Analyzing consumer shopping behavior from a large multi-category online store

Project Plan

Version 1.1

Revision History

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Version** | **Description** | **Author** |
| 09/11/2022 | 1.0 | Version 1 Project topic finalization, Dataset selection, Project deliverables | Saniya Lande, Shilpa Shivarudraiah, Neetu Rasinger Babu, Iqra Bismi |
| 11/20/2022 | 1.1 | Version 1.1 Design, Implementation, and documentation | Saniya Lande, Shilpa Shivarudraiah, Neetu Rasinger Babu, Iqra Bismi |

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. Organization 5

3.1 Project group 5

3.2 Customer 6

4. Development process 6

5. Deliverables 6

6. Project risks 6

7. Communication 7

7.1 Canvas 7

7.2 Git 7

8. Project plan 7

8.1 Time schedule 7

8.1.1 Remarks 7

8.2 Test plan 9

8.2.1 Testing Remarks 10

9. References 10

# Introduction

## Purpose of this document

The purpose of this document is to provide a detailed project description of the application called Analyzing consumer shopping behavior from a large multi-category online store, 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:

* Professor Andrew H. Bond
* Shilpa Shivarudraiah
* Saniya Lande, Neetu Rasinger Babu, Iqra Bismi

## Scope

This document defines the project plan of the Analyzing consumer shopping behavior from a large multi-category online store application. 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** |
| Project Name | Analyzing consumer shopping behavior from a large multi-category online store |
| Project Supervisor | Professor Andrew H. Bond |
| Project Leader | Shilpa Shivarudraiah |
| Team Member | Saniya Lande, Neetu Rasinger Babu, Iqra Bismi |
| Milestone | 11/20/2022 |
| Git | <https://github.com/ShilpaShivarudraiah/BigData_DATA228_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 | Saniya Lande |
| Product owner | Neetu Rasinger Babu |

### **Acronyms and abbreviations**

|  |  |
| --- | --- |
| **Acronym or**  **abbreviation** | **Definitions** |
| AWS | Amazon Web Services |
| ELT | Extract, Load, Transform |
| IAM | Identity & Access Management |
| EDA | Exploratory Data Analysis |
| CLI | Command Line Interface |

## References

1. <https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store>
2. <https://shoppingdashboard.s3.us-west-1.amazonaws.com/index.html>
3. <http://www.scrum.org/>
4. <http://kunagi.org/>

# Background and Objectives

Online market has shown tremendous growth and will be further increasing in the coming years. The number of consumers shopping online has also increased. Hence analyzing the consumer behavior and their satisfaction is an important parameter to be considered for further improving the online market trend. The main objective of this research is to analyze key parameters which are responsible for consumer’s online shopping behavior such as cost of a product, offers available on product, branding of a product etc.

# Architecture & High-Level Design

Diagram

Description automatically generated

# Organization

## Project group

|  |  |  |
| --- | --- | --- |
| **Name** | **Initials** | **Responsibility (roles)** |
| Shilpa Shivarudraiah | SS | Data Transformation, Athena Query, Documentation |
| Saniya Lande | SL | Data Visualization, Web Integration, Documentation |
| Neetu Rasinger Babu | NRB | Data Loading, Data Transformation, Documentation |
| Iqra Bismi | IB | Web Integration, Athena Query, Documentation |

## Customer

The target customers are listed below:

* E-Commerce platform managers
* Sales and Marketing Team
* Business Intelligence Team
* Developers

# Development process

The raw data which consisted of two files, one with data for the month of October and other for the month of November (in csv format) extracted from Kaggle was first loaded into the S3 bucket. The raw data had to be transformed. We used AWS Glue for ELT. Firstly, a crawler was created which crawls through all our data and creates a schema in the Data catalog. Once the schema is ready, we created a job in AWS Glue to do some transformations like changing the datatype of certain features, splitting the features into different columns, etc. Once the transformations are performed, we again crawl through the transformed data and the transformed data is stored in the S3. We then use Athena to query the data and to look at the schema of our data. The cloudwatch was used for monitoring the query-related metrics. The next step was analyzing the data through visualizations. We used Tableau Desktop for this purpose as it is free and user friendly and creates good interactive dashboards. We connected Athena to Tableau and extracted the data source to create visualizations and interactive dashboards. We then published dashboard to Tableau Public, generated an embed code which was then hosted on S3 bucket. We used IAM policies throughout to securely access to AWS resources used.

