AeroInsight Data Warehouse

Group Members

Shahzaib Ahmed CT-22042

Kumel Ahmed CT-22034

Muhammad Azlan Asim CT-22036

Wahaj Ahmed CT-22032

SUBMITTED TO

Ms. Mehar Fatima

Introduction

The AeroInsight Data Warehouse (DW) project aims to design and implement a comprehensive analytics platform for aviation-related data. By integrating information from flight records, airports, airlines, passenger satisfaction surveys, and operational performance metrics, the project establishes a unified environment for reporting and business intelligence. The goal is to transform raw aviation datasets into meaningful insights through systematic ETL processes, star-schema modeling, and visualization techniques.

This project enables organizations to perform accurate trend analysis, evaluate performance indicators, monitor operational efficiency, and support strategic decision-making. Through a combination of Power Query transformations, DAX-driven calculations, and Power BI dashboards, the system provides a scalable foundation for both current reporting needs and future analytical extensions such as machine learning and predictive analytics.

Project Background

The aviation industry in Pakistan is rapidly expanding, with increasing passenger volumes, growing airline operations, and rising expectations for service quality and operational performance. Every day, airlines, airports, and regulatory bodies generate large volumes of data. Despite the availability of such data, it typically resides in scattered sources, including CSV files, transactional systems, manual logs, and operational databases.

Because these datasets are fragmented and inconsistent in structure, organizations face significant challenges in conducting holistic analysis. Critical decision-making processes such as identifying delay causes, analyzing passenger sentiment, benchmarking airline performance, etc become difficult without a unified and standardized data framework.

OBJECTIVES

- Design and implement a star schema data warehouse for aviation analytics
- Develop an automated ETL pipeline for data extraction, transformation, and loading
- Ensure data quality and integrity through comprehensive validation
- Implementation of data quality and consistency by cleansing (duplicate removal, null handling, value normalization) plus referential integrity enforcement.
- BI Enablement: Stable KPIs (on-time %, satisfaction indices, delay minutes, distance metrics) exposed through dashboards.

Business Problem Addressed

Traveler Perspective

Travelers often struggle to choose the best airline and airport experience while balancing cost and reliability. Our platform solves this by:

- Review Visibility: Letting travelers search and compare airline ratings and reviews shared by other passengers
- Airport Comparison: Showing which airports in Pakistan offer the highest satisfaction levels based on real traveler feedback
- Fare Awareness: Providing insight into average fares across airlines and aircraft types to help find better value deals
- Delay Transparency: Displaying daily flight delays and their causes for each airline and city, enabling smarter and more reliable travel planning

Airline & Corporate Perspective

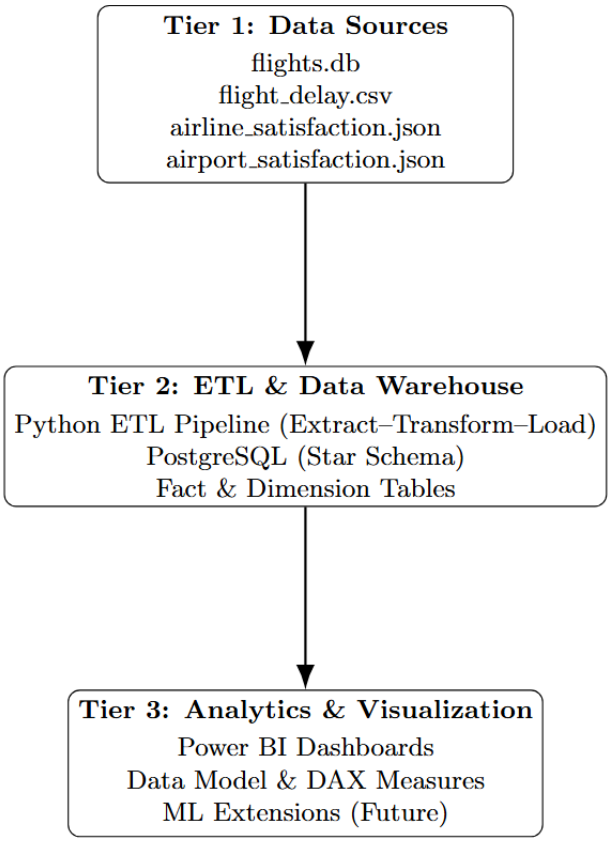
Airlines and aviation stakeholders need competitive intelligence on customer perception, pricing, and operational performance to stay ahead. Our platform solves this by:

