Міністерство освіти і науки України
Національний технічний університет
«Харківський політехнічний інститут»

Катедра комп’ютерної математики і аналізу даних

Технологія великих даних

Звіт до лабораторної роботи

Persistent layer design

Виконав:
ст. гр. КН–120 Р. Б. Питляр

НТУ «ХПІ»
Харків 2022

Зміст

1. Мета роботи

Отримання практичних вмінь з розробки та реалізації ER «сутність — зв'язок» діаграм.

2. План роботи

1. Ознайомитися з даними й підготувати структуру діаграми за посиланням https://www.kaggle.com/usdot/flight-delays?select=flights.csv.
2. Реалізувати ER «сутність — зв'язок» діаграму на датасеті за посиланняи.

3. Виконання лабораторної роботи

Ознайомившись з даними й підготувавши структуру діаграми за посиланням, реалізуємо ER діаграму за допомогою сервісу draw.io (рис. 1).

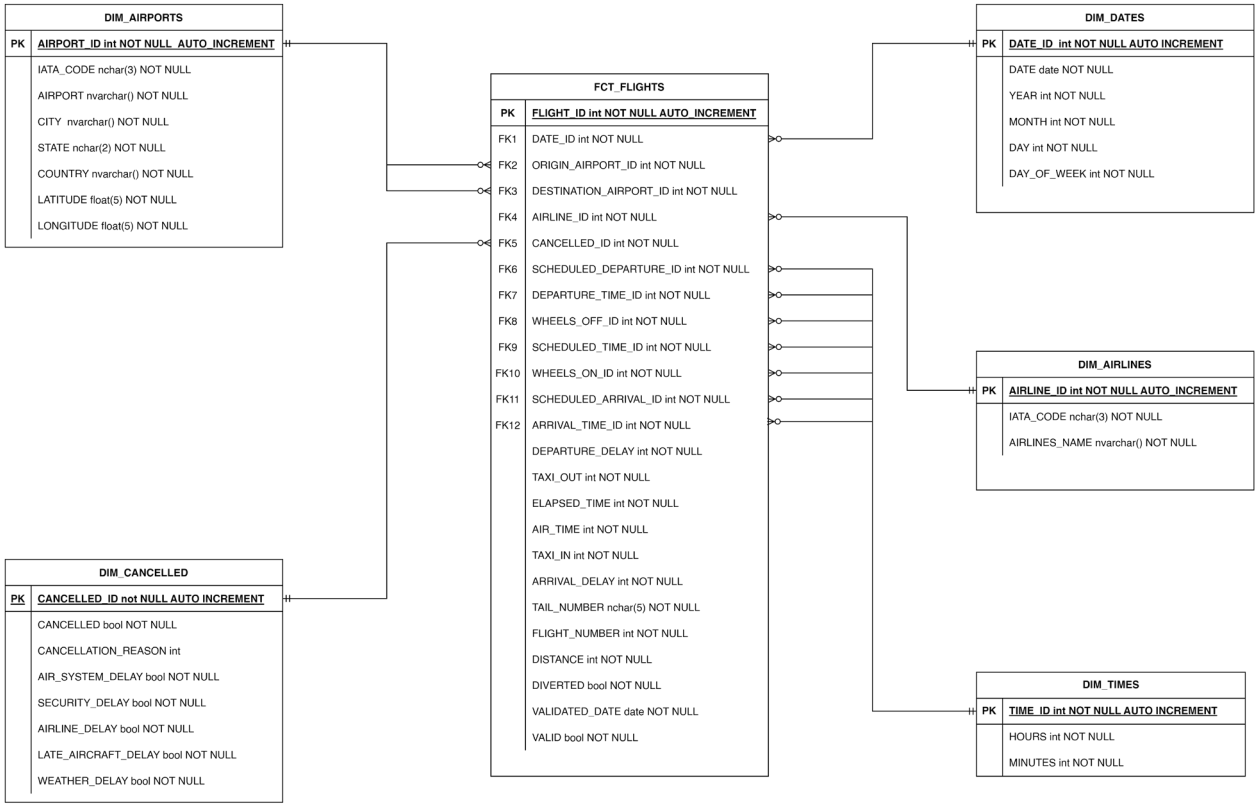

Рис. 1 ER діаграма

Базаданих складається з однієї «FACT» таблиці, та декількох «DIM» таблиць, що відповідає критеріям діаграм за приципом «Star-schema».

Таблиця «FCT_FLIGHTS» - головна таблиця, яка зберігає FK від інших таблиць та унікальні дані для польоту.

Таблиця «DIM_AIRPORTS» - таблиця з інформацією аеропортів.

Таблиця «DIM_AIRLINES» - таблиця з інформацією авіаліній.

Таблиця «DIM_DATES» - таблиця з датами.

Таблиця «DIM_CANCELED» - таблиця з інформацією про скасування рейсу.

Таблиця «DIM_TIMES» - таблиця з годинами.

Зазначені зв’язки:

Розглянемо зв’язок між таблицями «DIM_AIRPORTS» та «FCT_FLIGHTS».

Відомо, що для кожного польоту має бути зазначено «ORIGIN_AIRPORT_ID» та «DESTINATION_AIRPORT_ID», тобто кожен з «AIRPORT_ID» можливо буде задіяний декілька разів, але не кожний «AIRPORT_ID» з таблиці «DIM_AIRPORTS» може бути задіяний (можуть бути Airports, які не використовуються), тому відношення «1 mandatory to many optional».

Розглянемо зв’язок між таблицями «DIM_AIRLINES» та «FCT_FLIGHTS». Відомо, що для кожного польоту має бути зазначено «AIRLINE», тобто кожен з «AIRLINE_ID» можливо буде задіяний декілька разів, але не кожний «AIRLINE_ID» з таблиці «DIM_AIRLINES» може бути задіяний (можуть бути Airlines не літали в зазначений період), тому відношення «1 mandatory to many optional».

Анологічні зв’язки між «DIM_CANCELLED» та «FCT_FLIGHTS», «DIM_DATES» та «FCT_FLIGHTS» і «DIM_TIMES» та «FCT_FLIGHTS»„

Також було додані поля «VALID», «VALIDATED_DATE» до таблиці «FCT_FLIGHTS» для реалізації відстежування історію змін згідно з «SCD: type 2».

4. Висновки

Було проаналізовано датасет та реалізовано ER «сутність — зв'язок» діаграму, розгорнуто пояснено зв’язки між сутностями. Робота виконана в повному обсязі.