✈️ Airline Business Intelligence Database

Final Project Overview Document

Grace Polito – MSDS Capstone Project

Eastern University • 2025

---

1. Introduction & Problem Context

The Airline Business Intelligence Database is a full-stack analytical ecosystem built to model airline operations, commercial performance, loyalty engagement, and customer behavior. It integrates **real-world aviation data** (OpenFlights + U.S. BTS On-Time Performance) with **synthetically generated flights, bookings, and revenue**, producing a unified, analytics-ready warehouse.

Airlines depend on reliable, scalable BI systems to manage operational disruption, understand customer value, optimize networks, and monitor revenue performance. This project replicates that environment using PostgreSQL, Python, and advanced SQL analytics.

This document summarizes work completed across all six phases of development—from schema design through Python analytics and final presentation deliverables.

---

2. Data Sources & Business Domain

Real Sources

- **OpenFlights Airports & Airlines**
  ~7,700 airports and ~5,700 airlines loaded and normalized

  Capstone Phase 1

- **U.S. BTS On-Time Performance**
  22,595 records of historical delay metrics, standardized to 2024

Capstone Phase 2

Synthetic Sources

Generated using Python (`faker`, NumPy, SQLAlchemy):

- 5,000 flights
- 5,000 passengers
- 3,000 loyalty accounts
- 10,576 miles transactions
- 40,000 bookings
- 40,000 payments

Capstone Phase 2

Synthetic data is statistically realistic but deliberately time-shifted (including **future dates**) to:

- Fill complete annual cycles,
- Avoid missing values,
- Enable smooth visual trends,
- Ensure well-distributed training/analysis windows.

This is an explicit design choice to support reproducible BI workloads.

---

3. Schema Design & ERD (Phase 1)

Phase 1 established the relational backbone of the warehouse.

Capstone Phase 1

Schema Components

- Operational dimensions: airports, airlines, aircraft, routes, flights
- Commercial & customer tables: passengers, loyalty accounts, bookings, payments
- Performance table: BTS on-time metrics
- Full set of:
  - Primary keys
  - Foreign keys
  - NOT NULL, UNIQUE, CHECK constraints
  - ENUM types for controlled domain values

The ERD was exported as:
`docs/ERD_v1.pdf`

---

4. ETL Pipeline & Data Ingestion (Phase 2)

Phase 2 operationalized all real and synthetic data using structured ETL pipelines.

Capstone Phase 2

Key Achievements

- Complete ingestion of OpenFlights + BTS data
- Construction of synthetic fleets, passengers, bookings, and revenue
- Data loaded using Python ETL modules
- Referential integrity fully validated (**0 missing foreign keys**)

Row Count Verification

- Flights: 5,000
- Passengers: 5,000
- Loyalty Accounts: 3,000
- Miles Transactions: 10,576
- Bookings: 40,000
- Payments: 40,000

Visual proof:
`docs/pipeline_row_counts.png`

This phase provided the raw analytical content that later phases consume.

---

5. Data Cleaning, Standardization & Constraints (Phase 3)

Phase 3 transformed raw ingested data into a **clean, standardized, constrained** warehouse.

Capstone Phase 3

Cleaning & Normalization

- Uppercased all IATA/ICAO codes
- Normalized BTS delay fields
- Standardized flight statuses and timestamps
- Cleaned passenger names/emails
- Repaired invalid placeholders (`'', 'N/A', '\N' → NULL`)

Deduplication

- Airports deduplicated on IATA/ICAO
- Airlines deduplicated on IATA
- Booking duplicates removed and constrained
- Guaranteed one payment per booking

Constraints & Indexing

- Added strict PK/FK definitions
- Enforced UNIQUE constraints
- CHECK constraints on:
    - Latitude/longitude
    - Currency values
    - Delay logic
- BI-optimized indexes:
    - flights(flight_date, airline_id)
    - bookings(booking_date)
    - payments(paid_at)

Validation Results

All referential and domain checks passed.
Visual artifact: `docs/pipeline_quality_checks.png`

This phase ensured a **trustworthy analytical foundation** for all subsequent work.

---

6. Analytical SQL Development (Phase 4)

Phase 4 produced the **analytical intelligence layer**—15 advanced SQL queries powering operational, commercial, and loyalty insights.

