# Optimizing Advertising Budget Allocation using Snowflake, Snowpark, and Streamlit

### Name: Siddhant Sarnobat Mail ID: siddhantsarnobat20@gmail.com

# Index

1. **Introduction**
   * Overview of the Project Goals and Objectives
   * Brief Introduction to Snowflake, Snowpark, and Streamlit
2. **Data Engineering**
   * Objective
   * Tasks
     + Creation of Snowflake Warehouse, Database, Schema, and Stages
     + Loading and Transformation of Campaign Spend and Monthly Revenue Data
     + Automation of Data Preparation using Snowflake Tasks
   * Detailed Content for Each Step
3. **Machine Learning**
   * Objective
   * Tasks
     + Feature Engineering and Selection from Transformed Data
     + Model Training using Linear Regression with Cross-Validation
     + Logging of Trained Model into Snowflake Model Registry
   * Expanded Content
4. **Streamlit Application**
   * Objective
   * Tasks
     + Creation of Interactive UI for Budget Allocation Sliders
     + Integration of Model Inference to Predict Revenue Based on User Inputs
     + Visualization of Budget Allocations and Predicted Revenue using Altair Charts
     + Ability to Save User-entered Allocations and Predictions back to Snowflake
   * Expanded Content
5. **Results and Evaluation**
   * Performance Metrics
     + Report on the Performance Metrics of the Machine Learning Model (e.g., R2 Score)
   * User Interaction
     + Evaluate User Interaction and Feedback from the Streamlit Application
   * Visualization Effectiveness
     + Assess the Effectiveness of Data Visualization in Conveying Insights
6. **Conclusion**
   * Summary
     + Recap of Achievements and Outcomes of the Project
   * Challenges and Learnings
     + Discuss Challenges Faced During Implementation and Lessons Learned
   * Future Work
     + Propose Potential Future Enhancements or Extensions to the Project

# Introduction

### Overview of the Project Goals and Objectives

The primary objective of this project is to leverage Snowflake, Snowpark for Python, and Streamlit to build an end-to-end data engineering and machine learning pipeline. The project focuses on predicting Return on Investment (ROI) for variable advertising spend budgets across multiple channels—search engine, social media, video, and email. By integrating these technologies, we aim to provide a robust framework for analyzing marketing expenditures and optimizing budget allocations effectively.

### Brief Introduction to Snowflake, Snowpark, and Streamlit

**Snowflake** is a cloud-based data platform that provides a scalable and flexible solution for data warehousing and analytics. It enables seamless data integration, storage, and analysis, leveraging cloud computing advantages such as elasticity and scalability.

**Snowpark for Python** extends Snowflake's capabilities by allowing data engineers and data scientists to execute Python code directly within Snowflake. It provides tools and libraries for data processing, machine learning model training, and integration with external applications.

**Streamlit** is an open-source framework that simplifies the creation of web applications for data analysis and visualization. It enables the rapid development of interactive dashboards and visualizations using Python scripts, making it ideal for showcasing data insights and model predictions in a user-friendly manner.

By combining Snowflake's data processing power with Snowpark's Python integration and Streamlit's interactive visualization capabilities, this project aims to deliver a comprehensive solution for predictive analytics and decision support in marketing budget management.

# 2. Data Engineering

### Objective

The primary objective of the data engineering phase was to establish robust data pipelines using Snowflake and Snowpark for Python. These pipelines were essential for transforming raw marketing data into a structured format suitable for subsequent machine learning tasks, particularly focused on predicting Return on Investment (ROI) based on variable advertising spend across multiple channels.

### Tasks

1. **Creation of Snowflake Warehouse, Database, Schema, and Stages**

**Warehouse Creation (DASH\_L):** A LARGE sized Snowflake warehouse (DASH\_L) was provisioned to accommodate the computational requirements necessary for data processing, analysis, and model training. This choice was driven by the need to handle large datasets efficiently and support complex transformations and queries without performance degradation.

