**Process Flow Diagram & ETL Process Explanation**

**A screenshot of a computer

Description automatically generated**

**1. Introduction**

This project entails the development of an ETL pipeline using Apache Spark to process a merged dataset of Amazon reviews from three product categories: Electronics, Books, and Watches. The primary objective is to identify customers who engage with multiple categories and to derive comprehensive metrics that can inform cross-selling strategies. The enriched dataset includes detailed aggregates such as category-specific average ratings, review counts, an overall positive ratio, and a computed cross-sell score.

**2. ETL Process Explanation**

**a. Extraction**

In the extraction phase, the consolidated dataset is read from a TSV file containing the merged reviews. Apache Spark’s CSV reader is employed with parameters configured for tab separation (sep="\t"), header inclusion, and automatic schema inference. This ensures that the raw data is accurately loaded into a DataFrame with appropriate data types, establishing a robust foundation for subsequent transformations.

**b. Transformation**

The transformation stage is the core of the ETL process and consists of several key steps:

1. **Data Cleaning and Filtering:**  
   Records lacking critical fields—namely, review\_id, review\_date, star\_rating, and review\_body—are filtered out to ensure data integrity. Duplicate entries are removed based on the review\_id to avoid redundant information that could skew analytical results. Additionally, the review\_date field is converted from a string format to a Date type (assuming a yyyy-MM-dd format) to facilitate time series analysis.
2. **Sentiment Derivation:**  
   A new column, sentiment, is derived from the star\_rating. Reviews with a rating of 4 or higher are classified as "Positive," those with a rating of 3 as "Neutral," and those with a rating below 3 as "Negative." This categorization enables rapid segmentation of customer sentiment, which is essential for cross-category analysis.
3. **Customer-Category Aggregation:**  
   The data is then aggregated at the customer and category level. For each unique combination of customer\_id and category, key metrics are computed:
   * **Review Count:** The total number of reviews.
   * **Average Star Rating:** The mean rating for that category.
4. **Pivoting and Enrichment:**  
   The aggregated data is pivoted such that each product category (Electronics, Books, Watches) appears as a distinct set of columns. For instance, columns such as electronics\_review\_count and electronics\_avg\_rating are created. In parallel, overall metrics for each customer are computed:
   * **Total Reviews:** The sum of reviews across all categories.
   * **Weighted Average Rating:** A composite average rating calculated by weighting each category’s average rating by its review count.
   * **Total Categories Reviewed:** The count of distinct categories in which a customer has participated.
   * **Overall Sentiment:** Derived from the weighted average rating, indicating the dominant customer sentiment.
   * **Cross-Sell Score:** Defined as the product of the weighted average rating and the total number of categories reviewed, this metric highlights the cross-selling potential of customers.

**c. Loading**

The final phase involves writing the enriched, customer-level dataset to a TSV file for further analysis and visualization. This output dataset, replete with both granular category-level metrics and overall customer engagement indicators, is designed to be readily imported into visualization tools (such as Power BI or Tableau) for interactive dashboards and deeper analytical insights.

## 3. Conclusion

This ETL pipeline effectively cleans, enriches, and aggregates the raw Amazon reviews data to produce a comprehensive customer-level dataset. The final dataset—encompassing category-specific review counts, average ratings, overall sentiment, and a calculated cross-sell score—provides an in-depth view of customer behavior across multiple product categories. This robust data foundation is ideal for subsequent analytical visualizations and cross-selling strategy development.