Building a Data Ecosystem in GCP for Scalable Trend Analysis

Created by: Rhythm Billore

Introduction

Project Background & Motivation

- Inspired by an academic project focused on designing a database management system and conducting data analysis.
- The initial solution was constrained by limited tools, resulting in:
  - Lack of scalability and flexibility.
  - Challenges in handling multiple data sources and real-time data processing.

Project Extension & Objectives

- Recognized the limitations of the academic solution and reimagined it for a real-world scenario.
- Developed a scalable, end-to-end data pipeline using Google Cloud Platform (GCP) to overcome these challenges.

Solution Overview

- Implemented an integrated web portal for data upload.
- Leveraged cloud-native tools to automate data processing:
  - Google Cloud Storage for data storage.
  - Pub/Sub for event-driven data ingestion.
  - Implemented Cloud Functions, Google Cloud Run, Google Dataflow, and Composer for automated processing.
  - BigQuery for efficient data analysis.

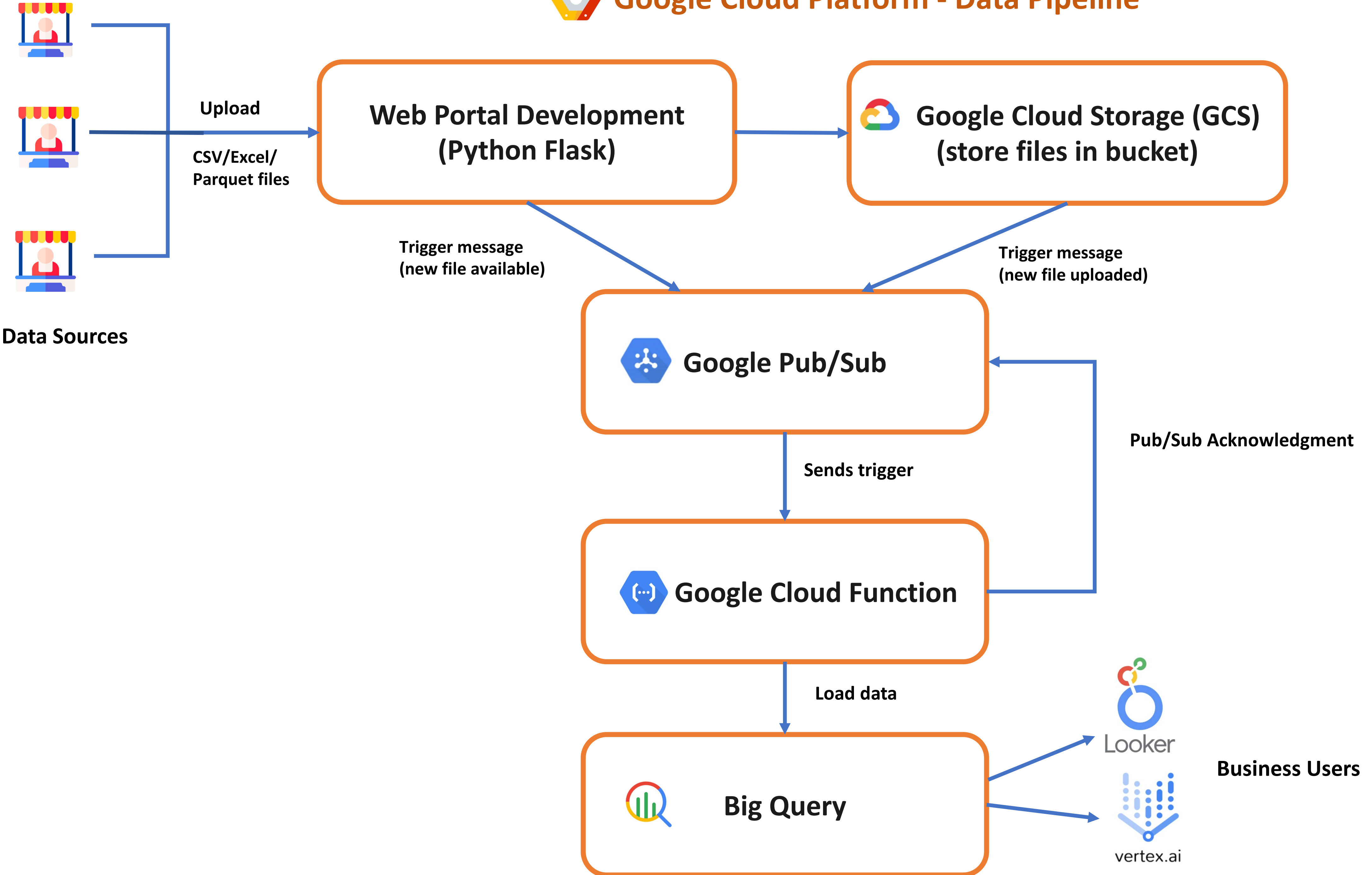
Key Benefits

- Created a robust data ecosystem enabling seamless data flow and real-time analysis.
- Scalable solution designed to meet the needs of multiple stakeholders.
- Provides business end-users with actionable insights through effective data visualization.

Deliverable

- Developed an interactive analysis dashboard to deliver in-depth insights on **Sales trends** and **customer behavior analysis**.
- Supports informed decision-making by leveraging comprehensive data analytics.

Google Cloud Platform - Data Pipeline



Data Sources

Upload

CSV/Excel/ Parquet files

Web Portal Development (Python Flask)

Google Cloud Storage (GCS) (store files in bucket)

Trigger message (new file available)

Trigger message (new file uploaded)

