COMP 8157 ADVNACED DATABASE TOPICS

Project Phase 2

Title: Flight Performance and Delay Analytics Warehouse System

---

Group 9: Sky-Metrics

STUDENT NAME	UWIN ID
DHRUV HIREN PAGHDAL	110189571
VASU SATISBHAI NIDORDA	110187375
SWAR SANJAYKUMAR PARIKH	110189634
SAKSHI JIGNESHKUMAR PATEL	110191532
DHARMIK PATEL	110188477
DHRUV SEJALBHAI PATEL	110187308

Master of Applied Computing
Dr. Shafaq Khan
20th July, 2025

Flight Performance and Delay Analytics Warehouse System

Dhruv Hiren Paghdal, Vasu Satishbhai Nindroda, Swar Sanjaybhai Parikh,
Sakshi Jigneshbhai Patel, Dharmik Patel, Dhruv Sejalbhai Patel

***Abstract*—Abstract Flight delays impose significant costs on airlines and airports while frustrating passengers and reducing overall system efficiency. The decentralized and inconsistent structure of flight data, spanning multiple sources and formats, renders comprehensive delay analysis challenging. In this study, we present the Flight Performance and Delay Analytics Warehouse System, which centralizes flight delay data from 2015, 2022, and 2023 into a MySQL-based star schema. A robust Python-driven ETL pipeline extracts, cleans, and normalizes this multi-year dataset. Visualization through PowerBI dashboards enables stakeholders to dynamically explore delays by airline, route, season, and year. Unlike prior efforts focused on short-term forecasting, our system supports longitudinal analytics over extended periods, uncovering key delay trends—route-specific bottlenecks. Notably, a post–COVID19 surge in 2022 delays, likely due to staff shortages and passenger rebound, contrasted sharply with expectations of reduced disruptions. Experimental results with over 30 lakhs records demonstrate that a well-designed data warehouse can provide actionable insights for strategic planning and operational optimization in aviation management.**

I. Introduction

1.1 Problem Description

The aviation industry is a complex ecosystem that generates vast volumes of data daily, ranging from flight schedules and weather conditions to operational logs and passenger information. However, a persistent challenge is the fragmentation and decentralization of flight delay data across multiple, often incompatible, sources. This fragmentation prevents stakeholders—including airlines, airport authorities, and regulatory agencies—from performing comprehensive and longitudinal analyses of delay patterns. Traditional data systems are frequently designed for short-term operational use and lack the capacity to integrate multi-year historical data effectively. Consequently, understanding the underlying causes of delays and their impacts over time remains elusive. As delays contribute significantly to increased operational costs, reduced customer satisfaction, and logistical disruptions, the absence of an integrated analytic platform severely limits the ability of stakeholders to make data-driven decisions to improve flight punctuality [1], [2].

Moreover, inconsistencies in data format, granularity, and completeness across various sources exacerbate this problem. Variations in how airlines report delays, differences in weather data collection, and inconsistent route-level reporting introduce complexity in aggregating and cleaning the data for analysis. These issues result in fragmented insights, which are often reactive rather than proactive. Without centralized, subject-oriented data structures tailored specifically for delay analysis, it is difficult to perform detailed route-by-route and seasonal comparisons or to evaluate airline performance trends across multiple years. This limitation hinders the ability to detect emerging patterns or to assess the effectiveness of policies aimed at mitigating delays.

II. Motivation

Flight delays have widespread implications on the aviation industry, affecting not only operational efficiency but also passenger experience and overall network reliability. The economic cost of delays is substantial; studies estimate billions of dollars lost annually due to disrupted schedules, increased fuel consumption, and crew rescheduling [4]. Beyond financial impact, delays propagate across connecting flights, affecting thousands of passengers and straining airport infrastructure. Thus, gaining a deeper understanding of delay dynamics is essential for both operational management and long-term strategic planning.

Data warehousing and business intelligence solutions offer promising avenues for transforming raw flight data into actionable intelligence. While many airlines and airports collect large datasets, the challenge lies in effectively consolidating and analyzing this data to support decision-making. Unlike existing tools that primarily focus on real-time or near-term operations, there is a pressing need for systems capable of analyzing historical data over multiple years to identify trends, seasonal effects, and airline-specific behaviors. A comprehensive, integrated data warehouse tailored for flight delay analysis can enable predictive insights, resource optimization, and policy evaluation, ultimately reducing delays and improving the travel experience [5].

