Flights Case Study

Presented by Catherine Sealey
CANDIDATE_00355

Executive Summary

Project Objective

The goal was to **transform raw US domestic flights data** into a clean, analysis-ready Snowflake model, enabling efficient insights and informed decision-making through Tableau integration.

Key Deliverables

Achievements included:

- Normalised data loaded from the flights.gz flat file
- A star schema with three dimension tables and one fact table
- Three calculated business metrics
- Continuous date dimension to enhance analytical capabilities
- Final reporting view that combines the star schema into clean, actionable data

Business Value

More reliable reporting through curated metrics, reduced manual data preparation, seamless Tableau integration, and a scalable foundation for future analytics.

Analysts can now explore flight delays, cancellations, and route performance with reliable filtering and grouping.

The date dimension enables trend analysis across various time periods.

Approach

This project follows a structured, best-practice approach to data modeling and analytics.

The process begins with a thorough analysis of the raw flight data to understand its structure and identify key business entities.

Dimensional modeling principles are applied to design a star schema with a central fact table and supporting dimension tables, ensuring the model is intuitive and efficient for BI tools like Tableau.

Data quality is prioritized by systematically identifying and repairing issues such as inconsistent naming, missing values, and type mismatches.

Each step, from data cleaning to schema design, is carefully documented to ensure transparency and reproducibility.

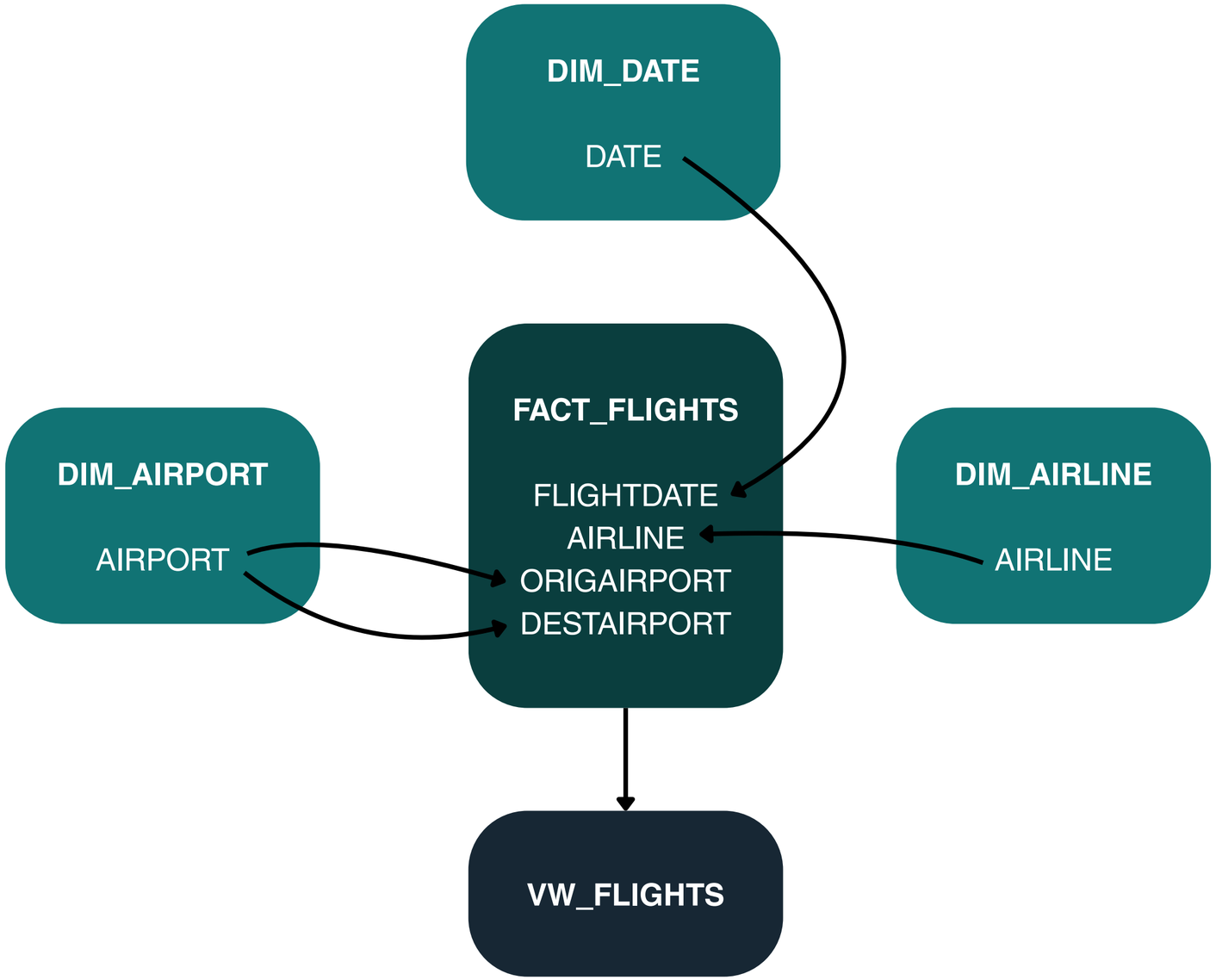
This approach results in a robust, extensible data model that supports accurate and insightful analysis.

