# SQL Querying

**Authors:** Ayman EL ALASS & Abderaouf KHELFAOUI

Objective: In this notebook, we perform analytical queries on our relational database `project_database.db`. We focus on **raw data extraction and aggregation** to derive relevent insights directly from the query results.

## Exemples

1. **Basic Retrieval:** Extracting specific high-delay incidents.
2. **Aggregation:** Ranking airlines by punctuality and reliability.
3. **Complex Analysis:** Identifying the "Black List" of flight routes using double joins.
4. **Cross-Domain Analysis:** Correlating weather conditions with flight cancellations.
5. **Database Evolution:** Modifying the schema to permanently store delay categories.
6. **Global Statistics:** Analyzing total flight volume and the busiest day of the year.
7. **Geographic Insights:** Identifying busiest airports and state-level traffic peaks.
8. **Delay & Cancellation Deep Dive:** Pinpointing the worst airline/airport combinations and cancellation hotspots.

In [19]:
```python
import sqlite3
import pandas as pd

# Connect to the database generated by main.py
db_name = "project_database.db"
conn = sqlite3.connect(db_name)
print(f"Connected to database: {db_name}")

# Helper function to execute SQL and return a readable DataFrame
def run_query(query):
    try:
        df = pd.read_sql(query, conn)
        return df
    except Exception as e:
        print(f"SQL Error: {e}")
```

```
Connected to database: project_database.db
```

## 1. Basic Data Retrieval (Sanity Check)

**Scenario:** A specific request from the Operations Center. We need to list details of flights operated by 'American Airlines Inc.' that faced extreme delays (> 4 hours) to identify patterns in specific airports.

In [20]:
```python
query_1 = """
SELECT
    f.flight_date,
    f.flight_number,
    f.origin_airport,
```

```
    f.dest_airport,
    f.dep_delay || ' min' as delay_minutes -- Formatting for readability
FROM FLIGHTS f
JOIN AIRLINES a ON f.airline_code = a.airline_code
WHERE a.airline_name = 'American Airlines Inc.'
  AND f.dep_delay > 240
ORDER BY f.dep_delay DESC
LIMIT 10;
"""

print(">>> Extreme Delays Report (American Airlines):")
display(run_query(query_1))
```

```
>>> Extreme Delays Report (American Airlines):
```

| | flight_date | flight_number | origin_airport | dest_airport | delay_minutes |
|---|---|---|---|---|---|
| **0** | 2015-01-03 | 1677 | MEM | DFW | 1380 min |
| **1** | 2015-01-04 | 1279 | OMA | DFW | 1255 min |
| **2** | 2015-01-01 | 2470 | BOS | DFW | 1190 min |
| **3** | 2015-01-05 | 1495 | EGE | DFW | 1190 min |
| **4** | 2015-01-05 | 970 | LAS | LAX | 1046 min |
| **5** | 2015-01-03 | 2281 | IND | DFW | 949 min |
| **6** | 2015-01-04 | 2208 | CLE | DFW | 925 min |
| **7** | 2015-01-03 | 291 | JFK | AUS | 891 min |
| **8** | 2015-01-03 | 45 | JFK | LAS | 886 min |
| **9** | 2015-01-02 | 1302 | RDU | DFW | 885 min |

## 2. Airline Performance Audit

**Exemple Question:** Which airlines are the most reliable? We calculate three key KPIs per airline:

1. **Volume:** Total flights.
2. **Punctuality:** Average departure delay.
3. **Reliability:** Cancellation Rate (%).

In [21]:

```
query_2 = """
SELECT
    a.airline_name,
    COUNT(f.flight_id) as total_flights,
    ROUND(AVG(f.dep_delay), 2) as avg_delay_min,
    SUM(CASE WHEN f.cancelled = 1 THEN 1 ELSE 0 END) as cancelled_count,
    ROUND(
        (CAST(SUM(CASE WHEN f.cancelled = 1 THEN 1 ELSE 0 END) as FLOAT) / COUNT
    2) || '%' as cancellation_rate
FROM FLIGHTS f
JOIN AIRLINES a ON f.airline_code = a.airline_code
GROUP BY a.airline_name
HAVING total_flights > 500 -- Filter to keep only major airlines
```

```
ORDER BY avg_delay_min ASC;
"""

print(">>> Airline Performance Matrix (Ranked by Punctuality):")
display(run_query(query_2))
```

```
>>> Airline Performance Matrix (Ranked by Punctuality):
```