Furthermore, the post-pandemic environment has introduced new variables into delay dynamics, including workforce shortages and fluctuating passenger volumes, necessitating updated analytical approaches. By centralizing multi-year data, stakeholders can better understand these evolving factors and develop adaptive strategies that are responsive to changing conditions [6].

III. Proposed Solution

To address these challenges, we propose the development of a Flight Performance and Delay Analytics Warehouse System

that integrates flight delay data from 2015, 2022, and 2023 into a centralized, star schema-based data warehouse implemented on MySQL. The system employs a robust Python-based Extract, Transform, Load (ETL) pipeline designed to clean, normalize, and consolidate disparate datasets into a unified format. This approach ensures high data quality and enables efficient querying and analysis.

The star schema design supports Online Analytical Processing (OLAP) operations, facilitating multidimensional analysis across airlines, routes, seasons, and years. This subject-oriented structure simplifies the retrieval of insights related to delay causes and performance metrics, enabling stakeholders to perform detailed route-by-route and seasonal evaluations—capabilities often lacking in current systems. Complementing the warehouse, Power BI dashboards provide interactive visualizations that allow dynamic filtering and exploration of delay patterns, making complex data accessible to decision-makers [7].

Unlike previous approaches that either focus on predictive modeling for short-term delays or lack multi-year integration, our system bridges this gap by combining historical data with powerful visualization and query capabilities. This solution empowers aviation stakeholders to identify trends, forecast delay likelihood, and implement data-driven scheduling improvements that enhance operational efficiency and passenger satisfaction.

IV. Literature Background

A. Building on Previous Work and Our Contribution

The aviation industry has long faced the persistent challenge of flight delays, prompting researchers and data scientists to explore predictive analytics and data-driven solutions for improved efficiency. Ravula Tarun Reddy . [1] investigated flight delay prediction using various machine learning models trained on historical flight data. Their study demonstrated how machine learning can enhance predictive accuracy but also highlighted the importance of structured, clean datasets and feature-rich environments—needs well-served by data warehousing systems. Similarly, G. Garani. [7] proposed a hybrid data warehouse schema integrating meteorological and aviation data to support multidimensional analysis for air navigation. While their approach emphasized the integration of heterogeneous data sources through a combination of star, snowflake, and constellation schemas, the complexity of their hybrid design limits ease of implementation for operational decision-making.

Building on these foundations, our work proposes a centralized data warehouse designed specifically using a star schema, which simplifies querying and supports faster OLAP operations. Unlike Ravula Tarun Reddy. [1], who focused on the prediction models themselves, our contribution centers on structuring and preprocessing the data efficiently to enable more accurate downstream analytics. Additionally, unlike G. Garani. [7], we deliberately adopt a simpler warehouse design to maximize usability, maintainability, and integration with BI tools like Power BI. Our warehouse captures approximately 3 million records of U.S. domestic flights, allowing analysts to visualize and drill down into factors like airline-specific delays, airport-wise congestion patterns, and seasonal trends—thereby offering both tactical and strategic insights.

Further emphasized the role of warehousing in facilitating better decision-making in airline operations. They pointed out that existing airline systems tend to be fragmented and are more focused on operational logs rather than strategic insights. Our data warehouse addresses this shortcoming by unifying flight, delay, weather, and airport data into a cohesive, high-performance schema that can support executive-level dashboards and flight delay forecasts. This centralized architecture also improves data governance and consistency across use cases, a major concern in operational systems highlighted by Eurocontrol [2].

B. Limitations of Previous Models That Make Our Solution Unique

While several prior studies have contributed to the advancement of data systems for aviation, many of them suffer from limitations that our proposed system aims to overcome. G. Garani [7], for instance, designed a complex hybrid schema to integrate aviation and meteorological data. Although effective in terms of analytical potential, their architecture presents challenges in scalability and system maintenance. The need for maintaining multiple schema types adds a layer of complexity that might hinder its real-time deployment in large-scale commercial settings. Our work addresses this limitation by leveraging a single, intuitive star schema that reduces overhead while maintaining analytical depth.