Data Schema

The model uses a simple star schema with FACT_FLIGHTS linked to airline, airport, and date dimensions.

This structure allows the creation of VW_FLIGHTS, a single, streamlined view of all key flight data.

Because star schemas are purpose-built for analytics and minimise complex joins, the result is faster queries, easier reporting, and a reliable foundation your team can use to make confident, data-driven decisions.



Key Calculated Metrics

Distance Grouping, Delay Flag and Next Day Arrival Indicator

Three calculated metrics can be found in the fact table and view. The table below highlights the importance of their inclusion and how they have been calculated.

Metric	What	Value	Example	Insight
DISTANCEGROUP Route Length Segmentation	Bins flights into 100-mile categories	Compare short-haul vs long-haul performance	94 → 0-100 miles, 274 → 201-300 miles, 2,580 → 2501-2600 miles	Are delays more common on short or long routes?
DEPDELAYGT15 On-Time Performance Flag	Binary flag (0/1) for departures >15 min late	Standard metric for late departures (DOT: 15 min threshold)	AVG(DEPDELAYGT15) = 0.32 → 32% of flights delayed	Which airline has the best on-time performance?
NEXTDAYARR Red-Eye Flight Indicator	Flags flights arriving next calendar day	Ensures accurate duration and delay calculations	Dep 11:30 PM → Arr 6:15 AM → NEXTDAYARR = 1	Avoids negative flight durations for overnight flights

Data Quality Challenges and Solutions

This section highlights the significant **data quality challenges** faced during transformation, including regex cleanup, type conversions, and generating a continuous date dimension, ensuring 100% success in data validation and accuracy.

The main challenges encountered were:

1. Airline Names with Codes and Notes
2. Airport Names with City/State
3. Date Format Challenge
4. Type Mismatches
5. Distance Data Quality
6. Columns with Incomplete Data

1: **Airline Names with Codes and Notes**

The raw data contained airline codes and historical merger notes, e.g., "America West Airlines Inc.: HP (Merged with US Airways 9/05…)".

This can cause cluttered reports and inconsistent displays.

In order to clean the airline names, the codes have been removed while preserving historical notes in a separate column. Regular expressions were used to standardize the names, resulting in cleaner, more consistent reporting.

2: **Airport Names with City/State**

Some airport fields included both city and state information, e.g., "AlbuquerqueNM: Albuquerque International Sunport", even though city and state were already in separate columns.

This created redundancy and complicated filtering in Tableau.

The airport names have been cleaned while keeping city and state in their own columns, enabling clearer visualizations and easier grouping.

Data Quality Continued

3: **Date Format Challenge**

Dates were stored as numbers (20020101), preventing date calculations and time-series analysis.

These have been converted to proper DATE types and a DIM_DATE table with attributes like Year, Quarter, Month, Day of Week, Weekend flag, and Week of Year.

This allows analysts to easily examine trends, such as weekend versus weekday delays.

4: **Type Mismatches in Delay Columns**

Delay columns were originally stored as integers (NUMBER(4,0)), which limited flexibility for calculations.

They have been converted to decimals (NUMBER(10,2)) for both departure and arrival delays.

This ensures the data type can accommodate future metrics, averages, or any calculations that produce fractional values, supporting more accurate analytics and reporting.

Data Quality Continued

5: **Distance Data Quality**

Distance data was stored as text, sometimes including the word "miles," preventing calculations or comparisons.

