Analysis of customer shopping behavior on multi category e-commerce platform

Project Plan

Version 1.1

Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Version** | **Description** | **Author** |
| 1st Oct 2021 | 1.0 | Initial Document- Finalized the dataset and deliverables |  |
| 12th Dec 2021 | 1.1 | Final Document- added all the required content |  |

Table of Contents

1. Introduction 4

1.1 Purpose of this document 4

1.2 Intended Audience 4

1.3 Scope 4

1.4 Definitions and acronyms 4

1.4.1 Definitions 4

1.4.2 Acronyms and abbreviations 5

1.5 References 5

2. Background and Objectives 5

3. Architecture & High-Level Design 6

4. Organization 6

*4.1* Project group 6

4.2 Customer 6

5. Development process 7

*6.* Deliverables 7

7. Project risks 7

8. Communication 7

8.1 Canvas 7

8.2 Git 7

9. Project plan 8

9.1 Time schedule 8

9.1.1 Remarks 8

9.2 Test plan 9

9.2.1 Testing Remarks 10

10. References 10

# Introduction

## Purpose of this document

The purpose of this document is to provide a detailed project description of the project called Analysis of customer shopping behavior on multi-category e-commerce platform, which is designed to help people buy better and more quality items. This document includes details about organization, roles, deliverables, project risks, time plans and financial plans.

## Intended Audience

This document shall be used in all phases of the project as a guideline. Intended audiences of this project are all project stakeholders:

* project supervisor
* project leader
* team members

## Scope

This document defines the project plan of the Analysis of customer shopping behavior on multi category e-commerce platform project. The overview includes objectives of the project, organization of the project team, development process that is going to be used during the project, assessment of possible risks, communication used between project stakeholders and project plan that includes time schedule and activity plan.

## Definitions and acronyms

### Definitions

|  |  |
| --- | --- |
| **Keyword** | **Definitions** |
| Analysis of customer shopping behavior on multi category e-commerce platform | The name of the project |
| Project Supervisor | A person in charge of supervising the project |
| Project Leader | A person in charge of organizing the team and communicating with the project supervisor |
| Team Member | An active member of the team responsible for making the job done |
| Milestone | A time in a project that marks the end of a project phase or the completion of an important deliverable. |
| Git | Version control system that will be used in this project |
| Scrum | An iterative and incremental agile software development method for managing software projects and product or application development |
| Kunagi | Web-based tool for integrated agile project management and collaboration based on Scrum |
| Scrum sprint | The basic unit of development in Scrum |
| Scrum master | Ensures the smooth working of the Scrum team and enforces Scrum practices |
| Product owner | Responsible for product management and its quality |

### Acronyms and abbreviations

|  |  |
| --- | --- |
| **Acronym or**  **abbreviation** | **Definitions** |
| ELT | Extract, Loat, Transform |
| RFM | Recency, Frequency and monetary |
| IAM | Identity and Access management |
| EDA | Exploratory Data Analysis |
| AWS | Amazon Web Services |

