Request for Proposal (RFP): Airbnb Data Cleaning, Processing, and Advanced Analysis

Overview

This RFP outlines a systematic approach to clean, process, transform, and analyze the Airbnb dataset to enhance data quality and derive actionable insights. The project is divided into several phases, each focusing on specific tasks and problem statements to ensure comprehensive data handling and analysis.

Problem Statements and Tasks

Phase 1: Onboarding and Initial Data Handling (Months 1-3)

Objective: Understand, clean, and augment the Airbnb dataset, and perform exploratory data analysis (EDA).

1. **Problem Statement 1: Comprehensive Data Profiling and Cleaning**
    - **RFP Task:** Conduct EDA on calendar.csv, listings.csv, and reviews.csv. Identify and visualize data quality issues.
    - **Instructions:**
        - Perform EDA to identify missing values, outliers, and anomalies using visualizations like heatmaps and box plots.
        - Implement data cleaning strategies, including imputing missing values, treating outliers, and standardizing data formats.
    - **Example:**
        - **Solution:** Use Python (Pandas) to conduct EDA and clean the data. Document the findings, strategies employed, and steps taken during the data cleaning process.
2. **Problem Statement 2: Data Augmentation and Synthesis**
    - **RFP Task:** Develop techniques to augment the dataset by creating synthetic data.
    - **Instructions:**
        - Develop machine learning models (e.g., GANs or SMOTE) to predict missing values and generate synthetic data points.
        - Validate the augmented data using statistical methods to ensure robustness.
    - **Example:**
        - **Solution:** Use Python libraries like Scikit-Learn to implement data augmentation techniques. Validate the synthetic data for accuracy and integrity.
3. **Problem Statement 3: Data Integration and Consolidation**
    - **RFP Task:** Integrate calendar.csv, listings.csv, and reviews.csv into a consolidated dataset.
    - **Instructions:**

        - Merge the datasets, resolve inconsistencies such as duplicate records, and remove redundant information.
        - Optimize the consolidated dataset for subsequent analysis by standardizing data types and normalizing values.
    - **Example:**
        - **Solution:** Utilize Python (Pandas) to merge datasets and clean any inconsistencies, ensuring data integrity.

Phase 2: Data Ingestion and Optimization (Months 4-6)

Objective: Develop efficient data ingestion pipelines and optimize data storage.

4. **Problem Statement 4: Scalable Data Ingestion Pipeline**
    - **RFP Task:** Design and implement a scalable data ingestion pipeline using Python, Pandas, and SQL.
    - **Instructions:**
        - Incorporate error handling, logging, and monitoring features for reliability.
        - Optimize the pipeline for performance, including batch processing and real-time data ingestion.
    - **Example:**
        - **Solution:** Create a data ingestion script in Python with batch processing capabilities, logging mechanisms, and real-time ingestion features.
5. **Problem Statement 5: Advanced Storage Optimization Techniques**
    - **RFP Task:** Implement storage optimization techniques such as data partitioning, indexing, and compression.
    - **Instructions:**
        - Apply storage optimization in a cloud-based data storage solution like AWS RDS or Google BigQuery.
    - **Example:**
        - **Solution:** Use tools like AWS RDS or BigQuery to partition, index, and compress data, improving storage efficiency and retrieval speed.

Phase 3: Data Transformation (Months 7-9)

Objective: Transform data to meet business requirements and enable advanced analytics.

6. **Problem Statement 6: Complex Data Transformation Workflows**
    - **RFP Task:** Design data transformation workflows using tools like Apache Spark or Apache Airflow.
    - **Instructions:**

        - Handle data aggregation, filtering, and enrichment processes to prepare for advanced analytics.
        - Ensure the workflows are scalable and efficient.
    - **Example:**
        - **Solution:** Develop workflows using Apache Spark to transform data, document each step, and validate performance on large datasets.