| | airline_name | total_flights | avg_delay_min | cancelled_count | cancellation_rate |
|---|---|---|---|---|---|
| **0** | Alaska Airlines Inc. | 2836 | 7.05 | 4 | 0.14% |
| **1** | Virgin America | 1036 | 7.93 | 2 | 0.19% |
| **2** | Hawaiian Airlines Inc. | 1328 | 8.52 | 2 | 0.15% |
| **3** | Delta Air Lines Inc. | 13156 | 10.94 | 19 | 0.14% |
| **4** | US Airways Inc. | 6968 | 11.02 | 62 | 0.89% |
| **5** | JetBlue Airways | 4688 | 18.43 | 16 | 0.34% |
| **6** | Skywest Airlines Inc. | 10475 | 19.93 | 418 | 3.99% |
| **7** | Southwest Airlines Co. | 21160 | 21.39 | 179 | 0.85% |
| **8** | Atlantic Southeast Airlines | 10810 | 21.55 | 421 | 3.89% |
| **9** | American Airlines Inc. | 9449 | 23.24 | 260 | 2.75% |
| **10** | United Air Lines Inc. | 8413 | 24.10 | 65 | 0.77% |
| **11** | Spirit Air Lines | 1813 | 27.29 | 18 | 0.99% |
| **12** | American Eagle Airlines Inc. | 6305 | 29.27 | 879 | 13.94% |
| **13** | Frontier Airlines Inc. | 1563 | 32.38 | 44 | 2.82% |

# 3. Route Analysis (Double Join)

**Context:** We want to identify the specific City-to-City connections that suffer from the worst delays.

**Technique:** We perform a **Double Join** on the `AIRPORTS` table (aliased as `origin` and `dest`) to retrieve readable city names instead of IATA codes.

In [22]:
```
query_3 = """
SELECT
    origin.city || ' -> ' || dest.city AS Route,
    COUNT(*) as Flight_Count,
    ROUND(AVG(f.dep_delay), 2) as Avg_Delay_Min,
    MAX(f.dep_delay) as Max_Delay_Min
FROM FLIGHTS f
```

```
JOIN AIRPORTS origin ON f.origin_airport = origin.iata_code
JOIN AIRPORTS dest ON f.dest_airport = dest.iata_code
GROUP BY f.origin_airport, f.dest_airport
HAVING Flight_Count > 20 -- Ignore rare charter routes
ORDER BY Avg_Delay_Min DESC
LIMIT 10;
"""

print(">>> Top 10 Routes with Highest Average Delays:")
display(run_query(query_3))
```

```
>>> Top 10 Routes with Highest Average Delays:
```

| | Route | Flight_Count | Avg_Delay_Min | Max_Delay_Min |
|---|---|---|---|---|
| **0** | San Juan -> Chicago | 21 | 74.67 | 358 |
| **1** | Chicago -> Mosinee | 23 | 63.83 | 299 |
| **2** | Columbus -> Chicago | 66 | 59.53 | 739 |
| **3** | Denver -> Palm Springs | 27 | 59.04 | 326 |
| **4** | Memphis -> Dallas-Fort Worth | 33 | 58.06 | 1380 |
| **5** | Lihue -> Los Angeles | 28 | 57.75 | 860 |
| **6** | New York -> Austin | 27 | 57.48 | 891 |
| **7** | Chicago -> Springfield | 22 | 56.77 | 331 |
| **8** | Aspen -> Chicago | 25 | 55.96 | 204 |
| **9** | Denver -> Colorado Springs | 37 | 55.73 | 200 |

# 4. Exemple of Weather Impact Querying

**Context:** For instance we can assume that to avoid heavy processing times, we analyze the correlation between wind and delays for a **single specific day** (January 1st, 2015).

**Technique:** We aggregate the average wind speed and average delay per airport for that day.

In [23]:
```
query_4 = """
WITH DailyWeather AS (
    SELECT
        airport_code,
        AVG(wind_speed) as avg_wind_speed
    FROM WEATHER
    WHERE date(reading_time) = '2015-01-01'
    GROUP BY airport_code
),
DailyFlights AS (
    SELECT
        origin_airport,
        AVG(dep_delay) as avg_dep_delay
    FROM FLIGHTS
    WHERE date(flight_date) = '2015-01-01'
    GROUP BY origin_airport
)
```

```
SELECT
    f.origin_airport,
    f.avg_dep_delay,
    w.avg_wind_speed
FROM DailyFlights f
JOIN DailyWeather w ON f.origin_airport = w.airport_code
ORDER BY f.avg_dep_delay DESC;
"""

df_result = run_query(query_4)
display(df_result)
```

