Metric Definition:

When do leads become a qualified opportunity with the pipeline? - Through touchpoints. Touchpoints - Leads interact with several different touchpoints, which include marketing efforts (such as events and webinars) or sales outreach (such as direct phone calls).

Objective:

Building an Attribution System to Determine: Which Touchpoints Source the Most Pipeline?

SQL Queries:

https://github.com/viraj-dhane/SQL_MarketingDataAnalysis/blob/main/SQL_MarketingDataAnalysis.sql

Modeling Assumption:

A successful source is one that has generated the highest pipeline revenue.

Input Data Schema:

Table - marketing_data

Column Name	**Data Type**
marketing_touchpoint_id (PK)	varchar(255)
channel_name	varchar(255)
contact_id	varchar(255)
marketing_touchpoint_date	datetime

Table - sales_outreach_data

Column Name	**Data Type**
sales_touchpoint_id (PK)	varchar(255)
channel_name	varchar(255)
contact_id	varchar(255)
sales_touchpoint_date	datetime

Table - contact_data

Column Name	**Data Type**
contact_id (PK)	varchar(255)
account_id	varchar(255)

Table - opportunity_data

Column Name	Data Type
opportunity_id (PK)	varchar(255)
account_id	varchar(255)
pipeline_amount	varchar(255)
opportunity_created_date	datetime
Sales_segment	varchar(255)

Attribution Model Consideration - First Touch

- Attribution window = 90 days (include touch points that happened up to 90 days before creation date)
- The first touch point that happened within 90 days before opportunity creation gets full credit for sourcing the opportunity.

Q. How did you structure your data table and why? What do you think are the important output dimensions?

Table - opportunity_touchpoints

Column Name	Data Type
opportunity_id (PK)	varchar(255)
account_id	varchar(255)
pipeline_amount	float
opportunity_created_date	datetime
sales_segment	varchar(255)
touchpoint_id	varchar(255)
touchpoint_type	varchar(255)
channel_name	varchar(255)
contact_id	varchar(255)
touchpoint_date	datetime

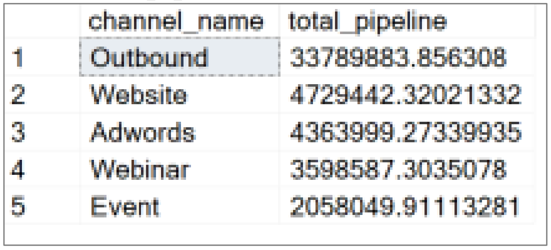
- The table includes all relevant identifiers (touchpoint, contact, account, opportunity) to ensure comprehensive coverage and traceability of interactions.
- ‘touchpoint_type’ helps in differentiating between marketing and sales efforts, which can be crucial for further analysis.
- ‘touchpoint_date’ and ‘opportunity_created_date’ provide necessary temporal information to calculate the attribution within the 90-day window.

Important Output Dimensions:

- Channel Name: Essential to identify which channels are most effective in generating pipeline revenue.
- Pipeline Amount: The key metric to measure the success of each touchpoint.
- Sales Segment: Important for understanding the effectiveness of channels across different market segments.

- Touchpoint Date: Provides insights into the timing and frequency of effective touchpoints.
- Touchpoint Type: Helps in differentiating and analyzing the impact of marketing vs. sales efforts.

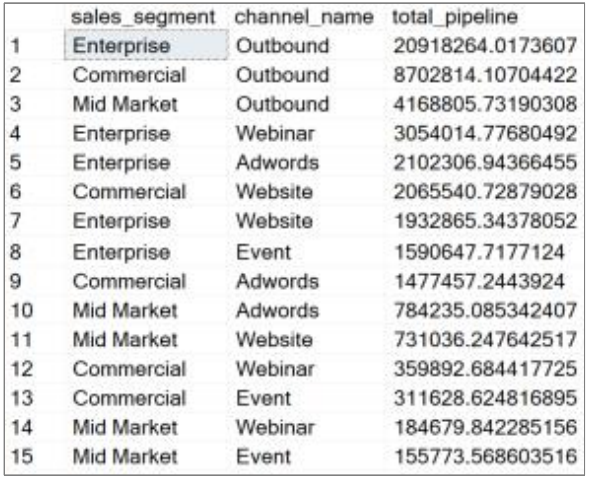
Q. Which channel sourced the most pipeline? How does this look by sales segment?



	channel_name	total_pipeline
1	Outbound	33789883.856308
2	Website	4729442.32021332
3	Adwords	4363999.27339935
4	Webinar	3598587.3035078
5	Event	2058049.91113281

The **Outbound** channel sourced the most pipeline, approximately **$33M** out of the total $48M mapped to marketing and sales outreach channels.

