# Year-long Data Engineering Project: A Journey of a Data Engineer

## Introduction

Welcome to your new role as a Data Engineer at a leading hotel chain. Over the next year, you will embark on a journey to manage, transform, and analyze hotel booking data. Your tasks will evolve from understanding and cleaning the dataset to creating advanced visualizations and providing strategic insights to drive business decisions. This storyline will guide you through the various phases of your project, providing problem statements and challenges that mirror real-world scenarios.

## Phase 1: Setting the Context (Months 1-3)

You have just joined the company and are eager to make your mark. The first task is to get familiar with the data and lay the groundwork for more complex analyses.

### Problem Statements:

1. \*\*Summary Statistics:\*\*  
 - Load the dataset and calculate summary statistics using Python (pandas). Highlight key attributes and their distributions using visualizations (histograms, box plots). Identify which distributions are normal and which are skewed. Provide a detailed report of your findings.

2. \*\*Missing Values:\*\*  
 - Check for any missing values in the dataset using SQL queries or Python (pandas). Describe the methods you would use for imputation and explain why you chose them. Example SQL Query to check for missing values: `SELECT \* FROM bookings WHERE column\_name IS NULL;` Use SQL's COALESCE or CASE WHEN functions for imputation. Document the steps and provide code snippets.

## Phase 2: Data Ingestion Strategies (Months 4-5)

Having cleaned the data, your next challenge is to establish efficient data ingestion processes to ensure smooth data flow into your systems.

### Problem Statements:

3. \*\*Exploratory Data Analysis (EDA):\*\*  
 - Conduct a comprehensive EDA on selected hotel booking attributes using Python (pandas, matplotlib, seaborn). Generate visualizations (histograms, scatter plots, box plots) to analyze key variables (e.g., stay duration, lead time, visitor types). Detect outliers using visual and statistical methods. Discuss how the insights gained from this EDA could influence business decisions. Provide a detailed report with visualizations and interpretations.

4. \*\*Guest Origins:\*\*  
 - Identify the countries guests are from. Calculate and visualize the percentage of total guests per country using Python (pandas, matplotlib). Highlight the top 3 countries. Round off the percentage to 2 decimal places. Provide a detailed report with visualizations.

## Phase 3: Data Transformation (Months 6-7)

With a robust ingestion strategy in place, you will now focus on transforming the data to derive meaningful insights.

### Problem Statements:

5. \*\*Monthly Visitors:\*\*  
 - Calculate the number of guests per month by hotel each year. Consider attributes like hotel, arrival\_date\_year, arrival\_date\_month, and total visitors (adults, children, babies). Drop the original columns after creating a copy. Use SQL queries to perform these calculations. Provide the SQL code and the resulting dataset.

6. \*\*Visitor Trends:\*\*  
 - Visualize the number of visitors to the hotel each month for 2015, 2016, and 2017 using line plots. Include weekend visitors and identify which year had the most weekend bookings. Use Python (matplotlib, seaborn) for visualizations. Provide a detailed report with visualizations and interpretations.

## Phase 4: Data Warehousing (Months 8-9)

Your task is to store the transformed data efficiently for future analysis and reporting.

### Problem Statements:

7. \*\*Room Type Rates:\*\*  
 - Analyze and visualize room type rates per night per person by hotel using seaborn bar plots. Perform comparative analysis of average prices paid per person per night per country. Use Python (pandas, seaborn) for analysis and visualization. Provide a detailed report with visualizations and interpretations.

8. \*\*Room Type Popularity:\*\*  
 - Visualize the most booked room types across countries. Perform a countplot analysis for reserved\_room\_type, sorted by values, with hue set to hotel. Use Python (seaborn) for visualizations. Provide a detailed report with visualizations and interpretations.

## Phase 5: Reporting and Visualization (Months 10-12)

Now, you will focus on creating visualizations and performing advanced analyses to provide strategic insights to the business.

### Problem Statements:

9. \*\*Market Segment Comparison:\*\*  
 - Compare market segments for hotels across countries. Visualize market segments for both cancelled and successful bookings. Use Python (seaborn, matplotlib) for visualizations. Provide a detailed report with visualizations and interpretations.

10. \*\*Correlation Analysis:\*\*  
 - Perform a heatmap correlation matrix analysis using Python (seaborn). Identify positive and negative correlations and infer their implications. Provide a detailed report with visualizations and interpretations.

11. \*\*Adult Accompaniment:\*\*  
 - Analyze and visualize the percentage of adults accompanying children to hotels across countries using a pie chart. Use Python (matplotlib) for visualization. Provide a detailed report with visualizations and interpretations.

12. \*\*Booking Cancellations:\*\*  
 - Visualize booking cancellations across years and countries. Identify cancellation patterns by hotel type and market segment. Use Python (seaborn, matplotlib) for visualizations. Provide a detailed report with visualizations and interpretations.

## Advanced Problem Statements:

13. \*\*Complex SQL Queries and Power BI Dashboard:\*\*  
 - Design SQL queries to analyze guest demographics and booking behaviors. Create a multi-dimensional Power BI dashboard with interactive elements to explore guest preferences and booking trends. Include visualizations like heat maps for geographical data, scatter plots for comparing trends, and histograms for distribution of booking durations. Provide a detailed report with SQL queries and dashboard screenshots.

14. \*\*Seasonal Trends Analysis:\*\*  
 - Develop a Python script to automate the extraction of seasonal trends and booking patterns. Identify peak and off-peak seasons and discuss strategic planning and pricing strategies. Use Python libraries (pandas, matplotlib, seaborn) for analysis and visualization. Provide a detailed report with code snippets and visualizations.

15. \*\*Final Analysis:\*\*  
 - Conclude with a comprehensive analysis on booking cancellation predictions, customer segmentation, satisfaction, and seasonality. Provide a detailed report summarizing your findings and insights.

# Additional Guidance and Resources

### Additional Guidance and Resources

1. \*\*Data Cleaning\*\*: Use Pandas library for data manipulation and cleaning. Refer to the Pandas documentation for comprehensive guides and examples.

2. \*\*Data Ingestion\*\*: Use Python scripts to automate data ingestion processes. Consider using Apache Airflow for orchestrating the data pipelines.

3. \*\*Data Transformation\*\*: Utilize Pandas and SQL for data transformation tasks. You can also explore libraries like Dask for handling large datasets.

4. \*\*Data Warehousing\*\*: Implement data warehousing using cloud solutions like Amazon RDS, Google BigQuery, or Snowflake. Refer to their respective documentations for setup and best practices.

5. \*\*Data Visualization\*\*: Use libraries like Matplotlib, Seaborn, or Plotly for creating visualizations. For more interactive dashboards, consider using tools like Tableau or Power BI.

6. \*\*Collaboration\*\*: Consider using version control systems like Git and platforms like GitHub or GitLab for collaborative work and code management.

7. \*\*Interim Feedback\*\*: Schedule periodic reviews and feedback sessions with your instructor or peers to stay on track and address any challenges early.

8. \*\*Advanced Analytics\*\*: Towards the end of the project, explore integrating machine learning models for predictive analytics to enhance your insights.

9. \*\*Documentation and Reporting\*\*: Maintain thorough documentation of your processes and findings. Prepare clear and concise reports to present your results effectively.