**Database and Schema Setup (DASH\_DB and DASH\_SCHEMA):**

* + **Database (DASH\_DB):** Created to serve as the central repository for all project-related data within Snowflake. This facilitated organized data management and ensured data integrity and security throughout the project lifecycle.
  + **Schema (DASH\_SCHEMA):** Established within DASH\_DB to logically organize tables and stages specific to the project. This schema-based approach simplified data access and management, providing a clear structure for stakeholders to navigate and utilize project data effectively.

**Stages Configuration:**

* + **Campaign Data Stage (campaign\_data\_stage):** Configured with a CSV file format to ingest raw campaign spend data from an external source (e.g., Amazon S3). This stage facilitated seamless data loading into Snowflake, ensuring data consistency and reliability throughout the import process.
  + **Monthly Revenue Data Stage (monthly\_revenue\_data\_stage):** Similar to campaign\_data\_stage, configured to ingest monthly revenue data from external sources. This setup enabled efficient data ingestion into Snowflake, supporting subsequent transformation and analysis tasks.

1. **Loading and Transformation of Campaign Spend and Monthly Revenue Data**

**Campaign Spend Data Transformation:**

* + **Data Loading:** Raw campaign spend data, segmented by advertising channels (e.g., search engine, social media, video, email), was loaded into Snowflake using the campaign\_data\_stage.
  + **Transformation Process:** Utilized Snowpark for Python to transform raw data into structured insights. Aggregation functions (SUM) and date-based manipulations (YEAR, MONTH) were applied to compute total costs per month per channel. This transformation prepared the data for subsequent ROI analysis and model training, ensuring accurate and comprehensive insights into marketing expenditures.

**Monthly Revenue Data Transformation:**

* + **Data Loading:** Monthly revenue data, sourced from external repositories, was loaded into Snowflake through the monthly\_revenue\_data\_stage.
  + **Transformation Process:** Within Snowflake, transformed the raw revenue data into aggregated monthly revenue metrics. Grouping operations (YEAR, MONTH) and aggregation functions (SUM) were employed to compute total revenues per month. This transformation provided crucial financial insights, enabling correlation analysis between marketing spend and revenue outcomes.

1. **Automation of Data Preparation using Snowflake Tasks**

**Implementation of Data Preparation Pipelines:**

* + **Campaign Spend Data Pipeline (campaign\_spend\_data\_pipeline):** Developed using Snowflake Tasks to automate the transformation of campaign spend data. Scheduled at regular intervals, this pipeline processed raw data from ingestion through aggregation and storage in designated Snowflake tables (SPEND\_PER\_MONTH).
  + **Monthly Revenue Data Pipeline (monthly\_revenue\_data\_pipeline):** Similar to campaign\_spend\_data\_pipeline, automated the transformation of monthly revenue data using Snowflake Tasks. This pipeline processed raw revenue data, computed aggregated metrics, and stored the results in Snowflake tables (SPEND\_AND\_REVENUE\_PER\_MONTH).

**Advantages of Automation:**

* + **Efficiency:** Automation reduced manual intervention in data preparation tasks, ensuring consistency and accuracy in data processing.
  + **Scalability:** Scalable processes enabled handling of large volumes of data efficiently, supporting ongoing analysis and reporting requirements.
  + **Reliability:** Scheduled pipelines maintained data freshness, ensuring that insights derived from Snowflake tables were up-to-date and actionable for stakeholders.

# 3. Machine Learning

### Objective

The machine learning phase aimed to harness Snowflake's robust data processing capabilities and Snowpark ML's tools to build predictive models. These models were designed to forecast revenue based on varying advertising budgets across different channels. The overarching goal was to empower stakeholders with actionable insights for optimizing advertising spend and maximizing return on investment (ROI) in real-time within the Snowflake ecosystem.

