OPTIMIZING FLIGHT DATA FOR ANALYTICAL INSIGHTS

(DATA MODELING FOR TABLEAU ANALYSIS)

By Cathrine Julie A Xavier

Business Context

- Flight data is large, complex, and unstructured.
- Challenge lies in data inconsistency, schema mismatch and data type discrepancies

Aim of Analysis

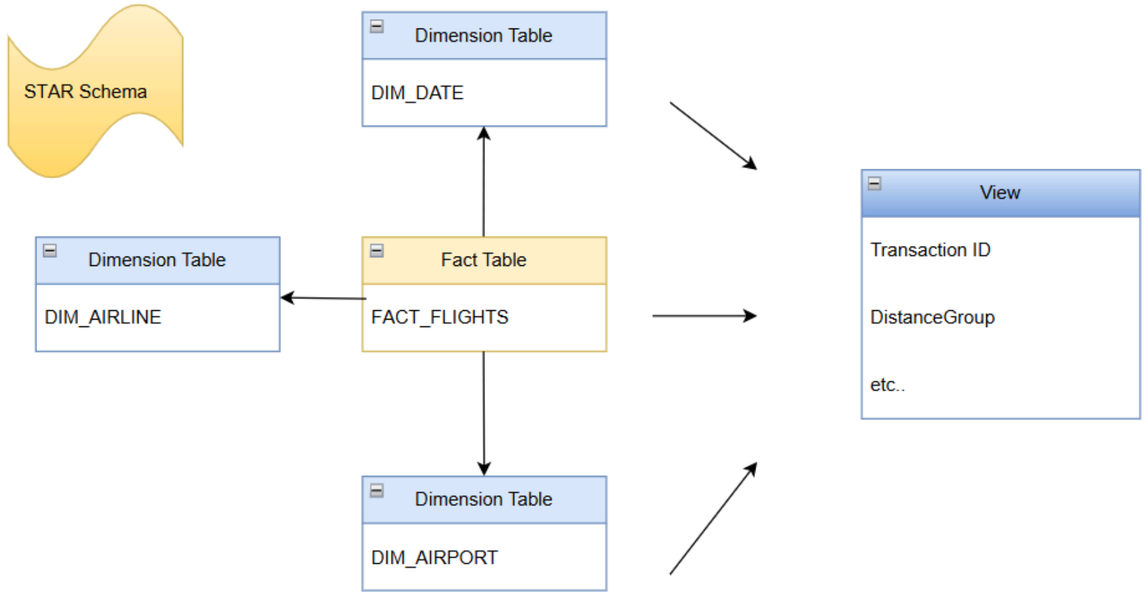
- Business-ready data to quickly identify trends like airline performance, recurring delays and seasonal variations in flight operations.

Source: Raw data from S3 (external stage)

Data Structuring & Normalization

Approach Taken: *Created Fact & Dimension tables in star schema format.*

- FACT_FLIGHTS stores transactional flight data.
- DIM_AIRLINE, DIM_AIRPORT, and DIM_DATE provide unique, contextual details, reducing redundancy.
- VW_FLIGHTS combines relevant data from Fact and Dimension tables into a single, analytics-ready structure.



Data Quality Checks & Preprocessing for Performance Optimization

Key Data Transformations

Transformation	*Before*	*After*
Date Format Fix	YYYYMMDD	YYYY-MM-DD
Data Cleaning	Incorrect naming conventions,entries and scattered data	DISTANCEGROUP categorizes flights into predefined ranges (0-100 miles, 101-200 miles, etc AIRLINENAME and AIRPORTNAME are cleaned for uniformity
Airline Name Cleanup	XYZ Airlines: Regional	XYZ Airlines
Columns	⛔ Not available	✅ Available for analysis New Columns (DISTANCEGROUP, DEPDELAYGT15, NEXTDAYARR) DEPDELAYGT15: Flags flights delayed by more than 15 minutes. NEXTDAYARR: Marks flights that land past midnight

Findings:

Feature Engineering:

- DEPDELAYGT15: Flags flights delayed by more than 15 minutes.
- NEXTDAYARR: Marks flights that land past midnight.
- DISTANCEGROUP: Groups flights by distance range.

