Phase 1: Dataset + Baseline Setup

1. Create 1 Final Table for downstream modeling [Completed]:

- Start with full 9M user reviews from Amazon Sports & Outdoors dataset to maximize interaction diversity and reduce cold-start risk.
- Dynamically determine top K products based on review volume — selected top 1237 products to hit a target of 1.5M reviews.
- Filter 1.5M reviews corresponding to these high-interaction products.
- Perform an inner join on parent_asin with product metadata table to enrich reviews with category, brand, and structural info.
- Split the joined table into 3 Parquet chunks of 500K rows each for memory-safe handling.
- Merge all 3 chunks into a single final_joined.parquet, used for all downstream tasks