7. **Problem Statement 7: Feature Engineering for Machine Learning**
    - **RFP Task:** Develop techniques for feature engineering to extract meaningful features from the dataset.
    - **Instructions:**
        - Extract temporal features from calendar.csv and textual features from reviews.csv.
        - Validate the effectiveness of these features for machine learning models.
    - **Example:**
        - **Solution:** Use Python to engineer features and test their impact on model performance. Document the process and findings.
8. **Problem Statement 8: Advanced Data Normalization and Encoding**
    - **RFP Task:** Implement data normalization techniques for machine learning model preparation.
    - **Instructions:**
        - Apply normalization techniques such as min-max scaling and Z-score normalization.
        - Use encoding methods like one-hot encoding, label encoding, and embeddings for categorical variables.
    - **Example:**
        - **Solution:** Utilize Python libraries like Scikit-Learn for data normalization and encoding, and document the effects on model performance.

Phase 4: Data Warehousing, Reporting, and Visualization (Months 10-12)

Objective: Store transformed data in a data warehouse and create advanced visualizations to support business decisions.

9. **Problem Statement 9: Enterprise Data Warehousing**
    - **RFP Task:** Design an enterprise-level data warehouse schema using a cloud-based solution.
    - **Instructions:**

        - Create star and snowflake schemas to support complex querying and analytics.
    - **Example:**
        - **Solution:** Use tools like Amazon Redshift or Google BigQuery to design and implement the data warehouse schema, and document the ETL process.

10. **Problem Statement 10: Advanced Business Intelligence Reporting**
    - **RFP Task:** Create advanced business intelligence reports and dashboards using tools like Power BI or Tableau.
    - **Instructions:**
        - Focus on key business metrics such as occupancy rates, revenue per listing, and customer satisfaction.
    - **Example:**
        - **Solution:** Develop interactive dashboards and reports, and document the design and the insights derived from the data.
11. **Problem Statement 11: Predictive Analytics and Machine Learning**
    - **RFP Task:** Develop predictive models using machine learning algorithms.
    - **Instructions:**
        - Use regression, classification, and clustering techniques to derive insights from the data.
    - **Example:**
        - **Solution:** Build and validate predictive models, iteratively improve their performance, and document the development process and outcomes.
12. **Problem Statement 12: Optimization and Revenue Analysis**
    - **RFP Task:** Perform advanced data analysis to identify optimization areas and potential revenue growth.
    - **Instructions:**
        - Use techniques like A/B testing, cohort analysis, and time series analysis to identify optimization areas.
    - **Example:**
        - **Solution:** Conduct data analysis, provide actionable recommendations for optimization and revenue growth, and document the findings.

Comprehensive Dataset Details

The Airbnb project involves the following datasets:

1. **calendar.csv**

- **Number of Rows:** 1,393,570
- **Number of Columns:** 7
- **Key Column Names:** listing_id, date, available, price, minimum_nights, maximum_nights, available_units

2. **listings.csv**
    - **Number of Rows:** 3,818
    - **Number of Columns:** 92
    - **Key Column Names:** id, listing_url, scrape_id, last_scraped, name, host_id, host_name, host_since, neighbourhood, latitude, longitude, room_type, price, number_of_reviews, review_scores_rating, instant_bookable, cancellation_policy
3. **reviews.csv**
    - **Number of Rows:** 84,849
    - **Number of Columns:** 7
    - **Key Column Names:** listing_id, id, date, reviewer_id, reviewer_name, comments, sentiment_score

Deliverables

- **Monthly Progress Reports:** Detailed reports outlining progress on each problem statement, including methodologies, challenges, and preliminary results.
- **Final Data Warehouse:** A fully populated and optimized data warehouse schema.
- **Predictive Models:** Trained and validated machine learning models with performance metrics.
- **BI Dashboards:** Interactive dashboards for exploring data and insights.
- **Comprehensive Final Report:** A detailed report summarizing the entire project, including methodologies, findings, and recommendations.