BÁO CÁO TỔNG HỢP DATA PROFILING - PROJECT 6

Dự án: Xây dựng Automated Data Pipeline cho hệ thống Glamira (41.4M Records)

Người thực hiện: Nguyễn Hữu Huy Anh

Ngày báo cáo: 30/12/2025

---

1. Mục tiêu báo cáo (Executive Summary)

Báo cáo này đánh giá chất lượng dữ liệu (Data Quality) và tính toàn vẹn (Integrity) của hệ thống sau khi thực hiện trích xuất từ MongoDB và nạp tự động vào Google BigQuery.

2. Thống kê quy mô dữ liệu (Data Volume Verification)

Dựa trên kết quả thực thi truy vấn kiểm tra, hệ thống đã nạp thành công toàn bộ dữ liệu từ nguồn vào các bảng đích với số lượng bản ghi như sau:

Tên bảng (BigQuery)	Số lượng bản ghi (Rows)	Định dạng nguồn
summary_final	**41,432,460**	Parquet
ip_locations	**3,239,628**	CSV
products_raw	**35,296**	JSONL

Row	table_name	row_count
1	products_raw	35296
2	ip_locations	3239628
3	summary_final	41432460

3. Đánh giá chất lượng dữ liệu (Data Quality Assessment)

Chúng tôi đã thực hiện các bài kiểm tra sâu về tính nhất quán và độ phủ của dữ liệu trong bảng summary_final:

- **Định danh người dùng:** Chỉ có **397** bản ghi thiếu email trên tổng số 41.4 triệu dòng, cho thấy tỷ lệ thu thập định danh khách hàng cực kỳ cao.
- **Định vị địa lý: 0** trường hợp thiếu địa chỉ IP. Mọi hành vi khách hàng đều có thể gắn với vị trí địa lý.
- **Tiền tệ (Currency):** Ghi nhận đầy đủ **85** loại tiền tệ khác nhau, phản ánh chính xác quy mô thương mại toàn cầu của Glamira.
- **Phạm vi giá (Price Range):** Giá trị dao động từ 0.0 đến 665,485.0, phù hợp với đặc thù sản phẩm trang sức cao cấp.

Row	missing_ip	missing_email	unique_currencies	min_price	max_price
1	0	397	85	0.0	665485.0

4. Phân tích tính kết nối và dữ liệu trống (Deep Insights)

A. Tỷ lệ khớp địa lý (Match Rate)

Kết quả JOIN giữa bảng hành vi người dùng và bảng danh mục địa lý cho thấy sự chính xác tuyệt đối:

- **Unique IPs in Summary:** 3,239,625
- **Unique IPs in Locations:** 3,239,625
- Match Rate: 100.0%
  => Kết luận: Hệ thống Pipeline đảm bảo không làm thất thoát dữ liệu địa lý trong quá trình chuyển đổi.

Row	unique_ips_in_su...	unique_ips_in_loc...	match_rate
1	3239625	3239625	100.0

B. Giải trình về dữ liệu NULL trong cột Price

Qua kiểm tra sâu (Deep Profiling) bằng SQL, chúng tôi ghi nhận:

- **NULL Price Count:** 41,386,675 dòng.
- **Positive Price Count:** 45,784 dòng.
- **Giải trình:** Phần lớn dữ liệu là các sự kiện tương tác (Clickstream) như xem trang, tìm kiếm nên không có giá trị tiền. Chỉ 45,784 dòng mang giá trị dương đại diện cho các hành vi thêm vào giỏ hàng hoặc giao dịch. Đây là cấu trúc dữ liệu **hợp lệ**.

Row	null_price_count	zero_price_count	positive_price_co...	total_rows
1	41386675	1	45784	41432460

5. Kết luận (Conclusion)

Hệ thống Pipeline tự động hóa đã hoàn thành xuất sắc nhiệm vụ:

1. Dữ liệu nạp vào BigQuery đạt độ chính xác **100%** về mặt logic địa lý.
2. Cấu trúc Schema (33 trường hành vi và 28 trường sản phẩm) đã được bóc tách và định dạng chính xác.
3. Hệ thống sẵn sàng cho các bước xây dựng báo cáo BI và phân tích hành vi khách hàng.