**Data Warehouse Business Process Documentation**

**1. Business Requirements & Process Description**

**Business Process:**

This data warehouse (DWH) is designed to track **customer interactions, complaints, and resolutions** in an airline service environment. It integrates multiple business processes, including **customer service interactions, complaints management, and airport operations**.

**Business Goals:**

* **Monitor customer interactions** with employees to measure satisfaction.
* **Track complaints resolution** process to identify delays and bottlenecks.
* **Improve employee performance** by evaluating response times.
* **Provide airport-specific insights** into customer service and complaints.

**2. Grain Definition**

Each fact table has a well-defined grain:

* **Interaction Fact**: One row per interaction between a customer and an employee.
* **Complaint Fact**: An accumulating snapshot fact table capturing the lifecycle of a complaint.

**3. Dimension Modeling**

**3.1. Employee\_Dim**

* **Grain**: One row per employee.
* **Attributes**:
  + **Employee\_ID**: Unique identifier for each employee.
  + **Age**: Employee's age.
  + **Gender**: Employee's gender (e.g., Male, Female).
  + **Role**: Job role or position within the organization.
  + **Status**: Employment status (e.g., Active, On Leave, Retired).

**3.2. Passenger\_Dim**

* **Grain**: One row per passenger.
* **Attributes**:
  + **Passenger\_ID**: Unique identifier for each passenger.
  + **Age Category**: Age group classification (e.g., Child, Adult, Senior).
  + **Gender**: Passenger's gender.
  + **Occupation**: Passenger's profession or job.
  + **Nationality**: Country of citizenship.
  + **Membership Status**: Whether the passenger is a frequent flyer or loyalty program member.
  + **Loyalty Tier**: Membership level in the airline's loyalty program (e.g., Silver, Gold, Platinum).
  + **City**: City of residence.
  + **Country**: Country of residence.

**3.3. Date\_Dim**

* **Grain**: One row per date.
* **Attributes**:
  + **Full Date**: The complete date (YYYY-MM-DD format).
  + **Day**: Day of the month (1-31).
  + **Month**: Numeric representation of the month (1-12).
  + **Month Name**: Full name of the month (e.g., January, February).
  + **Quarter**: Quarter of the year (Q1-Q4).
  + **Year**: Four-digit year.
  + **Weekday Name**: Name of the day (e.g., Monday, Tuesday).
  + **Is\_Weekend**: Indicator for weekends (1/0).
  + **Holiday Flag**: Indicator for public holidays (1/0).

**3.4. Airport\_Dim**

* **Grain**: One row per airport.
* **Attributes**:
  + **Airport Code**: Unique three-letter code for the airport (IATA code).
  + **Airport Name**: Full name of the airport.
  + **City**: City where the airport is located.
  + **Country**: Country where the airport is located.

**3.5. Complaint\_Category\_Dim**

* **Grain**: One row per complaint category.
* **Attributes**:
  + **Complaint Category ID**: Unique identifier for each complaint category.
  + **Category Name**: Name of the complaint category (e.g., Baggage Delay, Flight Cancellation).
  + **Default Resolution Time**: Standard time allocated for resolving complaints in this category.

**3.6. Degenerated Dimensions**

* **Complaint Channel**: Captures the source through which the complaint was made (e.g., Call Center, Website, Mobile App).
* **Interaction Channel**: Captures the medium used for customer interaction (e.g., Phone Call, Chat, In-person).

**4. Fact Table Design**

**4.1. Interaction\_Fact**

* **Grain**: One row per interaction.
* **Measures**:
  + **Satisfaction Score**: Customer satisfaction rating for the interaction.
  + **Response Time**: Time taken for the employee to respond.
  + **Interaction Time**: Duration of the interaction.
* **Foreign Keys**:
  + **Employee\_ID**: Links to Employee\_Dim.
  + **Passenger\_ID**: Links to Passenger\_Dim.
  + **Interaction Date**: Links to Date\_Dim.
  + **Response Date**: Links to Date\_Dim.

**4.2. Complaint\_Fact (Accumulating Snapshot)**

* **Grain**: One row per complaint that is updated over time.
* **Measures**:
  + **Satisfaction Score**: Customer satisfaction rating after complaint resolution.
  + **Complaint Status**: Current status of the complaint (e.g., Open, In Progress, Resolved).
* **Foreign Keys**:
  + **Complaint\_Category\_ID**: Links to Complaint\_Category\_Dim.
  + **Employee\_ID**: Links to Employee\_Dim.
  + **Passenger\_ID**: Links to Passenger\_Dim.
  + **Airport\_ID**: Links to Airport\_Dim.
* **Date Fields**:
  + **CreatedOn**: Date the complaint was logged.
  + **FirstResponseOn**: Date the first response was given.
  + **InvestigationStartOn**: Date the investigation started.
  + **InvestigationEndOn**: Date the investigation concluded.
  + **ResolvedOn**: Date the complaint was resolved.

**5. Slowly Changing Dimensions (SCD) Strategy**

* **Employee\_Dim**: SCD Type 1 (overwrites updates, maintaining only the latest information).
* **Passenger\_Dim**: SCD Type 2 (tracks historical changes, maintaining multiple records for the same passenger over time).

**6. Reporting & Analytics Considerations**

* **Customer Service Performance Dashboards**:
  + Employee response time vs. satisfaction score.
* **Complaint Resolution Analysis**:
  + Average time taken to resolve different complaint types.
  + Analysis of complaint lifecycle progression.
* **Passenger Insights**:
  + Loyalty tier impact on complaint frequency.
* **Airport-Specific Trends**:
  + Airports with the highest complaint rates.

**Conclusion**

This data warehouse enables efficient monitoring of customer interactions and complaints, leading to actionable insights for airline management. The model ensures scalability, supporting historical tracking, and dynamic analysis of service quality improvements.

**Table Distribution Strategy in Amazon Redshift**

Amazon Redshift uses **distribution styles** to determine how data is stored across nodes. Proper selection of distribution styles can **optimize query performance** by reducing data movement between nodes.

**Distribution Styles Used in the DDL Script**

**1️⃣ ALL Distribution (for Small Lookup Tables)**

Used for **dimension tables**:  
✅ **Employee\_Dim, Passenger\_Dim, Date\_Dim, Airport\_Dim, Complaint\_Category\_Dim**

**🔹 Reason:**

* These tables are relatively small and frequently joined with fact tables.
* **ALL** ensures each compute node has a **full copy**, eliminating the need for data redistribution during joins.

**2️⃣ KEY Distribution (for Fact Tables)**

✅ **Interaction\_Fact (DISTKEY: Employee\_ID)**  
✅ **Complaint\_Fact (DISTKEY: Passenger\_ID)**

**🔹 Reason:**

* Fact tables are **large** and frequently joined with dimension tables.
* **KEY** distribution ensures that matching rows in the **fact table** and **dimension table** are stored on the **same node**, improving join efficiency.
* **Employee\_ID** is used as the **distribution key** for Interaction\_Fact because **interactions are tracked by employees**.
* **Passenger\_ID** is used as the **distribution key** for Complaint\_Fact because **complaints are linked to specific passengers**.

**Why Not EVEN Distribution?**

* **EVEN** distribution randomly assigns rows across nodes.
* While good for independent workloads, it does **not optimize joins**, which are frequent in a DWH.

**Final Optimization Considerations**

* Choosing the right **DISTKEY** avoids costly **data redistribution** during joins.
* Small dimensions use **ALL** for quick lookup access.
* Large fact tables use **KEY** distribution for **performance tuning**.