Marketing Analytics Pipeline on AWS for Calendly KPI Reporting: Pipeline Documentation

Business Goal and Use Case

Objective

The primary objective of this project is to construct a robust Data Engineering pipeline leveraging AWS cloud technologies and services to ingest, process, and analyze marketing and Calendly event scheduling data. The pipeline is designed to deliver critical business insights and facilitate strategic decision-making informing marketing strategies through advanced analytics and visualization supported reporting.

Background and Context

For organizations that rely heavily on digital marketing campaigns across various channels (such as Facebook, YouTube, and TikTok) to drive customer engagement and lead generation, understanding the efficiency and effectiveness of marketing activities is paramount. Each marketing channel incurs substantial expenditures; thus, it's vital to clearly comprehend which channels offer optimal returns on investment (ROI) in terms of lead acquisition and eventual conversion into booked meetings.

Calendly—a platform extensively used within organizations to manage appointments and client interactions—generates a wealth of data related to scheduled meetings, cancellations, and attendee interactions. Leveraging this data alongside marketing spend information will provide a granular perspective on how marketing strategies translate directly into actual client engagements, appointments, and potential revenue streams.

Core Business Problems Being Addressed

The project targets specific business questions and problems that are critical to the marketing and operational strategy of organizations depending on digital marketing:

1. **Marketing Spend Optimization:**
    - How effective are current marketing channels in generating valuable leads?
    - What is the cost-effectiveness (Cost Per Booking or CPB) of each marketing channel?
    - Which marketing campaigns generate the most efficient lead conversions?
2. **Lead and Conversion Analysis:**
    - What are the volume and quality of leads generated across different marketing channels?

    - How does lead volume fluctuate over time, and what patterns or seasonal trends emerge?
    - How are scheduled meetings translating into potential business opportunities?

3. **Operational and Scheduling Efficiency:**

    - What insights can be obtained from scheduling patterns to optimize resource allocation?
    - How is the scheduling load distributed among employees, and are there instances of workload imbalance or potential burnout risks?

4. **Strategic Decision Making:**

    - Can the organization better align future marketing spend and calendar schedules based on insights derived from historical and real-time analytics?
    - How can analytics inform proactive adjustments to campaigns, enabling timely interventions to maximize performance?