## References

1. <http://www.scrum.org/>
2. <http://kunagi.org/>

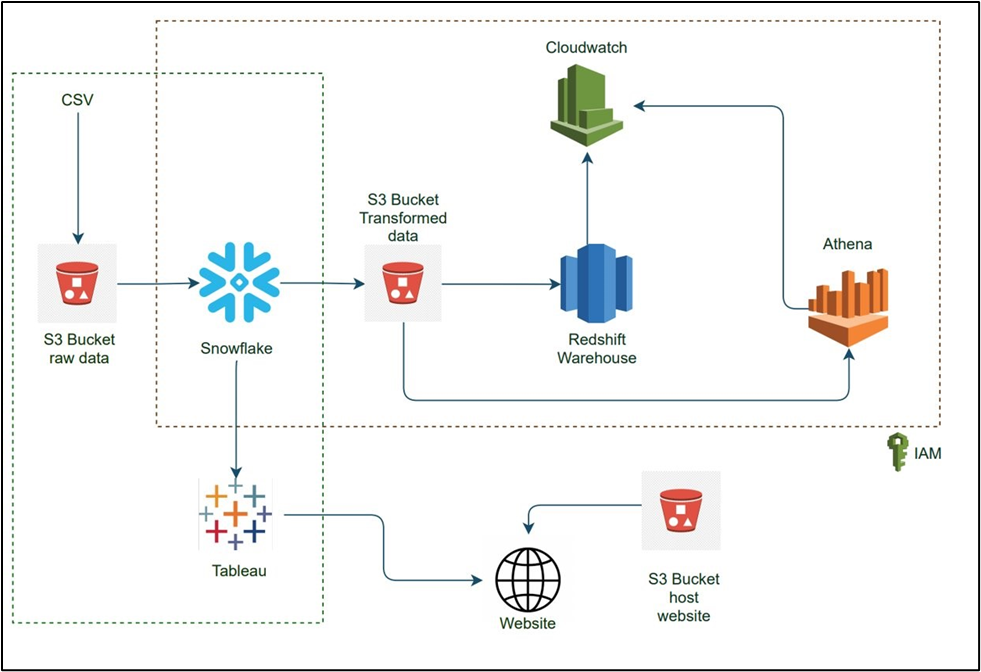
# Background and Objectives

In today’s world with the advancement in technology, e-commerce has become a vital part of the economy and retail industry. The continued expansion of e-commerce retail has driven consumers of all age groups to use e-retail as a part of their everyday life. The shift in retail sales from brick-and-mortar stores to digital was evident more than ever during the pandemic in 2020. With consumers having a myriad of options to choose from, it has become increasingly difficult for merchants and online shopping portals to engage customers. In such a scenario, it's important for them to deliver a good customer experience. This is where we can leverage the benefit of data that online shopping websites gather by analyzing and drawing meaningful insights. It helps to understand the customers better, recommend products accordingly and increase the conversion rate from a shopper to a customer.

We have the following objectives:

* + Use the data to find the conversion ratio from User viewing event to purchase event (work on strategies to improve this ration)
  + Identify brands which make the most/least revenue/ sales by count
  + Identify categories which are purchased the most
  + Identify the varying customer behavior on the website during time of the day/month
  + Use the data for calculating RFM

# Architecture & High-Level Design



We used IAM access for the project. We used one of the public data from Kaggle. The data was available in csv format. This data was uploaded to S3 bucket. The data from S3 was then fetched into Snowflake for ELT. Data was transformed in Snowflake. The transformed data was uploaded back into S3 in CSV format.

Data from S3 was uploaded into Redshift, where Redshift is our data warehouse. Chunks of data(categorical) was then unloaded from Redshift into S3. We connected Athena to S3 for querying purposes and used CloudWatch to monitor logs from Redshift and Athena.

The other part of the architecture is where we are performing visualization using the data from Snowflake, here Snowflake acts as our data warehouse. We connected Snowflake to Tableau to use this transformed data for visualization. We used S3 bucket to host a static website to upload the visualizations from Tableau to a web server.

# Organization

## Project group

|  |  |  |
| --- | --- | --- |
| **Name** | **Initials** | **Responsibility (roles)** |
| Snehal |  | EDA, Visualization |
| Sakshi Jain |  | Data Cleaning and Transformation, Visualization |
| Meghana |  | Querying in Athena, Visualization |
| Sakshi Tongia |  | Website Integration, Visualization |

## Customer

The target customers are listed below:

* Category managers of the e-commerce platform
* Marketing team
* Vendor managers
* Application developers

# Development process

We will perform initial Exploratory Data Analysis on the dataset using Python to check the dataset, columns present in the dataset and what all insights we can derive. Then used snowflake to perform EDA which consisted of removing/replacing missing values, removing duplicates, handling outliers, handling correlated columns. Performed data visualization to identify patterns in the data and generate insights related to segment customers based on amount spent, category of product purchased, time spent on the website, comparison in customer shopping behavior. Our project aimed to do an analysis for gaining insights into customer shopping behavior, not limited to customer segmentation. We connected Tableau to a static website which was hosted using S3. We also used Athena to perform query operations on the dataset.

# Deliverables

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **To** | **Output** | **Planned week** | **Promised week** | **Late +/-** | **Delivered week** | **Notes** |
|  | **Transformed data** | 27th Sep |  | **1** | 4th Oct |  |
|  | **Querying in Athena** | 1st Nov |  |  | 1st Nov |  |
|  | **Visualization in Tableau** | 8th Nov |  |  | 8th Nov |  |
|  | **Static Website** | 22th Nov |  | **1** | 29th Nov |  |

# Project risks

|  |  |  |
| --- | --- | --- |
| **Possibility** | **Risk** | **Preventive action** |
| High Project cost (15-30 GB data to be transformed in AWS Glue) | High project cost | Used Snowflake for Transformation |

# Communication

## Collaboration

## Git

All source code and finished documentation will be uploaded to Github repository.

