Phase 2: ETL Pipeline & Data Population

Phase 2 implemented the end-to-end ingestion pipeline for real and synthetic data. This included downloading, cleaning, transforming, validating, and loading all datasets into the production schema.

---

1. Data Acquisition & Quality Checks

OpenFlights Data

Downloaded from OpenFlights (airports + airlines), inspected for:

- Character encoding issues
- Missing latitude/longitude entries
- Duplicate airport IDs
- Non-standard country values

Columns were renamed, normalized, and loaded through `etl/load_openflights.py`.

BTS On-Time Performance (U.S. DOT)

Downloaded CSV slices, cleaned to retain:

- Reporting airline
- Airport
- Aircraft arrival/departure delays
- Delay cause indicators

Loaded through `etl/load_bts_performance.py` after:

- Column standardization

- Null removal
- Type coercion
- Airport code alignment

---

2. Synthetic Data Generation

Python scripts in the `etl/` directory created synthetic, statistically realistic data using **Faker**, **NumPy**, and **SQLAlchemy**.

Synthetic Flights (`synth_flights.py`)

- 5,000 flights generated
- Realistic departure/arrival windows
- Status distribution: Scheduled, Departed, Arrived, Cancelled, Diverted
- Delay logic based on BTS patterns

Synthetic Passenger & Loyalty Data

- **Passengers**: 5,000 individuals
- **Loyalty Accounts**: 3,000 accounts randomly assigned
- **Miles Transactions**: 10,576 earning/redeeming activities

Ensured referential integrity across all customer entities.

Synthetic Revenue (`synth_revenue.py`)

- 40,000 bookings (unique passenger–flight pairs)
- 40,000 payments
- Fare class + pricing distributions (Basic, Standard, Flexible, Business, First)

- Payment methods with status probabilities
- Enforced:
    - no duplicate bookings
    - cascading FK constraints
    - referential consistency

---

3. Production Loads & Integrity Validation

After running all ETL scripts, table row counts were:

Table	Rows
Airports	7,697
Airlines	5,733
Flights	5,000
BTS Performance	22,595
Passengers	5,000
Loyalty Accounts	3,000
Miles Transactions	10,576
Bookings	40,000
Payments	40,000

Referential Integrity

All foreign keys passed validation:

- **0 missing airline references**
- **0 missing airport references**

- **0 missing passenger or flight records**
- **0 orphaned bookings or payments**

This confirmed successful ETL execution and full schema alignment.

---

4. Pipeline Proof Output

A verification plot was generated via Jupyter and saved as:

```
docs/pipeline_row_counts.png
```

This demonstrates successful Phase 2 completion with validated row counts for all production tables.