МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»

Факультет інформатики та обчислювальної техніки

Кафедра інформатики та програмної інженерії

Практикум №1

з курсу «Аналіз даних в інформаційнних системах»

на тему: «Створення сховища даних»

|  |  |
| --- | --- |
| Викладач:  Олійник Ю.О. | Виконав:  студент 2 курсу Головня О.Р.  групи ІП-11 ФІОТ |
|  |  |

Київ-2023

**Мета роботи:** ознайомитись з підходами до створення сховищ даних.

**ЗАВДАННЯ**

Навчитися створювати процедури завантаження даних до сховища.

1. Самостійно обрати не менше 3-х джерел відкритих даних.
2. Спроектувати модель Stage зони для ETL процесів.
3. Спроектувати модель основного сховища за типом «зірка» або «сніжинка».
4. Створити ETL засоби:

завантажити дані до Stage зони

створити набір процедур/функцій для перетворення та завантаження даних до основного сховища (або створити засобами програмних ETL засобів). Передбачити можливість завантаження змінених та додаткових даних.

1. Завантажити дані до основного сховища даних.

Предметною областю лабораторної роботи є інтернет продажі окремих продуктів.

1. **Спроектувати модель Stage зони для ETL процесів.**

**Створення таблиці товару**

**CREATE TABLE `Products` (**

**`ProductKey` INT,**

**`ProductItemCode` text,**

**`Platform\_Name` text,**

**`SubCategory` text,**

**`ProductCategory` text,**

**`ProductColor` text,**

**`ProductSize` text,**

**`ProductLine` text,**

**`ProductModelName` text,**

**`ProductDescription` text,**

**`ProductStatus` text,**

**PRIMARY KEY (`ProductKey`)**

**);**

**Створення таблиці покупців**

**CREATE TABLE `Customers` (**

**`CustomerKey` INT,**

**`FirstName` text,**

**`LastName` text,**

**`FullName` text,**

**`Gender` text,**

**`DateFirstPurchase` DATETIME,**

**`CustomerCity` text,**

**PRIMARY KEY (`CustomerKey`)**

**);**

**Створення таблиці дат**

**CREATE TABLE `Calendar` (**

**`DateKey` text,**

**`Date` DATETIME,**

**`Day` text,**

**`Month` text,**

**`MonthShort` text,**

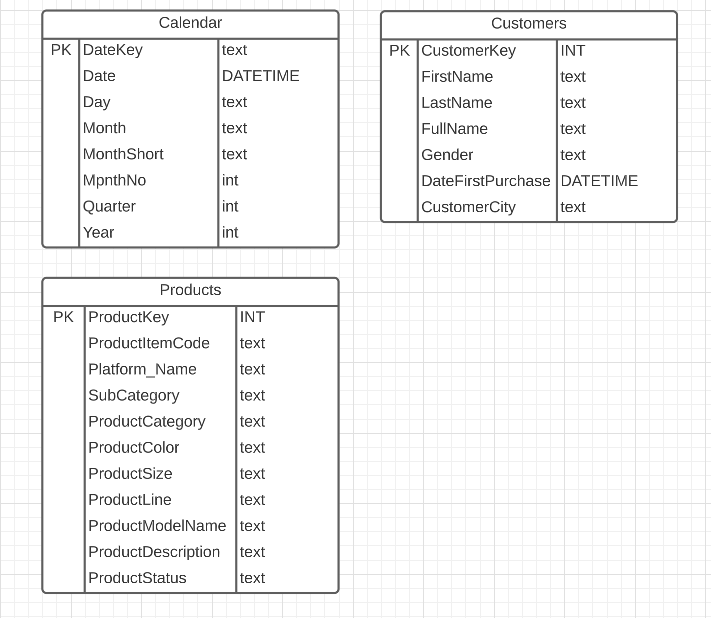
