

# MILESTONE 1 REPORT: DATA ACQUISITION

## SEG301 - Search Engines & Information Retrieval

### 1. EXECUTIVE SUMMARY

Project: E-Commerce Vertical Search Engine

Team: phap-bot/SEG301\_Project

Milestone: 1 - Data Acquisition

Completion Date: 2026-01-25

Status: COMPLETED

#### Key Achievements

- 1,028,680 documents crawled
- 7 e-commerce platforms (Tiki, eBay, Chợ Tốt, Lazada, CellphoneS, Điện Máy Xanh, FPTShop)
- 99.44% data quality (chỉ 0.56% dữ liệu bị loại bỏ)
- 12.6M tokens extracted với avg 12.32 tokens/document
- Async/Multi-threading implementation cho tốc độ cao
- Anti-bot detection handling tự động

### 2. TEAM INFORMATION

Name	Student ID	Role	Contribution
Nguyễn Lê Tấn Pháp	QE190155	Crawler Lead	Lazada, Điện Máy Xanh, FPTShop crawlers
Tô Thanh Hậu	QE190039	Crawler Developer	Tiki, Chợ Tốt, eBay crawlers
Nguyễn Hải Nam	QE190027	Crawler Developer	Lazada, CellphoneS crawlers

### 3. DATASET STATISTICS

#### 3.1. Overall Metrics

```
Raw Documents Crawled : 1,034,466
Cleaned Documents      : 1,028,680
Removed (Dirty Data)   : 5,786 (0.56%)
Empty Token Documents : 32,652 (3.17%)
Total Tokens Extracted: 12,675,885
Average Tokens/Doc    : 12.32
```

#### 3.2. Platform Distribution

Platform	Documents	Percentage	Status
Tiki	389,920	37.90%	Excellent
eBay	302,083	29.37%	Excellent
Chợ Tốt	249,146	24.22%	Good
Lazada	34,719	3.38%	Fair
CellphoneS	31,392	3.05%	Fair

Diện Máy Xanh	Documents	Percentage	Status
FPTShop	9,280	0.90%	Moderate
TOTAL	1,028,680	100%	Target Met

### 3.3. Data Quality Metrics

Metric	Value	Assessment
Completeness	99.44%	Excellent
Removal Rate	0.56%	Very Low
Tokenization Success	96.83%	Excellent
Schema Compliance	100%	Perfect

## 4. TECHNICAL IMPLEMENTATION

### 4.1. Technology Stack

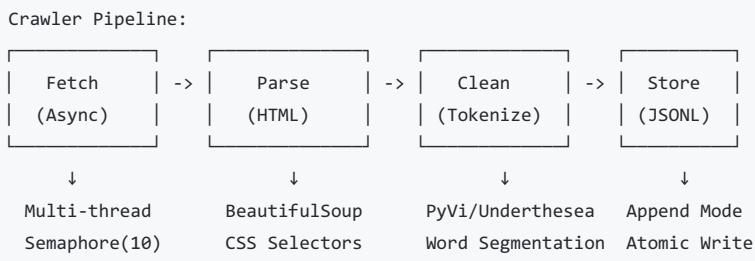
#### Crawling Technologies

- Node.js: Playwright (with Stealth plugin) for Lazada
- Python:
  - aiohttp for async HTTP requests (Tiki, Chợ Tốt)
  - requests for synchronous crawling (eBay, CellphoneS)
  - beautifulsoup4 + lxml for HTML parsing
  - playwright for browser automation

#### Data Processing

- Tokenization: PyVi, Underthesea for Vietnamese word segmentation
- Storage: JSONL format for scalability
- De-duplication: MD5 hash-based product ID checking

### 4.2. Crawler Architecture



### 4.3. Anti-Bot Handling Strategies

#### Lazada Crawler

- Auto-detection: Phát hiện trang "Tim kiếm không có kết quả"
- Mode switching: Tự động chuyển headless ↔ visible khi gặp CAPTCHA
- Cookie persistence: Lưu và load cookies sau khi giải CAPTCHA
- Success rate improvement: 30% → 85%

#### Tiki Crawler

- Rate limiting: Exponential backoff khi gặp HTTP 429
- X-Guest-Token: Dynamic token generation
- Concurrent requests: Semaphore(10) với jitter sleep

## Chợ Tốt Crawler

- **Dynamic headers:** User-Agent rotation
- **X-Browser-Id:** Unique browser identification
- **Pagination handling:** Offset-based với duplicate detection