- Competitive Benchmarking: Allowing airlines to analyze how their reviews and ratings compare to rival carriers in the same routes and markets
- Airport Performance Insights: Highlighting airport satisfaction rankings to guide route planning, partnerships, and service improvements
- Pricing Strategy Support: Providing average fare analytics by airline and aircraft type to refine pricing strategies and monitor market positioning
- Operational Performance Monitoring: Exposing detailed delay patterns and root causes by airline, date, city, and route, helping identify weaknesses and prioritize reliability initiatives

System Architecture

The system architecture follows a structured, multi-tier framework designed to handle aviation data from extraction to visualization. Our project uses a clear separation of layers so that each component which are data sources, ETL pipeline, database, and analytics tools works independently but remains fully integrated as part of the overall workflow.

The architecture is based on a three-tier design, where Tier 1 manages raw datasets, Tier 2 performs ETL processing and stores the transformed data in a PostgreSQL data warehouse, and Tier 3 handles analytics, reporting, and future machine-learning extensions. This layered approach ensures clean data flow, consistent transformations, efficient querying, and scalable performance for aviation-related analytics.



Tier 1: Data Sources

- flights.db : Detailed flight instances (times, distances, fares, capacity, status) extracted from.
- flight_delay.csv : Aggregated delay metrics by airline/date/airport (totals and cause-specific breakdowns).
- airline_satisfaction.json – Passenger airline service survey responses (ratings, delay experience, distance, age, class).
- airport_satisfaction.json– Passenger airport facility survey responses (service ratings, age).

Source Type	Implementation	Data Volume
OLTP System 1	flights.db relational database containing detailed flight instances including times, distances, fares, capacity, and status	Thousands of flight records across multiple routes and dates
OLTP System 2	Aggregated delay metrics loaded into flight_delay.csv with airline, date, and airport-level totals and cause-specific breakdowns	Daily delay metrics for worldwide airlines and Pakistani International airports
API Data	Passenger airline and airport service survey responses collected as JSON (airline_satisfaction.json & airport_satisfaction.json) including ratings, delay experience, distance, age, and class	Continuously growing review corpus retrieved on demand
Flat Files	Additional CSV extracts or Kaggle-based datasets if any used for benchmarking	Operational records used for modeling and benchmarking

Tier 2: ETL & Data Warehouse

ETL Pipeline:

1. EXTRACT:

- Read source data from different sources
- Validate file existence and accessibility
- Log extraction statistics such record counts, column counts

2. TRANSFORM:

- Data type standardization (dates, times, numeric fields)
- String cleaning and normalization (case, whitespace, label fixes)
- Column splitting and feature extraction (date/time parts, route keys)
- Derived metrics calculation (departure delay, aggregated delay minutes)
- Missing value handling (dropping invalid rows, reasonable imputations)
- Duplicate record detection and removal
- Surrogate key mapping (natural keys → dimension IDs)
- Outlier detection and handling (impossible or extreme values)
- Reshaping into fact and dimension tables for the Star Schema

3. LOAD:

- Bulk insert using psycopg2
- Batch processing
- Transaction management for atomicity
- Index creation post-load for performance

Data Warehouse:

Table: dim_airport

Column Name	Data Type
airport_id	character varying
airport_name	character varying
city	character varying
country	character varying

