GET YOUR PERFECT FIT

Harsimran Kaur, Karnik Kalani, Mounica Ayalasomayajula, Saumya Sinha, and Vidushi Bhati

Abstract—Future of retail depends on giving customers the products which are best suited for them.There has been a surge in online shopping since the onset of the pandemic. Leading brands have been witnessing the increase in their online sales.People find online shopping more convenient and thus, online stores are keeping very limited stock in their stores rather than websites as people can buy several clothes from all the available or newly launched outfits by HM.Therefore, providing consumers with such an e-commerce platform that captures the wants of the customers has become imperative. This project aims to suggest the customer their best possible outfit according to their body shape. Although, filters are available that provide customers outfits based on various parameters such as price, size, colour, and type. After evaluating the current HM Dataset, we came up with a plan to construct the architecture in such a way that customers can save their time by just providing a few details such as their age, size, body shape, colour preferences, skin tone, liking for specific fabrics while creating their accounts itself. Using these parameters users will be able to see only those outfits which will be best suited to them. Also, will take into consideration if the customer is suffering from any textile related skin condition to suggest clothes made from suitable fabrics.It will also evaluate various other factors for gaining useful insights about the products.

Keywords: H&M, Data Analysis, SQL, Python, Cloud, AWS, Jira, MySQL, MongoDB, Tableau, Jamboard, ETL, Spyder, Visual Studio

Project Github: *Link: https://github.com/DecipherData/data225semproj*

1 Introduction

Retail clothing industry has always been relevant and it continually strives to take all possible measures to gain a profitable edge in this highly competitive domain. Brands try to look at the granular level of what consumers want to retain their existing customers and attract the new ones. The interaction has increased directly or indirectly many folds with the advent of new technologies such as ML and AI. These technologies take into account the previous purchase patterns done by the customer and recommend clothes according to that. However, there is another aspect where a gap needs to be filled. Customer body shape and skin related conditions are not captured when customers register for some brand. Everybody is unique and not every style looks good on every body shape.Further, not all fabrics are suitable for everyone,so it is important for retailers to provide suggestions based on these factors.

1.1 Motivation

Fashion designers and stylists recommend their clients clothes that are a best fit for them based on their body shape and size. These skilled people are not accessible to everyone.The retailers only give suggestions to the customers based on their previous purchase patterns and sizes.If they were to give suggestions based on body types and further it by providing suggestions based on the skin conditions, many customers would save a lot of money and get their perfect fit. Also, customers may gain time by not spending time in trial rooms or if they have purchased an item online but was unsatisfactory after it arrived,then they need not waste time returning it.This will also benefit organizations who spend a considerable amount of money on logistics. Knowing the granular details of the customers may help brands in projecting what kind of clothes they need to stock up in their inventory and what kind they need to reduce manufacturing.

1.2 Literature Survey

To be able to provide suggestions according to the different body shapes, we must understand what types of body shapes exist. Idea of the body shapes are entirely subjective and arise from the societal standards that vary from culture to culture.
According to a study published in the International

Journal of Clothing Science and Technology[1],there are seven types of female body shapes.It may be possible that some body shapes may not fit into the categories mentioned above.But for the scope of this project we are focusing on these seven body types.

Body Types:

•Apple or Inverted Triangle:
This body shape has relatively smaller hips than their shoulders and bust. Range: (bust – hips) $\geq$ 3.6″ AND (bust – waist) $<$ 9″.

•Banana, straight, or rectangle:
This body shape has shoulders, waist, hip and busts of similar size.The body is somewhat rectangular looking. Range: (hips – bust) $<$ 3.6″ AND (bust – hips) $<$ 3.6″ AND (bust – waist) $<$ 9″ AND (hips – waist) $<$ 10″.

•Pear, spoon, bell, or triangle:
This body shape has narrow bust and shoulders but has hips on the heavier side. With slender arms and shoulders,weight is distributed in the leg. Range: If (hips – bust) $>$ 2″ AND (hips – waist) $\geq$ 7″ AND (high hip/waist) $\geq$ 1.193.