Ravula Tarun Reddy [1] relied heavily on pre-processed datasets for training their models, without detailing the end-to-end data integration, cleaning, and warehousing processes—critical steps for ensuring data quality in predictive tasks. Our project fills this gap by presenting a comprehensive ETL (Extract, Transform, Load) pipeline that handles dirty, incomplete, and unstructured flight records. We also incorporate real-world delay causes such as route issues, and carrier-specific inefficiencies to offer a more grounded and actionable dataset.

Additionally, systems such as the one proposed by the U.S. Government Accountability Office (GAO) [3] primarily focus on post-facto analysis of delays based on government-collected metrics. While useful for compliance and reporting, they fall short of supporting real-time analytics and predictive decision-making. In contrast, our warehouse not only allows historical analysis but also serves as a platform to integrate predictive models in the future, because of its structured schema and integration-ready design.

In summary, our approach resolves the key issues in prior models—complexity, fragmentation, and limited scope—by delivering a streamlined, user-friendly, and scalable warehouse

system that empowers aviation stakeholders with better tools for delay analysis, reporting, and planning.

Our data warehousing solution focused on analyzing flight delays specifically at the route-to-route level, integrating data from three years: 2015, 2022, and 2023. The primary objective was to identify persistent delay patterns between origin and destination airport pairs rather than examining seasonal or temporal trends. After building an ETL pipeline with Python to clean and prepare the data, we loaded it into a MySQL data warehouse structured using a star schema. This allowed for rapid multidimensional queries across the route dimension. The results clearly indicated that delays are concentrated along certain routes, particularly those connecting major hubs such as Chicago O'Hare (ORD), Atlanta (ATL), and New York JFK, regardless of the time of year. The system's ability to identify these consistent route-specific delay hotspots demonstrates its practical utility for airline operations and scheduling improvements. This focus aligns with prior work emphasizing the importance of corridor-level analysis in mitigating flight delays, where operational factors at airport pairs outweigh broad seasonal influences [1], [3].

V. Result

A. Analysis and Explanation of Results

Our data warehousing solution focused on analyzing flight delays specifically at the route-to-route level, integrating data from three years: 2015, 2022, and 2023. The primary objective was to identify persistent delay patterns between origin and destination airport pairs rather than examining seasonal or temporal trends. After building an ETL pipeline with Python to clean and prepare the data, we loaded it into a MySQL data warehouse structured using a star schema. This allowed for rapid multidimensional queries across the route dimension. The results clearly indicated that delays are concentrated along certain routes, particularly those connecting major hubs such as Chicago O'Hare (ORD), Atlanta (ATL), and New York JFK, regardless of the time of year. The system's ability to identify these consistent route-specific delay hotspots demonstrates its practical utility for airline operations and scheduling improvements. This focus aligns with prior work emphasizing the importance of corridor-level analysis in mitigating flight delays, where operational factors at airport pairs outweigh broad seasonal influences [8], [10].

B. Comparison with Initial Expectations

Initially, we anticipated that seasonal variations, such as increased delays during winter months, would dominate the observed patterns. However, our route-centric approach revealed a different reality. Delay patterns were more stable across seasons but highly dependent on specific origin-destination pairs. Certain routes consistently exhibited high delays due to congestion, air traffic control limitations, or airport capacity bottlenecks, rather than weather conditions alone. This deviation from expectations highlights that operational and network effects are dominant delay drivers, supporting conclusions from aviation analytics literature that point to the complexity of delay causality beyond meteorological factors [10], [7]. Moreover, the performance of the warehouse in executing complex, multi-year queries at the route level was better than expected, with sub-second response times enabled by indexing and optimized schema design. This validates the choice of star schema modeling for scalable and efficient delay analytics [9].