Channel sourcing most pipeline by sales segment -



	sales_segment	channel_name	total_pipeline
1	Enterprise	Outbound	20918264.0173607
2	Commercial	Outbound	8702814.10704422
3	Mid Market	Outbound	4168805.73190308
4	Enterprise	Webinar	3054014.77680492
5	Enterprise	Adwords	2102306.94366455
6	Commercial	Website	2065540.72879028
7	Enterprise	Website	1932865.34378052
8	Enterprise	Event	1590647.7177124
9	Commercial	Adwords	1477457.2443924
10	Mid Market	Adwords	784235.085342407
11	Mid Market	Website	731036.247642517
12	Commercial	Webinar	359892.684417725
13	Commercial	Event	311628.624816895
14	Mid Market	Webinar	184679.842285156
15	Mid Market	Event	155773.568603516

The **Enterprise** segment sourced the most pipeline, approximately **$20M** out of the total $48M. Out of the $33M sourced by the Outbound channel, **$20M** is attributed to the enterprise segment.

Q. What information do you need to know to understand the ROI (return on investment) of each channel?

- Cost of each marketing and sales channel.
- Conversion rates per channel

Profit = Total Pipeline Amount – Spend
ROI = (Profit / Spend) * 100

Q. 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?

- Ensure touchpoints are within the 90-day window relative to opportunity creation dates.
- Validate the uniqueness and consistency of IDs across tables.
- Check for any missing or null values in critical fields like pipeline amount and touchpoint dates. Verify data completeness, accuracy, and consistency by checking for missing values, comparing against reference data, and ensuring consistent formats.
- Ensure correct mapping between contacts, accounts, and opportunities.
- Maintain clear documentation of data lineage, ensure transparency in validation procedures, and keep stakeholders regularly informed about the status and any issues.
- Develop and execute unit and integration tests for ETL processes and implement robust data governance policies with routine audits and reviews.
- Continuously enhance data quality measures by actively seeking feedback from stakeholders.

ARCHITECTING A ROBUST ATTRIBUTION SYSTEM

Q. Ingest, transform and map interaction and opportunity data from several sources

Ingest Data: Collect interaction and opportunity data from multiple sources such as marketing platforms, sales tools.

- **Identify Data Sources:** Determine the various data sources, such as databases, APIs, and flat files (like CSV or JSON) etc.
- **Connection Setup:** Establish connections to these data sources using the appropriate tools and libraries (e.g. requesting libraries for APIs)
- **Data Extraction:** Extract data from these sources through methods like executing SQL queries, making API calls, or reading from files

Transform Data: Clean and preprocess the data to ensure consistency and accuracy.

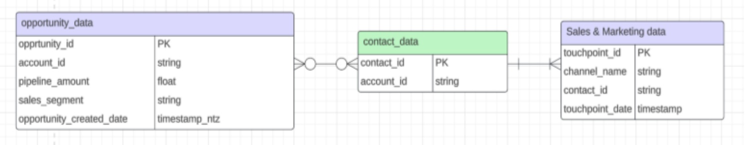
- **Data Cleaning:** Manage missing values, remove duplicate entries, and correct data types.
- **Normalization:** Convert data into a uniform format for consistency.
- **Aggregation:** Summarize the data when necessary, such as grouping and aggregating.
- **Join/Merge Data:** Combine data from multiple sources using common keys.

Map Data: Map interactions to opportunities based on relevant criteria such as time windows and account relationships.

- **Schema Mapping:** Establish a schema for the target data storage, such as a data warehouse schema.
- **Data Transformation Logic:** Align the transformed data with the target schema.
- **Load Data:** Insert the transformed and mapped data into the target storage system, such as a database, data warehouse, or data lake.

In the given case, observations for data model are as follows:

- An opportunity is linked to a single account, but an account can have multiple opportunities.
- A contact is associated with one account, but an account can have multiple contacts.
- A touchpoint is linked to a single contact, but a contact can have multiple touchpoints.



opportunity_data
opprtunity_id PK
account_id string
pipeline_amount float
sales_segment string
opportunity_created_date timestamp_ntz

contact_data
contact_id PK
account_id string

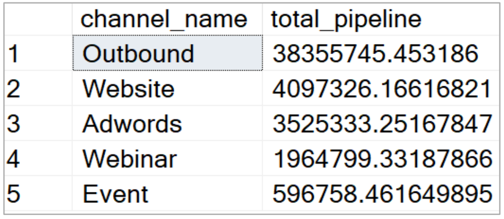
Sales & Marketing data
touchpoint_id PK
channel_name string
contact_id string
touchpoint_date timestamp

Q. Apply several different attribution models

Last Touch Attribution

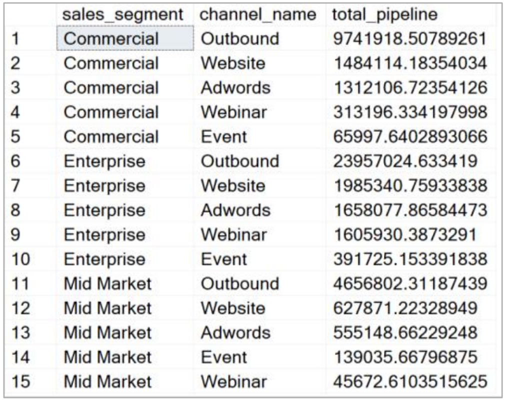
- Assigns all the conversion value to the final touchpoint before the conversion.
- Pros: Emphasizes the last interaction that led to the conversion.
- Cons: Ignores the impact of previous touchpoints throughout the customer journey.