•Hourglass, X shape, triangle, opposing, or facing inwards:
This body shape has nearly equal hips and bust with a relatively smaller well-defined waist. Range:(bust – hips) $\leq$ 1″ AND (hips – bust) $>$ 3.6″ AND (bust – waist) $\geq$ 9″ OR (hips – waist) $\geq$ 10″

•Top Hourglass:
This body shape has hips smaller than bust and has a well-defined waist. Range: (bust – hips) $>$ 1″ AND (bust – hips) $<$ 10″ AND (bust – waist) $\geq$ 9″

•Bottom Hourglass:
This body shape is generally an hourglass shape but with hips slightly larger than bust. Range: (hips – bust) $\geq$ 3.6″ AND (hips – bust) $<$ 10″ AND (hips – waist) $\geq$ 9″ AND (high hip/waist) $<$ 1.193

•Fabrics:
Clothing products are made of different materials.There are many people with sensitive skin or with other skin conditions such as eczema, dermatitis, Psoriasis. Cancer patients who go through radiations have sensitive skin and may develop rashes and sores from chemotherapy sessions.These patients if exposed to rough and bad quality fabric may aggravate their pain and discomfort.In our project we classify fabrics as good-for-skin (highest quality),medium quality and bad-for-sensitive-skin(low quality).

•Good-for-skin(highest quality) fabric examples are:
Silk, cotton, Linen, Flax, Hemp, Wool, cashmere, Crepe, Damask, Muslin, lace, satin, Chiffon, Organza.

•Medium quality fabric examples are:
Rayon, velvet, twill, acrylics, viscose, chenille

•Low quality examples are:
Spandex, polyester, nylon, leather, rubber, Tweed, Felt, Jute, Taffeta, Baize, Dyes, Anthraquinone

•Skin conditions can be categorized into :
1- severe, 2- medium and 3- no issues.

1.3 Functional Requirements

A functional requirement defines an input required or dependency for us to complete our task.
Following are the functional requirements of our project - Different attributes of data: customer, fabric, skin condition, product related.
A general perspective of available measurements are stereotyped and we want a bit of customization to be included in our body type declaration, for a customer to filter out product, based on many factors including fabric materials for their skin condition.

1.4 Methodology

We have collected data from different sources and also synthesized some data as per our functional requirements.
Once we drafted an initial version of our ER-diagram using draw.io , we extracted the required data and cleaned the data using python in Visual studio code by dropping the columns that were no required, removing the duplicates, using NumPy to Clean Columns, and checking for the missing data and assigning values accordingly for the data to be ready. The respective tables were then created and data was inserted into them using MySQL Workbench to get practical perspective on data. MongoDB cluster was used for storing and accessing the body measurement configurations.
Python connection was established to MySQL

server for developing few functions to implement logical decisions based on the SQL tables and data. Once the modifications and operations are done on data and it is ready, AWS S3 was used to keep our contact and AWS Redshift cluster with database was created. AWS GLUE was also used to create and load our tables and data. Analysis was done in Redshift. Queries were formed to get insights about what fabric related sales were and how many different products contained what styles and fabrics, what products were there for different body type measurements etc. To introduce bit of customization to these products we tried to recognize the parameters that were top influential. AWS S3 buckets are correspondingly used by Redshift tables for storage, and used AWS Glue for performing ETL process. Tableau was used for displaying the result by connecting it with our AWS Redshift database. For all this work to be version controlled and to be properly coordinated,GIT was used.

2 Project Walk-through

2.1 Data Sourcing and modeling

Data Sets:

- **Customer data:** Body Measurements Data
- **HM Data:** HM Data
- **Fabric data:** Synthesized by us from referring other links mentioned below in reference section.

About the Data:

HM data includes products and transactions data that is in the public domain. It is de-normalized and is obfuscated for the purpose of publishing online. Considering the data as the initial set of data, we need to model ER diagrams and generate a normalized data model used by the brand. Since, we have a requirement of customer size specific data and fabric details of the product. This data is not present in the data-set. Therefore, we had to take 2-3 external data-sets and merge them to get different body measurements of customers and fabric types of products. Also, since our idea is novel based on different styles, such kind of data is not present so we had to look for different style types and generate the data according to our requirement. We have made the data models to address the following use cases which can be generalized for any brand in the market.

Use Cases:
Registration and login is out of scope.