| | origin_airport | avg_dep_delay | avg_wind_speed |
|---|---|---|---|
| **0** | DEN | 29.416804 | 1.375000 |
| **1** | DFW | 22.316456 | 2.416667 |
| **2** | IAH | 14.187643 | 3.291667 |
| **3** | PHX | 12.929245 | 1.333333 |
| **4** | ORD | 11.043353 | 8.833333 |
| **5** | BOS | 9.355932 | 5.666667 |
| **6** | SFO | 9.248244 | 2.708333 |
| **7** | LAX | 8.199640 | 0.916667 |
| **8** | MIA | 7.330357 | 2.041667 |
| **9** | DTW | 6.750000 | 7.625000 |
| **10** | MSP | 6.191964 | 4.791667 |
| **11** | SEA | 4.872340 | 1.291667 |
| **12** | JFK | 4.680000 | 3.166667 |
| **13** | PHL | 4.085526 | 2.666667 |
| **14** | ATL | 2.782432 | 1.041667 |

# 5. Database Evolution

**Requirement:** To optimize future reporting, we need to persist the "Delay Category" directly in the database table, rather than calculating it every time.

**Actions:**

1. **ALTER TABLE:** Add a new column `delay_category`.
2. **UPDATE:** Populate this column based on the `dep_delay` value.

In [24]:
```
# 1. Add the column structure
try:
    conn.execute("ALTER TABLE FLIGHTS ADD COLUMN delay_category VARCHAR(20)")
    print("Schema Altered: Column 'delay_category' added.")
except sqlite3.OperationalError:
    print("Column 'delay_category' already exists.")
```

```
# 2. Populate the data
update_query = """
UPDATE FLIGHTS
SET delay_category = CASE
    WHEN dep_delay <= 0 THEN 'On Time / Early'
    WHEN dep_delay > 0 AND dep_delay <= 15 THEN 'Small Delay'
    WHEN dep_delay > 15 AND dep_delay <= 45 THEN 'Medium Delay'
    ELSE 'Major Delay (>45m)'
END;
"""
conn.execute(update_query)
conn.commit()
print("Data Updated: Categories populated.")

# 3. Verification Query
query_check = """
SELECT
    delay_category,
    COUNT(*) as flight_count,
    ROUND((CAST(COUNT(*) as FLOAT) / (SELECT COUNT(*) FROM FLIGHTS)) * 100, 1) |
FROM FLIGHTS
GROUP BY delay_category
ORDER BY flight_count DESC;
"""

print(">>> Verification: Distribution of new categories:")
display(run_query(query_check))
```

```
Column 'delay_category' already exists.
Data Updated: Categories populated.
>>> Verification: Distribution of new categories:
```

| | delay_category | flight_count | proportion |
|---|---|---|---|
| **0** | On Time / Early | 45965 | 46.0% |
| **1** | Small Delay | 22195 | 22.2% |
| **2** | Medium Delay | 17049 | 17.0% |
| **3** | Major Delay (>45m) | 14791 | 14.8% |

# 6. Global Statistics

In [25]:

```
# Total flights in 2015
query_total = "SELECT COUNT(flight_id) as 'Total Flights 2015' FROM flights;"
display(run_query(query_total))

# Busiest day of the year
query_busiest_day = """
SELECT flight_date, COUNT(flight_id) as number_of_flights
FROM flights
WHERE strftime('%Y', flight_date) = '2015'
GROUP BY flight_date
ORDER BY number_of_flights DESC
LIMIT 1;
"""
```

```
print(">>> Busiest day of 2015:")
display(run_query(query_busiest_day))
```

| | Total Flights 2015 |
|---|---|
| **0** | 100000 |

```
>>> Busiest day of 2015:
```

| | flight_date | number_of_flights |
|---|---|---|
| **0** | 2015-01-02 | 16741 |

# 7. Airport & Location Analysis

In [26]:
```
# Busiest Airport (Origin)
query_busiest_airport = """
SELECT f.origin_airport, a.airport_name, COUNT(f.flight_id) AS number_of_flights
FROM flights f
JOIN airports a ON f.origin_airport = a.iata_code
WHERE strftime('%Y', f.flight_date) = '2015'
GROUP BY f.origin_airport
ORDER BY number_of_flights DESC
LIMIT 1;
"""
print(">>> Busiest Airport:")
display(run_query(query_busiest_airport))

# Peak traffic day per State (Fixed to show ONLY 1 day per state)
query_state_peak = """
WITH DailyStats AS (
    SELECT
        a.state,
        f.origin_airport,
        a.airport_name,
        f.flight_date,
        COUNT(f.flight_id) AS number_of_flights,
        ROW_NUMBER() OVER (
            PARTITION BY a.state
            ORDER BY COUNT(f.flight_id) DESC, f.flight_date ASC
        ) as rang
    FROM flights f
    JOIN airports a ON f.origin_airport = a.iata_code
    WHERE f.flight_date BETWEEN '2015-01-01' AND '2015-12-31'
    GROUP BY a.state, f.origin_airport, a.airport_name, f.flight_date
)
SELECT
    state,
    origin_airport,
    airport_name,
    flight_date,
    number_of_flights
FROM DailyStats
WHERE rang = 1
ORDER BY state;
"""
```

```
print(">>> Peak Traffic Day per State (Unique Day):")
display(run_query(query_state_peak))
```

```
>>> Busiest Airport:
```