# Deliverables

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **To** | **Output** | **Planned week** | **Promised week** | **Late +/-** | **Delivered week** | **Notes** |
|  | Downloaded the data from Kaggle | Sept First Week | Sept First Week |  | 09/11/2022 |  |
|  | Loaded the data into S3 | Sept 3rd Week | Sept 3rd Week |  | 09/27/2022 |  |
|  | Transformed the data | Oct 1st Week | Oct 1st Week |  | 10/17/2022 |  |
|  | Querying Via Athena | Oct 4th Week | Oct 4th Week | +1 | 11/03/2022 |  |
|  | Data Visualizations | Nov 1st Week | Nov 1st Week | +1 | 11/10/2022 |  |
|  | Static Website | Nov 2nd  Week | Nov 2nd  Week | +1 | 11/18/2022 |  |
|  | Project Report & PPT | Nov 3rd Week | Nov 3rd Week |  | 11/20/2022 |  |

# Project risks

|  |  |  |
| --- | --- | --- |
| **Possibility** | **Risk** | **Preventive action** |
| Cost risk for using AWS services (15GB Data) | High Cost | Deleted the files and resources not in use. |

# Communication

We had weekly meetings over zoom call for project discussions and implementation.

## Collaboration

## Git

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

Repository URL: <https://github.com/ShilpaShivarudraiah/BigData_DATA228_Project>

# Project plan

## Time schedule

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Id** | **Milestone**  **Description** | **Responsible Dept./Initials** | **Finished week** |  |  |  | **Metr.** | **Rem.** |  |
|  |  |  | **Plan** | **Forecast** |  | **Actual** |  |  |  |
|  |  |  |  | **Week** | **+/-** |  |  |  |  |
| 1 | Topic finalization Finding Data Set | SS, SL, NRB, IB | Sept 1st Week |  |  | 09/11/2022 |  |  |  |
| 2 | Loading the data into S3 | NRB | Sept 3rd Week |  |  | 09/27/2022 |  |  |  |
| 3 | Data Transformation | SS, NRB | Oct 1st Week |  |  | 10/17/2022 |  |  |  |
| 4 | Querying in Athena | SS, IB | Oct 4th Week |  | +1 | 11/03/2022 |  |  |  |
| 5 | Data Visualizations using Tableau | SL | Nov 1st Week |  | +1 | 11/10/2022 |  |  |  |
| 6 | Web Integration and Static Website hosting in S3 | IB, SL | Nov 2nd Week |  | +1 | 11/18/2022 |  |  |  |

### **Remarks**

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

## Test plan

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Test No. | 001 | Phase: | 1 | Author: | Saniya Lande | Date: 11/08/2022 |
| Test Category: | | **System Test** | | |  |  |
| Software Product: | | Tableau | | | |  |
| Test Title: | | Test Connection Between Tableau and Athena | | | | |
| Test Purpose: | | Extract data from Athena to Tableau | | | | |
| Test Setup: | | Access Key ID and Secret Access Key | | | | |
| Prerequisites: | | Data Availability in Athena | | | | |
| Procedure: | | Setting up configuration between Athena and Tableau | | | | |
| Checks: | | N/A | | | | |
| Expected Results: | | Connection successful between Athena and Tableau by having data loaded in Tableau | | | | |
| Result: | | Data available in Tableau for analysis | | | | |
| Reason for Failure: | | N/A | | | | |
| Remarks: | | We used subset of data for our analysis in Tableau based extract option | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Test No. | 002 | Phase: | 2 | Author: | Neetu Rasinger Babu, Shilpa Shivarudraiah | Date: 11/03/2022 |
| Test Category: | | System Test | | |  |  |
| Software Product: | | S3, Glue, Athena | | | |  |
| Test Title: | | Connection between S3 and Glue, S3 and Athena | | | | |
| Test Purpose: | | Connect Glue to S3, Connect Athena to S3 | | | | |
| Test Setup: | | Output files is S3 | | | | |
| Prerequisites: | | Data Availability in S3 | | | | |
| Procedure: | | Setting up connection between S3 and Glue through crawlers, Setting up connection between Athena and S3 through crawlers, | | | | |
| Checks: | | Data Table availability in Athena for query | | | | |
| Expected Results: | | Transformed data available for querying | | | | |
| Result: | | Data Transformation and querying data successfully | | | | |
| Reason for Failure: | | N/A | | | | |
| Remarks: | | N/A | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Test No. | 003 | Phase: | 3 | Author: | Iqra Bismi | Date: 11/18/2022 |
| Test Category: | | System Test | | |  |  |
| Software Product: | | Tableau Public Server | | | |  |
| Test Title: | | Connection between Tableau public server and static website | | | | |
| Test Purpose: | | Host the tableau dashboard as a static website on S3 | | | | |
| Test Setup: | | Publishing the tableau desktop dashboard and tableau public server | | | | |
| Prerequisites: | | Dashboard available on tableau public server | | | | |
| Procedure: | | Creating dashboards on tableau desktop, publishing on tableau public server and hosting it as a static website on S3 | | | | |
| Checks: | | Dashboard available on tableau public servers | | | | |
| Expected Results: | | Dashboard available on web browser for public | | | | |
| Result: | | Dashboard available on web browser for public | | | | |
| Reason for Failure: | | N/A | | | | |
| Remarks: | | N/A | | | | |

