MILESTONE 1 REPORT: DATA ACQUISITION & PROCESSING

Course: SEG301 - Search Engines & Information Retrieval
Project: E-Commerce Vertical Search Engine
Team: phap-bot/SEG301_Project
Completion Date: January 25, 2026

---

1. EXECUTIVE SUMMARY

The primary objective of Milestone 1 was to establish a robust data foundation for our specialized search engine. We successfully acquired over **1,000,000 products** from 7 major e-commerce platforms. Our team implemented a high-performance multithreaded crawler system, an automated data cleaning pipeline, and a standardized schema architecture.

Key Achievements

- **Total Documents:** `1,028,126` cleaned records (exceeding project goals by 2.8%).
- **Multi-Platform Support:** 7 major entities (Tiki, eBay, Chợ Tốt, Lazada, CellphoneS, Điện Máy Xanh, FPTShop).
- **Data Integrity: 99.39% quality rate**, with only 0.61% (6,340 docs) filtered during cleaning.
- **Natural Language Processing: 24.4M tokens** extracted, averaging **23.76 tokens/document**.
- **Advanced Engineering:** Implemented Async/Multi-threading and sophisticated Anti-bot bypass mechanisms.

---

2. TEAM & CONTRIBUTIONS

Our team collaborated effectively across different platforms to ensure diverse data sources:

Name	Student ID	Role	Key Responsibilities
Nguyễn Lê Tấn Pháp	QE190155	Team Lead	Crawler logic for Lazada, Điện Máy Xanh, FPTShop
Tô Thanh Hậu	QE190039	Member	Crawler logic for Tiki, Chợ Tốt, eBay
Nguyễn Hải Nam	QE190027	Member	Crawler logic for Lazada, CellphoneS

---

3. DATA STATISTICS

3.1. Overall Metrics

Metric	Value
Raw Documents Crawled	1,034,466
Cleaned Documents	1,028,126
Total Tokens Extracted	24,429,834
Average Tokens/Doc	23.76

3.2. Platform Distribution

Platform	Count	Percentage
Tiki	389,699	37.90%
eBay	302,083	29.38%
Chợ Tốt	249,146	24.23%
Lazada	34,719	3.38%

Platform	Count	Percentage
CellphoneS	31,059	3.02%
Điện Máy Xanh	12,140	1.18%
FPTShop	9,280	0.90%

---

4. TECHNICAL ARCHITECTURE

4.1. Technology Stack

Languages & Runtimes

- **Python 3.8+**: Core logic, API integration, and data cleaning pipeline.
- **Node.js 18+**: Powering Playwright for complex JavaScript-heavy platforms (Lazada, Điện Máy Xanh).

Crawling Frameworks

- `aiohttp` & `asyncio` : Asynchronous crawling for Tiki and Chợ Tốt (maximized throughput).
- `Playwright` & `Selenium` : Browser automation for JS-heavy rendering and bot bypass.
- `httpx` : Modern HTTP client with async support utilized for eBay.
- `Requests` : Lightweight API communication for FPTShop.
- `BeautifulSoup4` & `lxml` : High-speed HTML structural parsing.

NLP & Text Processing

- `Underthesea` : Advanced word segmentation for Vietnamese text.
- `Regex` : Systematic removal of HTML tags, emojis, and special characters.

4.2. Data Processing Pipeline

1. **Extraction (Distributed Crawling):** Platform-specific crawlers gather raw data in diverse formats.
2. **Aggregation (`merge.py`):**
    - Merged 7 platform-specific files into a unified dataset.
    - Global deduplication using `(platform, product_id)` pairs.
    - Schema normalization across 11 key fields.
    - Missing value imputation (defaulting nulls to 0).
3. **Sanitization (`parser.py`):**
    - **Cleaning:** Removal of noise (script/style tags, UI artifacts like "opens in new window").
    - **Tokenization:** Vietnamese word segmentation and token array generation.
    - **Validation:** Final deduplication check and statistical logging.