A. Customer

- ➜ Assumption is that Customer has registered by providing relevant body measurements such as bust, hips, waist, high hip.
- ➜ Customer has given his/her skin condition level.
  - -Skin condition level 1 indicates: Severe skin conditions (eczema. psoriasis)
  - -Skin condition level 2 indicates: Mild skin conditions (dryness, itchiness)
  - -Skin condition level 3 indicates: No skin issues
- ➜ System should be able to evaluate the body type for each customer.
- ➜ Each body type may have one or more product styles associated with it.

B. Product available in online stores to its customers.

- ➜ Each product is divided into three categories.
  - Garment Upper Body
  - Garment Lower Body
  - Garment Full Body
- ➜ Each Product type is a part of one Product group. Product types can be Top, Blouse, Trouser, Skirt, Cardigan etc. (1-1)
- ➜ Each product can be of one product type. (1-1)
- ➜ Each product may have one or more colors associated with it. Also, each color may be associated with one or more products.(M-M)
- ➜ A product can be composed of multiple fabrics. Also, multiple products can be made from each fabric. (M-M)
- ➜ Each product may have one or more styles associated with it. Also, each style can be found in one or more products. (M-M)
- ➜ A style for a particular body should match the style associated with the product type.
- ➜ Each fabric should have a level indicating:
  - Fabric level 1: Highest quality (skin-level 1 can wear these)
  - Fabric level 2: Medium quality (skin-level 2 can wear these + fabric level 1)
  - Fabric level 3: Lowest quality (skin-level 3 can wear these + fabric level 1 + fabric level 2)