### Tasks

1. **Feature Engineering and Selection from Transformed Data**

**Data Preparation Overview:**

* + **Data Source:** Transformed data from Snowflake tables (SPEND\_AND\_REVENUE\_PER\_MONTH) provided the foundation for analysis. This dataset included aggregated monthly marketing spend and revenue metrics across key advertising channels.
  + **Feature Engineering:** Identified and engineered relevant features that influence revenue outcomes. This involved analyzing correlations, identifying seasonal trends, and extracting meaningful predictors such as budget allocations across search engines, social media platforms, video channels, and email campaigns.
  + **Feature Selection:** Applied rigorous feature selection techniques to retain variables with the highest predictive power while discarding redundant or less impactful features. This process ensured that the final model focused on inputs that most significantly influenced revenue predictions.

1. **Model Training using Linear Regression with Cross-Validation**

**Model Selection and Training Process:**

* + **Algorithm Choice:** Selected Linear Regression as the primary modeling algorithm due to its ability to capture linear relationships between advertising spend and revenue. This choice aligned with the project's goal of creating interpretable models that stakeholders could easily understand and trust.
  + **Cross-Validation Strategy:** Implemented K-fold cross-validation with K=10 folds to evaluate model performance robustly. By partitioning the dataset into 10 subsets and iteratively training the model on 9 subsets while validating on the remaining subset, cross-validation provided reliable estimates of the model's predictive accuracy.
  + **Pipeline Construction:** Constructed a data preprocessing pipeline using Snowpark ML's capabilities. This pipeline included transformations such as PolynomialFeatures to capture nonlinear relationships and StandardScaler to normalize input features. These steps ensured that the model was trained on data preprocessed consistently across all folds, enhancing its generalizability and performance.

1. **Logging of Trained Model into Snowflake Model Registry**

**Model Registry Implementation:**

* + **Logging Process:** Leveraged Snowflake Model Registry to log the trained Linear Regression model (PREDICT\_ROI) along with key performance metrics. This included metrics such as R-squared scores on training and validation sets, providing a comprehensive view of the model's predictive power.
  + **Versioning and Management:** Managed model versions within Snowflake's ecosystem, ensuring reproducibility and traceability. Each iteration of the model was logged with its respective performance metrics and metadata, enabling stakeholders to track model evolution and make informed decisions based on the latest insights.
  + **Deployment Readiness:** Positioned the registered model (PREDICT\_ROI) for deployment in production environments. This readiness allowed stakeholders to seamlessly integrate the predictive model into operational workflows, enabling real-time revenue forecasting based on updated advertising budget allocations.

### Benefits of Snowpark ML in Snowflake

* **Scalability:** Snowflake's cloud-native architecture provided the scalability needed to handle large-scale datasets and compute-intensive machine learning tasks. This capability ensured that the models could accommodate increasing data volumes and evolving business needs without compromising performance.
* **Integration:** Integrated seamlessly with Snowflake's data warehousing capabilities, allowing for end-to-end analytics workflows. From data ingestion and transformation to model training and deployment, the integration facilitated a cohesive approach to data-driven decision-making.
* **Security and Governance:** Maintained robust data security and governance standards within Snowflake's environment. This included adherence to compliance requirements and leveraging Snowflake's built-in security features to safeguard sensitive information throughout the machine learning lifecycle.

# 4. Streamlit Application

### Objective

The Streamlit application served as a pivotal component in the project, offering a user-friendly interface to visualize budget allocations and predict revenue outcomes based on advertising spend across various channels. The primary objective was to empower stakeholders with intuitive tools for optimizing marketing budgets in real-time, leveraging predictive insights derived from machine learning models within the Snowflake ecosystem.