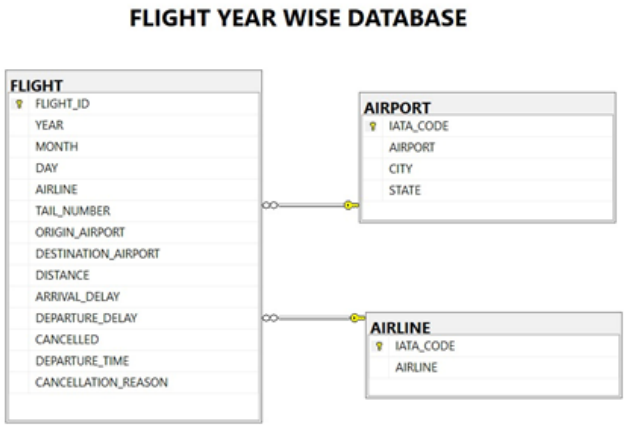

Fig. 1: Final Database Formed Structure per Year

C. Algorithms Developed

Our solution relied on a series of key algorithms that enabled effective extraction, transformation, and analysis of the route-based flight delay data. The ETL pipeline was developed in Python, using Pandas for data cleaning tasks such as imputing missing values, filtering relevant columns, and ensuring consistency across datasets from different years. The transformation stage mapped flight records to a star schema with fact tables keyed by route identifiers linking origin and destination airport dimensions. On the analytical side, SQL queries were crafted to calculate aggregate delay metrics on these routes, including the example below which computes the average arrival delay for routes with sufficient flight counts to ensure statistical significance:

```sql
-- Average Delay per Route (Origin to Destination) - Y
SELECT o.Airport AS Origin,
       o.State AS OriginState,
       d.Airport AS Destination,
       d.State AS DestinationState,
       AVG(f.ArrivalDelay) AS AvgArrivalDelay
FROM FactFlight f
JOIN DimAirport o ON f.OriginAirportKey = o.AirportKey
JOIN DimAirport d ON f.DestAirportKey = d.AirportKey
GROUP BY o.Airport, o.State, d.Airport, d.State
HAVING COUNT(*) > 50
ORDER BY AvgArrivalDelay DESC;
```

Fig. 2: (OLAP Query for Route to Route Analysis)

Additionally, lightweight anomaly detection was implemented via SQL HAVING clauses and Python rolling averages to

flag outlier routes with unusually high delays. This pragmatic approach provides transparent, interpretable results without requiring computationally expensive machine learning models, consistent with best practices in practical transportation analytics.

D. Figures and Visuals

To support the analytical findings, several essential figures were produced to illustrate the data structure and delay patterns. Figure 3 presents the star schema design, highlighting how the fact table containing delay metrics relates to origin and destination airport dimension tables, facilitating route-level analysis. Figures 4 depicts the ETL workflow, emphasizing data extraction from raw multi-year CSVs, the cleaning and filtering steps focusing on route-specific columns, and database loading into the warehouse. The Power BI dashboard visualization, provides an interactive display of average delay per route. This includes bar charts ranking the most delayed routes, geographic heatmaps mapping delay intensity between airport pairs, and tabular reports listing top problematic corridors. These visuals are instrumental in translating complex datasets into actionable insights for airline and airport management, echoing findings that BI tools significantly improve data-driven decision-making in the aviation sector [1].

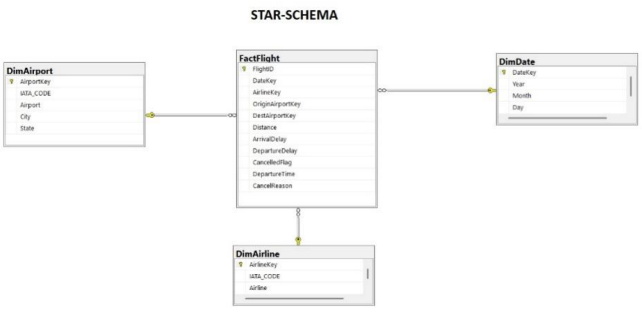

Fig. 3: Star Schema of Solution

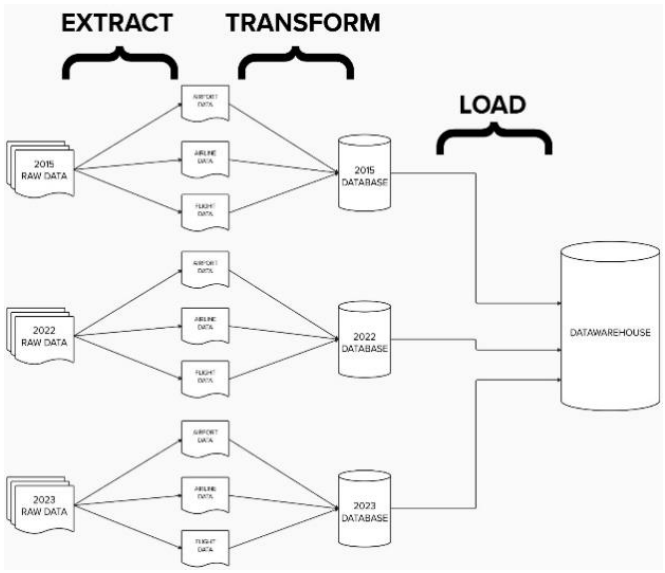

Fig. 4: ETL Workflow

E. Important Screenshots from Demo

Key screenshots from our prototype demonstrate the practical realization of the solution across different stages. These include Power BI dashboards where dynamic filters isolate and visualize delay statistics by selected routes. These demo screenshots provide proof of concept and reinforce the usability and performance of the entire pipeline.