Table: dim_route

Column Name	Data Type
route_id	integer
depart_airport	character varying
arrival_airport	character varying
route_key	character varying

Table: dim_date

Column Name	Data Type
date_id	integer
date	character varying
day	integer
day_of_week	character varying
month	integer
quarter	integer
year	integer

Table: dim_aircraft_model

Column Name	Data Type
aircraft_id	integer
aircraft_model_name	character varying
aircraft_company_name	character varying

Table: fact_flight

Column Name	Data Type
flight_id	integer
date_id	integer
route_id	integer
airline_id	integer
aircraft_id	integer
flight_number	character varying
scheduled_time	character varying
departure_time	character varying
status	character varying
distance_km	integer
duration_minutes	integer
seat_capacity	integer
fare_pkr	numeric

Table: fact_airport_satisfaction

Column Name	Data Type
airport_satisfaction_id	integer
airport_id	character varying
customer_age	integer
checkin_experience	numeric
cleanliness	numeric
shopping_options	numeric
food_options	numeric
wifi_quality	numeric
seating_availability	numeric
gate_location	numeric
satisfaction_rating	numeric

Table: dim_airline

Column Name	Data Type
airline_id	integer
airline_name	character varying
airline_country	character varying

Table: dim_class

Column Name	Data Type
class_id	integer
class_type	character varying

Table: fact_delay

Column Name	Data Type
delay_id	integer
date_id	integer
airline_id	integer
airport_id	character varying
total_flights	integer
total_departed	integer
total_cancelled	integer
total_delay_minutes	numeric
delay	character varying
weather_delay	numeric
nas_delay	numeric
late_aircraft_delay	numeric

Table: fact_airline_satisfaction

Column Name	Data Type
airline_satisfaction_id	integer
airline_id	integer
class_id	integer
customer_age	integer
flight_distance	integer
departure_delay	integer
inflight_wifi_service	numeric
food_and_drink	numeric
seat_comfort	numeric
inflight_entertainment	numeric
leg_room	numeric
inflight_service	numeric
cleanliness	numeric
satisfaction_rating	numeric

Tier 3: Analytics & Visualization

KPI:

- On-Time Performance (%)
- Average Delay per Flight (minutes)
- Average Airline Satisfaction Score
- Average Airport Satisfaction Score
- Average Fare for selective routes
- Airline Competitive Index
- Cancellation Rates
- Satisfaction Dimension
- Service Dimension Scores for Airlines
- and Airports

Data Quality and Validation

Quality Dimensions & Checks Implemented

- Completeness Validation
- Required field checks for core attributes (airline, airport, route, date, flight status, key satisfaction ratings).
- Cross-table completeness to ensure flights, delays, and surveys are all successfully linked to their dimensions.
- Coverage checks to confirm that all major Pakistani airlines, airports, and high-traffic routes are represented in the Star Schema.

Accuracy & Realness

- Value range validation (ratings restricted to 0–10, ages within realistic human bounds, distances and fares strictly > 0, delay minutes ≥ 0).
- Data type enforcement for dates, times, numeric measures, and categorical codes to avoid silent parsing errors.
- Business rule validation (e.g., departure airport ≠ arrival airport, cancelled flights not having valid actual departure times, distances consistent with route definitions).
- Realism checks by comparing distributions and typical ranges (e.g., average delays, fare levels, satisfaction scores) with publicly available aviation statistics and online references, ensuring the data "looks like" real-world behavior.

Consistency & Integrity

- Foreign key integrity between all fact tables and their dimensions to prevent orphan records.
- Temporal consistency checks so that flight, delay, and survey dates align with dim_date and obey logical time ordering (e.g., survey date not before flight date).
- Aggregate reconciliation: verifying that daily delay totals, flight counts, and summarized KPIs roll up correctly across facts and match expectations derived from source files.

