Customer Analysis Project

Prepared by:

Amany Ehab

Esraa Zidan

Malk Younis

Ahmed Refaat

Ahmed Mohamed

Zakaria Shaban

Moner Mohamed

Supervised by: Eng. Mohamed Gomaa

Course: Microsoft Data engineer

Table of Contents

---

1. Introduction

This project aims to analyze customer data by leveraging advanced data engineering techniques to extract valuable insights. The analysis focuses on understanding customer behavior patterns, predicting future purchasing trends, and identifying segments for targeted marketing campaigns.

---

2. Objective

The main objectives of this project are:

- **Customer Segmentation**: Group customers into meaningful segments based on their purchasing behavior and demographics.
- **Predictive Analytics**: Use machine learning models to predict customer lifetime value (CLV) and churn probability.
- **Business Insights**: Provide actionable insights to improve decision-making in marketing and sales strategies.

---

3. Data Sources

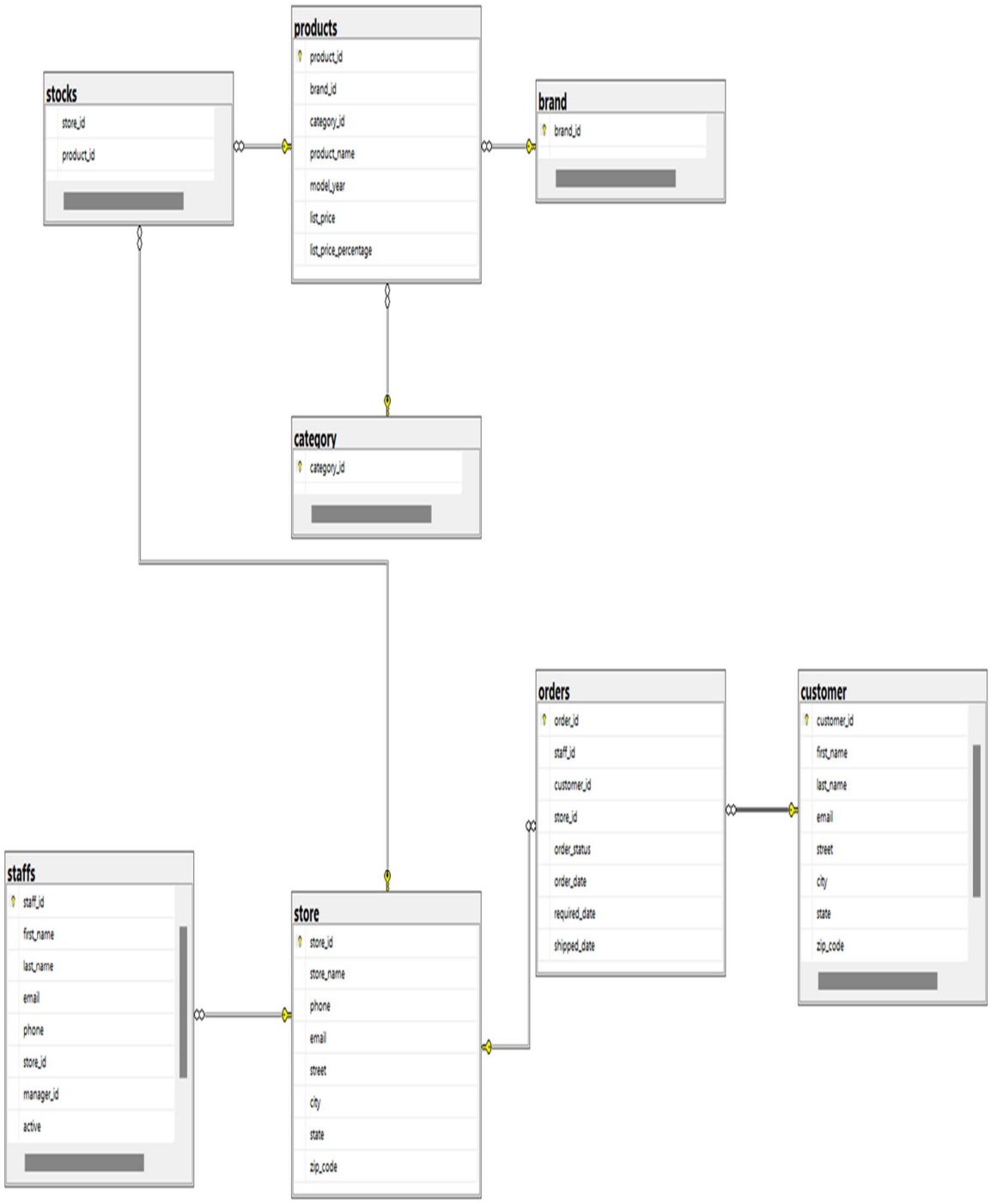
Internal Data

1. **Customer Transactions**: Historical data from customer purchases, including transaction timestamps, products, quantities, and payment methods.
2. **CRM Data**: Customer profiles, demographics, and loyalty program participation.
3. **Support and Feedback Data**: Customer interactions with support teams, including complaints, inquiries, and satisfaction ratings.

External Data

1. **Social Media**: Customer engagement with brand posts, sentiment analysis from comments and interactions.
2. **Market Data**: External market reports providing insights on industry trends, competitor activity, and pricing strategies.
3. **Web Analytics**: Clickstream data from the company's website, including browsing history, search queries, and product views.

ERD diagram


products
product_id
brand_id
category_id
product_name
model_year
list_price
list_price_percentage
stocks
store_id
product_id
brand
brand_id
category
category_id
orders
order_id
staff_id
customer_id
store_id
order_status
order_date
required_date
shipped_date
customer
customer_id
first_name
last_name
email
street
city
state
zip_code
staffs
staff_id
first_name
last_name
email
phone
store_id
manager_id
active
store
store_id
store_name
phone
email
street
city
state
zip_code

