#FASHION DATA STORE

## INTRODUCTION

This is a big data / business intelligence project which is part of the fulfillment of my master's degree.

It consist of 2 datasets anonymously sourced from a fashion data store and was used to provide actionable insight for an anonymous fashion store.

I used an extensive analytics framework encompassing big data analytics techniques. This involved the writing of HQL scripts in Hadoop to facilitate seamless data integration, utilizing Python for machine learning models, and leveraging Power BI for visualization. All of which facilitated the extraction of meaningful business intelligence insight within the fashion sector.

##AIMS AND OBJECTIVE

- To properly clean the dataset and perform EDA

- To split the p\_attribute column from JSON to regular column

- To integrate the two datasets together

- To identify customer behavior based on culture.

- To identify customer’s preferences based on color

- To ascertain which occasion clothing significantly contributed to the revenue.

- To build machine learning model which predicts ratings and price

## SKILLS DEMONSTRATED

-Crafting HQL/SQL scripts

-Understanding of Hadoop and the MapReduce ecosystem

-Python-based data cleansing, transformation, and preprocessing

-Utilizing Power BI for data visualization

## CONCLUSION AND RECOMMENDATION

This study focused on data integration and exploration within the fashion industry. The integration process merged two datasets using Hive and a left outer join approach, resulting in a unified table for analysis. Exploratory Data Analysis (EDA) unveiled insights into revenue influencers, customer preferences, and behaviors. Occasions and color impacts on revenue were investigated, with casual wear emerging as a major contributor. Cultural insights on saree fabrics were also revealed. Utilizing MapReduce for data analysis, brand prominence was examined, and machine learning techniques like linear regression and Random Forest were employed to predict ratings and outcomes with respectable accuracies. To enhance future analyses, the study suggests exploring advanced machine learning methods and consistently updating the dataset for relevant insights.