Capstone Phase 4

Analytical Coverage

- **CTEs:** busiest airports, on-time performance, monthly passengers, fare-class revenue
- **Window Functions:** delay ranking, cumulative revenue, CLV scoring, percent delayed
- **Recursive Queries:** airport connectivity graph
- **Aggregations:** payment success, worst routes, top loyalty members

Performance Evaluation

All queries tested with:

- `EXPLAIN`
- `EXPLAIN ANALYZE`
- Index usage validation
- Observed runtime: *all < 1.2s*

Artifacts

- `docs/phase_4_query_catalog.md`
- `docs/phase_4_notes.md`
- `docs/phase_4_analytics.png`

Phase 4 delivered the **BI-ready SQL layer** later consumed by Python.

---

7. Python Integration & Analytics (Phase 5)

Phase 5 operationalized SQL insights into Python for visual exploration and BI storytelling. The notebook:
`notebooks/04_python_analytics.ipynb`
performs SQL extraction → Pandas modeling → Matplotlib/Plotly visualization.

7.1 Database Integration

- Secure connection via `.env`
- Reusable helpers:
  - `get_engine()`
  - `get_df()`
- Centralized plotting defaults (Airline BI theme)

7.2 Analytical Python Functions

Mirroring Phase 4 SQL:

- Monthly revenue
- Revenue by fare class
- Payment success rate
- Airline delay ranking
- CLV distribution
- Top 10 loyalty customers
- Busiest airports

- Worst routes
- Network connectivity maps

7.3 Visualizations

Saved to `docs/`:

- Monthly Revenue Trend
- Revenue by Fare Class
- Flight Delay Distribution
- Airline Delay Ranking
- Payment Success Rates
- CLV Distribution
- Top 10 Customers
- Airport Network Map
- Busiest Routes Diagram

These visuals feed directly into the final presentation.

7.4 Note on Future-Dated Records

Some synthetic flights, bookings, and revenue records extend into **2025–2026**.
This is intentional and supports:

- Balanced annual seasonality
- Full-year delay & revenue trends
- Avoidance of sparsity in trend graphs
- Consistent business-cycle modeling

The dataset remains structurally valid and analytically consistent.

---

8. Key Business Insights

Across phases 4 and 5, several actionable insights emerged:

Operations

- Delay rates fluctuate seasonally (peaks in March/December).
- Certain routes exhibit extreme delay and cancellation behavior.

Network

- Many busiest airports are remote, low-volume nodes—indicating synthetic distribution vs real-world hubs.
- Network connectivity graphs reveal fragmented route structures.

Commercial

- Revenue is dominated by Basic, Standard, and Flexible fare classes.
- Monthly revenue exhibits stable trends with peaks in mid-year months.

Payments

- Payment success rates are low (~15% across all channels), revealing opportunity for commercial optimization.

Loyalty

- Top 5% of customers contribute disproportionally to CLV.
- Loyalty tiers may not align with earned miles (possible tier inflation).

These insights would support airline decisions regarding schedules, pricing, loyalty, and payment systems.

9. Challenges, Assumptions & Limitations

Challenges

- Integrating real BTS data with synthetic flight schedules
- Maintaining timestamp consistency across synthetic datasets
- Avoiding constraint violations when shifting dates

Assumptions

- Synthetic distributions approximate realistic airline behavior
- Payment logic and delay logic reasonably model commercial complexity

Limitations

- Synthetic data cannot perfectly replicate real operational patterns
- No real revenue or cancellation probability model applied
- Network maps may not reflect real aviation topology

These limitations were mitigated through controlled design choices and documentation.

10. Future Work & Enhancements

Potential expansions:

Modeling

- Advanced delay prediction using gradient boosting
- Market demand forecasting
- Customer churn modeling

Engineering

- Materialized views for BI workloads
- Airflow orchestration for repeatable ETL
- Docker containerization for deployment

Analytics

- Tableau / Power BI dashboards
- Real-time monitoring via streaming ingestion
- Geo-visualization of global flight patterns

---

11. Conclusion

This capstone project demonstrates a complete BI system lifecycle:

- **Schema engineering**
- **ETL pipeline construction**
- **Data cleaning & constraint enforcement**
- **Analytical SQL development**
- **Python analytics & visualization**
- **Final documentation & presentation**

The resulting Airline BI Database is a robust, scalable, analytics-ready environment suitable for operational BI dashboards, commercial insights, and advanced analytics.