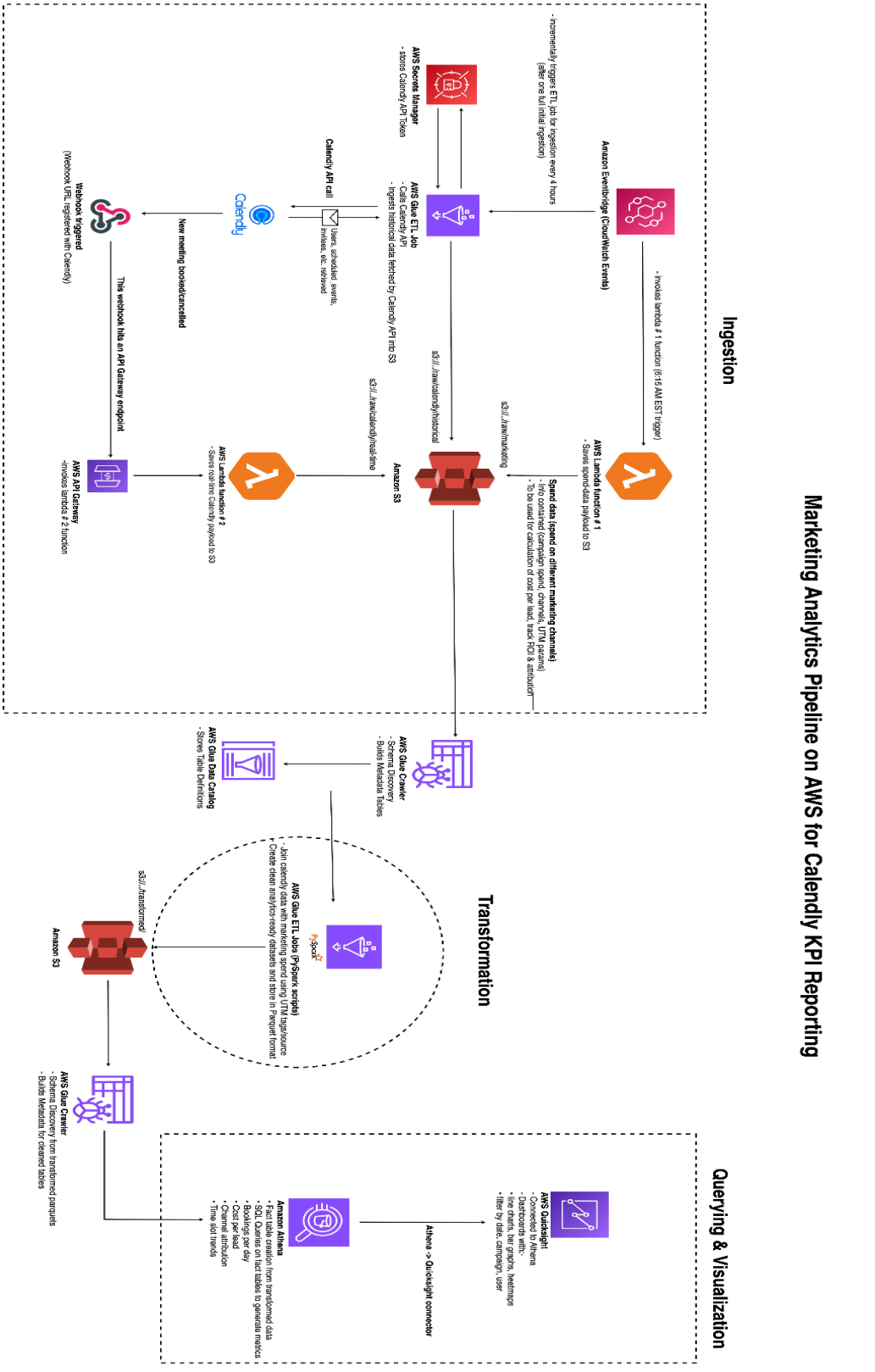
Proposed Approach and Justification

To effectively address these business questions, the proposed pipeline uses AWS cloud infrastructure to implement a modern hybrid ELT (Extract-Load-Transform) pipeline combined with traditional ETL (Extract-Transform-Load) steps. nitially, data is extracted from source systems and loaded directly into AWS S3 storage, forming raw data (Bronze tables). Subsequently, AWS Glue transforms this raw data into structured and clean datasets (Silver tables). Finally, Athena SQL queries perform additional transformations and analytics, generating analysis-ready datasets (Gold tables).

The choice of AWS and its suite of integrated services—such as Lambda, Glue, S3, Athena, API Gateway, Eventbridge and QuickSight—was driven by the following considerations:

- **Scalability and Performance:** AWS provides elastic, scalable services capable of handling high volumes of data without compromising on performance.
- **Cost Efficiency:** The pay-as-you-go pricing model allows for optimized cost control, particularly suitable for variable workloads and incremental growth.
- **Integration and Automation:** AWS services seamlessly integrate with each other, allowing automated workflows that minimize manual intervention and enhance reliability.
- **Advanced Analytics:** AWS Athena and Glue offer powerful, serverless solutions for data transformation and querying, making sophisticated analytics achievable without complex infrastructure management.
- **Real-Time Insights:** The integration of API Gateway and Lambda enables real-time data ingestion and analytics, ensuring insights are timely and actionable.

Pipeline Workflow and Technologies used



I. Overall Design

The pipeline adopts a Lambda Architecture, a robust and scalable design pattern that effectively integrates batch processing for historical data with real-time processing for webhook captured data. This architecture ensures comprehensive, accurate, and timely analytics by processing large volumes of historical data alongside continually streaming real-time data, thus providing both detailed historical insights and immediate responses to recent activities.

II. Data Ingestion

1. Lambda Architecture:

Lambda Architecture supports:

- **Historical Data (Batch)**: Managed using AWS Glue, capable of handling extensive volumes of data, lambda has a timeout period of 15 minutes, therefore a python script in Glue was used.
- **Real-time Data (Stream)**: Managed via AWS Lambda and API Gateway (hit by a calendly webhook whenever a meeting is booked/cancelled) to ensure timely data ingestion and immediate data availability for analytics.

2. Three-fold Data Ingestion Strategy:

A. Marketing Spend Data:

- **AWS EventBridge**: Scheduled daily at 6:15 AM, triggering a Lambda function.
- **Lambda Function**: Retrieves spend data from a public S3 bucket (`s3://{bucket_name}/calendly_spend_data/`) and stores it in `raw/marketing` within the project's S3 bucket.
- **Purpose**: Analyzing cost-efficiency metrics such as Cost Per Booking (CPB).

B. Historical Scheduled Events Data:

- **AWS Glue Job (`glue_calendly_ingestion.py`)**: Performs a comprehensive historical data ingestion from Calendly API (API token saved in AWS Secrets manager), storing data in `raw/calendly/historical`. Incremental updates are automated every 4 hours via AWS EventBridge.
- **Reasoning**: AWS Glue is preferred over Lambda due to Lambda's limited execution window, while Glue accommodates extensive data processing needs.