**`MpnthNo` int,**

**`Quarter` int,**

**`Year` int,**

**PRIMARY KEY (`DateKey`)**

**);**

****

**Скрипти створення основного сховища за типом «зірка»**

**Створення таблиці виміру дат**

**CREATE TABLE `DIM\_Calendar` (**

**`DateKey` int,**

**`Date` text,**

**`Day` text,**

**`Month` text,**

**`MonthShort` text,**

**`MpnthNo` text,**

**`Quarter` text,**

**`Year` text,**

**PRIMARY KEY (`DateKey`)**

**);**

**Створення фактової таблиці бюджету**

**CREATE TABLE `Fact\_Budget` (**

**`FactBudgetKey` int,**

**`DateKey` int,**

**`Budget` text,**

**PRIMARY KEY (`FactBudgetKey`),**

**FOREIGN KEY (`DateKey`) al `DIM\_Calendar`(`DateKey`)**

**);**

**Створення таблиці виміру товарів**

**CREATE TABLE `DIM\_Products` (**

**`ProductKey` int,**

**`ProductItemCode` text,**

**`Platform\_Name` text,**

**`SubCategory` text,**

**`ProductCategory` text,**

**`ProductColor` text,**

**`ProductSize` text,**

**`ProductLine` text,**

**`ProductModelName` text,**

**`ProductDescription` text,**

**`ProductStatus` text,**

**PRIMARY KEY (`ProductKey`)**

**);**

**Створення таблиці виміру покупців**

**CREATE TABLE `DIM\_Customers` (**

**`CustomerKey` int,**

**`FirstName` text,**

**`LastName` text,**

**`FullName` text,**

**`GenderKey` int,**

**`DateFirstPurchase` text,**

**`CustomerCity` text,**

**PRIMARY KEY (`CustomerKey`)**

**);**

**Створення фактової таблиці про інтернет-продажі**

**CREATE TABLE `Fact\_InternetSales` (**

**`IFactSalesKey` int,**

**`ProductKey` int,**

**`OrderDateKey` int,**

**`DueDateKey` int,**

**`ShipDateKey` int,**

**`CustomerKey` int,**

**`SalesOrderNumber` int,**

**`SalesAmount` int,**

**PRIMARY KEY (`IFactSalesKey`),**

**FOREIGN KEY (`OrderDateKey`) REFERENCES `DIM\_Calendar`(`DateKey`),**

**FOREIGN KEY (`ProductKey`) REFERENCES `DIM\_Products`(`ProductKey`),**

**FOREIGN KEY (`CustomerKey`) REFERENCES `DIM\_Customers`(`CustomerKey`),**

**FOREIGN KEY (`DueDateKey`) REFERENCES `DIM\_Calendar`(`DateKey`),**

**FOREIGN KEY (`ShipDateKey`) REFERENCES `DIM\_Calendar`(`DateKey`)**

**);**

**Створення таблиці виміру гендерів**

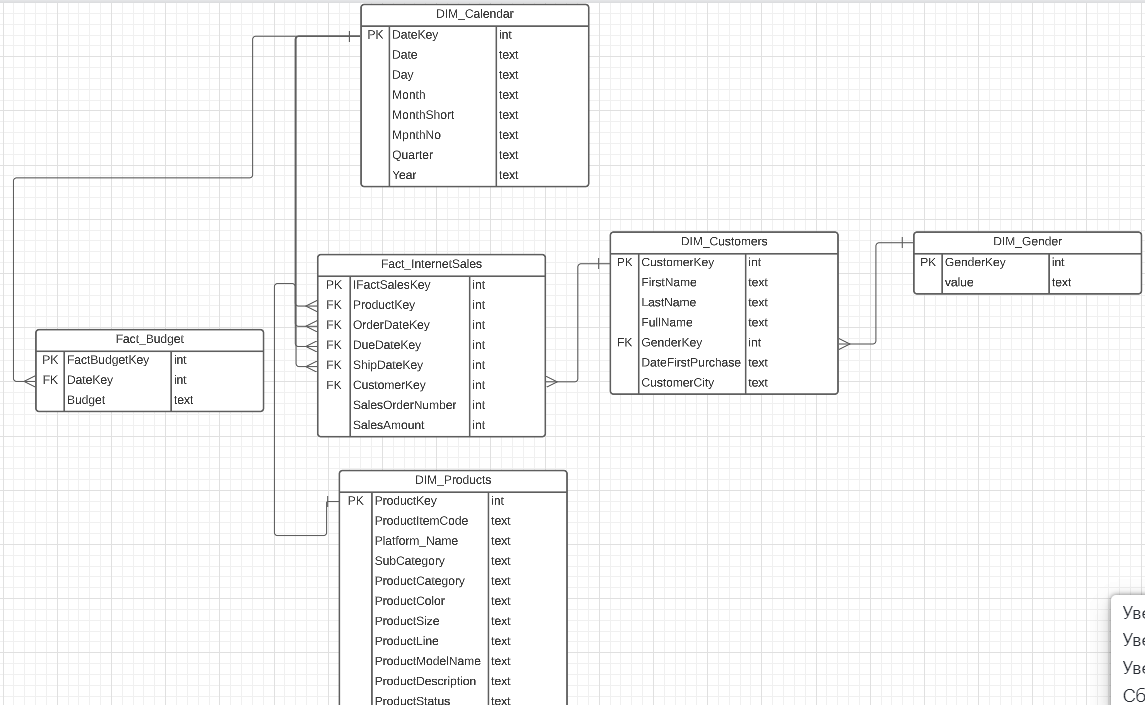