| | origin_airport | airport_name | number_of_flights |
|---|---|---|---|
| **0** | ATL | Hartsfield-Jackson Atlanta International Airport | 6009 |

```
>>> Peak Traffic Day per State (Unique Day):
```

| | state | origin_airport | airport_name | flight_date | number_of_flights |
|---|---|---|---|---|---|
| **0** | AK | ANC | Ted Stevens Anchorage International Airport | 2015-01-02 | 47 |
| **1** | AL | BHM | Birmingham-Shuttlesworth International Airport | 2015-01-05 | 40 |
| **2** | AR | LIT | Bill and Hillary Clinton National Airport (Ada... | 2015-01-02 | 35 |
| **3** | AS | PPG | Pago Pago International Airport (Tafuna Airport) | 2015-01-02 | 1 |
| **4** | AZ | PHX | Phoenix Sky Harbor International Airport | 2015-01-02 | 487 |
| **5** | CA | LAX | Los Angeles International Airport | 2015-01-02 | 640 |
| **6** | CO | DEN | Denver International Airport | 2015-01-02 | 658 |
| **7** | CT | BDL | Bradley International Airport | 2015-01-05 | 62 |
| **8** | DE | ILG | Wilmington Airport | 2015-01-01 | 1 |
| **9** | FL | MCO | Orlando International Airport | 2015-01-03 | 360 |
| **10** | GA | ATL | Hartsfield-Jackson Atlanta International Airport | 2015-01-02 | 1042 |
| **11** | GU | GUM | Guam International Airport | 2015-01-01 | 1 |
| **12** | HI | HNL | Honolulu International Airport | 2015-01-02 | 144 |
| **13** | IA | DSM | Des Moines International Airport | 2015-01-05 | 31 |
| **14** | ID | BOI | Boise Airport (Boise Air Terminal) | 2015-01-02 | 36 |
| **15** | IL | ORD | Chicago O'Hare International Airport | 2015-01-02 | 812 |
| **16** | IN | IND | Indianapolis International Airport | 2015-01-05 | 78 |
| **17** | KS | ICT | Wichita Dwight D. Eisenhower National Airport ... | 2015-01-02 | 24 |
| **18** | KY | CVG | Cincinnati/Northern Kentucky International Air... | 2015-01-02 | 67 |
| **19** | LA | MSY | Louis Armstrong New Orleans International Airport | 2015-01-04 | 118 |
| **20** | MA | BOS | Gen. Edward Lawrence Logan International Airport | 2015-01-05 | 305 |

| | state | origin_airport | airport_name | flight_date | number_of_flights |
|---|---|---|---|---|---|
| **21** | MD | BWI | Baltimore-Washington International Airport | 2015-01-02 | 253 |
| **22** | ME | PWM | Portland International Jetport | 2015-01-03 | 9 |
| **23** | MI | DTW | Detroit Metropolitan Airport | 2015-01-05 | 334 |
| **24** | MN | MSP | Minneapolis-Saint Paul International Airport | 2015-01-04 | 305 |
| **25** | MO | STL | St. Louis International Airport at Lambert Field | 2015-01-05 | 143 |
| **26** | MS | JAN | Jackson-Evers International Airport | 2015-01-02 | 24 |
| **27** | MT | BZN | Bozeman Yellowstone International Airport (Gal... | 2015-01-04 | 18 |
| **28** | NC | CLT | Charlotte Douglas International Airport | 2015-01-05 | 329 |
| **29** | ND | FAR | Hector International Airport | 2015-01-05 | 18 |
| **30** | NE | OMA | Eppley Airfield | 2015-01-05 | 52 |
| **31** | NH | MHT | Manchester-Boston Regional Airport | 2015-01-05 | 18 |
| **32** | NJ | EWR | Newark Liberty International Airport | 2015-01-05 | 318 |
| **33** | NM | ABQ | Albuquerque International Sunport | 2015-01-02 | 63 |
| **34** | NV | LAS | McCarran International Airport | 2015-01-02 | 390 |
| **35** | NY | LGA | LaGuardia Airport (Marine Air Terminal) | 2015-01-06 | 338 |
| **36** | OH | CLE | Cleveland Hopkins International Airport | 2015-01-06 | 102 |
| **37** | OK | OKC | Will Rogers World Airport | 2015-01-02 | 52 |
| **38** | OR | PDX | Portland International Airport | 2015-01-02 | 146 |
| **39** | PA | PHL | Philadelphia International Airport | 2015-01-02 | 197 |
| **40** | PR | SJU | Luis Muñoz Marín International Airport | 2015-01-03 | 102 |
| **41** | RI | PVD | Theodore Francis Green State Airport | 2015-01-05 | 34 |

