Київський національний університет імені Тараса Шевченка

Факультет комп'ютерних наук і кібернетики

Кафедра теорії та технологій програмування

Звіт

Лабораторна робота 1

Виконав студент 4 курсу групи ТТП-41

Бурлака Владислав

Domain short description

This is a countrywide car accident dataset that covers 49 states of the USA. The accident data were collected from February 2016 to March 2023, using multiple APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by various entities, including the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road networks. The dataset currently contains approximately 7.7 million accident records.

Content: This dataset was collected in real-time using multiple Traffic APIs. It contains accident data collected from February 2016 to March 2023 for the Contiguous United States.

Original dataset short description (fields, other info)

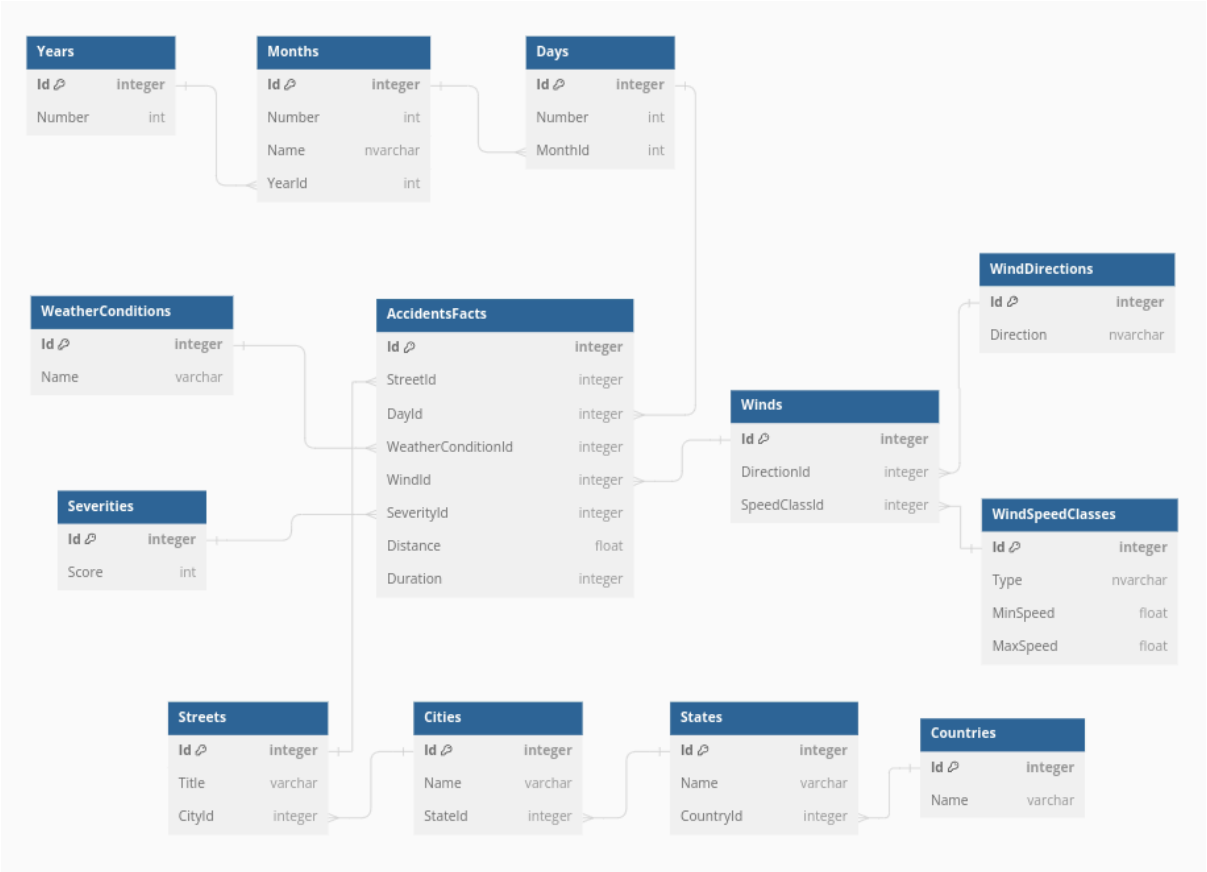
Raw data link:
https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents?resource=download

Columns:

- ID: This is a unique identifier of the accident record.
- SeverityId: Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay)
- Start_Time: Shows start time of the accident in local time zone.
- End_Time: Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow
- Start_Lat: Shows latitude in GPS coordinate of the start point.
- Start_Lng: Shows longitude in GPS coordinate of the start point.
- Distance(mi): The length of the road extent affected by the accident in miles.
- Street: Shows the street name in address field.
- City: Shows the city in address field.
- State: Shows the state in address field.
- Country: Shows the country in address field.
- Visibility(mi): Shows visibility (in miles).