**CREATE TABLE `DIM\_Gender` (**

**`GenderKey` int,**

**`value` text,**

**PRIMARY KEY (`GenderKey`)**

**);**

****

Таблиця 1 – Таблиця виміру дат dim\_calendar

|  |  |
| --- | --- |
| Назва поля | Опис поля |
| DateKey | Унікальний ідентифікатор запису дати |
| Date | Дата |
| Day | День |
| Month | Місяць |
| MonthShort | Місяць але скорочено |
| MonthNo | Номер місяця |
| Quarter | Квартал |
| Year | Рік |

Таблиця 2 – Таблиця виміру покупців dim\_customers

|  |  |
| --- | --- |
| Назва поля | Опис поля |
| CustomerKey | Унікальний ідентифікатор покупця |
| FirstName | Ім’я |
| LastName | Фамілія |
| FullName | Повне ім’я |
| GenderKey | Гендер |
| DateFirstPurchase | Дата купівлі |
| CustomerCity | Місто покупця |

Таблиця 3 – Таблиця виміру гендера dim\_gender

|  |  |
| --- | --- |
| Назва поля | Опис поля |
| GenderKey | Унікальний ідентифікатор гендера |
| value | Значення(Жін. Чол) |

Таблиця 4 – Таблиця виміру товару dim\_product

|  |  |
| --- | --- |
| Назва поля | Опис поля |
| ProductKey | Унікальний ідентифікатор запису дати |
| ProductItemCode | Код товару |
| Platform\_Name | Платформа |
| SubCategory | Саб-категорія |
| ProductCategory | Категорія |
| ProductColor | Колір |
| ProductSize | Розмір |
| ProductLine | Лінія |
| ProductModelName | Модель |
| { ProductDescription | Опис |
| ProductStatus | Статус |

Таблиця 5 – Таблиця фактів про бюджет fact\_budget

|  |  |
| --- | --- |
| Назва поля | Опис поля |
| FactBudgetKey | Унікальний ідентифікатор |
| DateKey | Дата |
| Budget | Сам бюджет |

Таблиця 6 – Таблиця фактів про інтернет купівлю fact\_internetsales

|  |  |
| --- | --- |
| Назва поля | Опис поля |
| FactSalesKey | Унікальний ідентифікатор факту |
| ProductKey | Ключ товару |
| OrderDateKey | Дата замовлення |
| DueDateKey | Дата оплати |
| ShipDateKey | Дата відправки |
| CustomerKey | Ключ покупця |
| SalesOrderNumber | Код замовлення |
| SalesAmount | Кількість продажу |

