Documentation: AWS Data Pipeline for Attribution Modeling & Performance Analysis

Overview

This document outlines the architecture, implementation, and optimization of an industrial-level data pipeline using AWS services. The pipeline processes sales, product, and customer data, applying attribution modeling, performance analysis, and strategic recommendations for a mobile app.

Objectives

- Extract data from MySQL and Microsoft databases.
- Store raw data in Amazon S3.
- Transform and process data using AWS Glue and PySpark.
- Query data using AWS Athena.
- Generate insights on total spend, impressions, conversion rates, and channel performance.
- Optimize advertising spend based on performance metrics.

Architecture

Data Flow

1. **Data Extraction:** AWS Lambda extracts data from MySQL and Microsoft databases.
2. **Storage:** Raw data is stored in an S3 bucket.
3. **Data Processing:** AWS Glue ETL jobs transform the data using PySpark.
4. **Data Querying:** AWS Athena is used to query transformed data.
5. **Insights & Reporting:** Data is analyzed to optimize ad spend and conversion strategies.

Technologies Used

- **AWS Lambda:** Extracts data from databases and uploads it to S3.
- **Amazon S3:** Central storage for raw and processed data.
- **AWS Glue:** ETL transformation using PySpark.
- **AWS Athena:** Query engine for S3 data.
- **Apache Airflow:** Orchestrates the workflow and schedules ETL processes.

Implementation Steps

1. Data Extraction

- Configure AWS Lambda to connect to MySQL and Microsoft databases.
- Extract relevant data fields related to spend, impressions, and clicks.
- Store extracted data in an S3 bucket in CSV/JSON format.

2. Data Ingestion & Transformation

- **Create a Glue Crawler** to automatically detect schema and create tables in the Glue Data Catalog.
- **Develop Glue ETL Jobs** using PySpark to clean and transform the data:

    - Convert raw data into Parquet format for efficient querying.
    - Aggregate data to compute key metrics (e.g., total spend, total impressions, conversion rates).
    - Join datasets to correlate spend with conversion performance.

3. Querying Data with Athena

- Use AWS Athena to run SQL queries on transformed data stored in S3.
- Compute key KPIs:
    - Impression-to-install ratio per channel.
    - Top-performing campaigns and channels.
    - Return on ad spend (ROAS) per channel.
    - Click-through and conversion rates.

4. Strategic Recommendations Based on Insights

Increase Investment in High-Performing Channels

- Google Ads has a strong conversion rate—allocate more budget to maximize ROI.
- Focus on **Facebook, Instagram, and Google Ads** for static content.

Reduce Spend on Low-Performing Channels

- Minimize investment in Twitter due to low performance.
- Reallocate funds from underperforming channels to those with higher conversion rates.

Optimization & Performance Tuning

AWS Glue Optimization

- Enable **Dynamic Frame Pruning** to improve performance on large datasets.
- Use **Partitioning and Bucketing** to enhance Athena query performance.
- Optimize PySpark jobs by **distributing workloads efficiently**.

Apache Airflow Optimization

- Configure **DAG scheduling** to optimize job execution time.
- Implement **failure notifications** for robust monitoring.
- Use **parallel processing** to speed up ETL tasks.

Conclusion

This AWS-based data pipeline enables efficient data extraction, transformation, and analysis for attribution modeling. The insights derived from the pipeline facilitate strategic ad spend allocation, improving overall campaign performance and ROI.