Non-numeric characters have been stripped, the column converted to a numeric type, and distance bins (DISTANCEGROUP) were created to segment flights.

This enables accurate analyses and easier comparisons across route lengths.

6: **Columns With Incomplete Data**

Some columns in the raw data are not relevant or have incomplete or missing data. These columns have been removed when creating the fact table:

1. TAILNUM - Too granular for strategic analysis
2. FLIGHTNUM - TRANSACTIONID serves as unique identifier
3. CRSDEPTIME - DEPDELAY captures the variance
4. TAXIOUT - Operational detail
5. WHEELSOFF - Operational detail
6. WHEELSON - Operational detail
7. TAXIIN - Operational detail
8. CRSARRTIME - ARRDELAY captures the variance
9. CRSELAPSEDTIME - Can derive if needed
10. ACTUALELAPSEDTIME - Can derive if needed
11. DIVERTED - Focus is on delays/cancellations

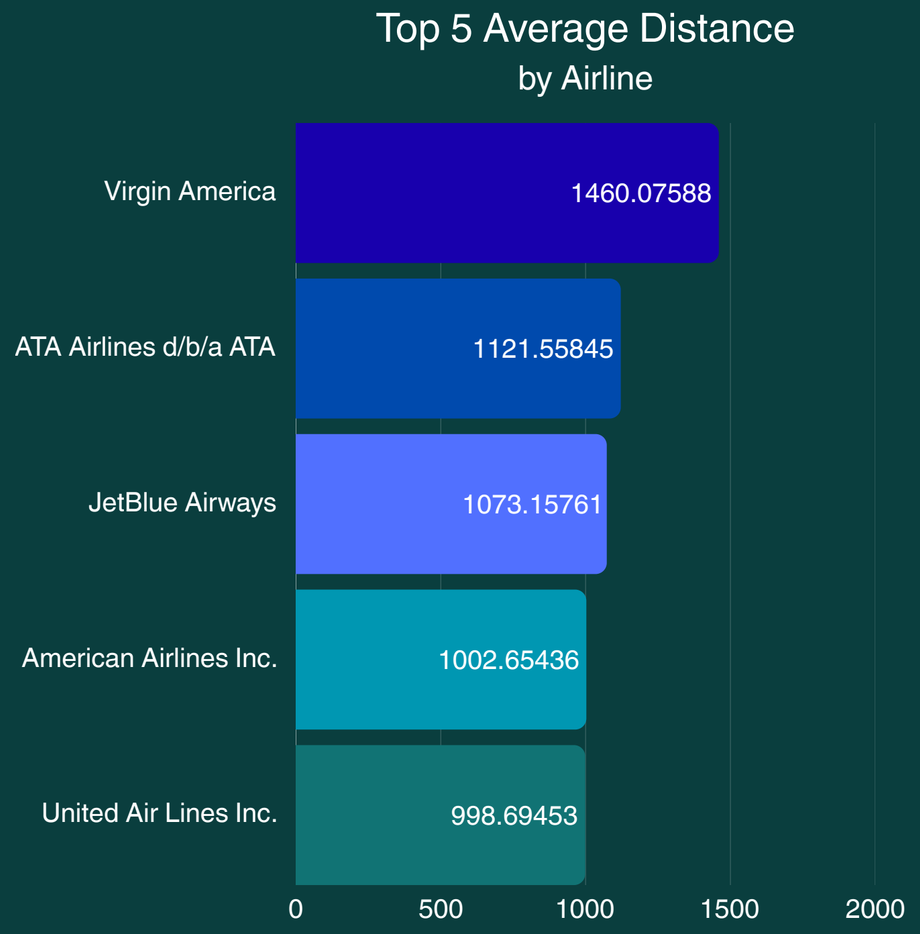
Business Output Example

Below is a demostration of an output from a basic query to return usable data. The table below shows a sample of total flights, average departure delay, delayed flights, cancelled flights and average flight distance by airline.

AIRLINENAME	TOTAL_FLIGHTS	AVG_DEPARTURE_DELAY	DELAYED_FLIGHTS	CANCELLED_FLIGHTS	AVG_DISTANCE
Southwest Airlines Co.	189985	8.74889036	31474	2550	570.09475485
Delta Air Lines Inc.	166601	6.89740497	20413	3303	763.99701082
American Airlines Inc.	139782	7.35415672	19086	3624	1002.65435464
United Air Lines Inc.	121804	8.82364606	18945	3316	998.69453384
US Airways Inc.	118261	5.70937277	14615	3399	633.36694261
Northwest Airlines Inc.	69196	5.09826659	7799	2795	711.79991618
ExpressJet Airlines Inc.	62443	9.23877201	10478	1834	469.71897571
...	...	...	...	...	...