1. **Скрипт для завантаження даних для Stage зони(тут тільки для 1 табл)**

**Для stage zone**

**LOAD** **DATA** **LOW\_PRIORITY** **LOCAL** **INFILE** 'C:\\Users\\C140~1\\DOWNLO~1\\SALESA~1\\FACT\_t.csv' **REPLACE** **INTO** **TABLE** `ad2`.`fact\_internetsales` **CHARACTER** **SET** **cp1251** **FIELDS** **TERMINATED** **BY** ',' **ESCAPED** **BY** '"' **LINES** **TERMINATED** **BY** '\r\n' **IGNORE** 1 **LINES** (`IFactSalesKey`, `ProductKey`, `OrderDateKey`, `DueDateKey`, `ShipDateKey`, `CustomerKey`, `SalesOrderNumber`, `SalesAmount`);

**LOAD** **DATA** **INFILE** 'C://sales.csv'

**INTO** **TABLE** Fact\_InternetSales

**FIELDS** **TERMINATED** **BY** ','

**IGNORE** 1 **ROWS**

(ProductKey, @var1, DueDateKey, ShipDateKey, CustomerKey, SalesOrderNumber, SalesAmount)

**SET** OrderDateKey = (**SELECT** DateKey **FROM** DIM\_Calendar **WHERE** **Date** = @var1)

**LIMIT** 58166;

**END** //

1. **Скрипт перенесення потрібної нам інформації з Stage зони у фактові таблиці та таблиці вимірів**

*Створення процедури для заповнення таблиці calendar:*

DELIMITER //

**CREATE** **PROCEDURE** fill\_table\_calendar()

**BEGIN**

**INSERT** **INTO** dim\_calendar(DateKey, **Date**, **Day**, **Month**, MonthShort, MonthNo, **Quarter**, **Year**)

**SELECT** **DISTINCT** DateKey,"Date","Day","MONTH",MonthShort,MpnthNo,"QUARTER","Year"

**FROM** stage.calendar;

**END** //

DELIMITER ;

*Створення процедури для заповнення таблиці customers:*

DELIMITER //

**CREATE** **PROCEDURE** fill\_table\_customers()

**BEGIN**

**INSERT** **INTO** dim\_customers(CustomerKey,FirstName,LastName,FullName,Gender,DateFirstPurchase,CustomerCity)

**SELECT** **DISTINCT** CustomerKey,FirstName,LastName,FullName,Gender,DateFirstPurchase,CustomerCity

**FROM** stage.customers;

**END** //

DELIMITER ;

*Створення процедури для заповнення таблиці products:*

DELIMITER //

**CREATE** **PROCEDURE** fill\_table\_products()

**BEGIN**

**INSERT** **INTO** dim\_products(ProductKey,ProductItemCode,Platform\_Name,SubCategory,ProductCategory,ProductColor,ProductSize,ProductLine,ProductModelName,ProductDescription,ProductStatus)

**SELECT** **DISTINCT** ProductKey,ProductItemCode,Platform\_Name,SubCategory,ProductCategory,ProductColor,ProductSize,ProductLine,ProductModelName,ProductDescription,ProductStatus

**FROM** stage.products;

**END** //

DELIMITER ;

*Створення процедури для оновлення табл gender:*

DELIMITER //

**CREATE** **PROCEDURE** update\_and\_insert\_gender()

**BEGIN**

*-- змінити значення 'Male' на '1'*

**UPDATE** dim\_customers **SET** Gender = '1' **WHERE** Gender = 'Male';

*-- змінити значення 'Female' на '2'*

**UPDATE** dim\_customers **SET** Gender = '2' **WHERE** Gender = 'Female';

*-- вставити значення 'Male' та 'Female' у таблицю dim\_gender*

**INSERT** **INTO** dim\_gender (GenderKey, **value**) **VALUES** (1, 'Male'), (2, 'Female');

*-- додати зовнішній ключ до dim\_gender*

**ALTER** **TABLE** dim\_customers **ADD** **FOREIGN** **KEY** (Gender) **REFERENCES** dim\_gender(GenderKey);

**END** //

DELIMITER ;