- Wind_Direction: Shows wind direction.
- Wind_Speed(mph): Shows wind speed (in miles per hour).
- Weather_Condition: Shows the weather condition (rain, snow, thunderstorm, fog, etc.)

Star scheme for OLAP warehouse



ETL description

Extract:

Download raw dataset from Kaggle:

https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents?resource=download

Transform:

Bash: cut -d ',' -f 1,3-7,10,12,13,15,17,25-27,29 US_Accidents_March23.csv > DatasetModified.csv

Initial changes using python:

```
import pandas as pd
import numpy as np

csv_file = 'DatasetModed.csv'
df = pd.read_csv(csv_file)

# Normalise ID column
df['ID'] = df['ID'].str.replace('A-', '')

# Normalise datetime columns
df['Start_Time'] =
pd.to_datetime(df['Start_Time'].str.replace(r'\.\d+', ''),
errors='coerce')
df['End_Time'] = pd.to_datetime(df['End_Time'].str.replace(r'\.\d+',
''), errors='coerce')

# Add new column
df['Duration'] = (df['End_Time'] - df['Start_Time']).dt.total_seconds()
df['Duration'] = df['Duration'].replace([np.inf, -np.inf, np.nan], 0)
df['Duration'] = df['Duration'].astype(int)

# Cut time values
df['Start_Time'] = df['Start_Time'].dt.date

# Drop the column
df.drop(columns=['End_Time'], inplace=True)

df.to_csv(csv_file, index=False)
```

Creating DB tables, filling with records and transforming data:

```sql
-- Create the Countries table
CREATE TABLE Countries (
    ID INT PRIMARY KEY,
    Name VARCHAR(50) NOT NULL
);

-- Populate the Countries table with distinct country names
INSERT INTO Countries (ID, Name)
SELECT ROW_NUMBER() OVER (ORDER BY Country) AS ID,
       Country AS Name
FROM AccidentsFacts
GROUP BY Country;
```

```sql
UPDATE AccidentsFacts
SET CountryId = (
    SELECT c.ID
    FROM Countries c
    WHERE AccidentsFacts.Country = c.Name
);
```

```sql
-- Insert all possible combinations of SpeedClass and Direction into the Winds table
INSERT INTO Winds (ID, SpeedClassId, DirectionId)
SELECT
    ROW_NUMBER() OVER (ORDER BY SC.ID, WD.ID) AS ID,
    SC.ID AS SpeedClassId,
    WD.ID AS DirectionId
FROM
    WindSpeedClasses AS SC
CROSS JOIN
    WindDirections AS WD;
```

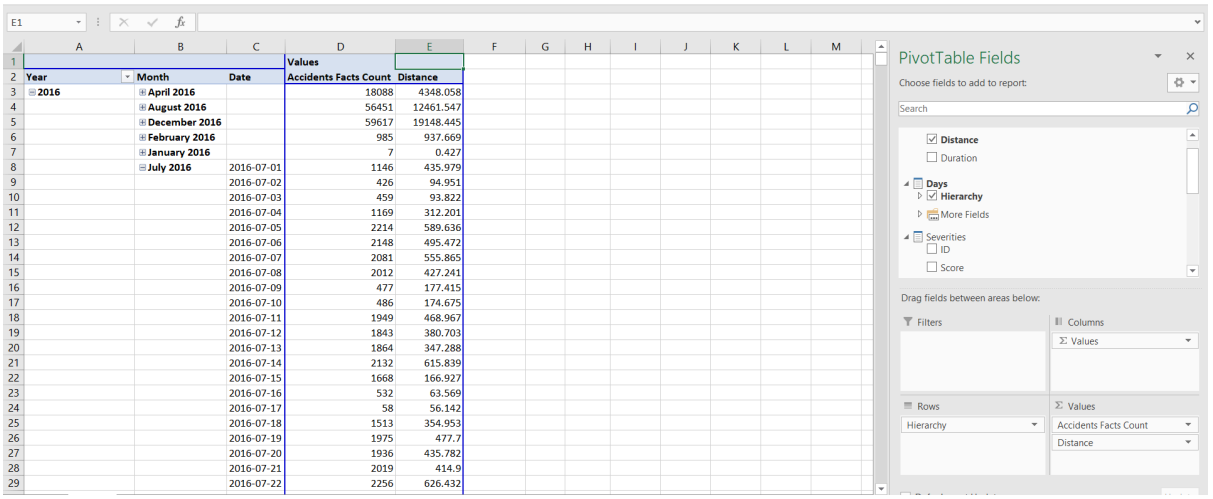
```sql
-- Add the WindId column to the AccidentsFacts table
ALTER TABLE AccidentsFacts
ADD WindId INT;

-- Update the WindId column with the corresponding ID values from the Winds table
UPDATE AccidentsFacts
SET AccidentsFacts.WindId = (
    SELECT TOP 1 W.ID
    FROM Winds AS W
    JOIN WindSpeedClasses AS SC ON W.SpeedClassId = SC.ID
    JOIN WindDirections AS WD ON W.DirectionId = WD.ID
    WHERE
        AccidentsFacts.Wind_Speed_mph BETWEEN SC.MinSpeed AND SC.MaxSpeed
        AND AccidentsFacts.Wind_Direction = WD.Direction
);
```