---

4. Data Pipeline Architecture

The data pipeline ensures efficient extraction, transformation, and loading (ETL) of data from multiple sources, ensuring consistency and quality for analysis.

4.1 Data Ingestion

- **Batch Ingestion**: Data is extracted periodically (e.g., daily or weekly) from databases, flat files, and APIs. This includes transaction logs and CRM data.
- **Real-Time Ingestion**: For real-time event processing, Apache Kafka is used to stream web analytics and social media data continuously.

4.2 Data Storage

- **Data Lake**: Raw data is stored in a distributed Hadoop cluster (HDFS) for scalability and flexibility. Data is stored in Parquet format for efficient querying.
- **Data Warehouse**: Processed and transformed data is stored in a columnar database (Google BigQuery or Amazon Redshift) for fast, ad-hoc queries.

4.3 Data Processing Framework

- **Apache Spark**: Used for distributed data processing and transformation. Spark jobs are scheduled via Apache Airflow to handle large datasets efficiently.

- **Workflow Orchestration**: Airflow manages the scheduling of Spark jobs, ensuring that data ingestion, transformation, and analysis are executed in the correct order.

4.4 Data Transformation

Data goes through various stages of transformation, including cleaning, normalization, and enrichment before it is ready for analysis.

---

5. Tools and Technologies

Data Ingestion and Processing

- **Apache Kafka**: Real-time data ingestion from web analytics and social media platforms.
- **Apache Nifi**: Automates data movement between systems, handling batch ingestion.
- **Python (Pandas, NumPy)**: Used for data cleaning, exploratory data analysis (EDA), and feature engineering.

Data Storage and Management

- **HDFS (Hadoop Distributed File System)**: For storing large volumes of unstructured and structured data.
- **Google BigQuery**: A scalable, high-performance data warehouse used for storing processed data for fast querying and analytics.

Data Analysis and Visualization

- **Apache Spark MLlib**: For building machine learning models on distributed datasets.
- **Jupyter Notebooks**: Used for data exploration and model prototyping.

- **Tableau / Power BI**: For creating interactive dashboards and visualizations.