Why It Matters:

- Preprocessed data reduces Tableau's computation load.
- Ready-to-use fields like DEPDELAYGT15, DISTANCEGROUP enable quick delay analysis.
- DIM_DATE supports trends like yearly/monthly flight patterns.
- Focus on specific KPIs such as flight delays, on-time performance etc

Curated Schema for Tableau

Business-Focused Insights

1. **On-time Performance** → Identifies delays (DEPDELAY, ARRDELAY), helping airlines improve schedules.
2. **Distance-Based Analysis** → Groups flights into distance categories, useful for pricing and operations.
3. **Peak Travel Trends** → FLIGHTDATE helps track seasonal or daily trends.
4. **Airport Performance** → Shows which airports experience more delays.

Why It Matters:

This data allows analysts to make smarter decisions. They can:

- Identify correlation between airline delays and specific airports
- See which airlines are the busiest
- Plan better routes by understanding flight distances

Before vs. After snapshots of raw vs. cleaned data

Before

	$1
1	TRANSACTIONID|FLIGHTDATE|AIRLINECODE|AIRLINENAME|TAILNUM|FLIGHTNUM|ORIGINAIRPORTCODE|ORIGAIRPORTNAME|ORIGINCITYNAME|ORIGINSTATE|ORIGINSTA
2	54548800|20020101|WN|Southwest Airlines Co.: WN|N103@@|1425|ABQ|AlbuquerqueNM: Albuquerque International Sunport|Albuquerque|NM|New Mexico|DAL|DallasT
3	55872300|20020101|CO|Continental Air Lines Inc.: CO|N83872|150|ABQ|AlbuquerqueNM: Albuquerque International Sunport|Albuquerque|NM|New Mexico|IAH|HoustonT
4	54388800|20020101|WN|Southwest Airlines Co.: WN|N334@@|249|ABQ|AlbuquerqueNM: Albuquerque International Sunport|Albuquerque|NM|New Mexico|MCI|Kansas C
5	54486500|20020101|WN|Southwest Airlines Co.: WN|N699@@|902|ABQ|AlbuquerqueNM: Albuquerque International Sunport|Albuquerque|NM|New Mexico|LAS|Las Vega

After

	TRANSACTION	DISTANCEGROU	DEPDELAYGT	NEXTDAYA	AIRLINENAME	ORIGINAIRPORTNAME	DESTAIRPORTNAME	FLIGHTDA	DEPTIM	ARRTIM	DEPDELA	ARRDELA	DISTANC
1	66326800	1001-1100 miles	FALSE	FALSE	Continental Air Lines Inc.	Cleveland-Hopkins International	Southwest Florida International	2004-01-24	00:28:50	00:33:33	-5	-7	1025 miles
2	66311600	1801-1900 miles	FALSE	FALSE	Continental Air Lines Inc.	Cleveland-Hopkins International	Luis Munoz Marin International	2004-01-24	00:15:20	00:23:43	-5	-17	1839 miles
3	68114500	101-200 miles	FALSE	FALSE	ExpressJet Airlines Inc. (1)	Cleveland-Hopkins International	Buffalo Niagara International	2004-01-25	00:25:16	00:26:50	-4	-12	192 miles
4	66328100	1001-1100 miles	FALSE	FALSE	Continental Air Lines Inc.	Cleveland-Hopkins International	George Bush Intercontinental/Houston	2004-01-25	00:24:08	00:28:24	2	14	1091 miles
5	66352700	2001-2100 miles	FALSE	FALSE	Continental Air Lines Inc.	Cleveland-Hopkins International	Los Angeles International	2004-01-25	00:28:25	00:30:52	-5	-21	2053 miles

THANK YOU