Phase 3: SQL Standardization, Deduplication, Constraints & Data Quality Validation

Phase 3 transformed the raw Phase-2–loaded database into an **enterprise-grade, analytics-ready relational warehouse**.
This phase focused entirely on **SQL-driven standardization, deduplication, key hygiene, constraint enforcement, and data validation**.

Deliverables for this phase included:

- `sql/03_dml_cleanup.sql` (full standardization + deduplication logic)
- `sql/04_constraints_indexes.sql` (constraints + indexes)
- `sql/05_validations.sql` (formal QA checks)
- `notebooks/02_data_quality_checks.ipynb` (profiling + documentation)
- Output artifacts:
    - `docs/pipeline_quality_checks.png`

---

1. Data Profiling & Sanity Checks

Before applying any transformations, a dedicated profiling notebook (**02_data_quality_checks.ipynb**) was created to inspect the state of all tables.

Row Counts (Pre-Cleanup)

Consistent with Phase 2 ETL results:

Table	Rows
airlines	1,108

airports	7,697
flights	5,000
flight_performance	22,595
passengers	5,000
loyalty_accounts	3,000
miles_transactions	10,576
bookings	40,000
payments	40,000

Null / Integrity Checks on Key Columns

The notebook evaluated all primary referential pathways:

- `airlines.iata_code`, `airports.iata_code`
- `flights.airline_id`, `flights.origin_airport_id`, `flights.destination_airport_id`
- `bookings.passenger_id`, `bookings.flight_id`
- `payments.booking_id`

All critical foreign key checks returned 0 missing values, confirming Phase 2's ETL maintained clean relationships.

Profiling Artifact

A summary "problem counts" table and a corresponding visual
`docs/pipeline_quality_checks.png`
were generated to document the pre-cleanup data landscape.

---

2. Standardization DML (sql/03_dml_cleanup.sql)

All cleanup operations were executed directly in SQL to ensure the database is self-maintaining and auditable.

Airports

- Trimmed whitespace across textual fields
- Normalized IATA/ICAO to upper case
- Converted placeholders (`'', ' ', 'N/A', '\N'`) → `NULL`
- Ensured valid numeric types for latitude/longitude
- Harmonized country strings where feasible

Airlines

- Trimmed and uppercased all carrier codes
- Normalized alternative spellings and placeholder values
- Removed invalid codes (`'', 'N/A', '\N'`)
- Applied consistent flags for "Unknown" or non-commercial records

Flights

- Standardized flight status values to match the enum (`Scheduled`, `Departed`, `Arrived`, `Cancelled`, `Diverted`)
- Ensured **`flight_date = DATE(scheduled_departure_utc)`** where missing
- Normalized `delay_cause` values
- Cleaned invalid timestamps and enforced type conformity

BTS Flight Performance

- Uppercased all carrier + airport codes

- Converted all delay/cause columns to numeric types
- Removed invalid or placeholder entries
- Validated year/month ranges (2024 only)

Passengers & Loyalty

- Trimmed and cleaned customer names
- Lowercased all emails for consistency
- Converted placeholder demographic fields to NULL
- Cleaned loyalty tier codes and enforced enum values

Throughout this process, row counts remained stable, verifying **no unintended deletions**.

---

3. Deduplication & Key Hygiene

After standardization, the same script performed deduplication where appropriate.

Airports

Deduplicated by:

- `iata_code`
- `icao_code`

Using window functions to retain:

- non-null latitude/longitude
- the most complete record

Airlines

- Deduplicated by `iata_code`
- Preferentially kept fully populated or active carriers

Passengers

- Checked for duplicates on `(first_name, last_name, email)`
- None were found (thanks to Phase 2 generation logic)

Bookings & Payments

- Enforced uniqueness on `(passenger_id, flight_id)`
- Ensured one payment per booking
- Verified no orphaned payments or bookings existed

All deduplication preserved referential integrity.

---

4. Constraints & Indexes (sql/04_constraints_indexes.sql)

Once the data was clean, constraints were added or re-applied.

Constraints Added

- **NOT NULL** constraints on all cleaned key fields
- **UNIQUE** constraints:
    - `airports(iata_code)`
    - `airports(icao_code)`
    - `airlines(iata_code)`

    - `flight_performance (airline_iata, airport_iata, year, month)`

- **CHECK constraints**:
    - Geographic bounds (lat/lon)
    - Pricing ranges
    - Delay ranges
- **FOREIGN KEYS**:
    - Flights → Airlines
    - Flights → Airports
    - Bookings → Flights
    - Bookings → Passengers
    - Payments → Bookings
    - Loyalty → Passengers
    - Miles transactions → Loyalty + Flights

Indexes Added

Designed for BI workloads:

- `flights(flight_date, airline_id)`
- `bookings(flight_id)`
- `bookings(booking_date)`
- `payments(paid_at)`
- `flight_performance(airport_iata)`
- Indexes on all FK columns

These dramatically improve aggregation and join performance in Phase 4.

---

5. SQL Validation & Integrity Checks (sql/05_validations.sql)

This script re-ran a full suite of QA checks:

Final Row Counts (Post-Cleanup)

All tables retained correct row counts.

Referential Integrity

All key checks returned **zero missing references**:

- No orphaned flights
- No missing airports/airlines
- No orphaned passengers, bookings, or payments

Business Rule Validations

Validated:

- `delay_minutes >= 0`
- `base_price_usd > 0`
- `amount_usd > 0`
- No duplicated IATA codes
- Loyalty accounts match exactly 3,000 unique passengers

BTS Consistency Validation

Confirmed:

- Year range = 2024 only
- 12 months fully represented
- 21 airlines and 357 airports included

Summary Output

Final notebook produced clear tables capturing:

- Delay cause totals
- Fare-class price distributions
- Payment method distributions
- Flights-with-bookings ratio (4,997/5,000 flights used)

---

6. Documentation & Proof Artifacts

All Phase 3 results were documented:

- Notebook: **02_data_quality_checks.ipynb**
- Cleanup scripts: **03_dml_cleanup.sql**, **04_constraints_indexes.sql**, **05_validations.sql**
- Visual artifact:
    - **docs/pipeline_quality_checks.png**

README updated with a new Phase 3 section summarizing:

> *"Phase 3 focused on SQL-based standardization, deduplication, and enforcing relational constraints across all production tables. The database was validated to be fully clean, normalized, and analytics-ready."*

Phase 3 Summary

Phase 3 successfully delivered a **clean, trustworthy, fully constrained analytical database**. The warehouse now exhibits:

- **Standardized** text and numeric formats
- **Deduplicated** core dimension tables
- **Strict PK/FK, UNIQUE, NOT NULL, and CHECK constraints**
- **Indexed** tables optimized for BI workloads
- **Validated** referential integrity and numeric ranges
- **Zero broken keys or missing relationships**

This foundation enables the advanced analytics, window functions, and BI insights planned for **Phase 4**.