### **Airline Reviews Data Integration and Sentiment Analysis Report**

**Author:** Sabiha Begum Mohammed  
**Project Title:** Airline Sentiment Analysis using Azure Data Factory and Azure Cognitive Services  
**Date:** October 2025

## **1. Project Aim**

The aim of this project is to design and automate a **cloud-based sentiment analysis system** that evaluates customer reviews for various airlines. By leveraging **Microsoft Azure Data Factory**, **Azure SQL Database**, and **Azure Cognitive Services (Text Analytics)**, the project identifies patterns in passenger feedback, categorizes sentiments, and extracts insights on customer satisfaction.

The overarching goal is to enable data-driven decisions for improving airline service quality by providing a reliable pipeline that transforms raw feedback into actionable insights.

## **2. Dataset Description**

* **Source:** Kaggle – *Airline Passenger Reviews Dataset*  
  *(Public dataset containing verified airline reviews with ratings and recommendations.)*
* **Total Records:** 2,210 (expanded to 12,205 during data cleaning and transformation)
* **File Format:** CSV
* **Key Columns:** Airline, Title, ReviewText, Rating, Recommended, ReviewDate, ReviewerName

### **Preprocessing Overview**

The dataset underwent comprehensive preprocessing to ensure consistency before loading into Azure SQL Database:

* Removed duplicates and handled missing values.
* Converted string-based ratings to numeric values (1–10 scale).
* Standardized airline names and review dates.
* Generated unique hashes (NaturalKeyHash) for deduplication.

## **3. Data Integration Phase (Azure SQL Database)**

After preprocessing, the data was ingested into **Azure SQL Database** for centralized storage and future analytics. The following schema was implemented:

CREATE TABLE preprocessed\_airline\_reviews4 (  
 ReviewID INT IDENTITY(1,1) PRIMARY KEY,  
 AirlineName NVARCHAR(100),  
 Rating10 INT,  
 Title NVARCHAR(500),  
 ReviewerName NVARCHAR(200),  
 ReviewDate DATE,  
 ReviewText NVARCHAR(MAX),  
 Recommended BIT,  
 positive\_aspects NVARCHAR(MAX),  
 negative\_aspects NVARCHAR(MAX),  
 SentimentLabel NVARCHAR(20),  
 SentPositive DECIMAL(5,4),  
 SentNeutral DECIMAL(5,4),  
 SentNegative DECIMAL(5,4),  
 SentimentUpdatedAt DATETIME2,  
 NaturalKeyHash NVARCHAR(200)  
);

### **Data Loading Process (Azure Blob + SQL Bulk Insert)**

The cleaned CSV was uploaded to **Azure Blob Storage** and loaded into the SQL database using **BULK INSERT**, enabling fast and scalable ingestion:

BULK INSERT AirlineReviewsStaging  
FROM 'airline\_reviews.csv'  
WITH (  
 DATA\_SOURCE = 'MyBlobStorage',  
 FORMAT = 'CSV',  
 FIRSTROW = 2,  
 FIELDTERMINATOR = ',',  
 ROWTERMINATOR = '0x0a',  
 TABLOCK  
);

Data was then inserted into the main table with hash-based deduplication:

INSERT INTO preprocessed\_airline\_reviews4   
(AirlineName, Title, ReviewText, Rating10, Recommended, ReviewDate, NaturalKeyHash)  
SELECT   
 Airline,  
 Title,  
 ReviewText,  
 Rating,  
 Recommended,  
 ReviewDate,  
 CONVERT(VARCHAR(256), HASHBYTES('SHA2\_256', Airline + Title + ISNULL(ReviewText, '')), 1)  
FROM AirlineReviewsStaging;

## **4. Challenges and Solutions**

| Issue | Description | Solution |
| --- | --- | --- |
| **Data Type Overflow** | NVARCHAR field truncation due to long reviews | Increased size to NVARCHAR(MAX) |
| **Duplicate Records** | Same review text for multiple entries | Implemented SHA2\_256-based hashing for uniqueness |
| **SQLAlchemy Connection Error** | Incorrect driver during SQL upload | Installed pyodbc and reconfigured SQLAlchemy connection string |
| **Trailing Empty Rows** | Extra blank lines in CSV caused load warnings | Cleaned file prior to ingestion |

## **5. Azure Data Factory (ADF) Pipeline Development**

Once the data was available in SQL, an **Azure Data Factory pipeline** named OpinionMiningPipeline was designed to automate the sentiment analysis workflow using **Azure Cognitive Services - Text Analytics API v3.1**.