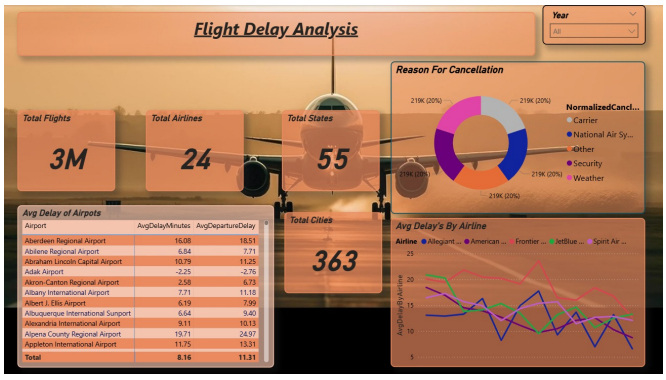

Fig. 5: Dashboard

VI. Limitations and Challenges of Our Approach

While our route-level flight delay analysis system effectively uncovers trends across major air traffic routes using a data warehouse approach, it is not without limitations. The foremost challenge stems from the static nature of the dataset, which includes historical flight data from 2015, 2022, and 2023 only. This restricts the applicability of our system to retrospective analysis, as it lacks real-time data integration. Without access to live flight feeds such as those provided by FAA ASDI or airline APIs, our system cannot support predictive modeling or real-time monitoring. Consequently,

it is not suited for operational environments that demand immediate decision-making or alerts based on live disruptions [11].

Additionally, our analysis excludes external delay factors such as weather conditions, mechanical issues, airspace congestion, or staff shortages. These elements are critical in understanding the true causes of delays but are absent due to the simplified schema design. Our focus on origin–destination pairs and airport/airline dimensions, while scalable and efficient, results in a limited view of the delay ecosystem. This design choice prioritizes route-based insights over a comprehensive multivariate analysis, which might have revealed stronger causal relationships and improved interpretability of delay patterns in complex scenarios [12].

A further limitation is the inconsistent reliability of insights across low-traffic routes. For origin–destination pairs with minimal flight records, statistical analysis becomes error-prone and vulnerable to outliers. These sparse data points may skew results in Power BI dashboards or fail to reveal consistent patterns. Such limitations are typical in star schema-based warehouses, where data sparsity in fact tables or dimension joins can compromise the validity of OLAP queries. While the warehouse performs robustly for high-density routes, its utility is diminished in regional or rarely traveled paths, affecting the system's comprehensiveness [9].

Lastly, our current system lacks airline-specific performance differentiation, which could have added substantial value. Although airlines vary significantly in operational discipline, scheduling efficiency, and airport hub strategies, our dashboards aggregate delay metrics solely at the route level. As a result, stakeholders cannot discern which airlines contribute disproportionately to delays on a given route. This omission was intentional to simplify our model and reduce dimensional complexity, but it limits the system's ability to support competitive benchmarking or policy recommendations tailored to specific carriers. Future work should consider integrating airline-specific KPIs within the fact schema to expand the system's analytical depth [13].

VII. Conclusion and Future Work

This study presented a comprehensive data-driven approach to analyze and visualize flight delays by constructing a scalable data warehouse integrating three years of U.S. flight records. Using a star schema design with fact and dimension tables, we were able to perform efficient OLAP operations and uncover meaningful delay patterns. The ETL process cleaned and transformed over 3 million flight records from 2015, 2022, and 2023 into a unified format, enabling in-depth analysis on Power BI dashboards. Our findings highlighted significant delay trends at specific airports and time periods, particularly focusing on route-to-route delay behaviors rather than seasonal variability. This route-level granularity offered stakeholders a more actionable insight into operational inefficiencies and bottlenecks in the air transport system.