C. Real-time Calendly Events Data:

- **Webhooks Integration**: Webhook triggers AWS API Gateway, invoking Lambda to ingest real-time booking/cancellation json payload into `raw/calendly/real-time`.

3. Raw Table Formation (Bronze Layer):

AWS Glue Crawlers establish initial schemas and metadata:

- `calendly-historical-crawler`
- `calendly-realtime-crawler`
- `marketing-spend-crawler`

III. Data Transformation (Silver Layer)

AWS Glue **PySpark** jobs structure and cleanse raw data into analytics-ready formats:

- **Marketing_spend_etl**:
  - Merges and deduplicates spend data.
  - Latest entries per date/channel retained.
  - Output stored as Parquet in `transformed/marketing_spend`.
- **Scheduled_events_historical_etl**:
  - Flattens nested JSON event data.
  - Computes meeting durations.
  - Stores processed historical events data in Parquet format at `transformed/scheduled_events`.
- **Webhook_invitee_events_etl**:
  - Processes real-time nested JSON.
  - Renames and reformats fields (timestamps, event details).
  - Clean data stored in Parquet format at `transformed/webhook_invitee_events`.

Silver Table Crawlers:

AWS Glue Crawlers catalog transformed Parquet files:

- `crawl-transformed-scheduled-event`
- `crawl-transformed-spend-data`
- `crawl-transformed-webhook-events-realtime`

IV. SQL Queries for Gold Tables (Athena)

Dimension Table:

- **event_type_dim**: Associates event_type URLs to channels (Facebook, YouTube, TikTok) simplifying channel attribution.

Fact Tables:

- **Scheduled_events_fact**:
    - In this transformation, the cleaned_scheduled_events (silver) dataset is converted into the scheduled_events_fact (gold) table by parsing and casting string-based timestamps into proper TIMESTAMP and DATE types, filtering for active events with non-null channel assignments, and enriching each record via a left join to the event_type_dim to incorporate channel and campaign attributes. The results are written in Parquet format to an S3 fact directory to ensure efficient storage, query performance, and consistency in downstream analytics.
- **Webhook_invitee_fact**:
    - The cleaned_webhook_invitee_events (silver) dataset is promoted to the webhook_invitee_fact (gold) table by selecting core invitee attributes (event and invitee identifiers, names, emails, timestamps), renaming the webhook_created_at field to real_time_ingestion_timestamp, and deriving a real_time_ingestion_date partition via date_trunc. Each record is enriched through a left join to event_type_dim to attach channel and campaign context, filtered to include only events with defined channels, and stored in Parquet format—partitioned by ingestion date—to optimize query performance and support real-time analytics.
- **Marketing_spend_fact**:
    - In this transformation, the cleaned_marketing_spend (silver) dataset is converted into the marketing_spend_fact (gold) table by casting the raw spend values to DOUBLE for consistent numeric analysis, truncating and casting the spend_date to DATE for daily granularity, and writing the output in Parquet format partitioned by spend_date in S3. This structuring ensures optimized storage, efficient querying, and clear temporal partitioning for downstream spend analytics.

V. Metrics Calculation (Athena SQL Queries)

After the fact tables (gold layer) is ready, crucial business metrics are calculated including:

- **Daily Calls Booked by Source**: Quantifies daily bookings per marketing channel.
- **Cost Per Booking (CPB)**: Evaluates cost-effectiveness across channels.
- **Booking Trends Over Time**: Visualizes trends and identifies seasonal effects.
- **Channel Attribution**: Measures campaign effectiveness through detailed ROI analysis.
- **Booking Volume by Time Slot/Day**: Supports optimized scheduling and resource allocation.
- **Employee Meeting Load**: Monitors employee workload distribution, aiding in operational efficiency.

VI. Data Visualization (QuickSight Dashboard)

Athena is connected with Amazon QuickSight so that direct visualization from the SQL queries is made possible. The QuickSight Dashboard offers interactive visual insights including but not limited to the following:

- **Booking Trends**: Line and area charts illustrate trends over time.
- **Cost Analysis**: KPIs and bar charts show CPB and expenditure insights.
- **Channel Performance**: Heatmaps and leaderboards rank channels by efficiency.
- **Time/Day Patterns**: Heatmaps/histograms reveal booking patterns.
- **Employee Workload**: Employee-specific workload charts highlight capacity and workload distribution.