2.2 ER diagram

2.2.1 Initial ER Diagram:

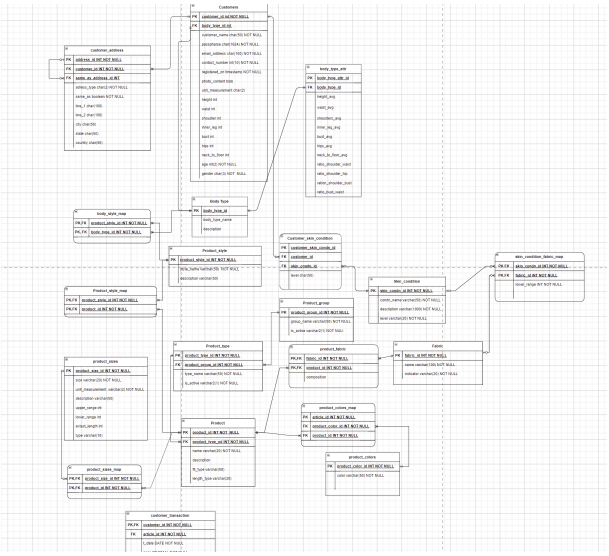

Fig. 1: Initial ER Diagram

2.2.2 Modified ER Diagram:

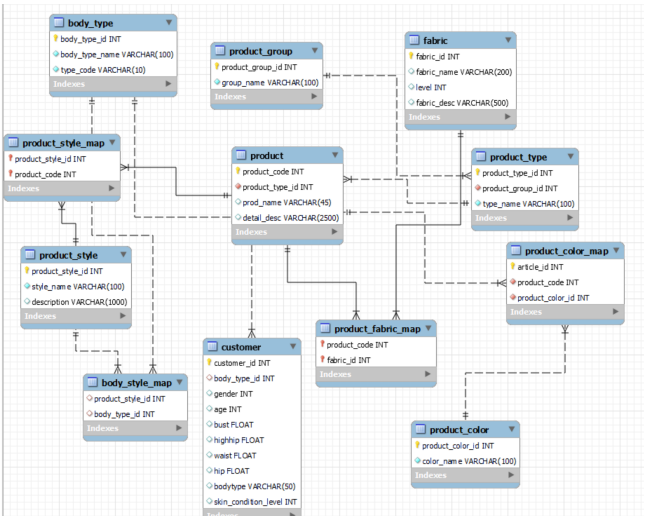

Fig. 2: Modified ER Diagram

2.2.3 Workflow Model:

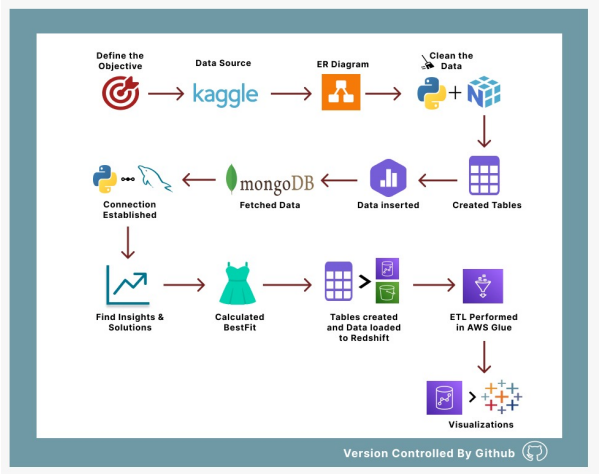

Fig. 3: Workflow

2.3 Implementation

➜ **use of MS-Excel for new data:** Data is gathered from various sources and also some new data about product styles which is unique to our problem was taken from websites and was generated using Excel sheets. Use of Vlookup,IF statement and String concatenations were done for generating insertion queries for this new data that mapped product styles with body shapes.Columns denote body type and rows denote product style. “Y” denotes which style is suited for which body. If any cell is Y, then insertion queries were generated which were used for populating database tables.

➜ **Database connection/API connectivity:** Data is gathered from various sources and also some new data about product styles which is unique to our problem was taken from websites and was generated using Excel sheets. Use of Vlookup,IF statement and String concatenations were done for generating insertion queries for this new data that mapped product styles with body shapes.Columns denote body type and rows denote product style. “Y” denotes which style is suited for which body. If any cell is Y, then insertion queries were generated which were used for populating database tables.

➜ **Connection with MYSQL workbench:** The configuration file for connecting with databases is added so that different sections for different databases can be made and utilized for connecting to different data sources.The configuration parser API parses the file using which connection to MySQL server is made.

➜ **Connection with MongoDB cluster:** Standalone MongoDB cluster is made that is running in the containerized docker instance. Database ProjectData225 is created with bodytype as collection. Standard body measurements parameters for each body type are kept here for referencing purposes later when they would be used for calculation of body shapes of the customers.

➜ **Other data in the form of CSVs were extracted, cleaned and loaded using**

Python: Since our main CSV file has so many columns which may be relevant to 'HM' but irrelevant to our problem hence, columns such as graphical color related, perceived related and index related were dropped. In our project, we are taking into account just three categories; Garment Lower body, Garment Upper Body and Garment Full body. Rest all the rows of product groups were dropped. There were places where columns had null values that were replaced by None in the description of the product.

➜ **Several Main Tables and junction tabled were populated using python:** Static data about product style,fabric, bodytype and other tables are inserted using scripts and python code. Other tables such as product, Productcolormap, Fabriccolormap etc are inserted after looking for which style and fabric is associated with which product by parsing a detail description column in the main article file.

➜ **Several Main Tables and junction tabled were populated using python:** Static data about product style,fabric, bodytype and other tables are inserted using scripts and python code. Other tables such as product, Productcolormap, Fabriccolormap etc are inserted after looking for which style and fabric is associated with which product by parsing a detail description column in the main article file.

2.4 Python code

Since our main article csv file does not have granular details about the product styles as it is denormalized.The general details about the product is given in the description column. So, we searched every product's description using regular expression to find if the product has any style associated with it.If there was any style associated with the product,then its product code was captured that was then used by other tables for making connections of product styles with body type and product.Similarly, since fabric composition was not given in our main article file, We populated fabric related to a particular product by searching product name and description of each row using regular expression to find which fabric is associated with which product.This is essential because fabric is associated with skin condition level of the customer who wants to buy a product.

- ❖ **Finding the body shape of the customer:** The customer measurements are taken and calculations are done using standard configuration taken from the MongoDB document. A customer will ideally fall into these seven categories of body shapes, if not then default is kept as hourglass.
- ❖ **Creating Queries to find the perfect fit for the customer:** This can be done by supplying the program with customer id (such as 44, 85, 40, 42, 183), the product type the customer is looking for(such as Top, Trousers, Skirt, Cardigan, Jacket) and skin conditions, if any(1- severe, 2-mild, 3-no issues). The program first calculates the body shape of the customer and then fetches the list of Product types with the fabric quality according to their skin condition level. Or if customer id is not available, Body shape can be given(such as rectangle,round,hourglass) along with product type and skin condition level to find the best possible outfit.

