Ryerson University

Department of Electrical, Computer, & Biomedical Engineering
Faculty of Engineering & Architectural Science

Course Title:	Fundamentals of Data Engineering
Course Number:	COE 848
Semester/Year (e.g.F2016)	W2021

Instructor:	Dr. Faezeh Ensan

Assignment/Lab Number:	1
Assignment/Lab Title:	Database Proposal

Submission Date:	January 26, 2021
Due Date:	January 26, 2021

Student LAST Name	**Student FIRST Name**	**Student Number**	**Section**	**Signature***
Shreekant	Vatsal	500771363	01	VS

**By signing above you attest that you have contributed to this written lab report and confirm that all work you have contributed to this lab report is your own work. Any suspicion of copying or plagiarism in this work will result in an investigation of Academic Misconduct and may result in a "0" on the work, an "F" in the course, or possibly more severe penalties, as well as a Disciplinary Notice on your academic record under the Student Code of Academic Conduct, which can be found online at: http://www.ryerson.ca/senate/current/pol60.pdf*

For the final project in the Data Engineering course, I propose to build a database of products offered on an e-commerce platform such as Amazon. The idea behind the database design on such a platform would be to enable the data engineer to gain insights and perform data analysis. This would allow the data scientists to make knowledgeable and informed decisions and predictions based on ML (Machine Learning) algorithms. The database would primarily document the product details, product analysis and customer trends on certain best-selling days. The product details along with the description and manufacturer would help categorize the products and sort them subsequently. The product rating would determine if the product is deemed successful amongst the client database and if they need to be added on the discount days. Furthermore, the highly rated products could lead to determining a popular manufacturer/seller. A successful manufacture/seller on the e-commerce platform would be incentivised by getting recommended to the customer. Another query of the database could determine which products are being sold to the customers on an almost daily basis. This would lead to the creation of daily essentials category and thus better organization of the shelves and more efficient inventory management. In order to filter out the unpopular products, the company can list them on sale days and the products not bought on the said day can be categorized as unpopular. Moreover, to overcome the challenges of trucking in supply chain, products can be categorized as either being shipped from the company's warehouse or the third-party seller. Products sold from the proximity of the warehouse to the customers can offer less than a day delivery and the estimated time of delivery from a vendor can be determined more logically. The high selling products in customer demand should be listed down in order to keep the inventory stacked only with the essential items that are in demand most of the time. It is essential for a growing business to have a recurring income. Building a query of products that offered through a subscription on the company's website is important as this will lead to periodic revenue streams. Customers are often exited to go shopping with their families before stat holidays such as 1st of July, Labor Day, Black Friday, Boxing Day etc. Maintaining a list of products that are strategically released before the listed days is essential to maintaining enough product inventory and not disappointing the customers. Finally, it is important to account for the outliers in the products inventory. At times there can be discrepancies on the website wherein although the ratings are high, but the product reviews are primarily unpleasant. The products that have been given bad reviews but are still rated high stars are important to have recognised and labelled for further audit. In summary, this database would help create insights into product details, history, reviews, ratings, manufacturer popularity, discount-worthy-products, and product outliers. To reiterate, this database would also render data scientists to extract meaningful analysis stemmed from ML models and provide the true picture of behind-the-scenes to the stakeholders involved.

Once the above-mentioned database has been built, I would like to retrieve the information from the database in many forms. This would include answer to questions like:

1. Products offered on the company website, their description, their category and their manufacturer.
2. Products listed by their ratings
3. Are highly rated products built by the same manufacturer?
4. The products that are bought by the customers on a daily basis.

5. In the event of a sale/discount day, which products are the least bought?
6. What products are offered through the company warehouse and not from the manufacturer?
7. What products are out of stock and in customer demand?
8. Products that offer a subscription service in lieu of a one-time purchase.
9. Products that are released 1 day before the stat holidays.
10. The products associated with concerning reviews and high rating.