*Створення процедури для заповнення факт бюджету(самостійно):*

DELIMITER //

**CREATE** **PROCEDURE** fill\_fact\_budget()

**BEGIN**

**INSERT** **INTO** fact\_budget(FactBudgetKey, DateKey, Budget)

**VALUES**(1,20200101,800000),

(3,20200301,1000000),

(4,20200401,1000000),

(5,20200501,1100000),

(6,20200601,1100000),

(7,20200701,1500000),

(8,20200801,1500000),

(9,20200901,1100000),

(10,20201001,1500000),

(11,20201101,1500000),

(12,20201201,2000000),

(13,20210101,800000),

(14,20210201,800000),

(15,20210301,1000000),

(16,20210401,1000000),

(17,20210501,1100000),

(18,20210601,1100000);

**END** //

DELIMITER ;



Рисунок 1 – dim\_calendar

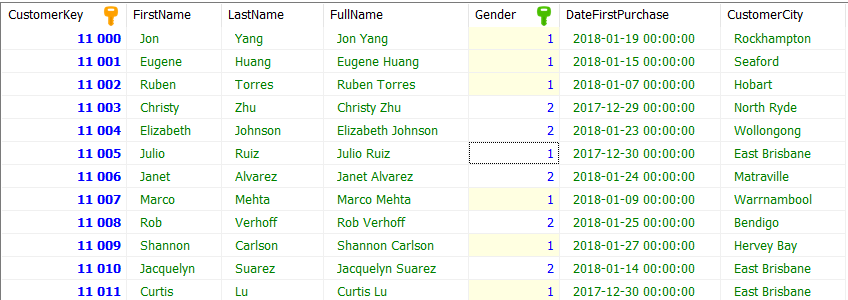


Рисунок 2 – dim\_customers

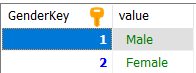


Рисунок 3 – dim\_gender

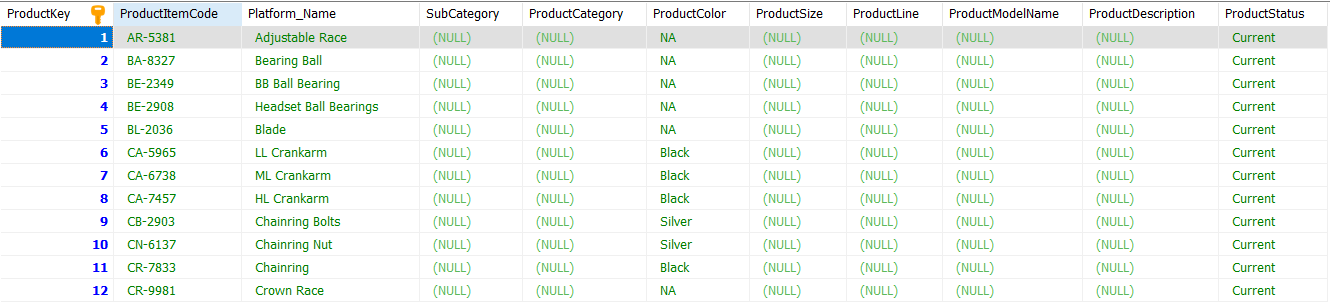


Рисунок 4 – dim\_product

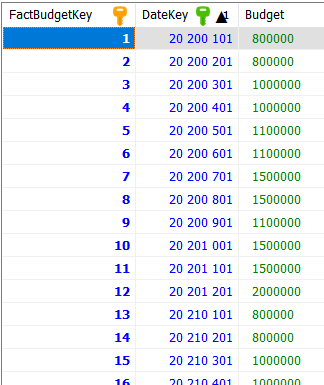


Рисунок 5 – fact\_budget

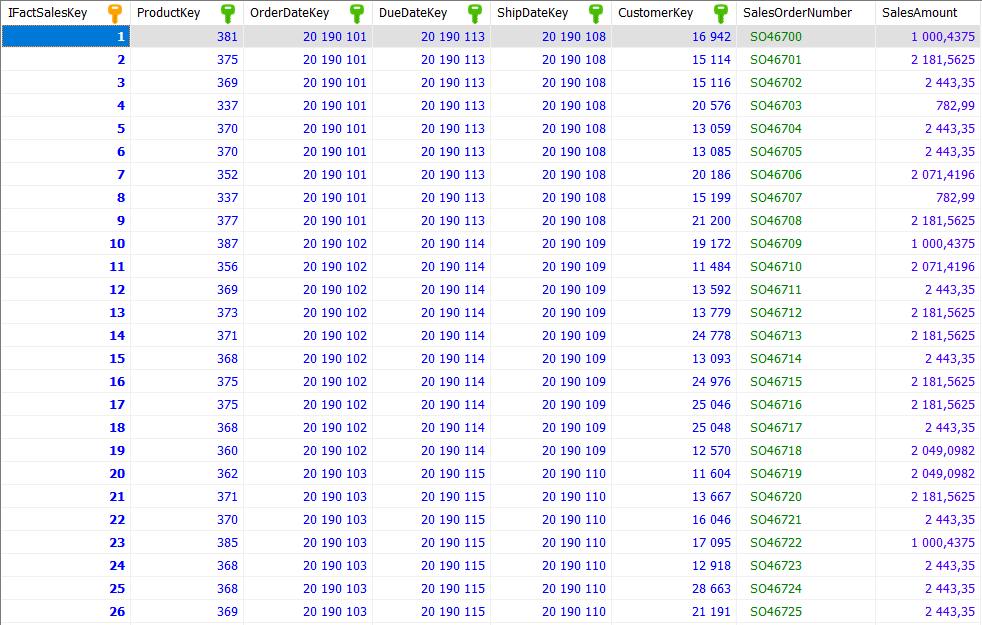


Рисунок 6 – fact\_internetsales(Близько 58 000 рядків)

**Реалізація SCD2:**

DELIMITER //

**CREATE** **FUNCTION** MergeCustomersSCD2(

CustomerKey **INT**,

FirstName **TEXT**,

LastName **TEXT**,

FullName **TEXT**,

GenderKey **INT**,