Conformity & Standardization

- Standardization of airline names, airport codes (IATA), and travel classes to a single canonical representation.
- Harmonized rating scales and units (e.g., minutes for delays, kilometers for distance, PKR for fares) across all tables.
- Enforcement of consistent route encoding (route_key such as KHI_LHE) and reuse of conformed dimensions for airlines, airports, routes, dates, and classes.

Data Quality and Validation

Reliability & Robustness

- Duplicate detection and removal based on natural keys (flight number + date, survey identifiers) to avoid double counting.
- Outlier detection and handling for extreme or impossible values, with suspicious records flagged or quarantined instead of silently accepted.
- Batch-level audit logging (row counts, rejection counts, source hashes) to support repeatable, trustworthy ETL runs and traceability of issues over time.

Expected Benefits

Enhanced Decision-Making

- Data-driven insights into airline performance, airport quality, fares, and delays.
- Proactive strategies for improving traveler experience and operational reliability.
- Better route and pricing decisions supported by consolidated KPIs and benchmarks.
- Clear visibility into which airlines and airports perform best for specific routes and traveler segments.

Improved Reporting Efficiency

- Query performance optimized through a Star Schema with conformed dimensions.
- Automated data preparation and refresh, significantly reducing manual Excel-based reporting.
- Self-service analytics in Power BI for both travelers (comparisons) and airline stakeholders (monitoring).
- Standardized reports and dashboards that can be reused across multiple stakeholders and use cases.

Data Consistency & Transparency

- Single, trusted source of truth for flights, delays, reviews, and satisfaction scores.
- Standardized definitions of metrics (on-time %, average delay, satisfaction indices, average fare).
- Clear lineage from raw files/APIs to facts and dimensions, with validation and audit logs.
- Easier communication between technical and business teams due to shared, documented data model.

Scalable & Analytics-Ready Architecture

- Schema and pipelines designed to incorporate new airlines, airports, and routes with minimal redesign.
- Feature-rich, ML-ready data layer supporting models for delay prediction, satisfaction scoring, and segmentation.
- Modular ETL that allows new data sources (e.g., weather, fuel prices, additional reviews) to be plugged in.

- Ability to migrate the same Star Schema to more powerful warehouse or cloud platforms in the future.

10.5 Value for Travelers & Airlines

- Travelers can quickly identify the best airports, airlines, and fare options based on real data rather than guesswork.
- Airlines gain competitive intelligence on rivals' performance and customer perception, supporting targeted improvements.

Both sides benefit from increased transparency, which encourages better service, more reliable operations, and fairer pricing

Conclusion

Our Project, The AeroInsight Data Warehouse project successfully demonstrates the complete lifecycle of a modern analytical platform, starting from fragmented aviation data and ending with decision-ready insights for both travelers and industry stakeholders. By integrating flight operations, delay statistics, fare information, and passenger satisfaction surveys into a unified star schema, the project overcomes the limitations of scattered CSV files and ad-hoc reports, providing a single, consistent source of truth for aviation analytics in the Pakistani context.

Through a carefully designed ETL pipeline, raw data is cleansed, standardized, and enriched before being loaded into a PostgreSQL-based warehouse. Data quality measures—such as duplicate removal, domain checks, referential integrity enforcement, and basic profiling—ensure that the resulting facts and dimensions are reliable enough to support trend analysis, benchmarking, and performance monitoring. The three-tier architecture (data sources, ETL & warehouse, analytics & visualization) keeps responsibilities clearly separated while maintaining a smooth end-to-end data flow.

On top of this foundation, Power BI dashboards and analytical queries translate technical structures into meaningful KPIs: on-time performance, delay minutes, cancellation rates, satisfaction indices, and route-level pricing metrics. Travelers gain transparent comparisons of airlines and airports, while airlines and airports gain competitive benchmarking, operational performance diagnostics, and pricing intelligence. Even within a lab-scale setup, AeroInsight DW shows how a well-modeled data warehouse can directly support real-world decision-making.