2.5 ETL

Steps Followed in AWS Redshift and AWS S3:

- Store the CSV files in S3 Bucket.
- Partitions are the folders in the S3 bucket.
- Created cluster with name data225-group6.
- Created tables in the database.
- Stored the csv file records to the database tables using Load Data UI located in the Query editor v2 or use "Copy" command.
- Using Load Data button UI in Query Editor which gives the "copy" command.

Steps followed in AWS Glue:
For ETL, we have used AWS glue. For which we have created a database and tables (using crawler) for the same. Jobs were created and provided a source folder of S3.

- Created a database in AWS glue.
- Added the tables using crawler from S3 bucket.
- Given crawler a name.
- Selected Data Store as S3 and Choose your file from S3 Bucket path.
- Selected the crawl data to be selected from existing account or another account.

- Selected the IAM role which has the policy access required for AWS Glue.
- We have selected frequency as “Run on demand” but hourly, monthly, weekly can also be selected.

Final Review:

- Crawler created successfully and executed.
- Table successfully created by crawler.
- Table schema and details after creation.
- Created a Connection to keep the credentials saved in case required.

Created a job for ETL:

- Alter the Schema of the table if required.
- Job is ready, we can alter the script accordingly and can also alter the file format before running the job.
- Table schema and details after creation.
- After running the job. Following files have been created and if records are more, the job separated one csv file into multiple csv files.

Same Process is repeated for transactions.csv file.

3 Data Analysis and Visualization

3.1 Insights according to our proposal

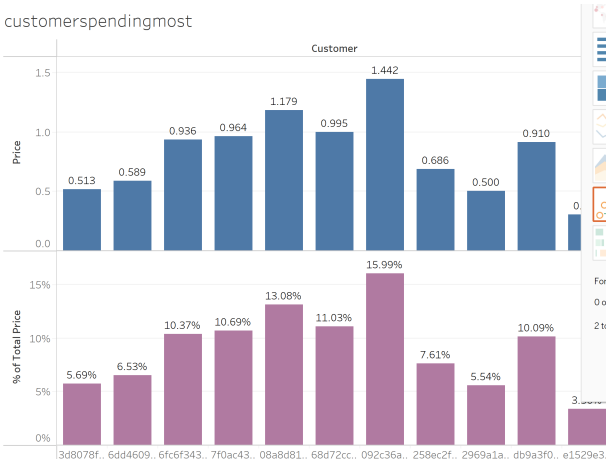

Fig. 4: ss

These are the list of the customers that has spend most amount of money on the purchase of items.The prices in the original data-set are scaled down between 0 and 1.This is a useful insight to know which customers have more spending power and teams may use different strategies to retain these customers.Also, provide incentives such as coupons, promotion codes to low spending customers to attract more sales.

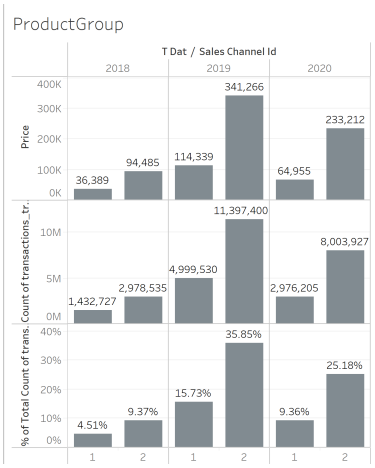

Fig. 5: Initial ER Diagram

The above figure shows the distribution of transactions between two sales channels across the years, 2018-2020.We do not have full yearly data for 2018.so, we will not analyze this year.Comparing between 2019 and 2020, sales were significantly dropped in both the channels.This may be due to the effect of pandemic on stores getting closed and people spending less.