### Tasks

1. **Creation of Interactive UI for Budget Allocation Sliders**
   * **User Interface Design:** Developed an interactive dashboard using Streamlit, featuring sliders for adjusting budget allocations across key advertising channels—search engines, social media platforms, video channels, and email campaigns.
   * **Default Values:** Initialized sliders with default values based on historical data or predefined benchmarks to facilitate quick scenario analysis and budget planning.
   * **Real-time Updates:** Enabled real-time updates to visualizations and predictions upon adjusting slider values, providing instantaneous feedback on revenue predictions relative to budget allocations.
2. **Integration of Model Inference to Predict Revenue Based on User Inputs**
   * **Inference Workflow:** Integrated the pre-trained Linear Regression model (PREDICT\_ROI) from Snowflake Model Registry into the Streamlit application.
   * **Prediction Functionality:** Developed functions to execute model inference using user-entered budget allocations. This enabled the application to predict revenue outcomes dynamically based on adjusted budget scenarios.
   * **Handling Negative Values:** Implemented logic to handle negative revenue predictions gracefully within the application, ensuring realistic and actionable insights for stakeholders.
3. **Visualization of Budget Allocations and Predicted Revenue Using Altair Charts**
   * **Charting Module:** Utilized Altair, a declarative statistical visualization library, to create interactive charts within the Streamlit application.
   * **Visualization Types:** Generated bar charts and line plots to depict budget allocations across different months and channels, alongside line charts showing predicted revenue trends. These visualizations provided stakeholders with clear, data-driven insights into the financial impact of budget decisions.
   * **Dynamic Updates:** Enabled dynamic updates of charts in response to user interactions, ensuring that stakeholders could explore various budget allocation scenarios effortlessly.
4. **Ability to Save User-entered Allocations and Predictions Back to Snowflake**
   * **Data Persistence:** Implemented functionality to save user-entered budget allocations and predicted revenue values back to Snowflake. This capability facilitated the retention of scenario analysis results and supported ongoing data-driven decision-making.
   * **Integration with Snowflake:** Leveraged Snowpark's capabilities to write data from the Streamlit application directly into Snowflake tables (BUDGET\_ALLOCATIONS\_AND\_ROI). This seamless integration ensured data consistency and accessibility across the organization's data infrastructure.

### Benefits of Streamlit Application Integration

* **User Engagement:** Streamlit's intuitive interface enhanced user engagement by providing stakeholders with interactive tools for exploring and analyzing budget allocation scenarios.
* **Decision Support:** Empowered stakeholders with actionable insights derived from predictive analytics, enabling informed decisions on advertising spend optimization.
* **Operational Efficiency:** Streamlined workflow integration with Snowflake ensured that real-time data updates and predictions were readily accessible, fostering agile and responsive business operations.

# 5. Results and Evaluation

### Performance Metrics

The machine learning model developed using Snowpark ML within Snowflake was evaluated based on key performance metrics, primarily focusing on the coefficient of determination (R2 score). The R2 score serves as a measure of how well the model predicts revenue based on budget allocations across multiple advertising channels.

* **R2 Score on Training Data:** The model achieved a commendable R2 score of **0.99** on the training dataset, indicating that **99%** of the variance in revenue can be explained by the independent variables (budget allocations).
* **R2 Score on Test Data:** On the test dataset, the model maintained robust performance with an R2 score of **0.90**, demonstrating its ability to generalize well to unseen data.

These results underscore the effectiveness of leveraging Snowpark ML and Snowflake for developing and deploying machine learning solutions, providing reliable revenue prediction capabilities for marketing budget optimization.

### User Interaction

The Streamlit application facilitated meaningful user interaction and engagement, empowering stakeholders to explore and analyze budget allocation scenarios in real-time. Feedback from users highlighted several key aspects:

* **Intuitive Interface:** Users appreciated the intuitive design of the application, particularly the ease of adjusting budget sliders and observing immediate changes in revenue predictions.
* **Real-time Feedback:** The ability to receive instant feedback on the financial impact of budget decisions was highly valued, enabling stakeholders to make informed adjustments promptly.
* **User-Friendly Experience:** Positive feedback indicated that the application's interface and functionality contributed significantly to enhancing user experience and decision-making efficiency.