Google Pub/Sub

Pub/Sub Acknowledgment

Sends trigger

Google Cloud Function

Load data

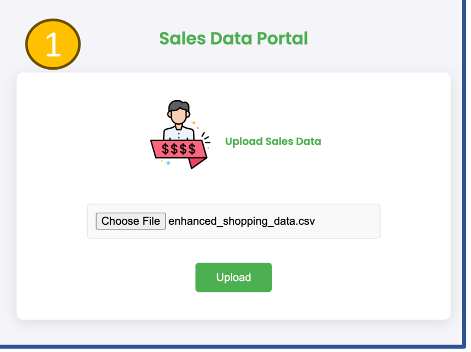
Big Query

Looker

vertex.ai

Business Users

1



Sales Data Portal

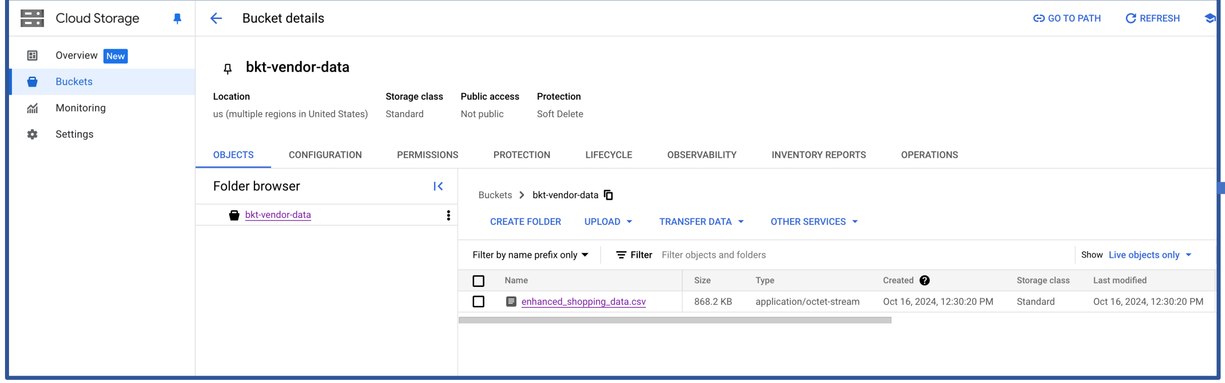
Upload Sales Data

Choose File enhanced_shopping_data.csv

Upload

2

GCS



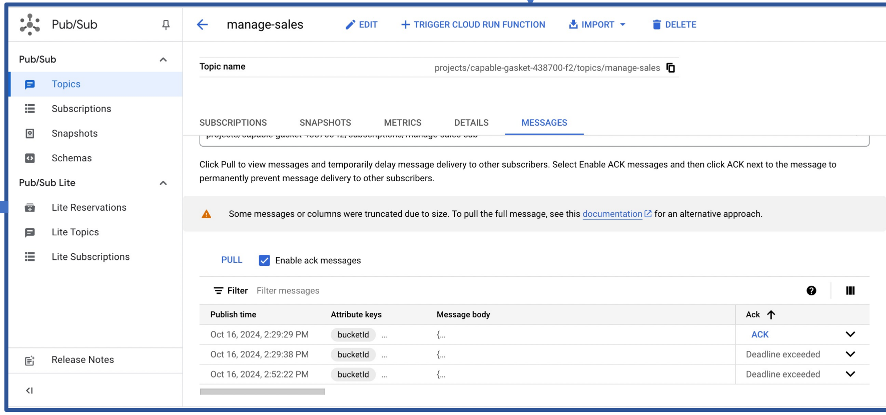
Bucket details

bkt-vendor-data

Location	Storage class	Public access	Protection
us (multiple regions in United States)	Standard	Not public	Soft Delete

□	Name	Size	Type	Created	Storage class	Last modified
□	enhanced_shopping_data.csv	868.2 KB	application/octet-stream	Oct 16, 2024, 12:30:20 PM	Standard	Oct 16, 2024, 12:30:20 PM

3 Pub/Sub



manage-sales

Topic name projects/capable-gasket-438700-f2/topics/manage-sales