DateFirstPurchase **TEXT**,

CustomerCity **TEXT**

) **RETURNS** **VARCHAR**(50)

**BEGIN**

**DECLARE** dt **DATETIME**;

**SET** dt = **NOW**();

*-- Перевірка чи запис існує в таблиці*

**IF** **EXISTS** (

**SELECT** 1 **FROM** `DIM\_Customers\_SCD2`

**WHERE** `CustomerKey` = CustomerKey

**AND** `ValidTo` = '9999-12-31 23:59:59'

) **THEN**

*-- Якщо запис існує, то встановити дату закінчення його актуальності*

**UPDATE** `DIM\_Customers\_SCD2`

**SET** `ValidTo` = dt

**WHERE** `CustomerKey` = CustomerKey

**AND** `ValidTo` = '9999-12-31 23:59:59';

**END** **IF**;

*-- Вставити новий запис в таблицю*

**INSERT** **INTO** `DIM\_Customers\_SCD2` (

`CustomerKey`,

`FirstName`,

`LastName`,

`FullName`,

`GenderKey`,

`DateFirstPurchase`,

`CustomerCity`,

`ValidFrom`,

`ValidTo`

) **VALUES** (

CustomerKey,

FirstName,

LastName,

FullName,

GenderKey,

DateFirstPurchase,

CustomerCity,

dt,

'9999-12-31 23:59:59'

);

*-- Повернути повідомлення про успішне виконання функції*

**RETURN** 'MergeCustomersSCD2 executed successfully';

**END** //

DELIMITER ;```

Ця функція створює новий запис в таблиці DIM\_Customers\_SCD2, що містить інформацію про клієнта, використовуючи підход SCD2 (Slowly Changing Dimensions Type 2).