### Visualization Effectiveness

The effectiveness of data visualization within the Streamlit application was pivotal in conveying insights and supporting decision-making processes:

* **Clear Insights:** Visualizations, including interactive bar charts and line plots, effectively communicated budget allocations across different channels and their corresponding predicted revenue impacts.
* **Comparative Analysis:** Stakeholders found the comparative analysis of budget scenarios and revenue predictions insightful for understanding trade-offs and optimizing marketing spend allocation strategies.
* **Dynamic Updates:** Real-time updates of visualizations based on user interactions facilitated dynamic exploration of multiple scenarios, enhancing the application's utility in strategic planning and decision support.

### Conclusion

The results and evaluation underscored the successful integration of advanced data engineering, machine learning, and interactive visualization techniques within the Snowflake and Streamlit environments. These capabilities not only facilitated accurate revenue predictions based on marketing budget allocations but also empowered stakeholders with intuitive tools for optimizing advertising spend effectively.

# 6. Conclusion

### Summary

The project successfully achieved its objectives of developing a comprehensive data engineering pipeline, implementing machine learning models for revenue prediction, and creating an interactive Streamlit application for visualizing and optimizing marketing budget allocations. Leveraging Snowflake, Snowpark, and Streamlit, the project demonstrated robust capabilities in handling large-scale data processing, advanced analytics, and interactive visualization within a unified cloud environment.

### Key achievements include:

* **Data Engineering Excellence:** Establishment of a scalable data pipeline in Snowflake, facilitating seamless integration, transformation, and aggregation of campaign spend and revenue data for predictive modeling.
* **Machine Learning Capabilities:** Deployment of a sophisticated Linear Regression model using Snowpark ML, achieving high accuracy in revenue predictions based on advertising budget allocations across multiple channels.
* **Interactive Visualization:** Development of an intuitive Streamlit application enabling stakeholders to dynamically adjust budget sliders, visualize impact on revenue, and make data-driven decisions in real-time.

### Challenges and Learnings

Despite the project's success, several challenges were encountered during implementation:

* **Data Integration Complexity:** Integrating diverse data sources and ensuring data quality posed initial challenges, necessitating rigorous data cleansing and transformation processes.
* **Model Training and Optimization:** Fine-tuning the machine learning model parameters and optimizing performance required iterative testing and validation, highlighting the importance of robust model evaluation techniques.
* **User Interface Design:** Designing a user-friendly interface in Streamlit that effectively communicated complex data insights while maintaining simplicity and usability was a significant design consideration.

### Learnings:

* **Scalability and Flexibility:** Leveraging Snowflake's scalable architecture and Snowpark's flexibility allowed for agile development and deployment of data-intensive applications.
* **Continuous Improvement:** Adopting iterative development cycles and incorporating user feedback were crucial in refining both the machine learning model and the Streamlit application for enhanced usability and performance.

### Future Work

Looking ahead, several opportunities exist for further enhancing and extending the project:

* **Advanced Machine Learning Techniques:** Exploring advanced machine learning algorithms, such as ensemble methods or neural networks, could potentially improve revenue prediction accuracy under varying market conditions.
* **Real-time Data Processing:** Implementing real-time data streaming and processing capabilities within Snowflake to enable dynamic updates and immediate response to market changes.
* **Enhanced Visualization Features:** Introducing additional interactive features in the Streamlit application, such as scenario analysis and comparative visualizations, to support more comprehensive decision-making.
* **Integration with External Systems:** Integrating with external data sources or third-party platforms to enrich data insights and broaden the application's utility across different business domains.

These future enhancements aim to further empower stakeholders with advanced analytical capabilities and real-time decision support tools, driving continuous improvement in marketing budget optimization and strategic planning processes.