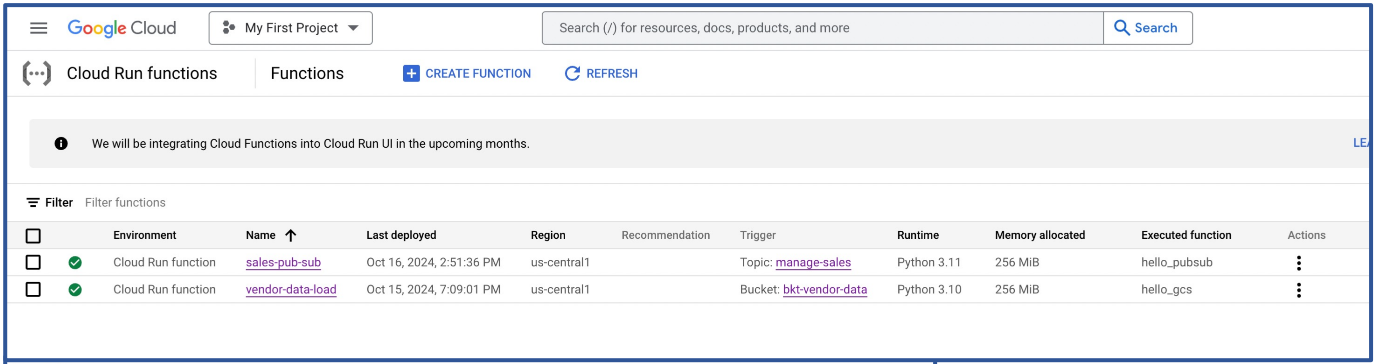
Click Pull to view messages and temporarily delay message delivery to other subscribers. Select Enable ACK messages and then click ACK next to the message to permanently prevent message delivery to other subscribers.

Some messages or columns were truncated due to size. To pull the full message, see this documentation for an alternative approach.

☑ Enable ack messages

Publish time	Attribute keys	Message body	Ack
Oct 16, 2024, 2:29:29 PM	bucketId ...	{...	ACK
Oct 16, 2024, 2:29:38 PM	bucketId ...	{...	Deadline exceeded
Oct 16, 2024, 2:52:22 PM	bucketId ...	{...	Deadline exceeded

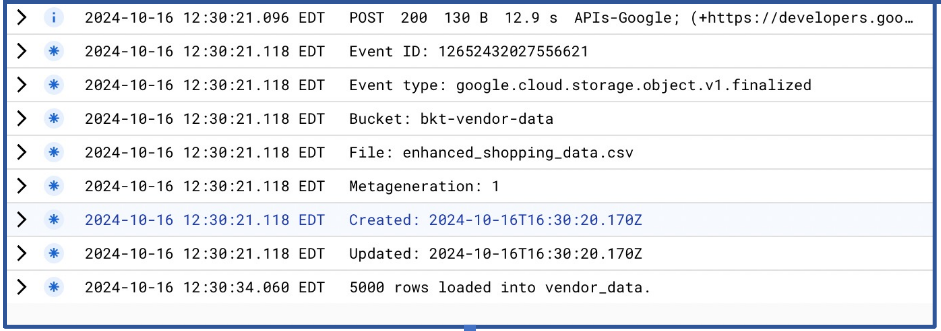
4 Cloud function



Cloud Run functions   Functions

We will be integrating Cloud Functions into Cloud Run UI in the upcoming months.