---

6. Data Ingestion and Processing

6.1 Batch Ingestion

- **Cron Jobs and Apache Airflow**: Used to schedule regular data pulls from databases and external APIs. Airflow manages the dependency and timing between different steps, ensuring data is ingested, transformed, and stored properly.

6.2 Real-Time Data Streaming

- **Apache Kafka**: Handles real-time data streams from customer interactions on the website and social media. This data includes customer clicks, time spent on pages, and sentiment analysis from social media comments.

6.3 ETL Process

- Data is extracted from various sources, loaded into HDFS (for raw storage) or directly into BigQuery (for structured storage).
- Spark processes the data in a distributed environment, transforming it into meaningful formats for analysis.

---

7. Data Transformation Process

7.1 Data Cleaning

- **Handling Missing Data**: Use techniques like mean imputation for numerical data or filling missing categorical data with the mode.

- **Removing Outliers**: Z-score or IQR-based methods are used to detect and remove extreme values that could skew the analysis.

7.2 Data Normalization

- **Standardization**: Normalize numerical features such as transaction amounts to ensure comparability across different customers.
- **Encoding**: Convert categorical variables into numerical form using one-hot encoding or label encoding for compatibility with machine learning models.

7.3 Feature Engineering

- **Derived Features**: Create new variables like "Average Order Value" and "Time Since Last Purchase" to enhance the quality of analysis.
- **Aggregation**: Aggregate customer transactions over time to compute total spend, frequency of purchases, and preferred product categories.

---

8. Data Storage and Management

Data Partitioning

- Data is partitioned by customer ID, date, and region to optimize query performance.

Data Archiving

- Older data is archived to cold storage (e.g., AWS S3 or Google Cloud Storage) for long-term storage and compliance purposes, but is still accessible for historical analysis.

---

9. Data Analysis and Machine Learning Models

9.1 Exploratory Data Analysis (EDA)

- Conducted an initial analysis to understand customer behavior patterns using Python (Pandas, Seaborn).
- Identified key trends such as seasonal spending and customer churn triggers.

9.2 Customer Segmentation

- **K-Means Clustering**: Customers were grouped into segments based on transaction frequency, average spend, and recency of purchases. This revealed key groups such as high-value customers, bargain shoppers, and churn-prone customers.

9.3 Predictive Models

- **Churn Prediction**: Logistic regression and Random Forest models were used to predict which customers are most likely to churn based on interaction frequency, time since last purchase, and customer service interactions.
- **Customer Lifetime Value (CLV)**: Regression models (Lasso and Ridge) were applied to predict the future value of customers, helping target high-value customers for retention campaigns.

9.4 Model Evaluation

- **Accuracy, Precision, Recall**: Evaluated the models using cross-validation to ensure robustness.
- **Confusion Matrix**: Used to assess model performance in churn prediction.

---

10. Results and Insights

The data analysis yielded several key insights:

- **Customer Segments**: We identified three main customer segments—high-value customers, discount-driven buyers, and churn-prone customers.
- **Churn Factors**: Customers with a low engagement rate and no purchase in the last six months had a 70% higher probability of churning.
- **Mobile vs Desktop**: Mobile users had a 25% higher conversion rate and spent 30% more on average than desktop users.

---

11. Monitoring and Performance

Monitoring Dashboard

- A live dashboard was created using Tableau, providing insights into customer engagement metrics, sales trends, and real-time churn risk for individual customers.

Model Performance Monitoring

- Deployed machine learning models are continuously monitored for performance degradation, with automated alerts triggering retraining when performance falls below a set threshold.

---

12. Future Enhancements

Data Enrichment

- Integrate additional data sources such as customer reviews and email engagement data for a more holistic view of customer behavior.

Real-Time Recommendations

- Implement a real-time recommendation engine to provide personalized product suggestions based on browsing history and purchase patterns.

---

13. Conclusion

The Customer Analysis project provided critical insights into customer behavior, segmentation, and churn prediction. By leveraging data engineering best practices and machine learning models, project can make more informed decisions to retain high-value customers, improve marketing efforts, and increase revenue.