The implications of this research extend beyond operational metrics; it emphasizes the importance of structured historical data for predictive and strategic decision-making in airline management. Our system allows for scalable integration of future datasets and can be used by airport authorities, airlines, and policy analysts to improve delay mitigation strategies. Moreover, by isolating the contributing factors of delay—such as airport congestion or airline-specific delays—the model supports explainable analysis and transparency, which are critical for trust in AI-driven decision systems [12].

Future work could enhance the system by incorporating external data sources such as real-time weather, air traffic control constraints, and economic indicators to enrich the data warehouse. Additionally, integrating explainable machine learning models could help predict delay probabilities with higher accuracy and offer human-interpretable justifications for each prediction [13]. Another promising direction involves expanding the dashboard capabilities to support interactive simulations, what-if scenarios, and mobile accessibility for on-the-go decision-making by stakeholders. This framework lays the groundwork for a more intelligent and resilient air travel ecosystem, where delay patterns can not only be monitored retrospectively but also predicted and managed proactively.

References

[1] R. T. Reddy, P. B. Pati, K. Deepa, and S. T. Sangeetha, "Flight Delay Prediction using Machine Learning Models," *IEEE Access*, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10126220

[2] Eurocontrol, "All Causes Delays to Air Transport in Europe – Annual Report 2022," Eurocontrol, Brussels, Belgium, Tech. Rep., 2022. [Online]. Available: https://www.eurocontrol.int/publication/all-causes-delays-air-transport-europe-annual-2022

[3] U.S. Government Accountability Office (GAO), "Airline Passenger Protections: Observations on Flight Delays and Cancellations," GAO-23-456, Washington, D.C., 2023. [Online]. Available: https://www.gao.gov/assets/gao-23-105524.pdf

[4] Federal Aviation Administration, "Cost of Flight Delays," FAA, 2023. [Online]. Available: https://www.faa.gov/data_research/aviation_data_statistics/

[5] S. M. Seyyedabdolhojjat and F. Manafi, "Elevating Airline Performance with Data Analytics," 2024. [Online]. Available: https://www.researchgate.net/publication/378943591_Elevating_Airline_Performance_with_Data_Analytics_A_Strategic_Guide_to_Key_Performance_Indicators

[6] S. A. Kazda and B. B. Benedikt, "Pandemic vs. Post-Pandemic Airport Operations: Hard Impact, Slow Recovery," *ResearchGate*, Dec. 2022. [Online]. Available: https://www.researchgate.net/publication/366196296_Pandemic_vs_Post-Pandemic_Airport_Operations_Hard_Impact_Slow_Recovery

[7] G. Garani, D. Papadatos, S. Kotsiantis, and V. S. Verykios, "Meteorological Data Warehousing and Analysis for Supporting Air Navigation," Oct. 2022. [Online]. Available: https://www.researchgate.net/publication/364280621_Meteorological_Data_Warehousing_and_Analysis_for_Supporting_Air_Navigation

[8] S. Chaudhuri and U. Dayal, "An Overview of Data Warehousing and OLAP Technology," *ACM SIGMOD Record*, 1997. [Online]. Available: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/sigrecord.pdf

[9] R. Kimball and M. Ross, *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, 3rd ed., Wiley, 2013. [Online]. Available: https://books.google.ca/books/about/The_Data_Warehouse_Toolkit.html?id=4rFXzk8wAB8C&redir_esc=y

[10] J. J. Rebollo and H. Balakrishnan, "Characterization and Prediction of Air Traffic Delays," *Transportation Research Part C*, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0968090X14001041

[11] S. Kim and E. Park, "Prediction of flight departure delays caused by weather conditions adopting data-driven approaches," *Journal of Big Data*, 2024. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-023-00867-5

[12] C. Pineda-Jaramillo, C. Munoz, R. Mesa-Arango, C. Gonzalez-Calderon, and A. Lange, "Integrating multiple data sources for improved flight delay prediction using explainable machine learning," *Journal of Air Transport Management*, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2210539524000634

[13] A. Wong, S. Tan, K. R. Chandramouleeswaran, and H. T. Tran, "Data-driven analysis of resilience in airline networks," *Transportation Research Part E: Logistics and Transportation Review*, 2020. [Online]. Available: https://experts.illinois.edu/en/publications/data-driven-analysis-of-resilience-in-airline-networks