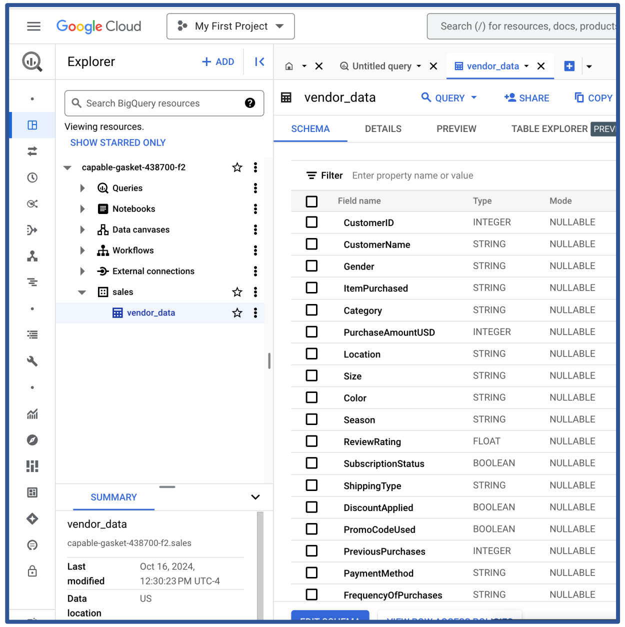
□		Environment	Name ↑	Last deployed	Region	Recommendation	Trigger	Runtime	Memory allocated	Executed function	Actions
□	✓	Cloud Run function	sales-pub-sub	Oct 16, 2024, 2:51:36 PM	us-central1		Topic: manage-sales	Python 3.11	256 MiB	hello_pubsub	⋮
□	✓	Cloud Run function	vendor-data-load	Oct 15, 2024, 7:09:01 PM	us-central1		Bucket: bkt-vendor-data	Python 3.10	256 MiB	hello_gcs	⋮



```
> i  2024-10-16 12:30:21.096 EDT   POST  200  130 B  12.9 s  APIs-Google; (+https://developers.goo…
> *  2024-10-16 12:30:21.118 EDT   Event ID: 12652432027556621
> *  2024-10-16 12:30:21.118 EDT   Event type: google.cloud.storage.object.v1.finalized
> *  2024-10-16 12:30:21.118 EDT   Bucket: bkt-vendor-data
> *  2024-10-16 12:30:21.118 EDT   File: enhanced_shopping_data.csv
> *  2024-10-16 12:30:21.118 EDT   Metageneration: 1
> *  2024-10-16 12:30:21.118 EDT   Created: 2024-10-16T16:30:20.170Z
> *  2024-10-16 12:30:21.118 EDT   Updated: 2024-10-16T16:30:20.170Z
> *  2024-10-16 12:30:34.060 EDT   5000 rows loaded into vendor_data.
```

5

Big Query



Viewing resources.

vendor_data

capable-gasket-438700-f2.sales

Last modified	Oct 16, 2024, 12:30:23 PM UTC-4
Data location	US

□	Field name	Type	Mode
□	CustomerID	INTEGER	NULLABLE
□	CustomerName	STRING	NULLABLE
□	Gender	STRING	NULLABLE
□	ItemPurchased	STRING	NULLABLE
□	Category	STRING	NULLABLE
□	PurchaseAmountUSD	INTEGER	NULLABLE
□	Location	STRING	NULLABLE
□	Size	STRING	NULLABLE
□	Color	STRING	NULLABLE
□	Season	STRING	NULLABLE
□	ReviewRating	FLOAT	NULLABLE
□	SubscriptionStatus	BOOLEAN	NULLABLE
□	ShippingType	STRING	NULLABLE
□	DiscountApplied	BOOLEAN	NULLABLE
□	PromoCodeUsed	BOOLEAN	NULLABLE
□	PreviousPurchases	INTEGER	NULLABLE
□	PaymentMethod	STRING	NULLABLE
□	FrequencyOfPurchases	STRING	NULLABLE