Total Pipeline Sourced by Each Channel



	channel_name	total_pipeline
1	Outbound	38355745.453186
2	Website	4097326.16616821
3	Adwords	3525333.25167847
4	Webinar	1964799.33187866
5	Event	596758.461649895

Pipeline Sourced by Each Channel by Sales Segment

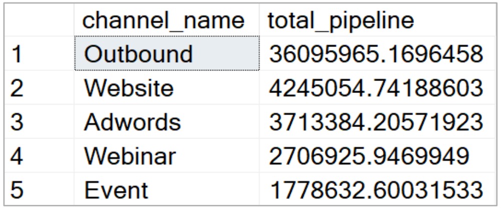


	sales_segment	channel_name	total_pipeline
1	Commercial	Outbound	9741918.50789261
2	Commercial	Website	1484114.18354034
3	Commercial	Adwords	1312106.72354126
4	Commercial	Webinar	313196.334197998
5	Commercial	Event	65997.6402893066
6	Enterprise	Outbound	23957024.633419
7	Enterprise	Website	1985340.75933838
8	Enterprise	Adwords	1658077.86584473
9	Enterprise	Webinar	1605930.3873291
10	Enterprise	Event	391725.153391838
11	Mid Market	Outbound	4656802.31187439
12	Mid Market	Website	627871.22328949
13	Mid Market	Adwords	555148.66229248
14	Mid Market	Event	139035.66796875
15	Mid Market	Webinar	45672.6103515625

Linear Attribution

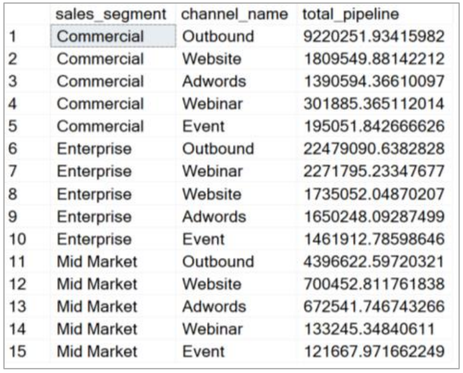
- Allocates the conversion value evenly among all touchpoints in the customer journey.
- Acknowledges the contribution of each interaction in the conversion process.
- Might overemphasize less important touchpoints while underestimating more significant ones.

Total Pipeline Sourced by Each Channel



	channel_name	total_pipeline
1	Outbound	36095965.1696458
2	Website	4245054.74188603
3	Adwords	3713384.20571923
4	Webinar	2706925.9469949
5	Event	1778632.60031533

Pipeline Sourced by Each Channel by Sales Segment



	sales_segment	channel_name	total_pipeline
1	Commercial	Outbound	9220251.93415982
2	Commercial	Website	1809549.88142212
3	Commercial	Adwords	1390594.36610097
4	Commercial	Webinar	301885.365112014
5	Commercial	Event	195051.842666626
6	Enterprise	Outbound	22479090.6382828
7	Enterprise	Webinar	2271795.23347677
8	Enterprise	Website	1735052.04870207
9	Enterprise	Adwords	1650248.09287499
10	Enterprise	Event	1461912.78598646
11	Mid Market	Outbound	4396622.59720321
12	Mid Market	Website	700452.811761838
13	Mid Market	Adwords	672541.746743266
14	Mid Market	Webinar	133245.34840611
15	Mid Market	Event	121667.971662249

Q. Pipe outputs into an easily accessible and usable system for end stakeholder consumption. End stakeholders include executive leadership, sales, marketing and finance.

Providing end stakeholders access to the table through a data warehouse like Snowflake. This approach simplifies maintaining table access and easily integrates the data with data visualization tools like Tableau or MS Power BI

Q. What data platforms will you use and what will your data stack look like? Why?

Utilize SQL queries for data extraction and summarization, or Python for scripting.

Use Airflow for orchestration, Snowflake as the data warehouse for stakeholders to access the data.

Q. Identify risks and vulnerabilities in your system.

- Inconsistent or incomplete data can lead to incorrect attribution results.
- Increasing data volume and complexity can impact performance.
- Changes in the data type of certain source data points could block the Airflow job and cause data delays.
- Choosing an inadequate server size for running Airflow could disrupt the pipeline.
- Costs may fluctuate based on data warehouse requirements.

Q. How will you maintain and scale the system?

- Regular Monitoring and Maintenance:
  Use monitoring tools to keep track of data pipelines and system performance.
  Schedule regular maintenance to apply updates and optimizations.
- Scalable Infrastructure:
  Leverage cloud services with auto-scaling capabilities to handle variable data loads.

- Continuous Improvement:
  Gather feedback from stakeholders and improve the system based on their needs.
  Implement version control and CI/CD (Continuous Integration/Continuous Deployment) for code and infrastructure changes.
- Training and Documentation:
  Provide comprehensive documentation and training for users and developers.
  Ensure stakeholders understand how to use the system and interpret the data accurately.