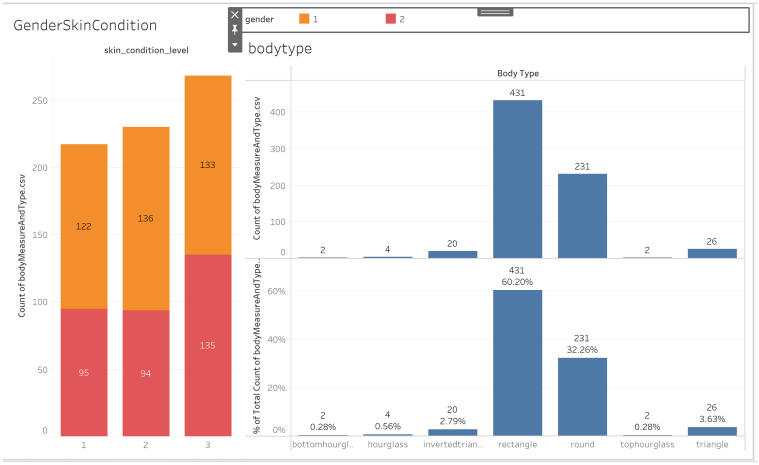

Fig. 6: Initial ER Diagram

The above right graph shows the distribution of body shapes in our customer database.Most of the customers are of type rectangle followed by round, hourglass customers are the least in count.

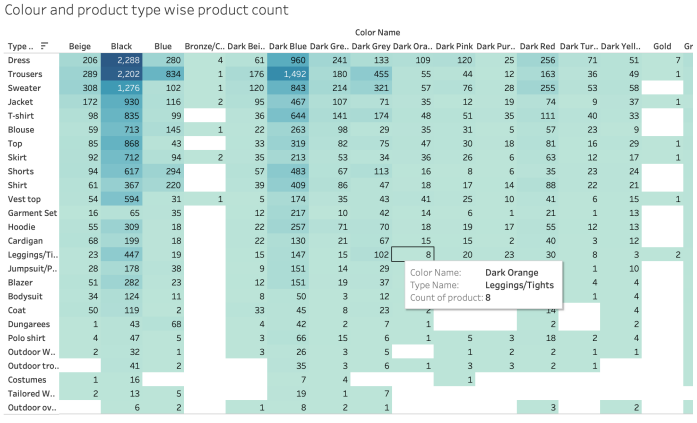

Fig. 6: Initial ER Diagram

Following are the different analysis which we performed to get insights:

1. Fabric Sales Per Year Per Month
2. The total number of products with distinct fabrics with their quality level associated with them.
3. Overall most sold fabric with total sales.
4. Most sold fabric with total sales Per Year Per Month
5. Inventory itemization based on product type, style name, fabric name, color name, count.
6. Top 20 items in inventory based on Product Type, Product Style, Fabric Name and Color.
7. List of names and number of product styles per body type.
8. Count of total customers per body type and then count of customers with different skin condition levels to get a percentage of exception we are targeting for.
9. List of Per day transaction running amount per product type for year(2020) and List all the quarters of the years with minimum sales. To basically target quarters with more fabric options or deals for people with skin conditions.

4 Tools and Technologies used

4.1 MySQL Server and MySQL workbench
4.2 MongoDB Cluster and MongoDB Compass
4.3 Visual Studio code
4.4 Spyder
4.5 WinMerge
4.6 SourceTree
4.7 AWS Redshift
4.8 AWS GLUE
4.9 AWS Sagemaker
4.10 Tableau
4.11 Latex
4.12 Grammarly
4.13 Wordpress for blog
4.14 MS Power Point Presentation
4.15 Draw.io

5 Significance to the real world

In the present day scenario, online shopping is like breathing. Though, there are services for customization in fashion industry, major part of the trend and available services are stereotyped on measurements. Our perspective on selecting a product in this industry is not only focusing on making customization part of regular trend but also on medical conditions of skin and thus handling a part of exception cube. Not everyone knows about styling and hiring a fashion designer may cost a lot of money to the customer. This can be avoided if they are given the options that suit their body shapes in a better fashion.

There are a fraction of customers who suffer from skin related diseases such as rash,eczema and their condition gets aggravated because of the poor quality of the fabric.Customer may not pay minute attention to fabric composition every time he or she shops. Giving them the option to select clothes based on their severity of skin condition may significantly relieve them from the external pain.

Further, this may be beneficial for the companies as well because they will have better understanding of their customers and will stock their inventory accordingly.In addition, the amount of money spent on logistic distribution when the items are returned often may be reduced.