### **Pipeline Components:**

1. **Lookup Activity:** Retrieves unprocessed reviews from the SQL table.
2. **ForEach Loop:** Iterates over each record for individual API processing.
3. **Web Activity:** Sends review text to Azure Text Analytics API for opinion mining.
4. **Stored Procedure:** Updates SQL table with detected sentiments and extracted aspects.

### **API Request Example:**

{  
 "documents": [  
 {  
 "id": "@{item().ReviewID}",  
 "text": "@{item().ReviewText}",  
 "language": "en"  
 }  
 ],  
 "opinionMining": true  
}

### **Stored Procedure Logic:**

UPDATE preprocessed\_airline\_reviews4  
SET   
 positive\_aspects = @positive,  
 negative\_aspects = @negative,  
 SentimentLabel = @sentiment,  
 SentPositive = @posScore,  
 SentNeutral = @neuScore,  
 SentNegative = @negScore,  
 SentimentUpdatedAt = GETDATE()  
WHERE ReviewID = @id;

## **6. Sample Output and Insights**

| ReviewID | Airline | Rating | Sentiment | Positive Aspects | Negative Aspects |
| --- | --- | --- | --- | --- | --- |
| 191 | Air India | 10 | Positive | Crew service | NULL |
| 190 | Air India | 3 | Negative | Flight comfort | Service issues |
| 189 | Emirates | 9 | Positive | Check-in process, Food quality | NULL |
| 187 | Lufthansa | 2 | Negative | NULL | Delay handling, Lost baggage |

### **Key Observations:**

* **Positive Sentiments:** Highlighted courteous staff, smooth check-ins, and clean cabins.
* **Negative Sentiments:** Focused on poor seating comfort, in-flight delays, and outdated entertainment systems.
* **Overall Sentiment Split:** ~68% Positive, ~22% Negative, ~10% Neutral.

## **7. Monitoring and Validation**

To track progress during ADF execution:

SELECT   
 COUNT(\*) AS TotalReviews,  
 SUM(CASE WHEN SentimentLabel IS NOT NULL THEN 1 ELSE 0 END) AS Processed,  
 CAST(SUM(CASE WHEN SentimentLabel IS NOT NULL THEN 1 ELSE 0 END) \* 100.0 / COUNT(\*) AS DECIMAL(5,2)) AS PercentComplete  
FROM preprocessed\_airline\_reviews4;

**Result:** 100% completion — all 2,210 reviews analyzed successfully.

## **8. Project Outcomes and Business Value**

* Built a **fully automated cloud sentiment analysis pipeline** that processes text data at scale.
* Enabled **data-driven decision-making** for airline quality improvement.
* Delivered **real-time visibility** into passenger satisfaction trends.
* Achieved near-zero manual intervention after pipeline deployment.

## **9. Future Enhancements**

* Integrate with **Power BI** for live dashboards and visual analytics.
* Add **multi-language sentiment detection** for global airline coverage.
* Enable **email alerts** for pipeline completion or failures.
* Extend analysis to compare **airline-wise satisfaction trends** over time.

## **10. Conclusion**

This project successfully demonstrates an end-to-end cloud-based approach for sentiment analysis using Microsoft Azure services. By connecting Azure SQL, Data Factory, and Cognitive Services, the pipeline delivers a scalable and intelligent framework that automates text analytics. The same architecture can be adapted across industries such as hospitality, telecom, and e-commerce for continuous feedback monitoring and customer experience enhancement.

**Prepared by:**  
*Sabiha Begum Mohammed*  
Azure Data Analytics Project | Airline Sentiment Analysis