### **Testing Remarks**

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

# References

<https://www.sciencedirect.com/science/article/pii/S2666518221000383>

[https://www.researchgate.net/publication/298739144\_Big\_data\_analytics\_in\_E- commerce\_a\_systematic\_review\_and\_agenda\_for\_future\_research](https://www.researchgate.net/publication/298739144_Big_data_analytics_in_E-%20commerce_a_systematic_review_and_agenda_for_future_research)

[https://d1wqtxts1xzle7.cloudfront.net/50409172/10.1.1.655.7217-with-cover-page- v2.pdf?Expires=1663392470&Signature=IZRvJYmHZ- hYlnQNB7L9s1wsj~P~5bZ3VnPa2Ow17EVdnnygeA218SM6DRsSa31jxPlNWvU5gFW Q7RankUa31-pR0O5ZCFAKaNjW5GpJmtFOpTIv25cn5xdrBfrs~8dsA0O6nLY-](https://d1wqtxts1xzle7.cloudfront.net/50409172/10.1.1.655.7217-with-cover-page-%20v2.pdf?Expires=1663392470&Signature=IZRvJYmHZ-%20hYlnQNB7L9s1wsj~P~5bZ3VnPa2Ow17EVdnnygeA218SM6DRsSa31jxPlNWvU5gFW%20Q7RankUa31-pR0O5ZCFAKaNjW5GpJmtFOpTIv25cn5xdrBfrs~8dsA0O6nLY-%204Q7KQXP1FmpK3RjihlFc9Y1lRpLLGWM5we-5oq-%20dWQ7XtQWFaEBh50Ka6UqGnyFfyaoe~NxSbuNJTD6~iIlawa6U1BkP0-NyFqAOxUTe-%20znC5EcyNN6C41fKCpMSQq~jyQ14fzcMUnuhcrpT1i-%20twf1h9nRf9oEHhDCVq~wICjuDrOKfZaDmmIG4NQLjlQk9nKUul6ftNX2gbA__&Key-%20Pair-Id=APKAJLOHF5GGSLRBV4ZA)

<https://help.tableau.com/current/pro/desktop/en-us/examples_amazonathena.htm>

<https://docs.aws.amazon.com/cli/latest/topic/s3-config.html>

<http://www.jait.us/uploadfile/2019/0823/20190823115117196.pdf>

<https://docs.aws.amazon.com/cli/latest/userguide/cli-configure-profiles.html>

<https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store>

<https://www.shopify.com/blog/marketing-analytics>

<https://ieeexplore.ieee.org/document/8073494>

**Image Source**: Google, Draw.io

<https://www.google.com/search?q=website+icon+images&tbm=isch&ved=2ahUKEwij5divzrb7AhXtmo4IHZIEDe0Q2-cCegQIABAA&oq=website+icon+images&gs_lcp=CgNpbWcQAzIECCMQJzIFCAAQgAQyBggAEAcQHjoECAAQQzoICAAQCBAHEB5QwwdYng1goBJoAHAAeACAAV2IAcwDkgEBNpgBAKABAaoBC2d3cy13aXotaW1nwAEB&sclient=img&ei=W-R2Y-PLIu21uvQPkom06A4&bih=714&biw=1536&rlz=1C1CHBF_enUS912US912#imgrc=27rmHmjzHIpI3M>

<https://www.google.com/search?q=s3+bucket+icon&rlz=1C1CHBF_enUS912US912&sxsrf=ALiCzsYRsacjmLplqXwBYMKc_bG0q9gjIw:1668736032502&source=lnms&tbm=isch&sa=X&ved=2ahUKEwiGrMOTzrb7AhXUI30KHV_uAzQQ_AUoAXoECAMQAw&biw=1536&bih=714&dpr=1.25#imgrc=HTL--CMhT7OOWM>

<https://www.google.com/search?q=tableau+image&rlz=1C1CHBF_enUS912US912&sxsrf=ALiCzsaTsl5T47u_3kEAhagdxKpFR2M5vg:1668736170109&source=lnms&tbm=isch&sa=X&ved=2ahUKEwiznJLVzrb7AhVzPH0KHc27A0oQ_AUoAXoECAMQAw&biw=1536&bih=714&dpr=1.25#imgrc=IC2ieWfwkth2fM>