При виклику функції потрібно передати значення всіх параметрів: CustomerKey, FirstName, LastName, FullName, GenderKey, DateFirstPurchase, та CustomerCity. Функція перевіряє, чи існує запис з таким же CustomerKey і датою закінчення актуальності '9999-12-31 23:59:59' в таблиці DIM\_Customers\_SCD2. Якщо такий запис існує, то дата його закінчення актуальності оновлюється на поточний момент часу, а новий запис додається до таблиці DIM\_Customers\_SCD2 з датою початку актуальності на поточний момент часу та датою закінчення '9999-12-31 23:59:59'.

****

**SELECT** MergeCustomersSCD2(11000, 'Jon', 'Yang', 'Jon Yang', 1, '2018-01-19 00:00:00', 'Rockhampton');

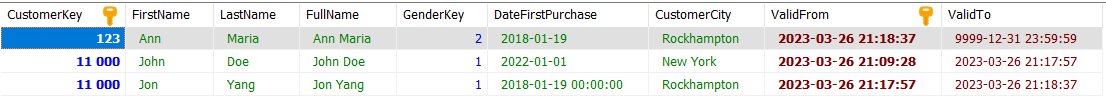
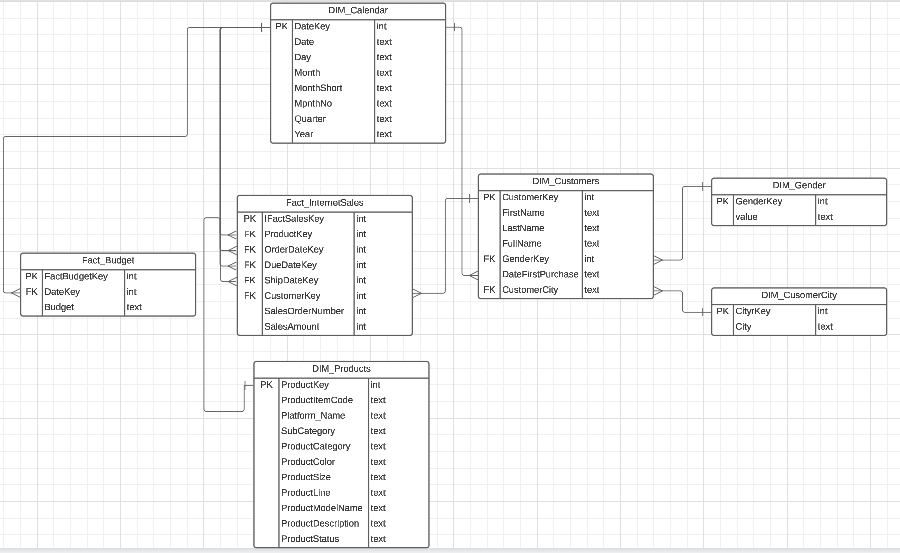


Рисунок 7 – dim\_customers\_scd2

**Висновок:** Виконуючу лабораторну роботу я обрав джерела даних, за ними спроєктував модель stage зони, спроєктував модель основного сховища за тим зірка, завантажив дані до stage зони, завантажив дані до основного сховища, деякі дані та таблиці змінив самостійно, реалізував **SCD2** для таблиці Клієнтів.

