E-commerce Project Data Cleaning and Validation Documentation

May 23, 2025

Executive Summary

This document outlines the data cleaning, validation, and analysis processes for an e-commerce database migration project on Azure Cloud. Key insights and recommendations are provided to enhance data quality, reporting accuracy, and business intelligence capabilities.

1 Introduction

The project involves migrating and managing an E-commerce database (`ecommerce_db`) on Microsoft Azure. The objectives include:
- Scalable and secure access.
- Advanced analytics and reporting using Power BI.
- Customer segmentation using machine learning.

2 Data Cleaning

2.1 Initial Inspection and Issues

- Missing values in `categories` table: 5 NULLs in `parent_id`.
- Significant issues in `discounts`: Missing foreign keys (`product_id`, `category_id`, `order_id`).
- Action Taken: Excluded `discounts` from analysis due to unusable data.

3 Key Insights

3.1 Customer Segmentation

Using TPOT, customers were segmented into:
- **Seekers**: Extensive browsing, add to wishlist, no purchase.
- **Quick Buyers**: Fast purchase decisions.
- **Hesitant**: Long sessions, no purchases.

3.2 Returns Analysis

- Some returned products received positive reviews, indicating issues like fit or expectations.
- Fake reviews detected from non-purchasers.

4 Recommendations

1. Improve discount data quality by enforcing foreign key constraints.
2. Add age and gender data for better customer segmentation.
3. Engage customers who returned products with positive reviews.
4. Implement mechanisms to detect and handle fake reviews.
5. Introduce delivery date and stock-in date fields for better logistics and sales timing insights.