Міністерство освіти і науки України

Національний технічний університет

«Харківський політехнічний інститут»

Катедра комп’ютерної математики і аналізу даних

**Технологія великих даних**

Звiт до лабораторної роботи

**Persistent layer design**

**Виконав:**

ст. гр. КН–120                                                Р. Б. Питляр

НТУ «ХПІ»

Харків 2023

**Зміст**

[**1. Мета роботи 3**](#_Toc130386667)

[**2. План роботи 3**](#_Toc130386668)

[**3. Виконання лабораторної роботи 4**](#_Toc130386669)

[**4. Висновки 6**](#_Toc130386670)

1. **Мета роботи**

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

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

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

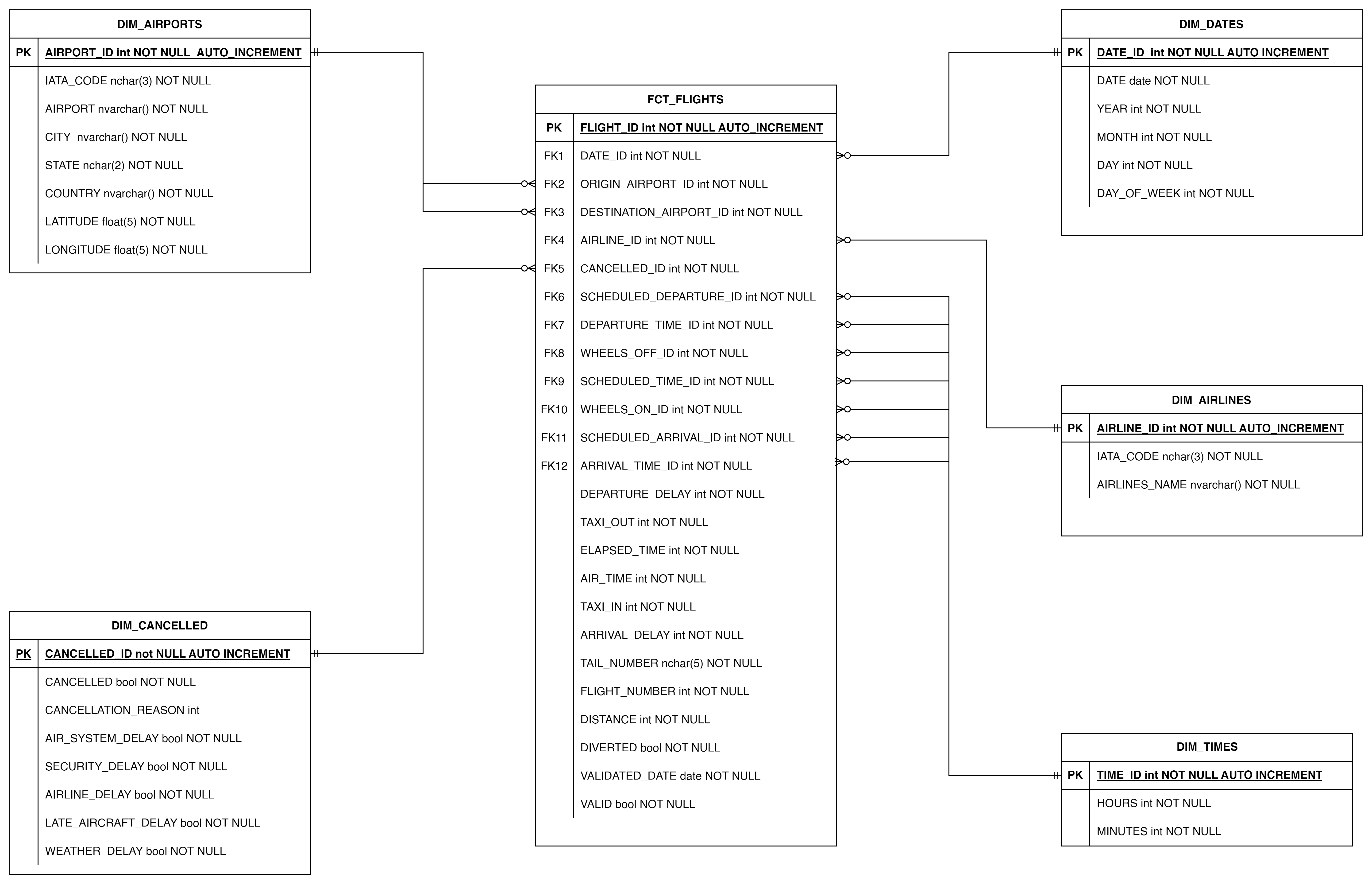


Рис. 1 ER діаграма

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».

1. Висновки

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