  
  
Рисунок 8 – Оновлена діаграма сховища (Додав вимір міст та з’єднав дату купівлі у Клієнта з Календарем)

**CREATE** **PROCEDURE** update\_customer\_city()

**BEGIN**

**CREATE** **TABLE** **IF** **NOT** **EXISTS** CustomerCityDimension **AS**

**SELECT** **DISTINCT** CustomerCity **FROM** DIM\_Customers;

**ALTER** **TABLE** CustomerCityDimension

**ADD** CustomerCityID **INT** **AUTO\_INCREMENT** **PRIMARY** **KEY**;

**UPDATE** dim\_customers **AS** c

**JOIN** CustomerCityDimension **AS** d **ON** c.CustomerCity = d.CustomerCity

**SET** c.CustomerCity = d.CustomerCityID;

**END**;

CALL update\_customer\_city();

ETL фактова таблиця:

DELIMITER //

**CREATE** **PROCEDURE** import\_sales\_data()

**BEGIN**

**INSERT** **INTO** Fact\_InternetSales

(ProductKey, OrderDateKey, DueDateKey, ShipDateKey, CustomerKey, SalesOrderNumber, SalesAmount)

**SELECT**

p.ProductKey,

c1.DateKey **as** OrderDateKey,

c2.DateKey **as** DueDateKey,

c3.DateKey **as** ShipDateKey,

cu.CustomerKey,

i.SalesOrderNumber,

i.SalesAmount

**FROM**

stage.Products p

**INNER** **JOIN** InternetSales i **ON** p.ProductID = i.ProductID

**INNER** **JOIN** stage.Customers cu **ON** i.CustomerID = cu.CustomerID

**INNER** **JOIN** stage.Calendar c1 **ON** i.OrderDate = c1.**Date**

**INNER** **JOIN** stage.Calendar c2 **ON** i.DueDate = c2.**Date**

**INNER** **JOIN** stage.Calendar c3 **ON** i.ShipDate = c3.**Date**;

**END** //

DELIMITER ;

Оновлене SCD:

ALTER TABLE `DIM\_Customers`

ADD COLUMN `current\_data` boolean;

DELIMITER $$

CREATE TRIGGER customers\_change\_trigger

BEFORE INSERT ON dim\_customers

FOR EACH ROW

BEGIN

UPDATE dim\_customers

SET current\_data = false

WHERE FullName = NEW.FullName

AND current\_data = true;

END$$

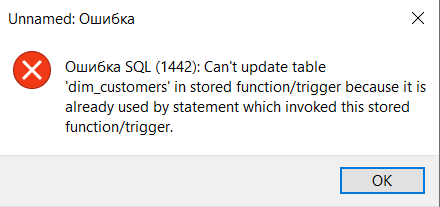
DELIMITER ;

**insert** **into** dim\_customers(FirstName, LastName, FullName,Gender,DateFirstPurchase,CustomerCity)

**values**

('Jon', 'Yang', 'Jon Yang',2,'2018-01-19 00:00:00',1);

**Юрій Олександрович, реалізовуючи SCD, я зіткнувся з такими помилками:**



Ця помилка пов'язана з обмеженнями безпеки в MySQL, які не дозволяють оновлювати таблицю в межах функції або тригера, якщо ця таблиця вже була використана в операції, яка запустила цю функцію або тригер.

У вашому випадку ви намагаєтеся оновити таблицю **dim\_customers** всередині тригера, але ця ж таблиця була використана в операції **INSERT**, яка спричинила запуск тригера. Це є порушенням безпеки, оскільки оновлення таблиці може призвести до непередбачуваних результатів в операції, яка викликала тригер.

**Я вимикав обмеження безпеки SQL\_SAFE\_UPDATES і це не допомогло**Проте, якщо використання SQL\_SAFE\_UPDATES вже вимкнене, то для усунення цієї помилки можна створити віддалений запит, виконавши його через зовнішній інтерфейс, наприклад, PHP або Python.  
  
**Із пропозицій стак-оверфлоу пропонують робити таблицю History(Як я по суті і зробив раніше) Тобто виконуючи лабу на якомусь Postgres – я це запросто реалізував би (Запитував у колег що вже здали)**