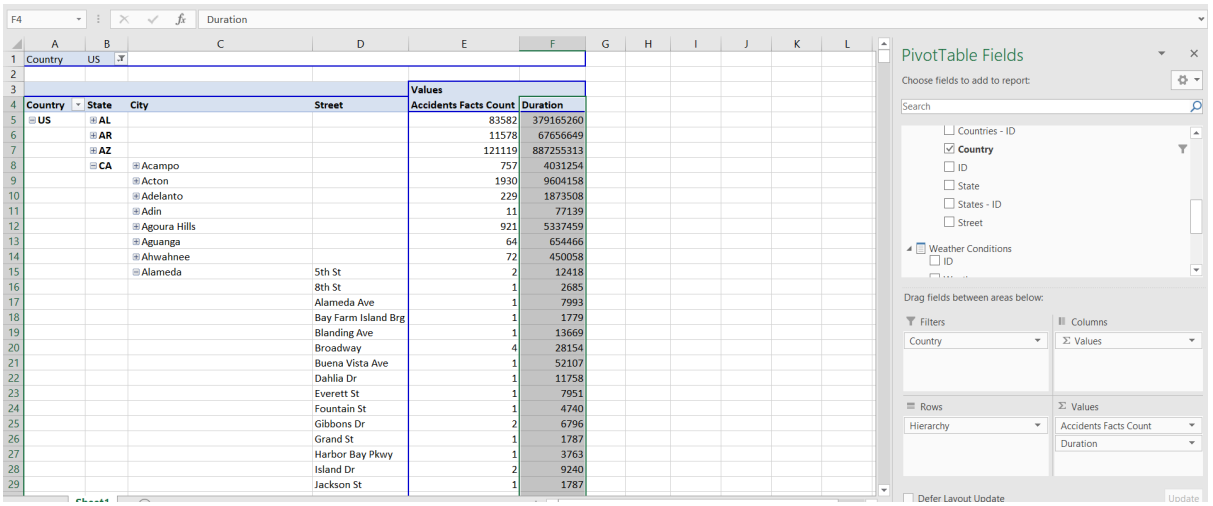
```sql
-- Insert into the Years table with sequential IDs starting from 1
INSERT INTO Years (ID, Number)
SELECT
    ROW_NUMBER() OVER (ORDER BY YEAR(CONVERT(DATE, StartTime, 120))) AS ID,
    YEAR(CONVERT(DATE, StartTime, 120)) AS Number
FROM AccidentsFacts
GROUP BY YEAR(CONVERT(DATE, StartTime, 120));
```

```sql
INSERT INTO Months (ID, Number, YearId, Name)
SELECT
    ROW_NUMBER() OVER (ORDER BY Y.ID, M.Number) AS ID,
    M.Number AS Number,
    Y.ID AS YearId,
    DATENAME(MONTH, DATEFROMPARTS(Y.Number, M.Number, 1)) + ' ' + CAST(Y.Number AS VARCHAR(4)) AS Name
FROM
    (
        SELECT
            YEAR(CONVERT(DATE, AF.StartTime, 120)) AS YearId,
            MONTH(CONVERT(DATE, AF.StartTime, 120)) AS Number
        FROM AccidentsFacts AS AF
        GROUP BY YEAR(CONVERT(DATE, AF.StartTime, 120)), MONTH(CONVERT(DATE, AF.StartTime, 120))
    ) AS M
JOIN
    Years AS Y ON M.YearId = Y.Number;
```

