**Data Attribution System: Approach, Assumptions, Observations, and Future Scope**

**Introduction**

This document outlines the approach undertaken to build a simple attribution system for understanding which touchpoints and channels are sourcing the most pipeline. The system adheres to the specified properties and assumptions, including a 90-day attribution window and a first-touch attribution model.

**Approach**

1. **Import Data:**
   * Load data from the provided CSV and Excel files for contacts, marketing touchpoints, sales touchpoints, and opportunities.
2. **Merge Marketing and Sales Data:**
   * Combine marketing and sales touchpoints into a single DataFrame, differentiating them by a touchpoint\_type column.
3. **Join Touchpoint Data with Contact and Opportunity Data:**
   * Merge the combined touchpoint data with the contact data on contact\_id.
   * Merge the result with the opportunity data on account\_id.
4. **Filter Touchpoints within 90 Days Before Each Opportunity Creation:**
   * Calculate the days between each touchpoint and the corresponding opportunity creation date.
   * Filter out touchpoints that fall outside the 90-day window.
5. **Identify the First Touchpoint for Each Opportunity:**
   * Sort the filtered touchpoints by opportunity ID and touchpoint date.
   * Group by opportunity ID to select the first touchpoint within the 90-day window for each opportunity.
6. **Calculate Sourced Pipeline:**
   * Attribute the pipeline amount of each opportunity to the corresponding first touchpoint.
7. **Create the Final Output Table:**
   * Select relevant columns and save the result to a CSV file for further analysis.

**Assumptions**

1. **Account-Level Attribution:**
   * Interactions on an account are related to opportunities on that account.
2. **First Touch Attribution Model:**
   * The first touchpoint within the 90-day period before the opportunity creation date gets full credit for sourcing the opportunity.
3. **Multiple Opportunities per Account:**
   * An account can have multiple opportunities, and each opportunity should be considered independently for touchpoint attribution.

**Observations**

1. **One Account ID through a Contact Can Have Multiple Opportunity IDs:**
   * This scenario was observed and handled by ensuring that touchpoints are filtered and attributed independently for each opportunity.

**Future Scope**

1. **Enhanced Attribution Models:**
   * Implement additional attribution models such as last-touch, linear, or multi-touch to provide a comprehensive analysis.
2. **Cost Data Integration:**
   * Incorporate cost data for each touchpoint or channel to enable ROI analysis.
3. **Real-Time Data Processing:**
   * Develop a real-time data processing pipeline to handle live data streams and provide up-to-date attribution analysis.
4. **Automated Data Validation:**
   * Implement automated data validation and quality checks to ensure the accuracy and reliability of the insights.

**Answers to Questions**

1. **Which channel sourced the most pipeline? How does this look by sales segment?**
   * **Channel sourcing the most pipeline:** The channel with the highest sum in channel\_pipeline.
   * **Pipeline by sales segment:** The segment\_channel\_pipeline table shows how the sourced pipeline is distributed across different sales segments and channels.

python

Copy code

channel\_pipeline = output\_df.groupby('channel\_name')['Sourced\_Pipeline'].sum().sort\_values(ascending=False)

print("Pipeline sourced by channel:")

print(channel\_pipeline)

segment\_channel\_pipeline = output\_df.pivot\_table(

values='Sourced\_Pipeline',

index='sales\_segment',

columns='channel\_name',

aggfunc='sum'

)

print("\nPipeline sourced by channel and sales segment:")

print(segment\_channel\_pipeline)

1. **What information do you need to know to understand the ROI (return on investment) of each channel?**
   * **Cost data:** The cost associated with each marketing or sales touchpoint/channel.
   * **Revenue data:** The actual revenue generated from opportunities sourced by each channel.
   * **Conversion rates:** The rate at which leads from each channel convert into opportunities and then into closed deals.
2. **How did you structure your data table and why? What do you think are the important output dimensions?**
   * **Structure:** The data table includes columns for touchpoint ID, channel name, contact ID, account ID, opportunity ID, touchpoint date, opportunity creation date, pipeline amount, sourced pipeline, and sales segment.
   * **Reasoning:** This structure ensures that all relevant information about the touchpoints and their associated opportunities is captured. The Sourced\_Pipeline column directly shows the pipeline attributed to each touchpoint.
   * **Important output dimensions:**
     + **Channel name:** Identifies which channel is sourcing the pipeline.
     + **Sales segment:** Breaks down the analysis by different sales segments.
     + **Pipeline amount and sourced pipeline:** Quantifies the value of the opportunities.
3. **This table is an important input into other data and business systems. What kind of data validations and checks would you implement to make sure that downstream stakeholders have confidence in the insights they are generating from this?**
   * **Date range checks:** Ensure touchpoints fall within the valid 90-day window before the opportunity creation date.
   * **Duplicate checks:** Ensure no duplicate touchpoints or opportunities.
   * **Null checks:** Ensure critical columns (e.g., channel\_name, opportunity\_id, pipeline\_amount) do not contain null values.
   * **Consistency checks:** Verify touchpoint dates are consistent with opportunity creation dates and that the touchpoints belong to the correct accounts.
   * **Aggregated data validation:** Cross-verify the aggregated pipeline amounts with the source data to ensure accuracy.
   * **Unit tests:** Implement unit tests for each function to validate expected transformations and calculations.