4.3. Platform-Specific Strategies (Ref: `ai_log.md`)

- **Lazada:** Developed custom logic to detect "No Results" pages. Implemented a `headless` to `visible` switch for manual CAPTCHA solving, followed by persistent cookie storage for seamless background operation.
- **Tiki:** Bypassed 403 errors by reverse-engineering internal **API v2**. Implemented dynamic **x-guest-token** acquisition and **Exponential Backoff** to prevent IP blacklisting.
- **Chợ Tốt:** Mitigated 429 (Too Many Requests) errors through sophisticated User-Agent rotation and duplicate ID detection to handle page shifting.
- **eBay:** Optimized memory usage via intelligent `seen_ids` management. Developed **Fallback Selectors** to handle UI A/B testing variations.
- **Điện Máy Xanh:** Deployed **Concurrent Deep Crawl** architecture, spawning parallel browser tabs to extract accurate ratings and detailed metadata from individual product pages.
- **FPTShop:** Replaced browser-based crawling with **Direct API requests**, resulting in order-of-magnitude speed improvements and higher data accuracy.
- **CellphoneS:** Solved image and price retrieval issues caused by **Lazy Loading** by targeting hidden attributes like `data-src` and `data-ks-lazyload` .

---

5. DATA SCHEMA

5.1. Unified JSON Structure

```
{
  "platform": "tiki",
  "product_id": "123456789",
  "product_name": "iPhone 15 Pro Max 256GB",
  "price": 29990000,
  "original_price": 34990000,
  "discount_percent": 14,
  "product_url": "https://tiki.vn/...",
  "image_url": "https://img.tiki.vn/...",
  "rating": 4.8,
  "review_count": 1234,
  "category": "Mobile Phones",
  "segmented_text": "iPhone 15 Pro Max 256GB",
  "tokens": ["iPhone", "15", "Pro", "Max", "256GB"]
}
```


5.2. Field Definitions

Field	Type	Description	Nullable
`platform`	String	Source platform identifier	No
`product_id`	String	Unique ID within the platform	No
`product_name`	String	Market-facing product name	No
`price`	Float	Current price (VND)	No
`original_price`	Float	Pre-discount price	Yes
`discount_percent`	Integer	Calculation of reduction %	Yes
`product_url`	String	Direct source link	No
`image_url`	String	Featured image link	No
`rating`	Float	Average star rating (0-5)	No
`review_count`	Integer	Total user feedback count	No
`category`	String	Product classification	No
`segmented_text`	String	Cleaned, segmented Vietnamese string	No
`tokens`	Array	Processed token list for indexing	No

---

6. CHALLENGES & SOLUTIONS

Throughout the 4-week development cycle, we navigated significant technical hurdles:

1. **Access Blocking (Rate Limit/403/429):**
   - *Problem:* Systems like Tiki/Chợ Tốt blocked IPs within minutes.
   - *Solution:* Implemented `Semaphores` for rate limiting and leveraged `X-Guest-Token` to mimic authentic user sessions.
2. **RAM Exhaustion:**
   - *Problem:* Loading 1M+ records into memory caused system crashes.
   - *Solution:* Adopted `JSONL` streaming with Python `generators`. Used a lightweight `set()` for IDs only, drastically reducing memory footprint.
3. **Data Inconsistency & Price Ranges:**
   - *Problem:* Platforms often display "placeholder" prices or ranges (e.g., 30k-100k).
   - *Solution:* Upgraded to a "Deep Crawl" model, visiting individual product detail pages to ensure price and inventory accuracy.
4. **Lazy Loaded Assets:**
   - *Problem:* Product images returned as base64 placeholders or empty strings.
   - *Solution:* Enhanced crawlers with `wait_until` conditions and targeted alternative DOM attributes (`data-src`) for actual asset links.
5. **Unstructured Vietnamese Text:**
   - *Problem:* Product names contained a mix of "Teencode," technical English, and Vietnamese.
   - *Solution:* Fine-tuned `Underthesea` configurations and pre-processed text with custom filters to normalize terminology before tokenization.

7. CONCLUSION

Milestone 1 has been successfully concluded with **1,028,126 documents** acquired from 7 major platforms. This foundational dataset provides the scale and quality necessary for the subsequent indexing and search optimization phases.

Milestone Highlights:

- **Goal Met:** 102.8% of the 1M document target achieved.
- **High Quality:** 99.39% data retention after rigorous cleaning.
- **Scale:** 24.4M processed tokens ready for inverted indexing.
- **Architected:** Standardized JSONL schema for cross-platform interoperability.

Report Prepared By: Team phap-bot/SEG301_Project
Last Updated: January 29, 2026