|  |
| --- |
| Data Pipeline for Data-Driven YouTube Campaign |
| July 12, 2024 |

# Overview

## Project background and description

|  |  |
| --- | --- |
| This project aims to develop a data pipeline to support a data-driven YouTube advertising campaign for our new product launch. The pipeline will leverage a cloud-based architecture to ingest and process data from various sources to inform video categorization, audience targeting, and campaign performance optimization. |  |

## Project scope

This project focuses on building a data pipeline using a cloud-based architecture. The pipeline will:

* **Extract data from:**
  + - YouTube Data API: Video comments, statistics (views, likes, dislikes, demographics)
    - Internal data sources: Customer demographics, product information
* **Transform data to:**
  + - Clean and standardize data from different sources
    - Analyze video comments for sentiment and product relevance using Natural Language Processing (NLP)
    - Enrich video data with audience insights from internal data
* **Load data into:**
  + - Data warehouse for further analysis and reporting
    - Campaign management platform for ad targeting

## High-level requirements

* Develop data pipelines to extract, transform, and load data from:
  + YouTube Data API
  + Internal data sources (customer, product data)
* Utilize Natural Language Processing (NLP) techniques for sentiment analysis of YouTube video comments.
* Integrate with a cloud-based data warehouse for data storage and further analysis.
* Integrate with a campaign management platform for audience targeting based on video categorization and audience insights.

## Deliverables

* Functional data pipelines for ingesting and processing data (refer to Figure 1: Data Pipeline Architecture).
* Data quality reports ensure data accuracy and completeness.
* Documentation outlining the data pipeline architecture and processes.

## Exclusions

|  |  |
| --- | --- |
| * This project excludes developing the YouTube ad creatives or managing the advertising campaign itself. |  |

## Implementation plan

The project will be implemented in phases:

* **Phase 1:** Design and develop data pipelines for YouTube Data API and internal data sources (2 weeks).
* **Phase 2:** Integrate NLP for comment analysis and data enrichment (1 week).
* **Phase 3:** Develop data pipeline integration with data warehouse and campaign management platform (1 week).
* **Phase 4:** Testing and deployment of the data pipeline (1 week).

## High-level timeline/schedule

* Project Kickoff: July 12, 2024
* Completion of Data Pipelines (Phases 1 & 2): July 24, 2024
* Integration with Data Warehouse & Campaign Platform (Phase 3): July 31, 2024
* Data Pipeline Testing & Deployment (Phase 4): August 7, 2024

## Data Pipeline Architecture (Figure 1)

A screenshot of a computer

Description automatically generated

**Figure 1: Data Pipeline Architecture**

A visual representation of the data pipeline architecture is included in Figure 1. The diagram depicts the flow of data as it progresses through various stages:

* **Source Systems:** Represent various data sources like YouTube Data API and internal databases.
* **Landing Area:** Stores raw data temporarily in an S3 bucket before processing.
* **Data Processing:** Utilizes services like AWS Glue and AWS Lambda to transform raw data. This might involve tasks like:
  + - Joining data from disparate sources.
    - Filtering and cleaning data.
    - Performing NLP sentiment analysis on comments.
    - Enriching video data with audience insights.
* **Cleansed/Enriched/Analytical Data:** Represents the transformed data ready for loading.
* **Data Warehouse:** Represents the data warehouse (e.g., Amazon Redshift) for storing and analyzing processed data.
* **Analytical Data Access:** Represents access to data in the warehouse for further analysis and reporting.
* **Data Flow:** Arrows depict the data flow throughout the pipeline.
* **Monitoring/Alert:** Represents monitoring and alerting tools to ensure smooth operation (e.g., Amazon CloudWatch).

**Benefits of a Data Pipeline**

This data pipeline offers several benefits for the YouTube campaign:

* **Automated Data Collection:** Streamlines data gathering from various sources, reducing manual effort and improving efficiency.
* **Improved Data Quality:** Ensures data consistency and accuracy through cleaning and transformation processes.
* **Data-Driven Targeting:** Enables precise audience targeting based on video categorization and audience insights.
* **Campaign Optimization:** Allows for ongoing monitoring and analysis of campaign performance for data-driven adjustments.

By implementing this data pipeline, we can gain valuable insights from YouTube data and internal data sources to ensure a successful and targeted launch campaign for our new product.

# Approval and Authority to Proceed

We approve the project as described above, and authorize the team to proceed.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Approved By | Project Manager | Date | Approved By | Data Engineer | Date |