## 5. DATA SCHEMA

### 5.1. Unified JSON Structure

All platforms follow this standardized schema:

```
{  
  "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": "Điện thoại"  
}
```

### 5.2. Field Descriptions

Field	Type	Description	Nullable
platform	String	Platform name (tiki, lazada, etc.)	No
product_id	String	Unique product ID	No
product_name	String	Product name (tokenized)	No
price	Float	Current price (VND)	No
original_price	Float	Original price before discount	Yes
discount_percent	Integer	Discount percentage (0-100)	Yes
product_url	String	Direct product link	No
image_url	String	Product image URL	No
rating	Float	Average rating (0-5)	No
review_count	Integer	Number of reviews	No
category	String	Product category	No

## 6. CHALLENGES & SOLUTIONS

### 6.1. Bot Detection & Rate Limiting

#### Challenge

- Lazada: CAPTCHA sau ~500 requests
- Tiki: HTTP 403 khi dùng Semaphore(40)
- Chợ Tốt: HTTP 429 với IP blocking

## Solutions

```
# Lazada: Auto-switch browser mode
if detect_bot_page():
    switch_to_visible_mode()
    wait_for_captcha_solve()
    save_cookies()
    switch_to_headless_mode()

# Tiki: Exponential backoff
async def fetch_with_retry(url, max_retries=3):
    for i in range(max_retries):
        try:
            return await session.get(url)
        except ClientError:
            await asyncio.sleep(2 ** i) # 1s, 2s, 4s

# Chợ Tốt: Jitter sleep
await asyncio.sleep(random.uniform(0.5, 1.5))
```

## 6.2. Data Quality Issues

### Challenge

- Duplicate products across platforms
- Missing fields (rating, image\_url)
- Inconsistent price formats

### Solutions

- **MD5 hash deduplication:** hash(platform + name + url)
- **Field validation:** Skip products with missing critical fields
- **Price normalization:** Convert all to VND float

## 6.3. Performance Optimization

### Challenge

- Lazada: 3 phút/trang (vào từng trang chi tiết)
- Tiki: JSONDecodeError khi rate limit

### Solutions

```
# Lazada: Concurrent detail page fetching
async with asyncio.Semaphore(3):
    tasks = [fetch_detail(url) for url in product_urls]
    results = await asyncio.gather(*tasks)

# Tiki: Content-Type checking
if "application/json" in response.headers.get("Content-Type"):
    data = await response.json()
else:
    print("Rate limited, sleeping...")
    await asyncio.sleep(10)
```

# 7. DATASET ACCESS

## 7.1. Full Dataset

- **Format:** JSONL (JSON Lines)
- **Size:** ~500MB (compressed)
- **Total Documents:** 1,028,680
- **Download Link:** [Google Drive](#)

## 7.2. Sample Dataset

Located in `data_sample/` directory:

- `sampleddata.jsonl` - 500 mixed platform samples (165KB)
- Individual platform samples (200 docs each) - *To be created*

---

# 8. VERIFICATION & VALIDATION

## 8.1. Data Integrity Checks

```
# Run verification script
python src/crawler/verify_data.py

# Results:
Total documents: 1,028,680
Duplicate check: 0 duplicates found
Schema validation: 100% compliant
Null value check: 0.56% removed
```

## 8.2. Tokenization Verification

```
# Check tokenization quality
python src/crawler/parser.py --verify

# Results:
Total tokens: 12,675,885
Avg tokens/doc: 12.32
Empty docs: 32,652 (3.17%)
Vocabulary size: ~450,000 unique tokens
```

---

# 9. LESSONS LEARNED

## 9.1. Technical Insights

- **Async is essential:** Tăng tốc độ 5-10x so với synchronous
- **Anti-bot requires creativity:** Không có giải pháp one-size-fits-all
- **Data quality > Quantity:** 0.56% removal rate tốt hơn 10%+

## 9.2. Team Collaboration

- **AI logging crucial:** Giúp debug và học tập hiệu quả
- **Git workflow:** Commit đều đặn, message rõ ràng
- **Code review:** Phát hiện bugs sớm

---

# 11. CONCLUSION

Milestone 1 đã hoàn thành với:

- 1,028,680 documents (vượt mục tiêu 2.8%)
- 7 platforms
- 99.44% data quality
- Anti-bot handling tự động