6 Lessons Learned

6.1 How data can be taken from different sources and extracted, Cleaned and merged according to the problem at hand.
6.2 Using the python connectivity with the MySQL server for generating functions and querying to determine an individual's body type and providing them with their desired product choices along with informing them regarding the fabric of the product and what all skin conditions is it suitable for.
6.3 Data cleaning in Spyder using python.
6.4 Using MongoDB for fetching the body measurement composition.
6.5 Using AWS cloud services such as AWS Redshift for data warehousing, AWS S3 for data storage, AWS Glue for our ETL process.
6.6 Connecting AWS Redshift with Tableau for visualizing our results.
6.7 We learned how to collaborate with each other, brainstorm and contribute in a timely manner.

7 Technical Difficulties and Resolutions

7.1 How data can be taken from different sources and extracted, cleaned and merged according to the problem at hand.

- Fabric
- Skin conditions
- Customers

7.2 Thus linking our synthetic data with the available data such as products was challenging.

8 Teamwork

Right from idea sharing and conceptualization stage of this project, every member of our team played their role. We used JIRA for our tracking and agile workflow. We voluntarily picked up our own and helped each other wherever it is needed. By following agile scrum methodology, we also balanced our work correspondingly. Even though our project resembles an iterative architectural workflow there is weekly cadence to let others in team know about what is going on present task and thus leading to group discussion and evolution on our approach down the lane.

9 Pair programming

We have collaborated extensively through Zoom, WhatsApp ,and in person for working together through the making of the project. Most of our SQL related work is directly done on AWS Redshift cluster with saved and shared Queries. For programming where we need to work with data modification and logic we used **Codeshare**(Python Code and SQL Code along with google docs.

10 Agile/Scrum

10.1 zoom links

- Meeting 1
- Meeting 2
- Meeting 3
- Meeting 4
- Meeting 5

10.2 Jamboard

- Link: "Click Here"

10.2 JIRA
Used JIRA to coordinate and distribute tasks among ourselves:

- Link: "Click Here"

11 Future Enhancements

We can further implement some features like:

11.1 Find solutions for analyzing needs for an efficient amount of supply according to customer's demand and thus, minimizing the excessive unnecessary product stocks.

11.2 The background about the customer's choice of styles and what majority of age group people like to wear could also be determined.

11.3 Better inferences by using other advanced tools and techniques.

11.4 Alert emails or notifications once the desired product is in stock.

11.5 This data-set will allow us to do a thorough analysis on user's behavior such as search history, click tracking and order history.

11.6 Additionally we can also draw inferences from trends based on holidays, seasons and festivals.

All of these insights can help us recommend products best suited to a customer's preferences and hence drive sales up.

12 Conclusions

12.1 This project covers all the basic features of any online shopping website.

12.2 This project is to help the customers by showing only those products which are meant for and are required by the customers based on their skin conditions and body type.

12.3 Filters provided to the customers give easy accessibility to get desired product along with fabric information.

12.4 Fulfilled our objective of learning RDBMS, Data warehouse, MySQL, ETL processes and implementing queries to create useful information and present them with the use of Data Visualization techniques.

References

[1] 2022. [Online]. Available: https://https://www.masterclass.com/articles/28-types-of-fabrics-and-their-uses28-different-types-of-fabric. [Accessed: 05- May- 2022].

[2] "The 12 Different Types of Fabric - Pico Cleaners — Blog", Pico Cleaners, 2022. [Online]. Available: https://https://www.masterclass.com/articles/28-types-of-fabrics-and-their-uses28-different-types-of-fabric. [Accessed: 05- May- 2022].

[3] "Types of Fabrics", AanyaLinen, 2022. [Online]. Available: https://www.aanyalinen.com/blogs/aanya-blog/types-of-fabrics. [Accessed: 05- May- 2022].

[4] ”Types Of Fabrics — Everything You Need To Know — Sewing 101”, Sewing.com, 2022. [Online]. Available: https://sewing.com/fabric-types-everything-you-need-to-know/. [Accessed: 05- May- 2022].
[5] 2022. [Online]. Available: https://www.kaggle.com/odins0n/hm256x256. [Accessed: 05- May- 2022].
[6] ”Apple Body Shape: A Comprehensive Guide — the concept wardrobe”, Theconceptwardrobe.com, 2022. [Online]. Available: https://theconceptwardrobe.com/build-a-wardrobe/apple-body-shape.. [Accessed: 05- May- 2022].
[7] ”Figma”, Figma, 2022. [Online]. Available: https://www.figma.com/file.. [Accessed: 05- May- 2022].