Repository URL: <https://github.com/Sakshijain3/Bigdata_SNOWFLAKE_PROJECT>

**Static Website**: https://us-west-1.console.aws.amazon.com/console/home?region=us-west-1

# Project plan

## Time schedule

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Id** | **Milestone**  **Description** | **Responsible Dept./Initials** | **Finished week** |  |  |  | **Metr.** | **Rem.** |
|  |  |  | **Plan** | **Forecast** |  | **Actual** |  |  |
|  |  |  |  | **Week** | **+/-** |  |  |  |
| 1 | Finding Dataset |  | 6th Sep |  |  | 6th Sep |  |  |
| 2 | Creation of Roles, S3 bucket |  | 14th Sep |  |  | 14th Sep |  |  |
| 3 | Load Data in Snowflake |  | 20th Sep |  |  | 20th Sep |  |  |
| 4 | Transform data in. Snowflake |  | 27th Sep |  | 1 | 4th Oct |  |  |
| 5 | Load transformed data in S3, Redshift |  | 25th Oct |  |  | 25th Oct |  |  |
| 6 | Querying in Athena |  | 1st Nov |  |  | 1st Nov |  |  |
| 7 | Visualization in Tableau |  | 8th Nov |  |  | 8th Nov |  |  |
| 8 | Host static website |  | 22th Nov |  | 1 | 29th Nov |  |  |

### Remarks

|  |  |
| --- | --- |
| **Remark Id** | **Description** |
| NA |  |

## Test plan

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Test No. | 001 | Phase: | 1 | Author: |  | Date: 20 Sept 2021 |
| Test Category: | | **System test** | | |  |  |
| Software Product: | | Tableau, Snowflake | | | |  |
| Test Title: | | Test connection between Tableau, Snowflake | | | | |
| Test Purpose: | | Access data from Snowflake in Tableau | | | | |
| Test Setup: | | Credentials to access Snowflake | | | | |
| Prerequisites: | | Data available in Snowflake | | | | |
| Procedure: | | NA | | | | |
| Checks: | | NA | | | | |
| Expected Results: | | Data is available in Tableau | | | | |
| Result: | | Data is available in Tableau | | | | |
| Reason for Failure: | | NA | | | | |
| Remarks: | | NA | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Test No. | 002 | Phase: | 2 | Author: |  | Date: 2nd Nov 2021 |
| Test Category: | | **System test** | | |  |  |
| Software Product: | | Athena, Redshift, S3 | | | |  |
| Test Title: | | Connection between S3 and Athena | | | | |
| Test Purpose: | | Connect Athena to S3 | | | | |
| Test Setup: | | Output files in S3 | | | | |
| Prerequisites: | | Data available in Redshift/S3 | | | | |
| Procedure: | | Check logs in Cloudwatch to monitor Redshift | | | | |
| Checks: | | If Redshift is up and running | | | | |
| Expected Results: | | Data can be queried from Athena | | | | |
| Result: | | Data can be queried from Athena | | | | |
| Reason for Failure: | | NA | | | | |
| Remarks: | | NA | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Test No. | 003 | Phase: | 3 | Author: |  | Date: 1 Dec 2021 |
| Test Category: | | **System test** | | |  |  |
| Software Product: | | Tableau Public server | | | |  |
| Test Title: | | Test connection between Website, Tableau public server | | | | |
| Test Purpose: | | Establish connection between Website, Tableau public server | | | | |
| Test Setup: | | Visualizations are available in Tableau public server | | | | |
| Prerequisites: | | Visualizations are available in Tableau | | | | |
| Procedure: | | Kept extracting data in Tableau desktop, viewed changes on the public server | | | | |
| Checks: | | Maintain consistency | | | | |
| Expected Results: | | Visualizations visible on the website | | | | |
| Result: | | Visualizations visible on the website | | | | |
| Reason for Failure: | | NA | | | | |
| Remarks: | | NA | | | | |

### Testing Remarks

|  |  |
| --- | --- |
| **Remark Id** | **Description** |
| NA |  |

# References

* https://www.kaggle.com/mkechinov/ecommerce-behavior-data-from-multi-category-store?select=2019-Nov.csv
* https://towardsdatascience.com/scalable-efficient-big-data-analytics-machine-learning-pipeline-architecture-on-cloud-4d59efc092b5
* https://help.tableau.com/current/pro/desktop/en-us/examples\_snowflake.htm
* <https://docs.aws.amazon.com/IAM/latest/UserGuide/tutorials.html>
* <https://docs.aws.amazon.com/athena/latest/ug/work-with-data-stores.html>