Sales Data Analysis - 2024

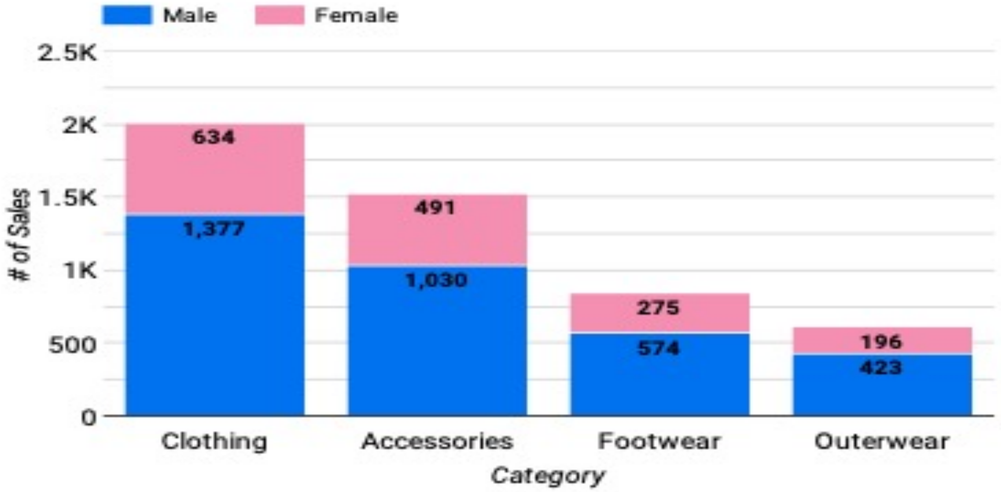
# of Sales
5000

States Covered
50

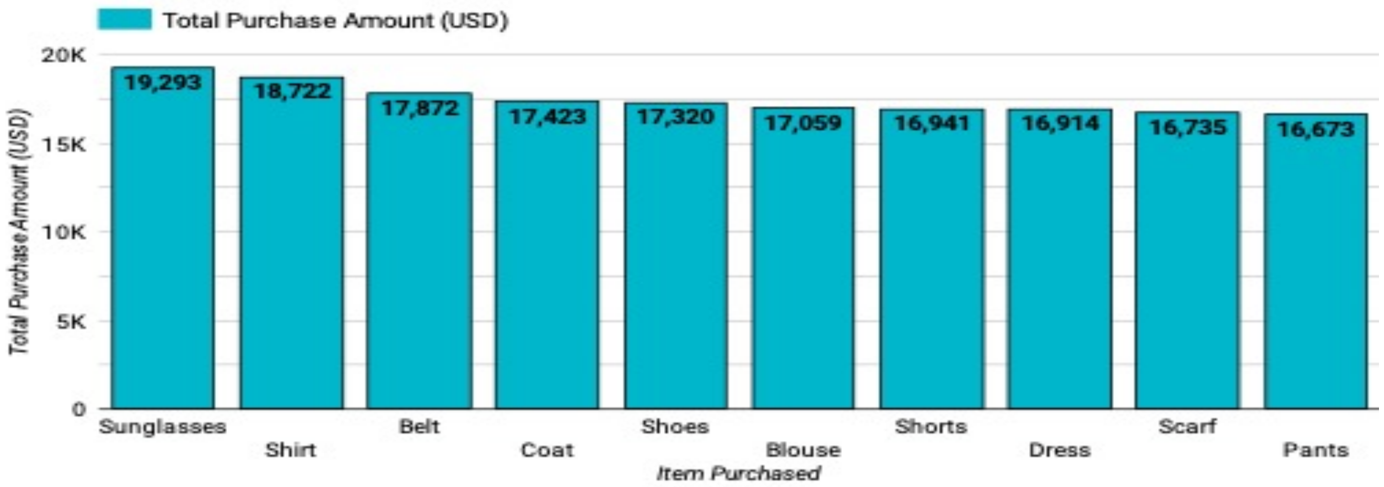
Net Revenue
402.2K

# of Active Subscribers
1,343

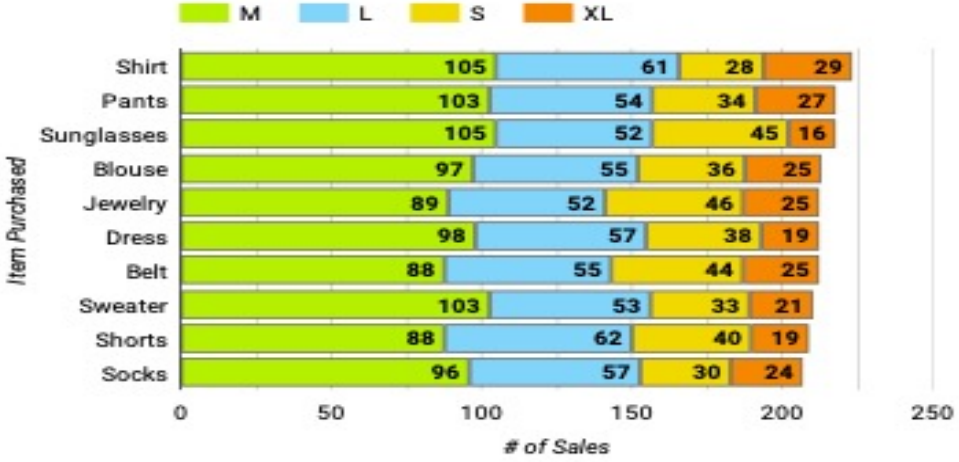
Sales Across Categories



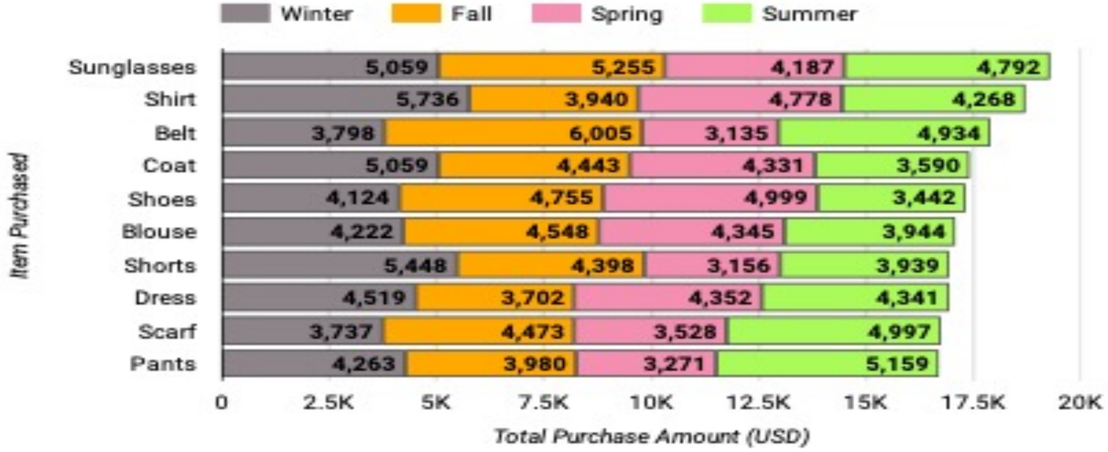
Top 10 Selling Items



Sales Trend by Size

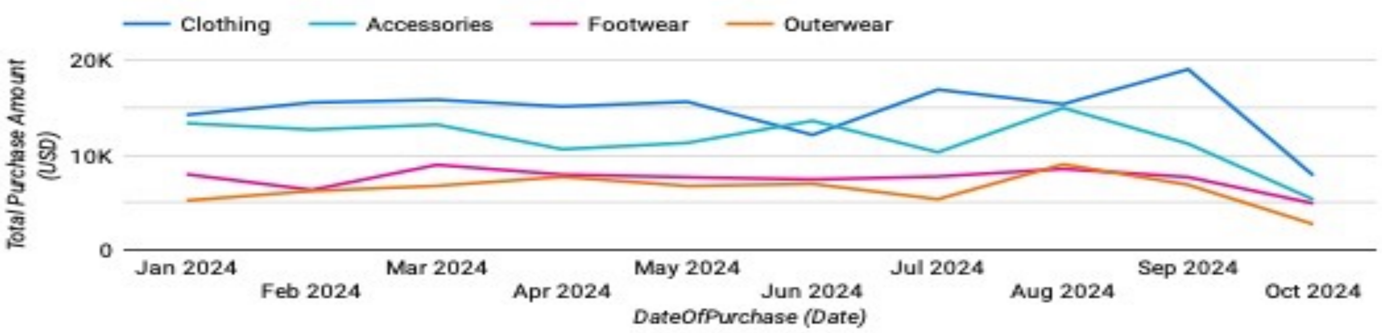


Season-wise Sales Trend

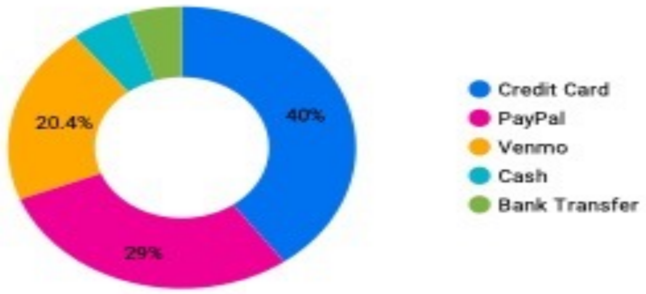