| | state | origin_airport | airport_name | flight_date | number_of_flights |
|---|---|---|---|---|---|
| **42** | SC | CHS | Charleston International Airport/Charleston AFB | 2015-01-02 | 37 |
| **43** | SD | FSD | Sioux Falls Regional Airport | 2015-01-05 | 21 |
| **44** | TN | BNA | Nashville International Airport | 2015-01-02 | 149 |
| **45** | TX | DFW | Dallas/Fort Worth International Airport | 2015-01-02 | 809 |
| **46** | UT | SLC | Salt Lake City International Airport | 2015-01-02 | 328 |
| **47** | VA | DCA | Ronald Reagan Washington National Airport | 2015-01-06 | 229 |
| **48** | VI | STT | Cyril E. King Airport | 2015-01-03 | 23 |
| **49** | VT | BTV | Burlington International Airport | 2015-01-02 | 12 |
| **50** | WA | SEA | Seattle-Tacoma International Airport | 2015-01-02 | 322 |
| **51** | WI | MKE | General Mitchell International Airport | 2015-01-02 | 88 |
| **52** | WV | CRW | Yeager Airport | 2015-01-04 | 8 |
| **53** | WY | JAC | Jackson Hole Airport | 2015-01-03 | 19 |

# 8. Delays and Cancellations

In [27]:
```
# Average Delay per Airline per Airport
query_delay_airline_airport = """
SELECT ar.airline_name, f.origin_airport, ROUND(AVG(f.dep_delay), 2) AS avg_dep_
FROM flights f
JOIN airports ap ON f.origin_airport = ap.iata_code
JOIN airlines ar ON f.airline_code = ar.airline_code
WHERE strftime('%Y', f.flight_date) = '2015'
GROUP BY ar.airline_name, f.origin_airport
ORDER BY avg_dep_delay DESC
LIMIT 10;
"""
print(">>> Top 10 Highest Average Delays (Airline/Airport):")
display(run_query(query_delay_airline_airport))

# Top 3 Airports with most cancellations
query_cancelled_airports = """
SELECT origin_airport, airport_name, COUNT(cancelled) AS cancelled_count
FROM airports ap
JOIN flights f ON ap.iata_code = f.origin_airport
WHERE cancelled = 1
```

```
GROUP BY origin_airport, airport_name
ORDER BY cancelled_count DESC
LIMIT 3;
"""
print(">>> Top 3 Airports for Cancellations:")
display(run_query(query_cancelled_airports))
```

```
>>> Top 10 Highest Average Delays (Airline/Airport):
```

| | airline_name | origin_airport | avg_dep_delay |
|---|---|---|---|
| **0** | Skywest Airlines Inc. | SYR | 159.00 |
| **1** | Delta Air Lines Inc. | LIH | 152.50 |
| **2** | United Air Lines Inc. | CMH | 135.00 |
| **3** | Frontier Airlines Inc. | ILG | 130.20 |
| **4** | Atlantic Southeast Airlines | HDN | 125.00 |
| **5** | Hawaiian Airlines Inc. | JFK | 118.00 |
| **6** | Skywest Airlines Inc. | LGA | 118.00 |
| **7** | United Air Lines Inc. | ICT | 105.67 |
| **8** | Spirit Air Lines | STT | 99.50 |
| **9** | Skywest Airlines Inc. | ALB | 97.80 |

```
>>> Top 3 Airports for Cancellations:
```

| | origin_airport | airport_name | cancelled_count |
|---|---|---|---|
| **0** | DFW | Dallas/Fort Worth International Airport | 335 |
| **1** | ORD | Chicago O'Hare International Airport | 304 |
| **2** | DEN | Denver International Airport | 120 |