

# Building a Scalable and Reliable Data Pipeline on AWS

Rahul Reddy Parupati  
Nivedha Balakrishnan

## Problem Statement

In an Ecommerce company, the Data Engineering team has been given two requirements:

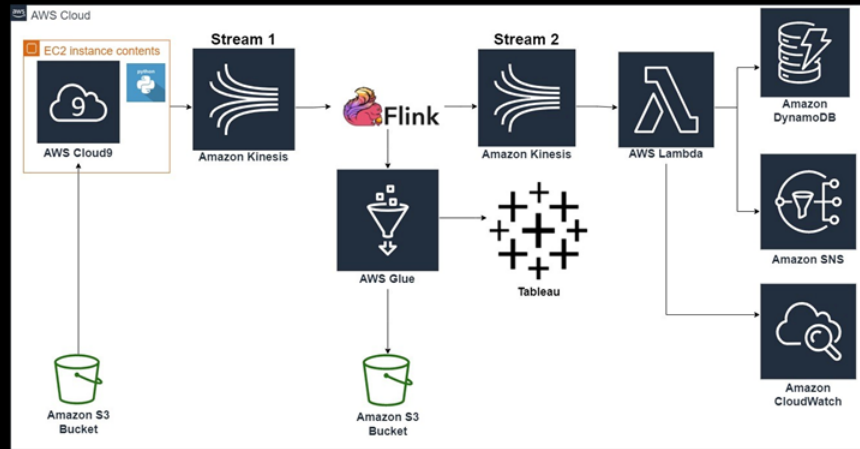
Marketing Team: Dashboard that helps them to analyze user shopping journal and behaviour in real time.

Data Science Team: Detect anomalies on the website in the real-time.

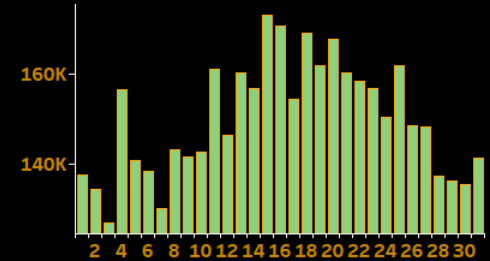
- Building a data pipeline for simultaneous data ingestion, processing, and visualization that can detect website anomalies or intrusions
- Deployment of pipeline to a production environment where regular monitoring and maintenance are required.
- Real-time user shopping journey analysis with Tableau.

## Goal

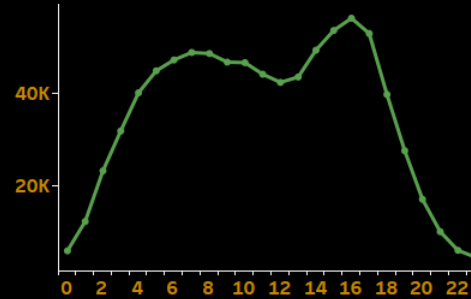
## System Architecture



## # of Unique visits



## # of unique visits for a day



## Event Type

All

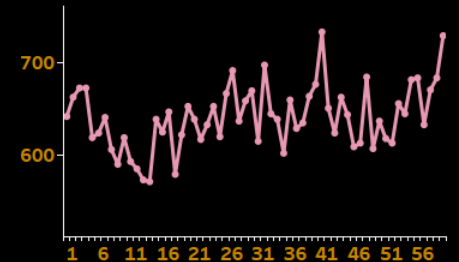
## Day of Event Time

5

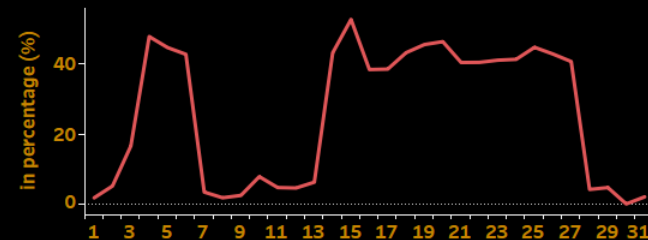
## Hour of Event Time

All

## # of unique visits for an hour



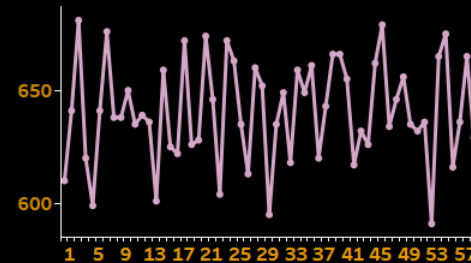
## Cart Abandment Rate



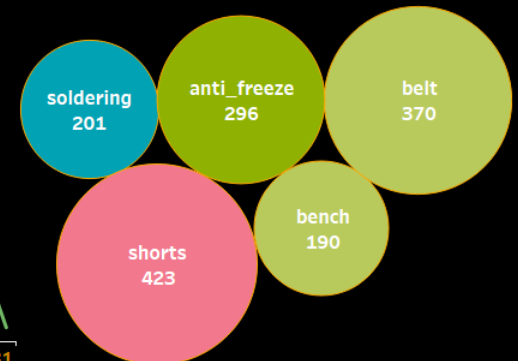
## Minute

All

## # of unique visits for a minute



## Bottom 5 Products



## Key Insights

- The purchase conversion rate is peak during 14th, 16th and 17th of the month, the least is absorbed during the first and the last day of the month.
- The cart abandonment rate is pretty high during 3rd and 4th week of the month.
- The visitors are usually high during second and third week of the month.

## Conversion Rate