Pivot Tables

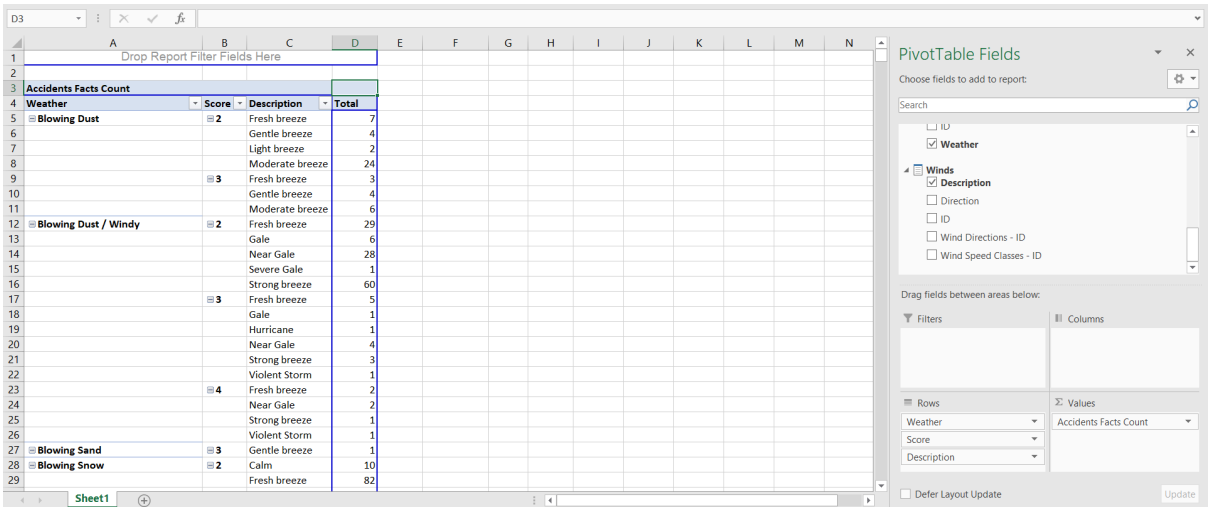


Year	Month	Date	Accidents Facts Count	Distance
2016	April 2016		18088	4348.058
	August 2016		56451	12461.547
	December 2016		59617	19148.445
	February 2016		985	937.669
	January 2016		7	0.427
	July 2016	2016-07-01	1146	435.979
		2016-07-02	426	94.951
		2016-07-03	459	93.822
		2016-07-04	1169	312.201
		2016-07-05	2214	589.636
		2016-07-06	2148	495.472
		2016-07-07	2081	555.865
		2016-07-08	2012	427.241
		2016-07-09	477	177.415
		2016-07-10	486	174.675
		2016-07-11	1949	468.967
		2016-07-12	1843	380.703
		2016-07-13	1864	347.288
		2016-07-14	2132	615.839
		2016-07-15	1668	166.927
		2016-07-16	532	63.569
		2016-07-17	58	56.142
		2016-07-18	1513	354.953
		2016-07-19	1975	477.7
		2016-07-20	1936	435.782
		2016-07-21	2019	414.9
		2016-07-22	2256	626.432



Country: US

Country	State	City	Street	Accidents Facts Count	Duration
US	AL			83582	379165260
	AR			11578	67656649
	AZ			121119	887255313
	CA	Acampo		757	4031254
		Acton		1930	9604158
		Adelanto		229	1873508
		Adin		11	77139
		Agoura Hills		921	5337459
		Aguanga		64	654466
		Ahwahnee		72	450058
		Alameda	5th St	2	12418
			8th St	1	2685
			Alameda Ave	1	7993
			Bay Farm Island Brg	1	1779
			Blanding Ave	1	13669
			Broadway	4	28154
			Buena Vista Ave	1	52107
			Dahlia Dr	1	11758
			Everett St	1	7951
			Fountain St	1	4740
			Gibbons Dr	2	6796
			Grand St	1	1787
			Harbor Bay Pkwy	1	3763
			Island Dr	2	9240
			Jackson St	1	1787



Drop Report Filter Fields Here

Accidents Facts Count

Weather	Score	Description	Total
Blowing Dust	2	Fresh breeze	7
		Gentle breeze	4
		Light breeze	2
		Moderate breeze	24
	3	Fresh breeze	3
		Gentle breeze	4
		Moderate breeze	6
Blowing Dust / Windy	2	Fresh breeze	29
		Gale	6
		Near Gale	28
		Severe Gale	1
		Strong breeze	60
	3	Fresh breeze	5
		Gale	1
		Hurricane	1
		Near Gale	4
		Strong breeze	3
		Violent Storm	1
	4	Fresh breeze	2
		Near Gale	2
		Strong breeze	1
		Violent Storm	1
Blowing Sand	3	Gentle breeze	1
Blowing Snow	2	Calm	10
		Fresh breeze	82

Multi-Dimensional Storage