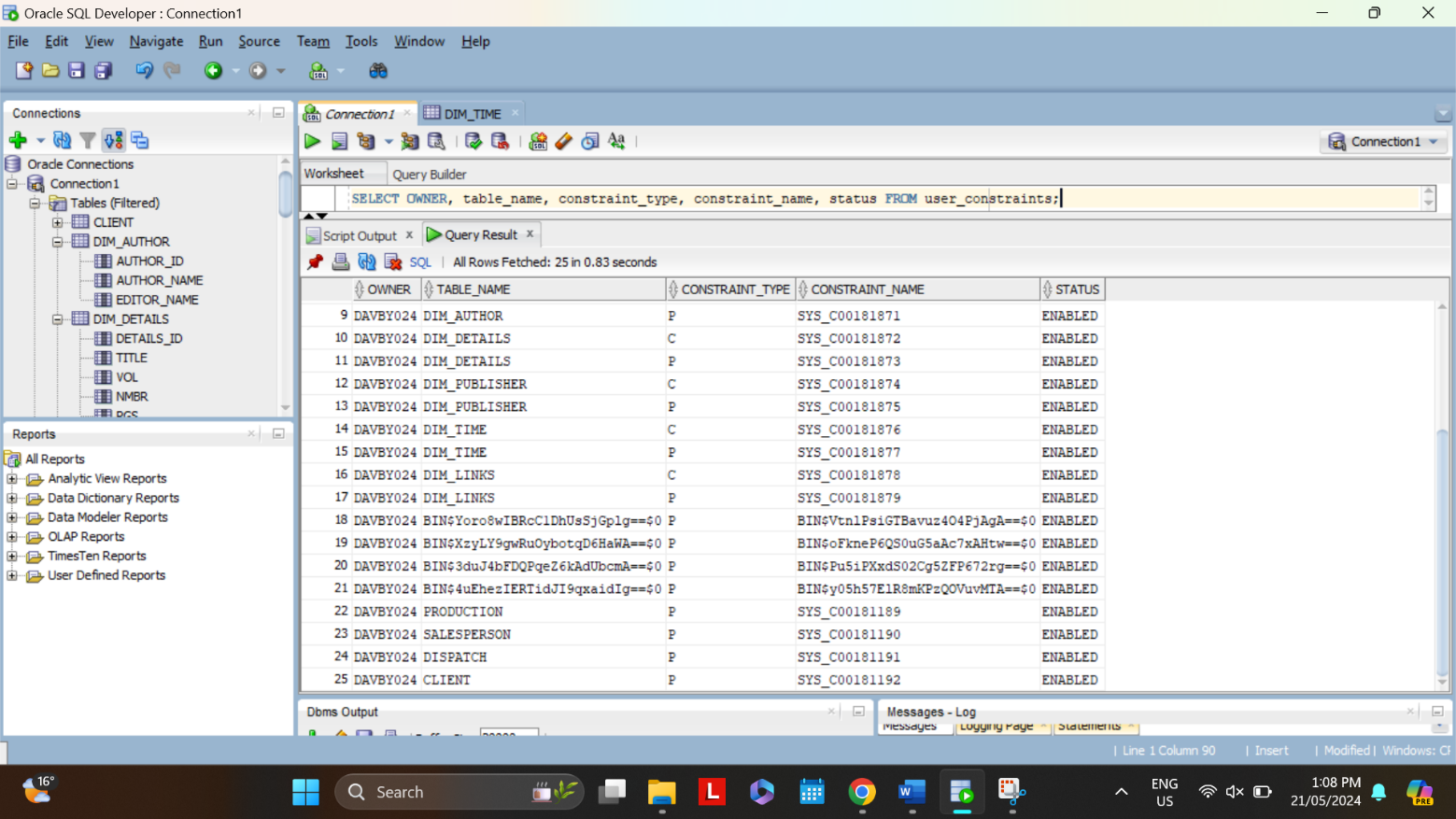
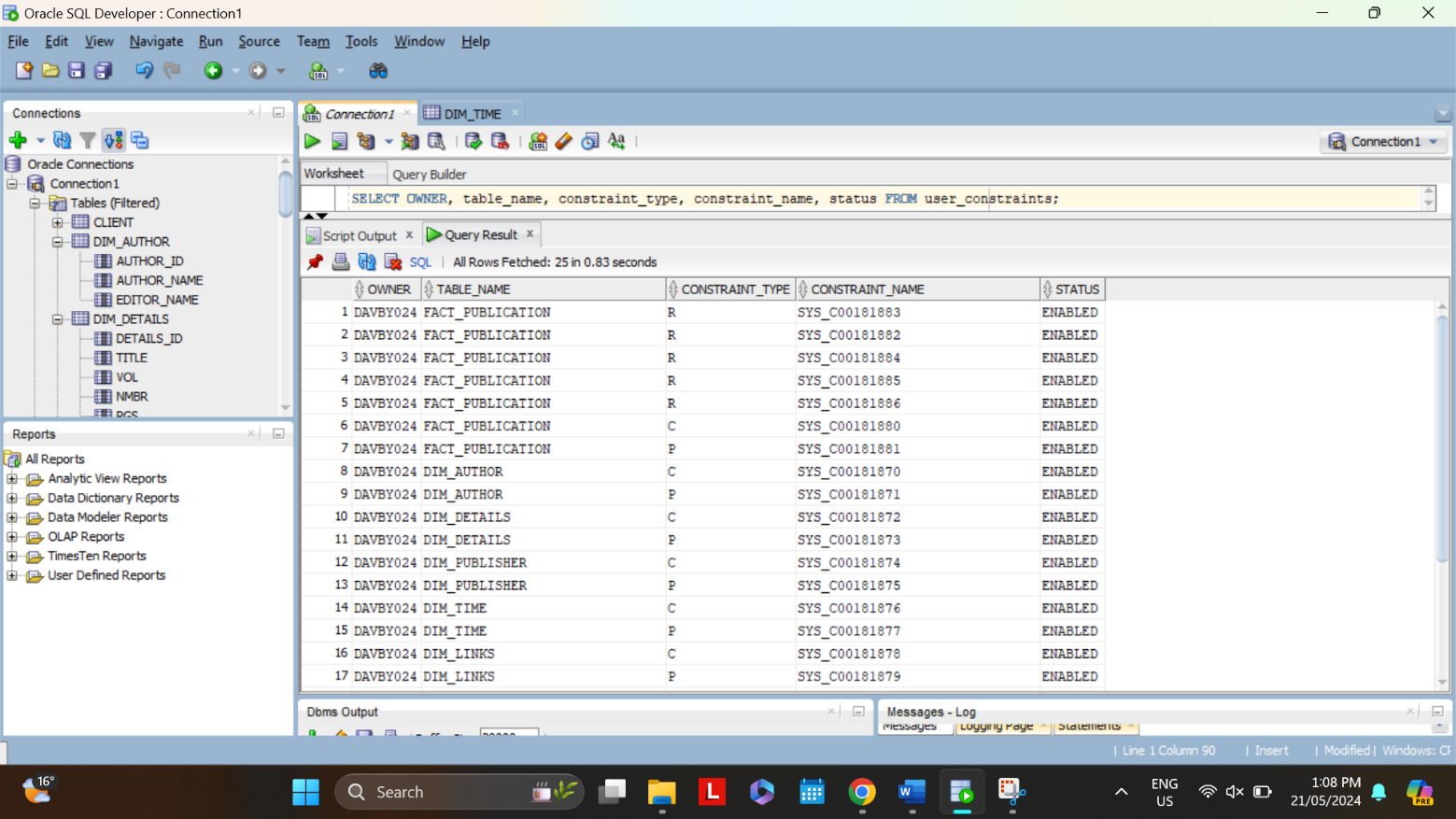
Dim\_Time:



Dim\_Links:



Table constraint summary (screenshot)



# **Task 3**

Summarize transformations and data uploading to oracle.

(note that here, you summarize the steps, detailed code should appear in transform?.\* files.)

transform1.xql:

Aim:

Input:

Processes:

Output:

Data tables (screenshots)

Justification of value and relationship preservation

Examples:

Values:

One-to-many relationships:

Dataset for an analysis.

Analysis:

SQL:

Screenshot:

# **Task 4**

1. Example of data entries in the data source.
2. The data ready for integration (a small table two records)
3. Assumptions and processes to extract and transform data from a) to b).

Assumptions

Processes

1. The purpose of new data. Does it extend an existing dimension or add a new dimension? Present the new table schema.
2. The way in which the new data records will be linked/matched to existing records.
3. Automation challenges.