Customer Behavior analysis - 2024

Month-on-Month Purchase Trends by Category (2024)

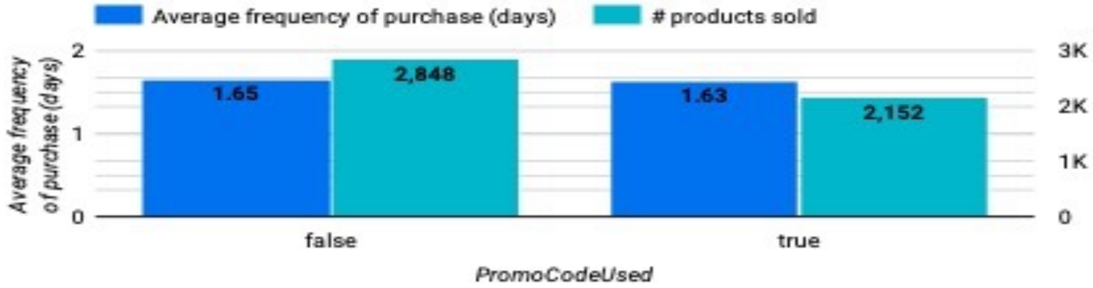


Transaction Share by Payment Method

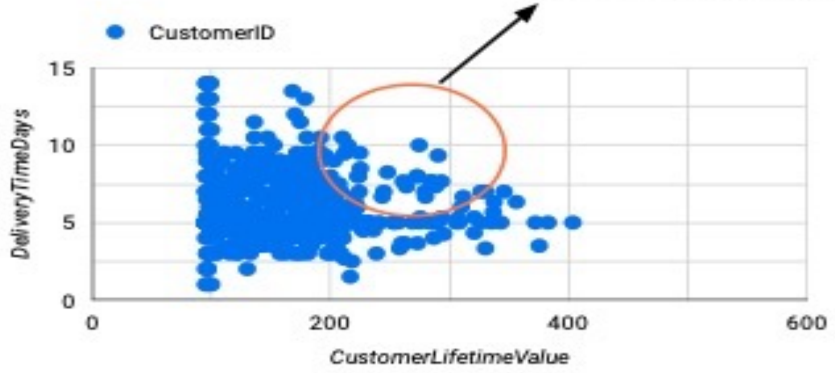


Promo Code Usage vs. Purchase Frequency

(Promo codes applied on 85 / 295 days)



Reduce Delivery Times for High CLV Customers to Prevent Loss



	Loyalty Segment	SubscriptionStatus	Current Status	Conversion Target	No. of Customers	Frequency of purchase (in days)	Avg Purchase Amount (USD)	CustomerLifetimeValue
1.	Low Loyalty Points	false	One-Off Buyers	High Loyality	3,135	1.62	59.9	68.2
2.	Low Loyalty Points	true			1,141	1.68	59.1	65.07
3.	Medium Loyalty Points	false			273	1.63	160.9	236.83
4.	High Loyalty Points	false	Loyal Non-Subscribers	Subscribers	247	1.7	256.2	333.24
5.	Medium Loyalty Points	true			105	1.76	154.8	241.54
6.	High Loyalty Points	true	Subscribers	Retain	95	1.39	246.9	335.02

1 - 8 / 8 < >