Airline Insights

Using the results from the demonstration table, we can create clear and actionable business intelligence visuals. Below are two examples of how this data can be effectively visualised.



Recommended Next Steps

01

Dashboards

Integrate the data model with Tableau to begin building dashboards and reports tailored to business needs.

02

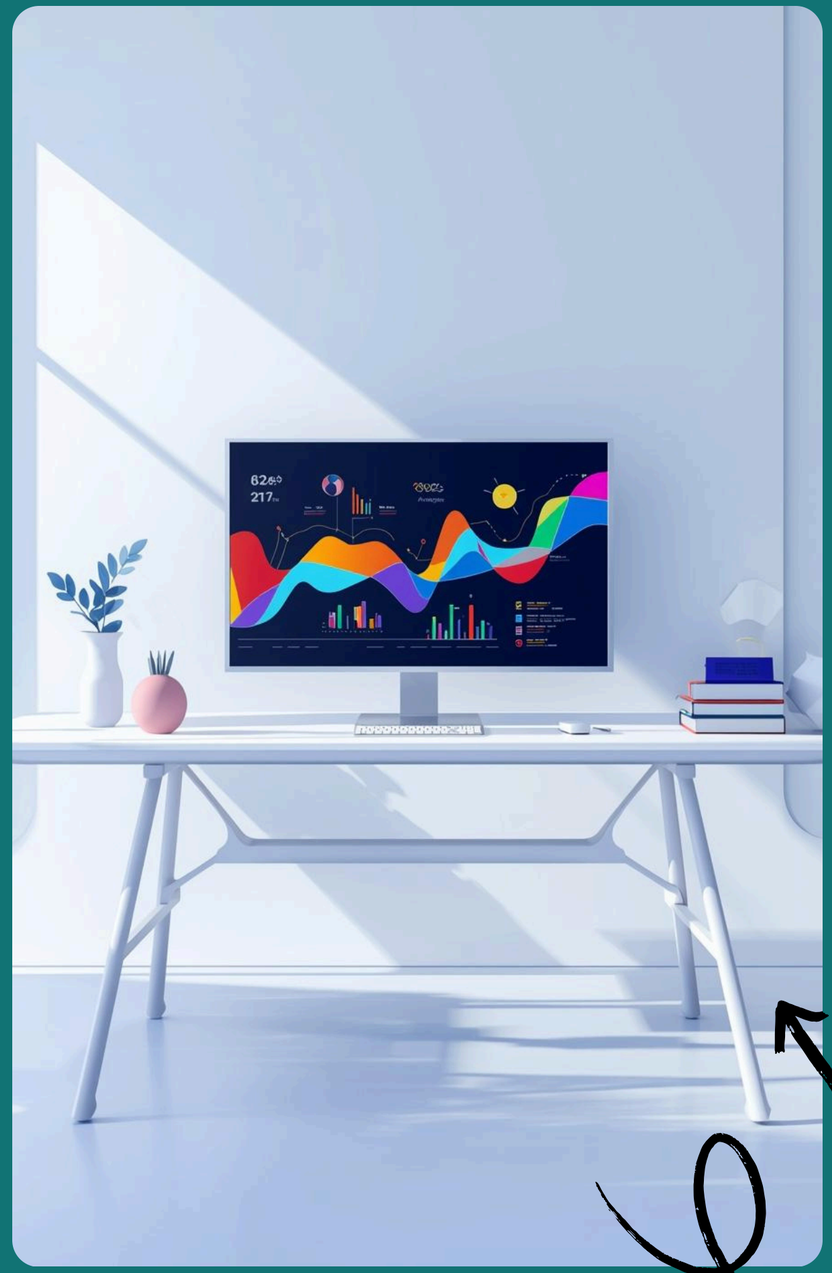
New Dimensions

Consider adding additional dimensions (e.g., DIM_PLANE, DIM_CREW) to further enhance analytical capabilities.

03

Documentation Training

Training for end users using the project documentation for future data model development.



Conclusion

Thank you for your attention!